Rapid Divergence of Gene Duplicates on the *Drosophila melanogaster* X Chromosome

Kevin Thornton and Manyuan Long*†*

*Committee on Genetics and †Department of Ecology and Evolution, University of Chicago

The recent sequencing of several eukaryotic genomes has generated considerable interest in the study of gene duplication events. The classical model of duplicate gene evolution is that recurrent mutation ultimately results in one copy becoming a pseudogene, and only rarely will a beneficial new function evolve. Here, we study divergence between coding sequence duplications in *Drosophila melanogaster* as a function of the linkage relationship between paralogs. The mean K_a/K_s between all duplicates in the *D. melanogaster* genome is 0.2803, indicating that purifying selection is maintaining the structure of duplicate coding sequences. However, the mean K_a/K_s between duplicates that are both on the X chromosome is 0.4701, significantly higher than the genome average. Further, the distribution of K_a/K_s for these X-linked duplicates is significantly shifted toward higher values when compared with the distributions for paralogs in other linkage relationships. Two models of molecular evolution provide qualitative explanations of these observations—relaxation of selective pressure on the duplicate copies and, more likely, positive selection on recessive adaptations. We also show that there is an excess of X-linked duplicates with low K_s, suggesting a larger proportion of relatively young duplicates on the *D. melanogaster* X chromosome relative to autosomes.

Introduction

It is generally thought that the fate of most duplicate genes will be degeneration into pseudogenes because of recurrent deleterious mutations (Haldane 1933; Fisher 1935; Lynch and Conery 2000), and only rarely will duplicates gain new function. However, the genomes of many eukaryotes contain a large fraction of duplications which appear to have avoided degeneration (Rubin et al. 2000; Long and Thornton 2001). Alternative theoretical models of the fates of gene duplications have shown that duplicates may persist if they are advantageous to the organism (Clark 1994), gain an advantageous novel function (Walsh 1995), or subfunctionalize (Force et al. 1999; Lynch and Force 2000). With respect to sequence evolution, the first two processes are adaptive, whereas subfunctionalization is a neutral process. To date, however, the general role of positive selection in gene family evolution remains unclear, with the exceptions of gene families involved in immune responses (Hughes and Nei 1992; Stahl et al. 1999) or male reproductive functions (Wyckoff et al. 2000), and the few examples of very young duplications resulting in chimeric proteins with novel expression patterns (Long and Langley 1993; Nurminsky et al. 1998; Wang et al. 2000), indicating a role for positive selection on novel functions. Here, we study the rates of divergence between gene duplicates in the *Drosophila melanogaster* genome (Adams et al. 2000) with respect to genome location, revealing significantly faster divergence between X-linked duplicates, indicative of positive selection on X-linked duplicates. We also show that there is an excess of X-linked duplicates with low K_s, suggesting a larger proportion of relatively young duplicates on the *D. melanogaster* X chromosome relative to autosomes.

Key words: gene duplication, *Drosophila*, sex chromosomes, adaptation, selection, nonsynonymous, synonymous.

Address for correspondence and reprints: Manyuan Long, Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, Illinois 60637.
E-mail: mlong@midway.uchicago.edu.

Mol. Biol. Evol. 19(6):918–925. 2002

Methods

Identification of paralogs in *D. melanogaster*

We identified paralagous gene pairs using Release 2 (October 2000) of the *Drosophila* Genome Project (Adams et al. 2000) using methods similar in principle to those of Lynch and Conery (2000). Sequence files were obtained directly from www.fruitfly.org, and positional information for each gene was parsed directly from the sequence headers. We used the FASTA33 package (Pearson 1990) to identify paralogs from the peptide sequence annotation in an all-by-all comparison. In the initial FASTA33 searches, we rejected all alignments with amino acid identity less than 0.30 or an aligned region less than 35 amino acids in length (or both). We parsed the resulting output, purging for redundancy and requiring two-way hits to identify paralogs. For the purpose of further analysis, we required that amino acid identity be $\geq$0.50 in the region aligned by FASTA33 because we found K_a to be close to saturation for alignments with identity $<$0.50, suggesting that the alignments might not reflect true paralogy.

We also generated a second restricted data set consisting only of gene families of size 2 with a maximum K_s of $\leq$0.5. This was done in order to avoid problems of both phylogenetic nonindependence which arise when analyzing multigene families and the possibility of serious errors in estimating K_a and K_s between highly diverged genes (Comeron 1995). For the data sets discussed here, we ignored genes on the small fourth chromosome and focused on chromosomes X, 2, and 3 because there are very few genes on the fourth chromosome involved in duplication events.

