Excess of Amino Acid Substitutions Relative to Polymorphism Between X-Linked Duplications in *Drosophila melanogaster*

*Kevin Thornton**[1] *and Manyuan Long**

*Department of Ecology and Evolution and the Committee on Genetiecs, University of Chicago

We have obtained sequence polymorphism data from 13 genes belonging to 5 gene families in *Drosophila melanogaster* where the K_a/K_s between copies is greater than 1. Twelve of these 13 loci are X-linked. In general, there is evidence of purifying selection in all families, as inferred both from levels of silent and replacement variation and insertion/deletion variation, suggesting that the loci are likely functional. Shared polymorphisms indicative of gene conversion between paralogs are rare among the X-linked families, in contrast to available data from autosomal duplicates. McDonald-Kreitman tests between duplicates reveal an excess of amino-acid fixations between copies in the X-linked families, suggesting that the divergence between these loci was driven by positive selection. In contrast, available data from autosomal duplicates show a deficit of fixations, consistent with gene conversion being a strong homogenizing force.

Introduction

The evolutionary forces governing the fates of duplicated loci is a topic of persistant interest in evolutionary biology as it bears on the question of how novel functions arise. Generally speaking, most duplicated loci should eventually become nonfunctional pseudogenes as a result of the accumulation of deleterious mutations (Haldane 1933; Fisher 1935; Nei and Roychoudhury 1968). Neutral processes can preserve duplicate loci if functional domains of the loci are independently mutable (Force et al. 1999; Lynch and Force 2000; Lynch et al. 2001). Alternatively, positive selection can preserve duplicates which evolve benefical functions (Clark 1994; Walsh 1995).

Several recent studies have highlighted the importance of genomic location in the molecular evolution of duplicate loci. In *Drosophila melanogaster*, X-linked duplicates have roughly twice the K_a/K_s between copies as autosomal duplicates (Thornton and Long 2002), and there is an excess of autosomal duplicate loci derived by retrotransposition of cDNA from X-linked loci (Betrán et al. 2002), a pattern that is also observed in mammals (Emerson et al. 2004). These observations of X/autosome differences have been interpreted in light of the theoretical result that the dominance of adaptive mutations influences the relative rates of evolution of X-linked and autosomal loci (Charlesworth et al. 1987). Genomic location has also been implicated in affecting rates of synonymous divergence between *amylase* gene copies in *Drosophila* (Zhang and Kishino 2004*a*). Finally, there is a tendency for duplicates in regions of low recombination in the *Saccharomyces cerevisiae* genome to exhibit higher rates of amino acid evolution than their paralogs in regions of high recombination (Zhang and Kishino, 2004*b*). These last two studies implicate an effect of weak selection on the divergence between paralogs.

In this study, we examine nucleotide variation in an African sample of *D. melanogaster* in several families of X-linked duplicates that we have previously identified as having rapidly diverged (Thornton and Long 2002). We are interested in addressing four features of polymorphism and divergence in this class of genes. First, we examine evidence for selective constraint to ask whether patterns of polymorphism are consistent with the duplicates being pseudogenes. Second, we assess the impact of gene conversion as a homogenizing force among these loci. Third, we test the neutral mutation hypothesis by comparing polymorphism and divergence between copies at replacement and synonymous sites (McDonald and Kreitman 1991), in an effort to test the theory that positive selection and gene conversion can have opposite effects on the divergence of paralogs, and that selection may need to be quite strong to overcome the homogenizing force of gene conversion (Innan 2003*b*). Finally, as the synonymous divergence between the loci studied is low, we use the annotation of the *Drosophila pseudoobscura* genome, the most closely related genome to *D. melanogaster* currently available, to ask if there is evidence that some of the duplication events post-dated the divergence of the two species. We find strong evidence of selective constraint in the X-linked gene families, and only weak evidence for conversion between copies. Additionally, there is a significant of excess amino acid fixations between copies relative to polymorphism among the X-linked duplicates in the sample, suggesting that positive natural selection has acted on these loci after duplication, and that selection has been a stronger long-term force than ectopic conversion. The comparison with *D. pseudoobscura* suggests that many of these loci are of relatively recent origin.

We have also compiled all polymorphism data from duplicate loci in *D. melanogaster* that we are aware of, and provided two new data sets from autosomal duplicates, and tested the neutral mutation hypothesis. In contrast to the X-linked loci, there is a deficit of fixations among these autosomal loci, consistent with the expectation of concerted evolution and the theory that strong selection is needed to escape ectopic gene conversion (Innan 2003*b*).

Materials and Methods

Genes Sequenced

We gathered polymorphism data from the subset of duplicate genes identified by Thornton and Long (2002) as having K_a/K_s greater than 1 between copies (table 1).

[1] Present address: Department of Molecular Biology and Genetics, Cornell University 227 Biotechnology Building, Ithaca, NY 14853

Key words: Drosophila, gene duplicates, positive selection, pseudogenes, polymorphism, concerted evolution.

Mol. Biol. Evol. 22(2):273–284. 2005
doi:10.1093/molbev/msi015
Advance Access publication October 13, 2004

Table 1
Duplicate Pairs in *D. melanogaster* with $K_a > K_s$ (Thornton and Long 2002)

Gene 1	Chromosome	Gene 2	Chromosome	K_a	K_s
CG10102	2R	CG12505	2R	0.091	0.08622
CG11461	3R	CG12206	X	1.249	1.029
CG13732	3L	CG15644	X	0.356	0.342
CG15644	X	CG15645	X	0.227	0.175
CG15644	X	CG18620	X	0.011	0.009
CG15645	X	CG18620	X	0.227	0.223
CG18256	X	CG18352	3L	0.055	0.05
CG2532	X	CG2885	X	0.121	0.05
CG2885	X	CG9807	X	0.174	0.082
CG6997	X	CG6999	X	0.311	0.222
CG9123	X	CG12608	X	0.026	0.021
CG11941	X	CG11942	X	0.263	0.236
CG11941	X	CG12700	X	0.084	0.045

Based on previous results (Thornton and Long 2002), we are interested in examining linked duplications, as they are the most common type present in the *D. melanogaster* genome. One striking feature of table 1 is that there is only one pair of linked, autosomal duplicates with K_a/K_s greater than 1, once pairs with $K_s > 1$ are excluded (CG10102 and CG12505). The K_a/K_s ratios for these genes are not significantly greater than 1, suggesting that they may have diverged because they are unconstrained at amino acid sites. We also updated the gene duplication database of Thornton and Long (2002) for Release 3 of the Drosophila Genome Project (Celniker et al. 2003) and confirmed that the results presented there held for the new assembly and annotation. Functional information for these loci, mined from FlyBase (http://www.flybase.org), is summarized in table 2.

In addition to the set of genes with $K_a/K_s > 1$ (12 of 13 of which are on the X chromosome), we are also interested in the evolutionary history of autosomal duplicates. The autosomal pair CG10102/CG12505 (table 1) was originally included in our study, but we were unable to cleanly amplify CG10102 after making attempts with multiple primers and primer combinations. We obtained data from two pairs of tandem autosomal duplicates, which were chosen using a random-number generator from the set of linked, autosomal duplications with the same range of synonymous divergence as the X-linked gene families (table 1). The two pairs chosen were the *amylase* duplication, which has been the subject of molecular evolutionary studies in several *Drosophila* species (Shibata and Yamazaki 1995; Okuyama et al. 1996; Inomata et al. 1997; Araki, Inomata, and Yamazaki 2001; Inomata and Yamazaki 2002), and CG11466/CG17875, which are members of the *Drosophila* cytochrome p450 gene family (Tijet, Helvig, and Feyereisen 2001). The CG17875 locus is missing a canonical heme domain and is believed to be a pseudogene (Tijet, Helvig, and Feyereisen 2001).