Divergence Analysis

We aligned the peptide sequences for each paralagous pair using Clustal W 1.81 (Thompson et al. 1994)

and then aligned the coding sequences (CDS) for each paralagous pair, using amino acid alignments as guides. We then calculated the number of amino acid replacement substitutions per site (K_a) and the number of synonymous substitutions per site (K_s) for each pair of aligned CDS sequences using Li's (1993) method as implemented in the GCG 10.1 software package. We also calculated K_a and K_s for each pairwise comparison by the maximum likelihood method implemented in PAML 3.01d (Yang 2000). For the likelihood method, we used the F3×4 method to calculate equilibrium base compositions for each pairwise comparison, which corrects for codon usage bias (Dunn, Bielawski, and Yang 2001). Results obtained from the likelihood and Li's method were very similar, so we report the results from the Li method here for simplicity.

In general, assuming the strict neutrality of silent substitutions, a $K_a/K_s < 1.0$ indicates selective constraint on amino acids (but does not rule out positive selection), whereas $K_a/K_s > 1.0$ is often taken as evidence for strong positive Darwinian selection. A $K_a/K_s = 1.0$ is the expectation under a strictly neutral model of molecular evolution (Kimura 1983) and should be observed for unconstrained sequences such as pseudogenes. In the extreme case of a pair of duplicates where one gene maintains its original function and the other copy is a pseudogene, the pair's K_a/K_s ratio could be as low as 0.5. We therefore take 0.5 as conservative criterion to test if both copies of the gene duplicates are functional. We used a simple form of the sign test (Sokal and Rohlf 1995) to test the null hypothesis that K_a/K_s values are equally likely to be less than 0.5 or greater than 0.5. In this test, the distribution of the ratio should follow binomial distribution which can be approximated by the normal distribution $C = (X - 0.5N)/(0.5 \times \sqrt{N})$ where X is the number of duplicate pairs with $K_a/K_s < 0.5$, and N is the total number of duplicate pairs.

Statistical Procedures

In order to assess the statistical significance of the observed mean K_a/K_s value between X-linked duplicates, we employed a random resampling procedure similar to the standard bootstrap (Sokal and Rohlf 1995, pp. 823–825). If the entire data set consists of m values and a subset of size n has a mean $K_a/K_s = \bar{x}_{obs}$, we randomly chose n values from the set of m (with resampling) and calculated the mean, $\bar{x}$. The resampling was repeated 10^5 times, and the corresponding one-tailed P-value is estimated from the fraction of runs where $\bar{x} \geq \bar{x}_{obs}$. In the case where we ask if the mean is significantly lower than expected, the P-value is the proportion of runs where $\bar{x} \leq \bar{x}_{obs}$. We applied the one-sided Kolmogorov-Smirnov two sample test (Sokal and Rohlf 1995, p. 434) to test for differences between two empirical distributions, calculating P-values using the R software package (Ihaka and Gentleman 1996).

Software and Data Availability

The programs written to automate and parse the FASTA33 searches, alignments, and calculations of K_a/K_s were written in perl. The resampling method was implemented in C. These programs, and a MySQL database of the computational results, are available from the authors upon request.

Results

We first show that most of the gene duplicates have a K_a/K_s ratio lower than 0.5 (with average $K_a/K_s = 0.2803$). Using the sign test (see *Methods* for a description), we have 1,677 pairs that have K_a/K_s ratio less than 0.5 and only 164 pairs equal or larger than 0.5. These numbers yield a C-value of 35.26, indicating the probability of the null hypothesis is very low ($P \ll 0.001$). Thus, in general, both members of the duplicate pairs we analyzed are subject to selective constraint, indicating that both copies are functional. Furthermore, on the X chromosome, 70 pairs out of 107 have a K_a/K_s ratio lower than 0.5, yielding a C-value of 3.20 ($P < 0.01$), suggesting that at least most of the pairs are functional in both copies of the duplicate. These observations prompted further analysis.

Figure 1 shows the distributions of K_a/K_s for four different kinds of duplicate pairs: X-linked and autosomal-linked (fig. 1*a*) and unlinked pairs (X-autosome or chromosome 2/3; fig. 1*b*). Qualitatively, the distributions for linked autosomal (fig. 1*a*), unlinked X-autosome, and unlinked autosomal (fig. 1*b*) look identical, with most of the mass centered near $K_a/K_s \approx 0.27$. The mean values for these three distributions fall in the narrow range of 0.2581–0.2740 (table 1). However, the distribution of K_a/K_s for X-linked duplicates has less mass around $K_a/K_s \approx 0.27$ than the other three distributions and more mass where $K_a/K_s > 0.50$ (fig. 1*a*). The K_a/K_s distribution for X-linked duplicates is significantly shifted to the right when compared with the distribution for linked, autosomal paralogs ($P = 10^{-7}$). Further, the mean K_a/K_s between X-linked duplicates is 0.4701, nearly double that of all other duplicates in the *Drosophila* genome, regardless of linkage (table 1). The high mean K_a/K_s between X-linked duplicates and the shift in mass of the distribution of K_a/K_s implies that a subset of X-linked gene duplicate pairs have diverged much more rapidly from each other than have most gene duplicate pairs in *Drosophila*.

We used a bootstrap-like random resampling procedure to assess the significance of the observed pattern (see *Methods*). First, however, we purged all gene families with more than two members to eliminate problems of phylogenetic nonindependence. Secondly, we restricted the data set to pairs diverged by $K_s \leq 0.5$ because larger K_s values may suffer from serious errors in estimation (Comeron 1995). In this restricted data set, there are 120 duplicate pairs, and the subset of X-linked duplicates contains 23 pairs with a mean K_a/K_s of 0.4797 (table 1). Given the distribution of K_a/K_s from the 120 pairs, we estimate the probability of observing a subset of size 23 with a mean $K_a/K_s \geq 0.4797$ to be 0.0256. For the entire genome (including both nonindependent pairs and pairs with $K_s > 0.5$), there are 1,841 paralagous pairs, and the subset of X-linked duplicates con-

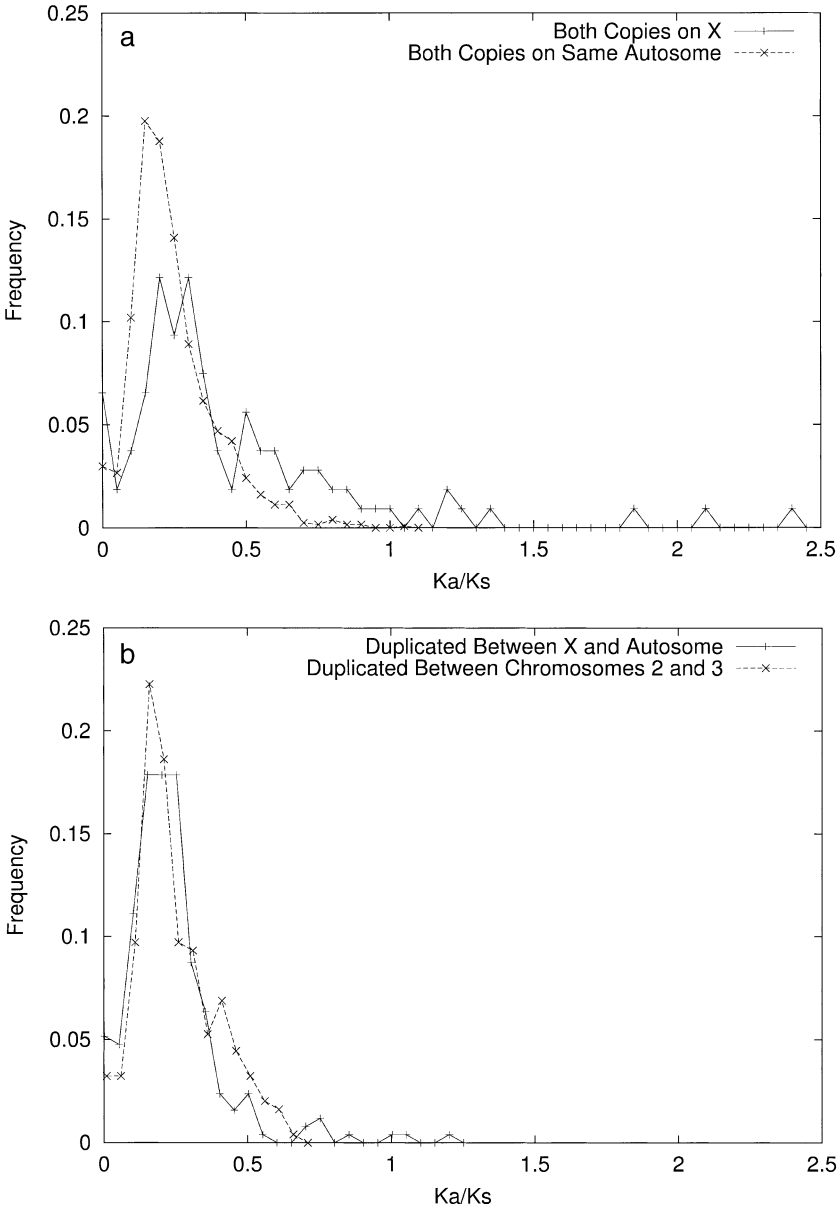

FIG. 1.—Frequency distributions of K_a/K_s between paralagous genes in *D. melanogaster*. K_a/K_s is binned into groups of size 0.05. *a*, The distribution of K_a/K_s between duplicates on the same chromosome. The distribution for X-to-X duplicates shows a long tail to the right and a higher mean value (table 1) than duplicates both linked to the same autosome. *b*, The distribution of K_a/K_s for unlinked duplicates, i.e., between the X and an autosome or between chromosomes 2 and 3, the major autosomes of *D. melanogaster*. Compared with (*a*), the two distributions plotted look nearly identical, with the exception of a small tail to the right (indicating rapidly divergening genes) for X-autosome duplicates.