We have also have mined all available data on gene duplications that we are aware of. These loci include the members of the *hsp* family (Bettencourt and Feder 2002) and the *attacin* family of immunity peptides (Lazzaro and Clark 2001). We excluded data from both the *esterase* family in *D. melanogaster* (Balakirev Balakirev, and Ayala 2002; Balakirev et al. 2003), because silent sites are saturated between copies, and from the *amylase* duplication (Araki, Inomata and Yamazaki 2001), because we sequenced *amylase* in this study. Data from the alcohol dehydrogenase (*Adh*) region of a Zimbabwe, Africa, sample of *D. melanogaster* encompassing the regions sequenced by Kreitman and Hudson (Kreitman and Hudson, 1991) were kindly provided by P. Andolfatto. As silent sites are saturated between *Adh* and *Adh-r*, we used a sequence from *D. teissieri* (GenBank: X54118) to restrict the analysis to ingroup-specific substitutions. The primary interest in including these data will be the application of McDonald and Kreitman (McDonald and Kreitman, 1991) tests (see below). An analysis of gene conversion among the *attacin* and *hsp* loci is found in Innan (2003*a*).

Drosophila Lines

Isofemale lines were sequenced from 10 lines from Zimbabwe, Africa (5 from Harare, 5 from Sengwa) (Begun and Aquadro 1993), obtained from M.-L. Wu. These lines have been maintained in the lab for over 10 years (greater than 200 generations), and we observed no heterozygous base calls at autosomal loci, allowing direct inference of haplotypes.

PCR, Sequencing, and Assembly

All polymerase chain reaction sequencing was done from genomic DNA extracted from single male flies. Gene-specific primers were designed by aligning the genomic regions and coding sequences (CDS) of gene families. Primer positions were then chosen to cover sites that differed among the different genes in the alignment, and the 3′ base of at least one primer was placed over a base that differentiated the genes in the *Drosophila* genome sequence. A list of all PCR and sequencing primers used is available from the authors.

Polymerase chain reaction conditions were optimized on a gradient thermocycler for each gene pair, and then each line was amplified at the optimum temperature in a 50 μl reaction. The PCR products were cleaned with Qiagen columns (Qiagen, Inc.) and sequenced in both directions using Big Dye 3.0 sequencing kits (PerkinElmer). Sequences were obtained with an ABI 3700 machine. Base calls from the sequencing reactions were done, and contigs were assembled, with the phred/phrap software package (Ewing et al. 1998; Ewing and Green 1998). Contigs were assembled individually for each fly line, and a multiple alignment was assembled using the mace software (W. Gilliland and C. Langley, manuscript in preparation; program available from http://ludwig.ucdavis.edu/mace/), and visualized with consed (Gordon, Abajian, and Green 1998). For assembly using mace, bases with quality scores less than 30 were initially called as missing data, and were only included back into the sequence if the base calls were easily made, with no ambiguity, by eye from chromatographs. After assembly, all polymorphisms were confirmed by visual inspection of chromatographs in consed. Sequences have been deposited in GenBank (accession

Table 2
Functional Annotation of Genes Sequenced

Gene	Synonymns	Flybase ID[a]	Domains	ESTs[b]	Miscellaneous
CG11941	*skpC*	FBgn0026175	*skp1-skp2* dimerization POZ	None	RNA pol II. May be lowly expressed and male-specific[c]
CG11942		FBgn0031074	*skp1-skp2* dimerization POZ	None	RNA pol II. transcription elongation factor
CG12700	*skpD*	FBgn0026174	*skp1-skp2* dimerization POZ	AT18217[d]	May be lowly expressed and male-specific[c]
CG15644[e]	CG18620 CG32584	FBgn0052584	None	None	
CG15645[h]		FBgn0030657	None	RE46906.5prime[d]	(Betrán *et al.* 2002)
CG13732		FBgn0037730	None	RH65158[d]	(Betrán *et al.* 2002)
CG2532	CG32671[f]	FBgn0052671	None	None	RAB small monomeric GTPase
	CG32670[g]	FBgn0052671	None	None	RAB small monomeric GTPase
CG2885		FBgn0030200	Ras GTPase, Rab subfamily P loop-containing nucleotide triphosphate hydrolase	None	RAB small monomeric GTPase
CG6997[h]	CG32708	FBgn0052708	RNP-1 (RNA rec. motif)	None	(Lasko 2000)
CG6999		FBgn0030085	RNP-1	None	(Lasko 2000)
CG9123		FBgn0030629	WD40 repeat	RE62683.5prime[d]	
CG12608[h]		FBgn0030630	WD40 repeat	LD30439[d]	
Amy(p)	CG18730	FBgn0000079	α-amylase	GH10266[d]	Calcium-ion binding
Amy(d)	CG17876	FBgn0000078	α-amylase	RH48856[d]	Calcium-ion binding
CG11466[h]	*Cyp9f2*	FBgn0038037	Cytochrome P450	GH26976[d]	Cytochrome P450
CG17875	CR17875 *Cyp9f3ψ*	FBgn0038034		RE06354[d]	Pseudogene

NOTE.—Horizontal lines separate gene families.

[a] http://flybase.bio.indiana.edu.

[b] Not all expressed sequence tags (ESTs) are listed. ESTs from the Drosophila Gene Collection (DGC) collection are preferentially listed as they are the results of high-quality sequencing efforts. The first two letters of the EST names designate the source of the cDNA library: AT (adult testes), GH (adult head), LD (mixed stage embryos, 0–22 hours), RE ("Riken embryo," mixed stage embryos, 0–22 hours), and RH ("Riken head," adult heads).

[c] T. Murphy, C. Kennedy, and G. Karpen. (1999) Abstract submitted to the 40*th Drosophila* Research Conference, Flybase Reference FBrf0106993.

[d] EST is from the DGC collection.

[e] CG15644 and CG18620 were merged into one gene, CG32584, between Release 2 and Release 3. Our sequence data, however, support the existence of two separate loci.

[f] Referred to as CG2532 (3′ exon) in this study, and is paralagous to CG2885.

[g] Referred to as CG2532 (5′ exon) in this study.

[h] Locus is present in the "freeze1" assembly of the *D. pseudoobscura* genome.

numbers AY752495–AY752639 and AY754313–AY754331), and alignments are available from K.T. upon request.

Analysis

Levels of Variation

Levels of nucleotide polymorphism were summarized using two statistics—Watterson's (1975) $\hat{\theta}_W$, which depends on the number of mutations in the data, and $\hat{\theta}_\pi$ (Tajima 1983), which depends on the mean number of pairwise differences in the sample. Under a neutral model, both values are unbiased estimators of $\theta = 4N_e\mu$, the population mutation rate. For these calculations, sites in the alignments with gaps were excluded. Sites with missing data were included by adjusting the sample size at each site and summing the value of the statistic across sites. For $\hat{\theta}_w$, sites with more than two alleles segregating were counted as $k - 1$ segregating sites for a site with k alleles. In other words, we used the inferred number of mutations at that site. There were only two sites in the data with more than two alleles (in genes CG13732 and CG15644 from the Zimbabwe sample). Under the assumption that males and females have the same effective population size (N_e), the N_e of an X-linked locus is $\frac{3}{4}$ that of an autosomal locus. Thus, to make parameter estimates comparable between chromosomes, X-linked values have been multiplied by $\frac{4}{3}$. We note that the observation of variation in male reproductive success in natural population (Charlesworth 2001) implies that $\frac{4}{3}$ may not be the appropriate correction. However, as we do not directly compare X and autosomal levels of variation in this study, our conclusions remain unaffected.

Silent and Replacement Polymorphism

The mean number of silent and replacement sites in the alignment were estimated using the method of Comeron (Comeron, 1995). Sites with more than two states segregating were excluded, which led to the exclusion of one single nucleotide polymorphism (SNP) in Zimbabwe, out of

259 total polymorphic sites. Partial codons at the end of the alignments were excluded because one does not know if the missing sites are variable or not. This led to the exclusion of one polymorphism at CG2885 in Zimbabwe. $\hat{\theta}_w$ and $\hat{\theta}_\pi$ were calculated for replacement sites, synonymous sites, noncoding sites, and noncoding plus synonymous sites. Noncoding sites do not distinguish 5′, 3′, or intron sequence because we do not always have data from each site class for each gene. All diversity estimates for X-linked loci are multipied by $\frac{4}{3}$ for comparison with the autosomes.

Inference of Ectopic Gene Conversion

When polymorphism data are obtained from duplicate loci, gene conversion between copies (ectopic gene conversion) is visually detectable if polymorphisms are shared between the loci. However, the absence of shared polymorphisms does not rule out an important role for gene conversion, because not all gene conversion events lead to shared polymorphisms (on singleton lineages, for example). We are therefore interested both in identifying conversion events and in estimating the rate of conversion between paralogs.