Table 1
Mean K_a/K_s Between Linked and Dispersed Paralogs in *Drosophila melanogaster*.

Duplicate Type	Mean (K_a/K_s)	Mean K_s	*n*
$X \rightleftharpoons X \ldots$	0.4701 (0.4797)	0.8934 (0.1957)	107 (23)
$A_i \rightleftharpoons A_i \ldots$	0.2697 (0.3044)	1.3356 (0.1625)	1,235 (87)
$X \rightleftharpoons A \ldots$	0.2581 (0.4789)	1.5369 (0.2240)	252 (7)
$A_i \rightleftharpoons A_j \ldots$	0.2740 (0.4333)	1.4921 (0.274)	247 (3)
All	0.2803 (0.3513)	1.3585 (0.1752)	1,841 (120)

NOTE.—Values in parentheses are obtained after removing all gene families of size >2 and all paralagous pairs diverged by $K_s > 0.5$. X refers to the X chromosome, and A refers to an autosome. The mean K_a/K_s for $X \rightleftharpoons X$ duplicates is 0.4701 when considering the whole genome. This value is nearly double the genome-wide mean of 0.2803. We use the $\rightleftharpoons$ notation to indicate that the direction of duplication is unknown

tains 107 pairs with a mean K_a/K_s of 0.4701 (table 1). We estimate the probability of this observation to be $<10^{-5}$. Thus, the observed pattern is highly significant. Because the whole-genome comparison includes duplicate pairs of all ages, it indicates that the rapid divergence between X-linked duplicates is a genome-wide pattern and is an ancient and ongoing process in *Drosophila*.

Table 1 shows that the mean K_s for X-linked duplicates is 0.8934, whereas the genomic mean is 1.3585, suggesting a large difference between X-linked and autosomal duplicates. The distributions of K_s between duplicate pairs are shown in figure 2. The histograms are plotted separately for the linkage relationships described