To detect ectopic gene conversion, we tallied the number of shared polymorphisms between pairs of duplicates. Alignments for these analyses were done using CLUSTALW (Thompson, Higgins, and Gibson 1994) and edited by eye to remove highly divergent regions with questionable alignments when necessary (all such regions were at the 5′ or 3′ ends of alignments).

In addition to counting shared polymorphisms, we also used the GENECONV program (Sawyer 1999), which implements extensions of Sawyer's (1989) method for detecting gene conversion. In our analysis, we ignore results on within-locus fragments, as GENECONV has been shown by simulation to be reasonably powerful in detecting reciprocal exchange events (i.e., classical crossing over with no gene conversion) (Posada and Crandall 2001), meaning it is likely that the analysis would confound within-locus conversion events with crossing over. For all analyses with GENECONV, we used the default settings, and applied the program to nucleotide alignments. GENECONV reports a P value associated with every conversion event detected. For simplicity, we report P values only for the most likely (least significant) fragment. The P values reported are from a permutation procedure which implicitly corrects for multiple comparisons (Sawyer 1989, 1999).

To estimate the population rate of ectopic gene conversion, we used the moment estimators of Innan (2003*a*), based on a simplified coalescent process for gene families of size 2. Recombination occurs between loci, not within them; ectopic conversion is assumed to be intrachromosomal; and there is no gene conversion between alleles within loci. A simplifying assumption is made that the tract length of conversion events is small enough that only one site is converted per event. The parameters estimated are $\hat{\theta}(=4N_e\mu)$, the population mutation rate, $\hat{R}(=4N_e r)$, the population recombination rate between the loci, and $\hat{C}(=4N_e c)$, the population rate of ectopic conversion. The expressions for the estimators are:

$$\hat{\theta} = \frac{\pi_w + 2D_{sum}}{2}, \tag{1}$$

$$\hat{C} = \frac{\pi_w - 2D_{sum}}{2(\pi_b - \pi_w)}, \tag{2}$$

and,

$$\hat{R} = \frac{\pi_w^2 + 4D_{sum}^2 - 4\pi_b D_{sum}}{2(\pi_b - \pi_w)D_{sum}}. \tag{3}$$

In these expressions, π_w is the average of the nucleotide diversity for the two loci, π_b is the mean number of pairwise differences between loci (only counted between chromosomes drawn from different individuals), and D_{sum} is a function of the number of shared polymorphisms in the alignment. Equations (1) through (3) correspond to equations 11 through 13 of Innan (2003*a*).

It can be seen from equation (2) that $\hat{C}$ can be greater than zero even when $D_{sum} = 0$ (which occurs when there are no shared polymorphisms). These estimators only return sensible values under certain conditions. Equation (2) will return negative values if the mean within-locus diversity is greater than the between locus diversity ($\pi_w > \pi_b$) and $D_{sum} = 0$. Also, the recombination rate between the loci is undefined when $D_{sum} = 0$ (eq. 3). These problems pose some practical issues both for data analysis and for investigation of the properties of the estimator by simulation.

A program was written to calculate the estimators in equations (1) through (3) from an alignment of polymorphism data from two duplicates. Gapped positions in alignments were not analyzed, and sites with more than two states in the alignment were excluded. In our program, when the sample sizes differ between the two loci, the actual number of comparisons between loci is kept track of when calculating π_b and is used instead of $n(n-1)$. Estimates for X-linked loci are scaled by $\frac{4}{3}$ to allow comparison with autosomes.

Divergence Between Duplicates

A variant of McDonald-Kreitman-style contingency tables (McDonald and Kreitman 1991) (the "MK" test) was used to test the null hypothesis that divergence between duplicate loci is a neutral process. In our application, the MK test was performed between alignments of polymorphism data from pairs of duplicated loci. For this test, we only considered changes at replacement and synonymous sites (i.e., intron and untranslated region (UTR) sites were not considered).

To perform the MK test, coding regions were extracted from the alignment, and coding regions of duplicate pairs were aligned using CLUSTALW (Thompson, Higgins, and Gibson 1994). To ensure accurate alignment, we made use of alignments of the coding sequences (CDS) from the genome we had previously generated, using peptide alignments as guides (Thornton and Long 2002). The alignments required minimal manual editing, with the exception of the removal of highly divergent regions from the 5′ end of some alignments (particularly for the CG15644/CG15645/CG13732/CG18620 family). This removal is conservative with respect to testing for an excess of amino acid

replacements in these data, as it reduces the number of amino acid fixations in the alignment with little effect on the number of polymorphisms.

In our application of the MK test, we analyze either the most closely related pair of genes, or the pairs for which we know the ancestral/derived relationship of the duplicates (CG15645/CG13732, Betrán, Thornton, and Long. [2002]). The exception to these conditions is the inclusion of CG11941/CG11942, to ensure that every gene sequenced appears at least once in the analysis. There are two reasons for restricting the number of comparisons in this fashion. First, for families of size greater than two, we do not know the phylogeny of the gene family, which would result in fixed differences being counted multiple times when all possible comparisons are made. Second, when analyzing all but the most closely related pairs, the number of fixed amino acid differences increases dramatically, meaning that removing such pairs is conservative for our purpose.

Gene Discovery in *Drosophila pseudoobscura*

The pairwise K_s between all X-linked duplicates in this study falls in the range of 0.01–0.25 (table 1), the upper value of which is slightly larger than the mean value of 0.18 between *D. melanogaster* and *D. yakuba* (Takano 1998). We are therefore interested in whether there is evidence that the duplication events leading to these loci are relatively recent. Whole-genome sequence data are now avaliable for *D. pseudoobscura*. We made use of the *D. melanogaster*/*D. pseudoobscura* orthology assignments generated by FlyBase (http://www.flybase.org) (Brian Bettencourt, personal communication) based on the freeze1 assembly available at Baylor College of Medicine (http://www.hgsc.bcm.tmc.edu/projects/drosophila/).

Software Availability

All analyses of polymorphism data were performed using custom software. All SNP analysis programs were implemented in C++, based on a common library (Thornton 2003). The program to extract CDS from GenBank files was written in perl using routines from the bioperl libraries (http://www.bioperl.org). Source code for all programs is available from K.T.

Results

Silent and Replacement Variation

The majority of the loci considered in this study are gene models predicted during the annotation of the *Drosophila* genome (Adams et al. 2000), and because the K_a/K_s values are not significantly greater than 1, suggests that it is possible that several of the predicted loci are actually pseudogenes rather than functional loci. It should be noted that pseudogenes are rare in *D. melanogaster* (Misra et al. 2003), likely owing to their rapid elimination from the genome by accumulating deletions (Petrov, Losovskaya, and Hartl 1996; Petrov et al., 1998). Although duplicate loci in *D. melanogaster* generally show evidence of selective constraint, as inferred from divergence between copies (Thornton and Long 2002), the synonymous divergence between some of the pairs studied here is less that 0.10, roughly the mean divergence at synonymous sites between *D. melanogaster* and *D. simulans* (i.e., Table 1 of Takano [1998]). These genes may be destined for future loss in the absence of ongoing purifying selection.

The null hypothesis that these genes are pseudogenes leads to testable predictions. First, insertion/deletion (indels) polymorphism should be very informative, and indels should have lengths that are multiples of 3 and in-frame within the coding sequence of functional genes, whereas pseudogenes (or null-alleles) are more likely to bear indels that result in frame-shift mutations. Second, if these genes are pseudogenes, then mutations at silent and replacement sites should occur with equal frequency per site, leading to similar levels of variation observed in the two site classes, i.e., $\hat{\theta}_{\pi,N} \approx \hat{\theta}_{\pi,S}$, where the subscripts N and S refer to replacement and synonymous sites, respectively. We focus the analyis on levels of diversity ($\hat{\theta}_{\pi}$), rather than on Watterson's $\hat{\theta}_W$, because the former is affected by the frequency of segregating variation, which is more informative for our purposes than simply counting the total number of variants. A summary of the total polymorphism data from each locus is presented in table 3.