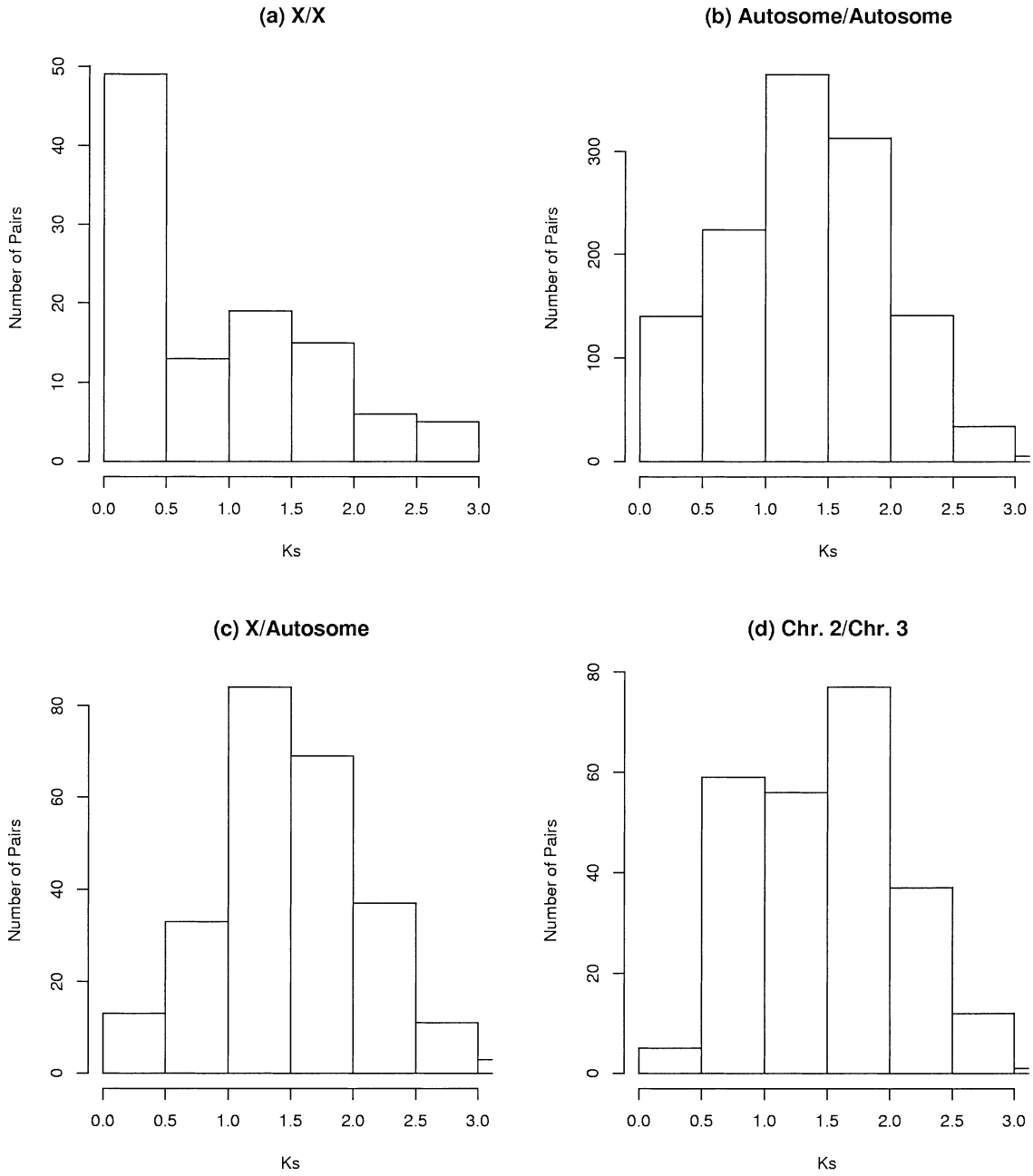

FIG. 2.—Frequency distributions of K_s between paralagous genes in *D. melanogaster*. *a*, The distribution of K_s between duplicates on the X chromosome, showing a large fraction of pairs with low K_s. *b*, The distribution of K_s for linked, autosomal duplicate pairs. *c*, The distribution of K_s for X-autosome duplications. *d*, The distribution of K_s for duplications between chromosomes 2 and 3.

earlier and in figure 1. The distribution of K_s for X-linked duplicates is significantly shifted toward lower values (fig. 2*a*) compared with the other distributions ($P < 10^{-11}$).

Discussion

We have studied the distribution of K_a/K_s between duplications of coding sequence in the *D. melanogaster* genome. On average, X-linked duplicate pairs have a K_a/K_s nearly double that of all other duplicate pairs in the genome (fig. 1 and table 1). Moreover, we found evidence for a significant acceleration of K_a/K_s for X-autosome duplicates, which one would predict to be somewhere in between the K_a/K_s for X-X and autosome-autosome duplications. Consistent with this prediction, the distribution of K_a/K_s for X-autosome duplicates shows a tail to the right (fig. 1*b*), and the difference between the distributions is significant ($P = 0.0355$) even though only 9 duplicate pairs make up the tail. Further, it is likely important to consider the direction of duplication. If most duplication events are from the X to the autosome, and the X-linked locus is constrained, then only the derived autosomal copy is free to evolve rapidly, resulting in a mean K_a/K_s for X-autosome duplicates close to the genome average. Because

the data obtained from the *D. melanogaster* genome are all pairwise, we cannot infer which copy is ancestral. Resolution of this issue is only possible with extensive sequencing of duplicate gene pairs from outgroup species.

In order to explain the rapid divergence between X-linked paralogs, we need to consider the following possible hypotheses: First, it is possible that we have analyzed an unknown number of X-linked pseudogenes, inflating the K_a/K_s between X-linked duplicates. Secondly, K_a/K_s may be higher between X-linked duplicates either because of higher constraint at silent sites on the X, relative to autosomal loci, or because of a relative lack of constraint on X-linked replacement sites. Finally, K_a/K_s may be accelerated by fixation of amino acid changes under selection.

X-linked Pseudogenes

Pseudogenes are nonfunctional duplications of coding sequence and as such are expected to evolve with high K_a/K_s because of an absence of purifying selection. Thus, the acceleration of K_a/K_s on the X chromosome could be because of including pseudogenes in the analysis. Because we used sequence annotation obtained directly from the genome project web site, we would only have analyzed pseudogenes if they had been misdiagnosed as functional loci. We find this explanation unlikely for three reasons. First, it would imply that X-linked pseudogenes are disproportionately misdiagnosed as functional genes, which seems implausible. Secondly, the pseudogene argument would predict that the mean K_s between X-linked paralogs would be much smaller than the genome average because pseudogenes are rapidly eliminated from *Drosophila* genomes (Petrov, Lozovskaya, and Hartl 1996). To test this, we again restricted our analysis to independent pairs diverged by $K_s \leq 0.5$. For this data set, the mean K_s between X-linked duplicates is 0.1957, which is actually higher than the genomic mean but not significantly high ($P = 0.2610$). Finally, the K_a/K_s is less that 0.50 for most of the X-X pairs, suggesting selective constraint on amino acid substitutions.