The deletions that are observed in coding and noncoding regions are listed in table 4. There are many fewer deletions observed in the coding regions than in the noncoding regions. Assuming that each indel results from a unique mutational event, we can calculate the per-site heterozygosity in the sample from indels ($\hat{\theta}_{\pi,indel}$). For noncoding regions, the mean $\hat{\theta}_{\pi,indel}$ is 0.0025. In coding regions, $\hat{\theta}_{\pi,indel}$ is 0.00079, showing that there is an order of magnitude less sample diversity from indels in the coding regions compared to noncoding regions. Deletions in coding regions are generally in-frame and multiples of 3 in length. The exceptions are deletions in CG17875, which is believed to be a pseudogene (Tijet, Helvig, and Feyereisen 2001), and one allele of CG6997 in the sample.

Many of the indels in noncoding regions are found in CG13732 and are located in the remnants of the gene's poly-A tail. CG13732 arose by retroposition of the coding portion of CG15645, which is X-linked, to 3L, and is expressed in the testis of adult males (Betrán, Thornton, and Long 2002). Taken together, the data in table 4 suggest that the (predicted) exon regions of these loci are under more constraint with regard to length variation than the noncoding regions.

Levels of variation at replacement and silent sites are presented in table 5. We are particularly interested in levels of silent versus replacement variation in the set of duplicates with high K_a/K_s. Considering only loci from that set for which polymorphisms are observed in the exons, $\hat{\theta}_{\pi,N}$ is less than $\hat{\theta}_{\pi,S}$ 11/13 loci ($P = 0.02$, sign test). This result indicates that there is purifying selection against aminoacid replacement mutations, further suggesting functional constraint on these genes.

CG17875 is believed to be a cytochrome p450 pseudogene (Tijet, Helvig, and Feyereisen 2001). In our sample, one of the alleles bears a 5 bp insertion (relative to CG11466), and another a 9 bp deletion. The insertion allele disrupts the coding sequence, and translating the data set reveals one polymorphic stop codon. Thus, this

Table 3
Summary of Variation in Zimbabwe

Gene	Chrom.	*n*	Length	Length (no gaps)	*S*	No. Singletons	No. Mutations	$\hat{\theta}_W$	$\hat{\theta}_\pi$
CG11941	X	10	299	299	4	2	4	0.006	0.005
CG11942	X	7	851	851	17	11	17	0.011	0.09
CG12700 (5′)	X	9	602	584	15	10	15	0.013	0.011
CG12700 (coding)	X	10	258	258	2	2	2	0.004	0.002
CG12608	X	9	1392	1392	7	3	7	0.001	0.002
CG9123	X	9	819	819	9	5	9	0.005	0.005
CG15644	X	10	1085	1068	20	15	21	0.010	0.008
CG15645	X	10	1140	1123	22	16	22	0.009	0.007
CG13732	3L	9	1369	1333	28	21	29	0.015	0.006
CG18620	X	8	1204	1193	14	11	14	0.006	0.005
CG2532 (5′ exon)	X	10	526	526	7	3	7	0.006	0.005
CG2532 (3′ exon)	X	10	482	482	3	2	3	0.003	0.002
CG2885	X	10	626	626	19	12	19	0.014	0.011
CG6997	X	9	561	549	6	5	6	0.005	0.004
CG6999	X	8	821	821	7	7	7	0.005	0.003
CG11466	3R	9	872	872	11	6	11	0.005	0.004
CG17875	3R	9	799	778	20	11	20	0.018	0.008
Amy(p)	2R	9	696	696	10	3	10	0.010	0.006
Amy(d)	2R	9	891	891	38	14	38	0.016	0.016

(pseudo)gene appears to be segregating several potentially functional alleles. We show below that gene conversion is occurring between CG17875 and its functional paralog, CG11466.

Detection and Rates of Ectopic Gene Conversion

A common method employed to detect conversion between paralogs is to identify shared polymorphisms between copies. However, an absence of shared polymorphisms does not imply that the rate of ectopic conversion is zero. For example, a mutation on a branch leading to a fixed difference between copies can be converted to the ancestral state on a singleton lineage. Thus, there is a discrepancy between the values of the summary statistic and the value of the parameter of interest. A similar issue arises in inferring recombination events at a single locus—the number of pairs of sites showing all four gametes (Hudson and Kaplan 1985) may be zero, but one can still estimate non-zero values of the population recombination rate by using more of the information present in the data, such as the amount of pairwise linkage disequilibrium in the sample (i.e., Hudson [2001]). We will therefore be interested in the summary of the data, the number of shared polymorphisms, and whether or not estimates of the rate of ectopic conversion are compatible with 0.

Table 4
Insertion and Deletion Polymorphism in Coding and Noncoding regions

	Coding			Noncoding		
Gene	Position	Length	Frequency	Start	Length	Frequency
CG12700				143	1	1/9
				182	2	2/9
				183	13	6/9
				428	2	1/9
				429	1	1/9
CG13732				47	2	1/9
				48	1	1/9
				255	12	2/9
				312	1	1/9
				1105	9	3/9
				1110	2	1/9
				1111	1	4/9
				1119	12	3/9
CG15645				2	13	1/10
				199	4	1/10
CG18620				156	2	1/10
				699	9	1/10
CG15644	806	3	7/10	418	5	1/10
CG6997	54	12	1/9			
CG17875	75	5	7/9	457	7	2/9
	70	14	1/9			

When only SNPs are considered, there are very few shared polymorphisms in the entire data (table 6). The exceptions to this pattern are the autosomal duplicates (CG11466/CG17875 and *Amy(p)*/*Amy(d)*) and two X-linked pairs (CG18620/CG15644 and CG2532 (3′ exon)/CG2885). The general observation for X/X duplications in these data is that the number of fixed differences between loci is much larger than the number of shared and private polymorphisms. The numbers of shared polymorphisms per site were 0.001 and 0.009, for X-linked and autosomal gene families, respectively. Note that there are shared polymorphisms between CG17875, a putative pseudogene (Tijet, Helvig, and Feyereisen 2001), and CG11466 (table 6), which is believed to be functional.

We also used GENECONV (Sawyer 1989, 1999) to detect tracts of similar sequence between duplicates. When used on the sequence data analyzed in table 6, GENECONV only detected significant fragments between loci in four cases: a fragment of length 10 between CG6999 and CG6997 ($P = 0.002$ for the least significant fragment), several fragments ranging from 82 bp to 167 bp in length between CG2532 (3′ exon) and CG2885 ($P = 0.037$ for the least significant fragment), several fragments 300–420

Table 5
Silent and Replacement Variation in Zimbabwe

	Replacement				Synonymous				Noncoding				Synonymous + noncoding			
Gene	S	L	$\hat{\theta}_W$	$\hat{\theta}_\pi$	S	L	$\hat{\theta}_W$	$\hat{\theta}_\pi$	S	L	$\hat{\theta}_W$	$\hat{\theta}_\pi$	S	L	$\hat{\theta}_W$	$\hat{\theta}_\pi$
CG11941	3	239.55	0.006	0.006	1	57.44	0.008	0.005	0	0	0.000	0.000	1	57.44	0.008	0.005
CG11942	2	394.26	0.003	0.003	4	106.73	0.020	0.017	11	350	0.017	0.015	15	456.73	0.018	0.015
CG12700	2	261.76	0.004	0.002	0	64.97	0.000	0.000	15	473	0.016	0.013	15	537.97	0.014	0.012
CG12608	0	788.38	0.000	0.000	4	216.61	0.009	0.010	3	389	0.004	0.003	7	605.61	0.006	0.006
CG9123	4	618.25	0.003	0.003	5	170.07	0.014	0.012	0	30	0.000	0.000	5	200.07	0.012	0.010
CG15644	0	375.22	0.000	0.000	5	90.17	0.026	0.029	14	601	0.012	0.009	19	691.17	0.013	0.011
CG15645	1	349.65	0.001	0.001	3	85.34	0.017	0.009	18	687	0.012	0.009	21	772.34	0.013	0.009
CG13732	8	565.03	0.005	0.004	5	130.96	0.014	0.013	14	634	0.008	0.006	19	764.96	0.009	0.007
CG18620	1	419.33	0.001	0.001	3	97.73	0.016	0.010	10	668	0.008	0.007	13	765.73	0.009	0.007
CG2532 (5′ exon)	3	395.75	0.004	0.004	3	109.44	0.013	0.009	0	2	0.000	0.000	3	111.44	0.013	0.009
CG2532 (3′ exon)	1	386.11	0.001	0.001	2	93.88	0.010	0.009	0	0	0.000	0.000	2	93.88	0.010	0.009
CG2885	14	458.33	0.014	0.011	3	120.67	0.012	0.008	2	41	0.023	0.025	5	161.67	0.015	0.012
CG6997[a]	2	334.86	0.003	0.002	1	88.13	0.006	0.004	2	139	0.007	0.007	3	227.13	0.007	0.005
CG6999	5	528.89	0.005	0.003	2	127.89	0.008	0.006	0	122	0.000	0.000	2	249.89	0.004	0.003
CG11466	4	630.63	0.002	0.002	6	155.52	0.014	0.013	1	61	0.006	0.004	7	216.52	0.012	0.010
CG17875[b]	11	581.45	0.007	0.006	6	138.21	0.016	0.014	3	61	0.018	0.016	9	199.21	0.017	0.015
Amy (p)	1	543.29	0.001	0.001	9	152.03	0.022	0.023	0	0	0.000	0.000	9	152.03	0.022	0.023
Amy (d)	8	665.69	0.005	0.004	30	190.80	0.058	0.059	0	0	0.000	0.000	30	190.80	0.058	0.059

[a] Analysis excludes one allele with an out-of-frame 12-bp deletion that disrupts the coding region and eliminates the predicted start codon.
[b] Excludes one allele with a 5-bp insertion.

bp in length between the *amylase* genes in ($P = 0.0316$ for the least significant fragment), and several fragments ≈ 320 bp in length between CG11466 and CG17875 ($P = 0.0405$ for the least significant fragment).

We have also estimated the parameters of Innan's model (Innan 2003*a*) (table 7). The numbers of sites compared for this analysis are given in table 6. The parameters estimated are $\hat{\theta} = 4N\mu$, the population mutation rate; $\hat{C} = 4Nc$, the population rate of ectopic conversion; and $\hat{R} = 4Nr$, the population rate of recombination between the two loci (the model assumes no intralocus recombination of any sort). The per-locus estimates of $\hat{C}$ are generally an order of magnitude lower than $\hat{\theta}$ (table 7) in our sample. Estimating $\hat{R}$ usually returned negative values, because D_{sum} was often zero or negative (table 7, Eq. 3).

Table 6
Shared and Private Polymorphisms Between Duplicates

Gene 1	Gene 2	No. Sites	Shared	Fixations	Private
CG11941	CG11942	283	0	60	8
CG11941	CG12700	270	0	15	4
CG11942	CG12700	594	0	95	13
CG12608	CG9123	819	0	18	13
CG15644	CG13732	384	0	73	8
CG15644	CG15645	372	0	61	7
CG15644	CG18620	622	5	1	13
CG15645	CG13732	616	0	90	19
CG15645	CG18620	372	0	60	6
CG2532 (3′ exon)	CG2885	473	1	44	16
CG6997	CG6999	285	0	53	9
CG11466	CG17875	751	4	2	21
Amy(p)	*Amy(d)*	696	9	0	28

Many of the X-linked duplicate pairs in our sample show no shared polymorphisms (table 6) and the values of $\hat{C}$ are low (table 7). As noted above, it is possible for there to be no shared polymorphisms in the sample, but the data may not be compatible with a model where $C = 0$, raising the possibility that the moment estimator is unlikely to be useful in discriminating conversion models from no-conversion models. To address this, we ran coalescent simulations of a gene family of size 2 with no conversion over a range of

Table 7
Estimates of Ectopic Gene Conversion Parameters

Gene 1	Gene 2	π_w	π_b	D_{sum}	$\hat{\theta}$	$\hat{C}$[d]	$\hat{R}$
11941	11942	2.052	61.921	0	1.026	0.017	NA[b]
11941	12700	0.963	17.580	0	0.482	0.029[a]	NA[b]
11942	12700	3.29	97.3	0	1.647	0.017	NA[b]
9123	12608	3.112	24.562	0	2.574	0.073[a]	NA[b]
15644	13732	1.855	77.198	0	0.923	0.012	NA[b]
15644	15645	1.639	62.7	0	0.819	0.013	NA[b]
15644	18620	4.429	6.028	0.361	2.576	1.159[a]	1.460
15645	18620	1.033	61.417	0	0.517	0.009	NA[b]
15645	13732	2.639	96.617	0	1.319	0.014	NA[b]
2532 (3′ exon)	2885	3.319	52.644	0.0222	1.682	0.033[a]	0.681
6997	6999	1.783	56.734	0	0.892	0.016	NA[b]
11466	17875	4.639	12.3288	0.288	2.607	0.264[a]	1.732
Amy(d)	*Amy(p)*	8.583	11.096	0.342	4.634	1.572[a]	34.251
Attacin-A[a]	*Attacin-B*	8.93	31.41	−0.03	4.47	0.20[a,d]	NA[b]
hsp70 Aa[b]	*hsp70 Ab*	6.41	6.38	1.69	4.90	NA[c]	NA

[a] Indicates estimate of C is incompatible with $C = 0$ at the 2.5% level (see text for details).
[b] Value undefined, $D_{sum} = 0$.
[c] Value undefined, see Innan (Innan, 2003*a*) for details.
[d] From Innan (Innan, 2003*a*).

mutation parameters similar to those in table 7, which were estimated from our data (data not shown). To be conservative with regard to rejecting a model with $C = 0$, for each pair, we asked if $\hat{C}$ is greater than the largest value for the 97.5th percentile value obtained from all simulations, which was the case for 6 of the 13 values estimated (4/11 of the X-linked pairs and 2/2 autosomal pairs, labeled in table 7). The gene pairs with shared polymorphisms (table 6) exhibit higher values of $\hat{C}$ than pairs with no shared polymorphisms, as expected (table 7).

Divergence Between Duplications

To study the evolutionary forces underlying the degree of divergence between these duplicate loci, we conducted McDonald-Kreitman tests (McDonald and Kreitman 1991) on alignments of polymorphism data from the coding regions of the genes we sequenced. First, however, we explored the effect that gene conversion between paralogs has on rejecting the neutral model using the MK test.

We simulated contingency tables for the MK test under the standard neutral model using a modification of Innan's (2003*a*) program. The purpose of this model was to examine the distribution of P values for the test under the null model when applied to data sampled from duplicate loci. Genealogies were simulated according to Innan (2003*a*). Replacement mutations were placed on the tree with rate θ_R, conditional on the length of the tree, and synonymous changes occur at rate θ_S. The simulation output was parsed into the cell entries for the MK test, and P values were calculated as two-tailed Fisher's exact tests in R (R Development Core Team 2004). The distribution of P values for four different rates of gene conversion ($C = 4N_ec = 0.01$, 0.1, 0.5, and 1) were obtained in this fashion from 10^4 simulated histories. Figure 1 shows results for the case where $n = 10$ alleles sampled from each locus, $\theta = 4N_eu = 5$ at synonymous sites, $\theta = 2.5$ at replacement sites, and there is no recombination between loci. Figure 1 reports two summaries for each value of $4N_ec$. First is the fraction of times the MK test was applicable (i.e., at least one fixation was present in the simulated sample). We report this because, in practice, an MK test would not be applied if no fixations were observed. Second, we report the rejection rate of the test, *conditional on the test being applicable*. In other words, the first quantity is the probability of observing any fixations at all, given the parameters, and the second is the probability of rejecting the neutral mutation hypothesis at $\alpha = 0.05$, given the observation of at least one fixation.

The important feature of figure 1 is that, as the rate of conversion increases, the distribution of P values becomes highly right-skewed, due to a lack of fixations between copies. Biologically, this occurs because branches that would lead to fixations between copies in the case of no gene conversion lead to shared polymorphisms as the conversion rate increases. In addition, the probability that the MK test can be applied becomes small as $4N_ec$ increases (down to 10% in figure 1). From the point of view of hypothesis testing (and rejecting the constant-rate neutral model), the deficit of fixations results in very few tests having P values less than $\alpha = 0.05$. For practical purposes, it is clear from figure 1 that there will be very little power to reject the neutral mutation hypothesis when conversion is strong, consistent with Innan's (2003*b*) result that, in the face of gene conversion homogenizing sequences, selection has to be strong enough to allow the accumulation of fixations. In our simulations, we found that increasing the rate of recombination between the two loci gives results very similar to figure 1 (data not shown).