Constraint on Silent Sites

In *Drosophila*, patterns of codon usage bias have been interpreted as evidence for weak selection on silent sites (Akashi 1995). Genes on the X chromosome of *D. melanogaster* show higher codon bias than autosomal loci (Comeron, Kreitman, and Aguade 1999), and Powell and Moriyama (1997) observed an inverse relationship between codon bias and K_s, as expected from the weak selection hypothesis. It is therefore possible that K_a/K_s is accelerated for X-linked duplications because of stronger constraint on silent sites at X-linked loci. Table 1 shows that the mean K_s for X-linked paralogs is 0.8934, substantially lower than the genomic mean of 1.3585, seemingly consistent with the codon bias hypothesis. However, we do not believe that codon bias is a good explanation for the results for two reasons. First, the excess codon bias on the *D. melanogaster* X chromosome is rather slight (Comeron, Kreitman, and Aguade 1999), and silent divergence between *D. melanogaster* and its close relative *D. simulans* is not reduced on the X chromosome relative to autosomes (Bauer and Aquadro 1997). More important, however, is the possibility that the relationship between K_s and codon bias observed by Powell and Moriyama (1997) is a consequence of using measures of divergence that do not properly account for the compositional bias of the sequences, as pointed out by Dunn, Bielawski, and Yang (2001). Using the maximum likelihood method of the PAML package (Yang 2000), Dunn, Bielawski, and Yang (2001) were only able to recover an inverse relationship between K_s and codon bias by not accounting for compositional bias (per coding position) in the estimates. Thus, if the main reason why we observe an acceleration of K_a/K_s on the X chromosome is simply the result of codon bias, then repeating the analysis using the likelihood method (Yang 2000) and correcting for codon bias should eliminate the evidence for acceleration between X-linked paralogs. However, repeating the analysis using PAML (Yang 2000) and correcting for compositional bias (see *Methods*) did not change any of the qualitative patterns shown in figures 1 or 2 and table 1, suggesting that codon bias alone cannot explain the increased K_a/K_s between X-linked duplicates.

As an alternative to codon bias, the distribution of K_s in figure 2*a* suggests that there is a higher proportion of young duplicate pairs on the X chromosomes compared with the autosomes. We hypothesize that either the rate of gene duplication on the *D. melanogaster* X chromosome is higher than for the autosomes or that there has been a recent burst of duplication on the X. Both these hypotheses provide an explanation for the excess of X-linked duplicate pairs with low K_s (fig. 2*a*).

Slightly Deleterious Substitutions

It is possible that gene duplicates experience less selective constraint with slightly deleterious mutations. Under this model, their fates will be governed by drift rather than selection. The two critical parameters to consider are the effective population size of the X relative to autosomes and the dominance of the weakly deleterious mutations (Charlesworth, Coyne, and Barton 1987). This relaxed constraint model qualitatively provides an explanation for the distributions of K_a/K_s between duplicate pairs (fig. 1). However, this model has two limitations as a general explanation for the data. First, the relative difference in constraint between the X and the autosomes would have to be about 1.74-fold (0.4701/0.2697, table 1), whereas a maximum of a 1.2-fold difference is expected (this maximum occurs when weakly deleterious alleles are fully dominant) (Charlesworth, Coyne, and Barton 1987). It is known that most deleterious mutations are recessive rather than dominant (Crow and Temin 1964; Mukai et al. 1972; Crow and Simmons 1983). The effect of slightly deleterious recessives is to slow the rate of substitution on the X relative to autosomes (Charlesworth, Coyne, and Barton 1987; McVean and Charlesworth 1999), a prediction in-

compatible with our observations. Secondly, relaxation of constraint should apply to all duplicate pairs, and so the mode of the distribution of K_a/K_s on the X should be near the mean, which is not what we observe (fig. 1). Rather, figure 1 shows a large mode centered near the genomic mean and a large tail of pairs with accelerated K_a/K_s.

Adaptive Models

One possible adaptive reason why X-linked duplicates should diverge rapidly is that selectively favorable mutations may generally be fully or partially recessive and fix via transmission in the heterogametic sex (males in *Drosophila*). Charlesworth, Coyne, and Barton (1987) have shown that X-linked genes will evolve more rapidly than autosomal loci, provided that most adaptive mutations are recessive or partially recessive. The relative increase in the evolutionary rate of X-linked loci occurs because the adaptive alleles are sheltered from positive selection on autosomes but are fully expressed in males when X-linked, resulting in a higher fixation probability for X-linked mutations. To date, the dominance of adaptive changes remains an open question (Charlesworth, Coyne, and Barton 1987). Although it is believed that most adaptive changes should be dominant (Haldane 1924, 1927), some adaptive phenotypes in inbreeding plant species have been shown to be recessive (Charlesworth 1992; Bradshaw et al. 1995, 1998), whereas insecticide resitance, an adaptation found in outcrossing insect species, is generally dominant (see Charlesworth, Coyne, and Barton 1987; Orr and Betancourt 2001 and references therein).