Our sample of X-linked gene duplicates shows an excess of amino acid fixations in 6 out of 7 comparisons (table 8). The number of fixations in table 8 differs from that in table 6 because the former only considers coding regions while the latter considers the entire alignment of the duplicated region and does not distinguish amino acid from synonymous/noncoding fixations. One comparison (CG9123/CG12608) is significant at the $\alpha = 5\%$ level, but not at $\alpha = 0.0033$, which represents an experiment-wide Bonferonni correction for applying 15 tests (table 8). When all X-linked comparisons are pooled, there is a significant excess of amino acid fixations after correcting for multiple tests. The pooled analysis remains significant, even after removal of the most significant comparison (CG9123/CG12608), suggesting that the observation is a general feature of these loci rather than a property of one extreme duplicate pair.

Table 8 also show results of MK tests for the two autosomal duplicate pairs we sequenced, plus all the available data from the literature that we are aware of. In contrast to the X-linked duplications, the autosomal loci show a deficit of fixations, in particular amino acid fixations.

Data Mining in *D. pseudoobscura*

The duplicates that we sequenced fall in a K_s range of 0.01 to 0.25 (table 1), which is much less than the median K_s between *D. melanogaster* and *D. pseudoobscura* (1.79, S. Richards, K. Thornton, A. Clark, R. Nielsen, unpublished data), suggesting that several of these duplicates may have arisen along the *D. melanogaster* lineage. The six loci (out of 17 total) that are found in the assembly of the *D. pseudoobscura* genome are labeled in table 2. In comparison with our database of gene duplicates in *D. melanogaster*, the duplicate pairs sequenced in this study appear to be young tips of gene familes conserved between the two species.

Discussion

Selective Constraint

Several lines of evidence point to the action of purifying selection on duplicated loci in *D. melanogaster*. Genome-wide studies of divergence between duplicated loci have shown that $K_a/K_s < 1$ for most duplicate pairs in eukaryotic genomes (Lynch and Conery 2000; Kondrashov et al. 2002), and K_a/K_s is usually less than 0.5 between duplicates in *D. melanogaster*, indicating strong selective constraint (Thornton and Long 2002). A pertinent question here is whether X/X duplications in *D. melanogaster*, which tend to be relatively highly diverged at the

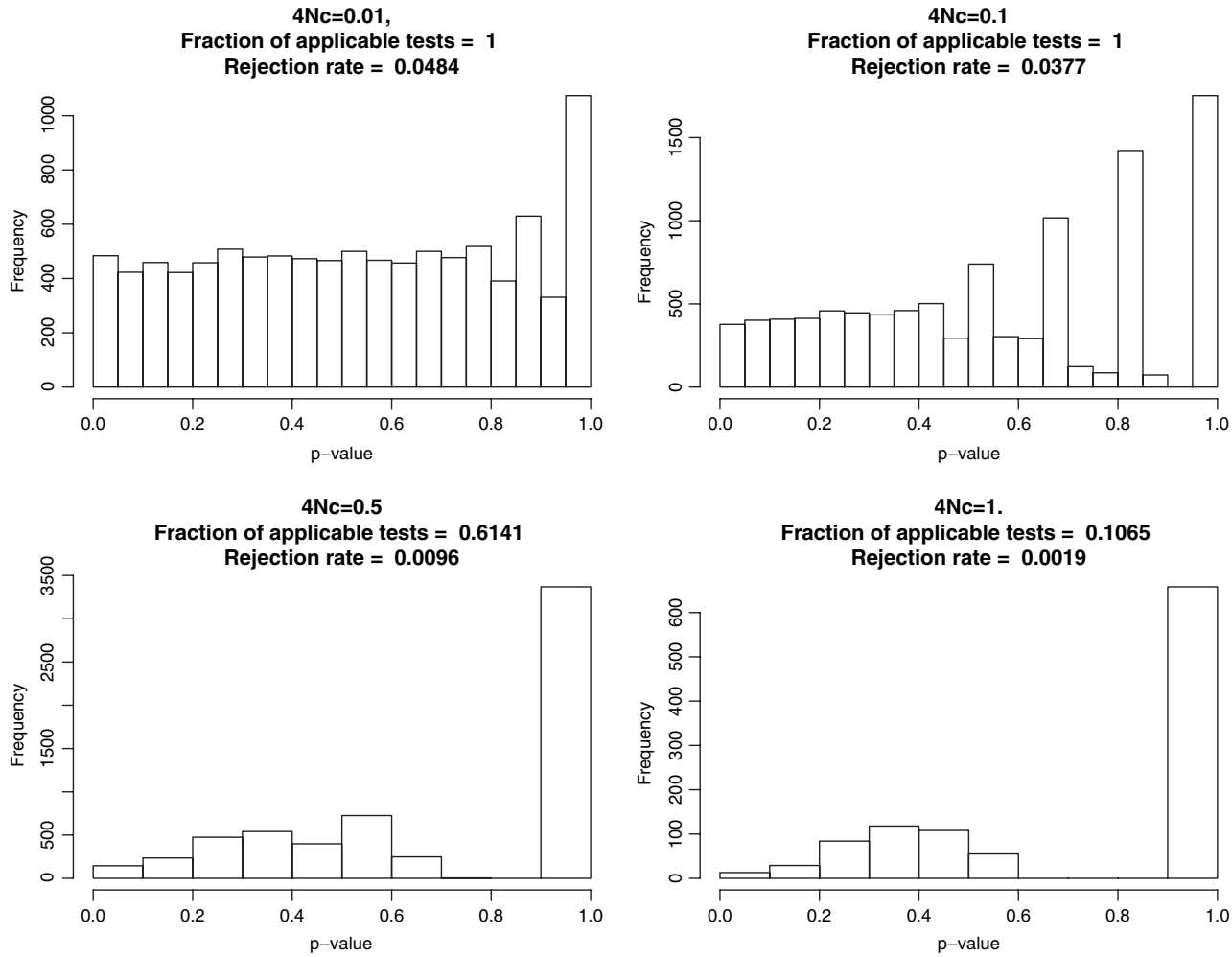

FIG. 1.—Simulated distribution of *P* values for the McDonald-Kreitman test between duplicate loci.

amino acid level (Thornton and Long 2002) show evidence of relaxation of selective constraint and/or are pseudogenes. The current polymorphism data, however, provide three lines of evidence that these X/X duplications are currently evolving under selective constraint. First, levels of silent and replacement variation fall well within values reported for single-copy genes in African populations of *D. melanogaster* (Tables 3 and 4 of Andolfatto [2001]). Second, there are fewer insertion deletion mutations in coding than in noncoding regions (table 4). Third, diversity from amino acid variation is much less than synonymous variation (table 5).

There are, however, signs that some duplicate genes are less constrained than others. Length variation that disrupts open reading frames is observed in CG6999 and CG17875, and a polymorphic stop codon was found in CG17875. Of these three loci, only CG6999 is a member of the set of highly diverged X/X duplications, and CG17875 is believed to be a pseudogene (Tijet, Helvig, and Feyereisen 2001). If we assume that the pattern of polymorphism at CG17875 is typical of a pseudogene, then the null allele in CG6999 in Zimbabwe may lead us to classify CG6999 as a potential pseudogene. However, many of the alleles of CG17875 are intact, reinforcing the

Table 8
McDonald-Kreitman Tables for Duplicate Loci

Class	Gene 1	Gene 2	Fixed N	Fixed S	Poly N	Poly S	*P*
High K_a/K_s	CG11941	CG11942	51	16	5	4	0.231
	CG11941	CG12700	13	2	4	1	1
	CG12608	CG9123	29	9	4	9	0.006
	CG15644	CG18620	0	0	1	5	NA[a]
	CG15645	CG13732	36	17	4	6	0.150
	CG2532 (3′ exon)	CG2885	36	7	13	4	0.712
	CG6997	CG6999	46	16	8	3	1
Pooled			211	67	39	32	0.001
Autosomal	CG11466	CG17875	0	1	11	10	1
	Amy(p)	*Amy(d)*	0	0	5	26	NA[a]
	hsp70Aa	*hsp70Ab*	0	0	11	35	NA[a]
	hsp70Ba	*hsp70Bb*	0	0	15	35	NA[a]
	attacin A	*attacin B*	6	10	14	47	0.336
	Adh	*Adh-r*	14	46	4	14	1
Pooled			20	56	44	131	0.875

[a] Contingency table not testable when a row is filled with zero counts

point that one should be cautious in declaring *Drosophila* genes as pseudogenes, as loci originally believed to be pseudogenes are sometimes later shown to have patterns of molecular evolution inconsistent with the expectation for a pseudogene (Long and Langley 1993; Begun 1997). In addition, *Amy(d)* is a duplicate gene known to be present in many *Drosophila* species (Aquadro et al. 1991; Inomata and Yamazaki 2002), and it is less constrained than *Amy(p)* (table 5). Araki, Inomata, and Yamazaki (2001) observed several deletions leading to null alleles in a Kenyan sample (but not in Japan). Segregating null alleles have also been observed in the *D. melanogaster esterase* gene family (Balakirev Balakirev, and Ayala 2002; Balakirev et al. 2003) and at *Attacin A* (Lazzaro and Clark 2001) The observation that most of the lesions and stop codons observed were found in autosmal duplicates (table 4; Araki Inomata, and Yamazaki [2001], Balakirev et al. [2002, 2003]) may lead us to conclude that the X/X duplicates surveyed in this study may be under more selective constraint than the autosomal duplicates for which data are available, possibly because deleterious alleles will be eliminated more quickly from the X chromosone because of hemizygosity in males.

The Role of Ectopic Gene Conversion

Gene conversion has long been recognized as a potentially powerful force of "concerted evolution," a process with the long-term effect of retarding sequence divergence (Ohta 1981; Nagylaki and Barton 1986; Ohta 1987*a*; 1987*b*; Nagylaki 1988). In *Drosophila* species, the *amylase* duplication is a well-studied example of such concerted evolution. The coding sequences of *Amy(p)* and *Amy(d)* are more similar within than between species, while the 5′ regulatory sequences have diverged (Shibata and Yamazaki 1995). In addition, sequence polymorphism studies reveal shared polymorphisms between the coding sequences of the two loci (Araki, Inomata, and Yamazaki 2001). Shared polymorphisms have also been found between *Attacin* A and B genes (Lazzaro and Clark 2001), which are involved in immune responses, as well as within and between the *hsp70* gene clusters (Bettencourt and Feder 2002). As discussed above, samples from duplicate loci that show shared polymorphisms require a gene genealogy involving ectopic gene conversion (assuming that every mutation falls at a previously unmutated site).

In our sample of duplicates with high K_a/K_s between copies, nonparametric methods to detect gene conversion provide little evidence for ectopic conversion between copies. First, very few shared polymorphisms are observed (table 6). Second, few tracts of similar sequence were found using GENECONV. However, not all gene conversion will be detected as shared polymorphisms in a sample, leading us to estimate the ectopic exchange rate using available methods (Innan 2003*a*). The estimates of C ($= 4N_e c$) are generally very low for this set of duplicates. Table 7, and the results of our simulation suggest that we can rule out a model where $C = 0$ for 4 of 11 X-linked duplicate pairs at the 2.5% level, before correcting for multiple tests.

The methods employed to make inference on gene conversion make very different assumptions about the gene conversion process. Sawyer's (1989, 1999) method assumes that tracts of conversion are large enough to result in runs of similarity between copies, whereas Innan's (2003*a*) method assumes that conversion affects only one mutation per event. In *D. melanogaster*, if average diversity is 0.01 per site, this would correspond to an average tract length of 100 base pairs. It is unclear in *D. melanogaster* what the tract length is between non-allelic loci, nor is it clear how sensitive Innan's model is to violations of this assumption. Despite these differences between approaches, results from both approaches agree in that they do not suggest high rates of gene conversion between copies in this data set. For the X-linked duplicates that appear to be undergoing ectopic exchange (labeled in table 7), it is unlikely that violations of the infinite sites are sufficient to explain the observation. First, the K_s between copies are all low (table 1), and only one of the loci has a single site with more than two states segregating (CG15644, the inferred number of mutations is greater than the number of segregating sites, table 3).

Divergence Between Duplicates

We have applied the McDonald-Kreitman (1991) test to two different samples of duplicate loci. The first consists of (mostly) X-linked duplicates with $K_a/K_s > 1$ between copies (table 1). The second set consists of tandem autosomal duplicates taken from our own data and from the literature. The patterns of polymorphism relative to divergence are very different in these two sets of loci, with a significant excess of amino acid replacements in the former, and a deficit of fixations in the latter (table 8).

In the case of the X-linked duplicates, our data suggest that they are currently evolving under purifying selection (table 5), with relatively little statistical support for gene conversion between copies, and a historical excess of amino acid replacements between copies. Two alternative models are compatible with this excess of amino acid fixations. First, a new function could have evolved relatively quickly after duplication, and the substitutions accumulated under positive selections. Second, if the duplicate genes were redundant in function, substitutions could accumulate as a consequence of relaxation of purifying selection. In that case, the duplicates may have been preserved if, for example, an environmental change occurred resulting in a change in selective pressure (i.e., Kimura [1983, pp. 104–113]) leading to current selective constraint. However, fixations between copies are neutral under the latter model, and it seems unlikely that the high K_a/K_s between copies (table 1) would have fortuitously accumulated on X-linked genes in the face of the gene conversion that is expected to occur between young, highly similar, tandem duplications. Rather, we argue that selection was strong enough, relative to gene conversion, to drive the divergence of these loci at the amino acid level.

The autosomal duplicates that have been studied to date show a rather different pattern. The deficit of divergence among these loci (table 8) is consistent with the expectation for duplicate pairs evolving under rather

strong gene conversion (i.e., fig. 1). In fact, estimates of the ectopic conversion rate of these autosomal loci are higher than for the genes in our X-linked families (table 7). Additionally, patterns of molecular evolution at *amylase* in *Drosophila* suggest strong, long-term, concerted evolution (Shibata and Yamazaki 1995; Inomata, Tachida, and Yamazaki 1997), and tracts of gene conversion are readily visible within and between *hsp70* gene clusters (Bettencourt and Feder 2002). In fact, all previous studies of nucleotide variation at duplicate loci in *D. melanogaster* (Araki, Inomata, and Yamazaki 2001; Lazzaro and Clark 2001; Bettencourt and Feder 2002; Balakirev, Balakirev, and Ayala 2002; Balakirev et al. 2003) and *D. pseudoobscura* (King 1998) that we are aware of have documented evidence for gene conversion between paralogs. Given that many of the duplicate pairs studied by previous authors are ancient (King 1998; Bettencourt and Feder 2002; Balakirev, Balakirev, and Ayala 2002; Balakirev et al. 2003), the extent to which gene conversion affects the distribution of K_s between duplicates within genomes remains an important question. A recent analysis of duplicates from several yeast genomes has found widespread evidence for gene conversion, suggesting that the distribution of K_s between duplicates does not conform in general to the assumptions of the molecular clock (L. Gao and H. Innan, personal communication).

Conclusions

We have studied polymorphism and divergence between highly diverged X-linked duplicates in the *D. melanogaster* genome (Thornton and Long 2002). Patterns of both single-nucleotide and insertion/deletion polymorphism suggest that purifying selection is currently acting at these loci, and that there is rather little evidence for recent gene conversion. However, there is a significant excess of amino acid divergence between copies, implicating the action of positive selection after duplication. These data are consistent with the hypothesis that natural selection must be stronger than gene conversion in order for duplicates to diverge (Innan 2003*b*). By contrast, autosomal duplications that have been studied to date show several shared polymorphisms and very little divergence between copies, suggesting that selective pressure leading to divergence between copies has been weaker than the force of gene conversion.

Acknowledgments

The authors acknowledge Bruce Lahn and Steve Dorus for sequencing assistance. We thank J. J. Emerson, Bret Payseur, Peter Andolfatto, Aida Andres, Richard Hudson, and three anonymous reviewers for feedback provided during the course of the work and preparation of the manuscript. We also thank J. W. O. Ballard and M-L. Wu for providing the Drosophila lines used in this study. K.T. was supported by National Institutes of Health Training Grant in Genetics and Regulation and a GAAN Training Grant. M.L. was supported by grants from the National Institutes of Health (R01 GM65429-01) and National Science Foundation Career Award (MCB-0238168).