The theory of Charlesworth, Coyne, and Barton (1987) is a model in which selection occurs on new adaptive mutations, rather than on standing variation in a population at mutation-selection balance (Orr and Betancourt 2001). The classical model of gene duplications posits that purifying selection is relaxed for a short period after duplication, allowing the accumulation of fixed substitutions in the duplicated genes and that these substitutions can be used for later adaptive evolution, i.e., when environmental conditions change (see for example, Kimura 1983, pp. 104–113). Because the classical model is a neutral model, there should be a corresponding increase in amino acid polymorphism at the duplicate loci, resulting in an accumulation of amino acid variation early in the evolution of duplicate genes. Many of the amino acid changes that accumulate during the period of relaxed negative selection should be partially recessive, in the genetic sense that they create partial loss of function alleles. Positive selection may then later act on the standing variation accumulated at duplicate loci, selecting for these recessive changes. However, if selection acts on standing variation rather than on new mutations, the fixation probabilities of the adaptive mutants are essentially independent of dominance, and X-linked loci will evolve more slowly than autosomal loci, assuming equal degrees of selection (Orr and Betancourt 2001). Thus, the standing variation model would predict that X-linked duplicates diverge more slowly than autosomal duplicates, a pattern opposite to what we observe (table 1 and fig. 1). In the light of the above considerations, we conclude that selection on recessive adaptations is the most general model that accounts for the high average K_a/K_s observed between X-linked duplicates (table 1 and fig. 1).

The hypothesis that gene duplicates in *D. melanogaster* are subject to selection on recessive adaptations has two implications. First, it argues that adaptations at X-linked duplicate loci may be recessive on average, supporting the theory of Charlesworth, Coyne, and Barton (1987). Secondly, it provides new insights into the evolutionary fates of gene duplicates. We argued earlier that our data are more compatible with selection on new mutants, rather than on standing variation. Therefore, we suggest that, given that a duplicate gene does not degenerate into a pseudogene, several adaptive substitutions are required to guarantee its survival. This suggestion is more compatible with adaptive models of gene duplicate evolution (Clark 1994; Walsh 1995) than with the classical model (Kimura 1983) or other neutral models (Force et al. 1999; Lynch and Force 2000).

Recent analyses of polymorphism data in *Drosophila* have found significantly reduced amino acid (Andolfatto 2001) and silent (Begun and Whitley 2001) polymorpshism on the X, relative to autosomes in *D. melanogaster* and *D. simulans*, respectively. Both these patterns are consistent with the ''faster X'' observations seen here for duplicate genes and are also consistent with selection for adaptive recessive substitutions. One cannot, however, rule out a strong effect of deleterious recessives in reducing amino acid polymorphism on the *D. melanogaster* X chromosome (Andolfatto 2001). It is important to note that most of the genes studied by Andolfatto (2001) and Begun and Whitley (2001) are single-copy loci, and it is likely that the dynamics of single-copy genes and duplicates differ substantially. In general, single copy genes should evolve under purifying selection to maintain protein function. Duplicate loci, on the other hand, may in general degenerate quickly into pseudogenes (Haldane 1933; Fisher 1935). The fate of those duplicates that survive degeneration can be resolved by positive selection for improved or new functions or neutral processes such as subfunctionalization.

Conclusions

We have shown that the rate of divergence between X-linked duplicates in *D. melanogaster* is significantly higher than the rate of divergence between other linked and dispersed duplicates. Positive selection on adaptive recessives can explain the observed pattern. However, the degree of faster X evolution we observe does not seem well explained by a simple relaxation model, where an unrealistic assumption of dominant deleterious mutation is invoked. Also, the distributions of silent divergence between duplicates in *D. melanogaster* suggest that many X-linked duplicates may be young, as either the rate of duplication of the X is higher than the au-

tosomes or that there has been a recent burst of duplication events on the *D. melanogaster* X chromosome.

Acknowledgments

This work was supported by an NIH training grant to K.T. and an NSF Grant and a Packard Fellowship to M.L. We thank Dick Hudson for discussions about statistical testing. We also thank Peter Andolfatto, Brian Charlesworth, Eli Stahl, Chung-I Wu, Jerry Coyne, and Thom Nagylaki for helpful discussions. The manuscript also benefited from the comments of an anonymous reviewer.

References

ADAMS, M. D., S. E. CELNIKER, and R. A. HOLT et al. (195 co-authors). 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185–2195.

AKASHI, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in Drosophila DNA. Genetics **139**:1067–1076.

ANDOLFATTO, P. 2001. Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18**:279–290.

BAUER, V. L., and C. F. AQUADRO. 1997. Rates on DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. **14**:1252–1257.

BEGUN, D. J., and P. WHITLEY. 2001. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. PNAS **97**: 5960–5965.

BRADSHAW, H. D., K. G. OTTO, B. E. FREWEN, J. K. MCKAY, and D. W. SCHEMSKE. 1998. Quantitative trait loci affecting differences in floral morphology between two species of monkeyflower (Mimulus). Genetics **149**:367–382.