Literature Cited

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle (194 co-authors). 2000. The genome sequence of Drosophila melanogaster. Science **287**:2185–2195.

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

Aquadro, C. F., A. L. Weaver, S. W. Schaeffer, and W. W. Anderson. 1991. Molecular evolution of inversions in *Drosophila pseudoobscura*—the *Amylase* gene region. Proc. Natl. Acad. Sci, U.S.A. **88**:305–309.

Araki, H., N. Inomata, and T. Yamazaki. 2001. Molecular evolution of duplicated amylase gene regions in Drosophila melanogaster: evidence of positive selection in the coding regions and selective constraints in the cis-regulatory regions. Genetics **157**:667–677.

Balakirev, E. S., E. I. Balakirev, and F. J. Ayala. 2002. Molecular evolution of the Est-6 gene in Drosophila melanogaster: contrasting patterns of DNA variability in adjacent functional regions. Gene **288**:167–177.

Balakirev, E. S., V. R. Chechetkin, V. V. Lobzin, and F. J. Ayala. 2003. DNA polymorphism in the β-*esterase* gene cluster of *Drosophila melanogaster*. Genetics **164**:533–544.

Begun, D. J., and C. F. Aquadro. 1993. African and North-American populations of Drosophila melanogaster are very different at the DNA level. Nature **365**:548–550.

Begun, D. J. 1997. Origin and evolution of a new gene descended from *alcohol dehydrogenase* in *Drosophila*. Genetics **145**: 375–382.

Betrán, E., K. Thornton, and M. Long. 2002. Retroposed new genes out of the X in Drosophila. Genome Res. **12**:1854–1859.

Bettencourt, B. R., and M. E. Feder. 2002. Rapid concerted evolution via gene conversion at the Drosophila hsp70 genes. J. Mol. Evol. **54**:569–586.

Celniker, S. E., D. A. Wheeler, B. Kronmiller, J. W. Carlson, A. Halpern, S. Patel, M. Adams, M. Champe, S. P. Dugan, E. Frise (32 co-authors). 2003. Finishing a whole genome shotgun: release 3 of the Drosophila melanogaster euchromatic genome sequence. Genome Biol. **3**:0079.1–0079.14.

Charlesworth, B. 2001. The effect of life-history and mode of inheritance on neutral genetic variability. Genet. Res. **77**:153–166.

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. U. S. A. **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.

Emerson, J. J., H. Kaessmann, E. Betran, and M. Long. 2004. Extensive gene traffic on the mammalian X chromosome. Science **303**:537–540.

Ewing, B., and P. Green. 1998. Basecalling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8**:186–194.

Ewing, B., L. Hillier, M. Wendl, and P. Green. 1998. Basecalling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8**:175–185.

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.

Gordon, D. C., C. Abajian, and P. Green. 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8**:195–202.

Haldane, J. B. S. 1933. The part played by recurrent mutation in evolution. Am. Nat. **67**:5–19.

Hudson, R. R. 2001. Two-locus sampling distributions and their application. Genetics **159**:1805–1817.

Hudson, R. R., and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA-sequences. Genetics **111**:147–164.

Innan, H. 2003*a*. The coalescent and infinite-site model of a small multigene family. Genetics **163**:803–810.

———. 2003*b*. A two-locus gene conversion model with selection and its application to the human *RHCE* and *RHD* genes. Proc. Natl. Acad. Sci. U.S.A **100**:8793–8798.

Inomata, N., and T. Yamazaki. 2002. Nucleotide variation of the duplicated *amylase* genes in *Drosophila kikkawai*. Mol. Biol. Evol. **19**:678–688.

Inomata, N., H. Tachida, and T. Yamazaki. 1997. Molecular evolution of the *Amy* multigenes in the subgenus *Sophophora* of *Drosophila*. Mol. Biol. Evol. **14**:942–950.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, U.K.

King, L. M. 1998. The role of gene conversion in determining sequence variation and divergence in the Est-5 gene family in *Drosophila pseudoobscura*. Genetics **148**:305–315.

Kondrashov, F. A., I. B. Rogozon, Y. I. Wolf, and E. V. Koonin. 2002. Selection in the evolution of gene duplications. Genome Biol. **3**:0008.1–0008.9.

Kreitman, M., and R. R. Hudson. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-Dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127**:565–582.

Lasko, P. 2000. The *Drosophila melanogaster* genome: translation factors and RNA binding proteins. J. Cell Biol. **150**: F51–F56.

Lazzaro, B. P., and A. G. Clark. 2001. Evidence for recurrent paralogous gene conversion and exceptional allelic divergence in the Attacin genes of *Drosophila melanogaster*. Genetics **159**:659–671.

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.

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

McDonald, J. H., and M. Kreitman. 1991. Adpative protein evolution at the *Adh* locus in *Drosophila*. Nature **351**:652–654.

Misra, S., M. A. Crosby, C. J. Mungall, B. B Matthews, K. S. Campbell, P. Hradecky, Y. Huang, J. S. Kaminker, G. H. Millburn, S. E. Prochnik (30 co-authors). 2003. Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. **3**:0083.1–0083.22.

Nagylaki, T. 1988. Gene conversion, linkage, and the evolution of multigene families. Genetics **120**:291–301.

Nagylaki, T., and N. Barton. 1986. Intrachromosomal gene conversion, linkage, and the evolution of multigene families. Theor. Popul. Biol. **29**:407–437.

Nei, M., and A. Roychoudhury. 1968. Probability of fixation of nonfunctional genes at duplicate Loci. Am. Nat. **107**: 362–372.

Ohta, T. 1981. Genetic-variation in small multigene families. Genet. Res. **37**:133–149.

———. 1987a. A model of evolution for accumulating genetic information. J. Theor. Biol. **124**:199–211.

———. 1987b. Simulating evolution by gene duplication. Genetics **115**:207–213.

Okuyama, E., H. Shibata, H. Tachida, and T. Yamazaki. 1996. Molecular evolution of the 5′-flanking regions of the duplicated *Amy* genes in *Drosophila melanogaster* species subgroup. Mol. Biol. Evol. **13**:574–583.

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

Petrov, D. A., Y. C. Chao, E. C. Stephenson, and D. L. Hartl. 1998. Pseudogene evolution in *Drosophila* suggests a high rate of DNA loss. Mol. Biol. Evol. **15**:1562–1567.

Posada, D., and K. A. Crandall. 2001. Evaluation of methods for detecting recombination from DNA sequences: Computer simulations. Proc. Natl. Acad. Sci. U. S. A. **98**:13757–13762.

R Development Core Team. 2004. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.

Sawyer, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.

———. 1999. GENECONV: a computer package for the statistical detection of gene conversion. Department of Mathematics, Washington University in St. Louis, http://www.math.wustl.edu/ sawyer.

Shibata, H., and T. Yamazaki. 1995. Molecular evolution of the duplicated amy locus in the *Drosophila*-melanogaster species subgroup—concerted evolution only in the coding region and an excess of nonsynonymous substitutions in speciation. Genetics **141**:223–236.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437–460.

Takano, T. S. 1998. Rate variation of DNA sequence evolution in the Drosophila lineages. Genetics **149**:959–970.

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.

Thornton, K. 2003. libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics **19**:2325–2327.

Thornton, K., and M. Long. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Mol. Biol. Evol. **19**:918–925.

Tijet, N., C. Helvig, and R. Feyereisen. 2001. The cytochrome P450 gene superfamily in *Drosophila melanogaster*: annotation, intron-exon organization and phylogeny. Gene **261**:189–198.

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

Watterson, G. A. 1975. Number of segregating sites in genetic models without recombination. Theor. Popul. Biol. **7**:256–276.

Zhang, Z., and H. Kishino. 2004*a*. Genomic background drives the divergence of duplicated amylase genes at synonymous sites in Drosophila. Mol. Biol. Evol. **21**:222–227.

———. 2004*b*. Genomic background predicts the fate of duplicated genes: evidence from the yeast genome. Genetics **166**:1995–1999.

Michael Nachman, Associate Editor

Accepted October 4, 2004