BRADSHAW, H. D., S. M. WILBERT, K. G. OTTO, and D. W. SCHEMSKE. 1995. Genetic-mapping of floral traits associated with reproductive isolation in monkeyflowers (Mimulus). Nature **376**:762–765.

CHARLESWORTH, B. 1992. Evolutionary rates in partially self-fertilizing species. Am. Nat. **140**:126–148.

CHARLESWORTH, B., J. A. COYNE, and N. H. BARTON. 1987. The relative rates of evolution of sex-chromosomes and autosomes. Am. Nat. **130**:113–146.

CLARK, A. G. 1994. Invasion and maintenance of a gene duplication. Proc. Natl. Acad. Sci. USA **91**:2950–2954.

COMERON, J. M. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41**:1152–1159.

COMERON, J. M., M. KREITMAN, and M. AGUADE. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151**:239–249.

CROW, J. F., and M. J. SIMMONS. 1983. The mutation load in *Drosophila*. Pp. 1–35 *in* M. ASHBURNER, H. L. CARSON, and J. N. THOMPSON JR., eds. The genetics and biology of *Drosophila*, Vol. 3c. Academic Press, London.

CROW, J. F., and R. G. TEMIN. 1964. Evidence for the partial dominance of recessive lethal genes in natural populations of *Drosophila*. Am. Nat. **98**:21–33.

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

FISHER, R. A. 1935. The sheltering of lethals. Am. Nat. **69**: 446–455.

FORCE, A., M. LYNCH, F. B. PICKETT, A. AMORES, Y. L. YAN, and J. POSTLETHWAIT. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**: 1531–1545.

HALDANE, J. B. S. 1924. A mathematical theory of natural and artificial selection, Part I. Trans. Camb. Philos. Soc. **28**:19–41.

———. 1927. A mathematical theory of natural and artificial selection, Part II. Trans. Camb. Philos. Soc. **28**:838–844.

———. 1933. The part played by recurrent mutation in evolution. Am. Nat. **67**:5–19.

HUGHES, A. L., and M. NEI. 1992. Maintenance of MHC polymorphism. Nature **355**:402–403.

IHAKA, R., and R. GENTLEMAN. 1996. R: a language for data analysis and graphics. J. Comput. graphical statistics **5**(3): 299–314.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

LI, W. H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

LONG, M., and K. THORNTON. 2001. Gene duplication and evolution. Science **293**:1551a.

LONG, M. Y., and C. H. LANGLEY. 1993. Natural-selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science **260**:91–95.

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

LYNCH, M., and A. FORCE. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics **154**: 459–473.

MCVEAN, G. A. T., and B. CHARLESWORTH. 1999. A population genetic model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. **74**:145–158.

MUKAI, T., S. I. CHIGUSA, L. E. METTLER, and J. F. CROW. 1972. Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster*. Genetics **72**:335–355.

NURMINSKY, D. I., M. V. NURMINSKAYA, D. DE AGUIAR, and D. L. HARTL. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature **396**:572–575.

ORR, H. A., and A. J. BETANCOURT. 2001. Haldane's sieve and adaptation from the standing genetic variation. Genetics **157**:875–884.

PEARSON, W. R. 1990. Rapid and sensitive sequence comparison with fastp and fasta. Methods Enzymol. **183**:63–98.

PETROV, D. A., E. R. LOZOVSKAYA, and D. L. HARTL. 1996. High intrinsic rate of DNA loss in *Drosophila*. Nature **384**: 346–349.

POWELL, J. R., and E. N. MORIYAMA. 1997. Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA **94**: 7784–7790.

RUBIN, G. M., M. D. YANDELL, and J. R. WORTMAN et al. (50 co-authors). 2000. Comparative genomics of the eukaryotes. Science **287**:2204–2215.

SOKAL, R. R., and F. J. ROHLF. 1995. Biometry. 3rd edition. W. H. Freeman and Company.

STAHL, E. A., G. DWYER, R. MAURICIO, M. KREITMAN, and J. BERGELSON. 1999. Dynamics of disease resistance polymorphism at the Rpm1 locus of *Arabidopsis*. Nature **400**: 667–671.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

WALSH, J. B. 1995. How often do duplicated genes evolve new functions. Genetics **139**:421–428.

WANG, W., J. M. ZHANG, C. ALVAREZ, A. LLOPART, and M. LONG. 2000. The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in *Drosophila melanogaster*. Mol. Biol. Evol. **17**:1294–1301.

WYCKOFF, G. J., W. WANG, and C. I. WU. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature **403**:304–309.

YANG, Z. 2000. PAML: phylogenetic analysis by maximum likelihood. Version 3.0. University College London, London.

ANTONY DEAN, reviewing editor

Accepted February 5, 2002