Retroposed New Genes Out of the X in *Drosophila*

Esther Betrán,[1] Kevin Thornton,[2] and Manyuan Long[1,2,3]

[1]Department of Ecology and Evolution; [2]Committee on Genetics, The University of Chicago, Chicago, Illinois 60637, USA

New genes that originated by various molecular mechanisms are an essential component in understanding the evolution of genetic systems. We investigated the pattern of origin of the genes created by retroposition in *Drosophila*. We surveyed the whole *Drosophila melanogaster* genome for such new retrogenes and experimentally analyzed their functionality and evolutionary process. These retrogenes, functional as revealed by the analysis of expression, substitution, and population genetics, show a surprisingly asymmetric pattern in their origin. There is a significant excess of retrogenes that originate from the X chromosome and retropose to autosomes; new genes retroposed from autosomes are scarce. Further, we found that most of these X-derived autosomal retrogenes had evolved a testis expression pattern. These observations may be explained by natural selection favoring those new retrogenes that moved to autosomes and avoided the spermatogenesis X inactivation, and suggest the important role of genome position for the origin of new genes.

[The sequence data from this study have been submitted to GenBank under accession nos. AY150701–AY150797. The following individuals kindly provided reagents, samples, or unpublished information as indicated in the paper: M.-L. Wu, F. Lemeunier, and P. Gibert.]

New genes that originated by various molecular mechanisms are an essential component in understanding the evolution of genetic systems (Long 2001). These mechanisms include the classic mechanism of duplication (Ohno 1970), exon shuffling (Gilbert 1978), retroposition (Brosius 1991), and gene fusion through deletions or recruitment of new regions (Nurminsky et al. 1998), or a combination of these mechanisms (Long and Langley 1993; Begun 1997; Nurminsky et al. 1998). Despite the progress in recent years (Long 2001), little is known about the general pattern of new gene origination, because of the challenge to identify new genes in adequate numbers for pattern analysis.

There is increasing evidence, fortunately, that retroposition, which generates new genes in new genomic positions via reverse transcription of mRNA from a parental gene, is important for the origin of new gene functions (Brosius 1999). In mammalian systems, a classic example is the human retrogene *Pgk-2* with male specific function (McCarrey and Thomas 1987). *Pgk-2* is autosomal (chromosome 19) whereas the parental copy *Pgk-1* is X-linked. *Pgk-2* evolved late spermatogenesis-specific expression. This new expression pattern is related to the fact that late spermatogenesis cells are the only ones that do not express *Pgk-1* because of male germline X inactivation (McCarrey 1994). Subsequent analyses of retroposed genes in mammalian genomes suggested that retroposition had efficiently sown the seeds of evolution in genomes (Brosius 1991). Among invertebrate systems, *Drosophila* genomes have been found containing a number of young genes recently created by retroposition. For example, the *sphinx* gene in *Drosophila melanogaster* and the *jingwei* gene in the *Drosophila yakuba* clade were created within 2–3 Myr by retroposition from parental genes encoding *ATP synthase* and *alcohol dehydrogenase*, respectively (Long and Langley 1993; Long et al. 1999; Wang et al. 2000, 2002). In general, recently completed genome sequences in humans

(Lander et al. 2001; Venter et al. 2001) and *Drosophila melanogaster* (Adams et al. 2000) contain new genes created by retroposition which provide opportunities to examine the pattern of origin of new genes.

We investigated the pattern of new genes created by retroposition in the *Drosophila* genome. New retroposed gene copies are identified by examining hallmarks of retroposition (Li 1997): (1) one member of the pair is intronless in the coding region of sequence similarity (new copy), whereas the other has introns (parental copy); (2) one of them contains a polyA tract (new copy), if both copies are intronless; (3) the new copy may still be flanked by short duplicate sequences. The analyses of these *Drosophila* retrogenes (analysis of expression, substitution, and population genetics) revealed that these genes are functional. The study of the direction of retroposition showed a surprising asymmetric pattern. There is a significant excess of retrogenes that originate from the X chromosome and retropose to autosomes. These retrogenes evolved a testis expression pattern. We discuss possible explanations and conclude that these observations may be explained by natural selection favoring those new retrogenes that moved to autosomes and avoided the spermatogenesis X inactivation. Our results support the important role of genome position in new genes evolution.

RESULTS AND DISCUSSION

We have identified, from the annotated genes in the *D. melanogaster* genome, all pairs of homologs (70% amino acid identity or more) that are located on different chromosomes with hallmarks of retroposition (Table 1). Twenty-four young paralogous pairs fulfilled these criteria: 23 pairs in which the new copy lost the introns (*CG12628*, one of the 23, is additionally flanked by short repeats), and one pair with no introns in either copy but with the new copy retaining a degenerated poly-A tract (*CG 12324/Rp515A*). Interestingly, *CG12628*, which seems to be the youngest of the described retrogenes, is the only one that retains the direct repeats, a hallmark of the recent insertion event. Some other retrogenes also retained a degenerated poly-A tract: *CG12628*, *CG10174*,

Table 1. Young Retroposed Genes in the *Drosophila melanogaster* Genome Compared to Its Parental Genes

#	New genes			Parental genes					
	Locus	Position	Expression	Locus	Position	Expression	Gene type	K_A/K_S	K_S
1	*CG12628*	2L_40D	—	*Mgst1*	**X_19E**	GH/LP	Glutathion transferase	0.5370	0.03015
2	*CG12324*	2R_47C	LP	*RpS15A*	**X_11E**	AT/GM	Ribosomal protein	0.2488	0.04035
3	*CG10174*	2L_36F	at	*Dntf2*	**X_19E**	GH/LP	Transporter	0.3426	0.16811
4	*CG13732*	3L_74C	at	*CG15645*	**X_13E**	a	Unknown	0.7951	0.19015
5	*CG4960*	3R_96F	AT	*CG8331*	2R_50E	GM/GH/LP	Membrane protein	**0.3630**	0.32681
6	*Act5C*	X_5C	LD/LP/GH	*Act57B*	2R_57B	AT/GM/HL	Actin	**0.0744**	0.33619
7	*CG17856*	3R_98C	at	*CG3560*	**X_14B**	GH/LP	NAD dehydrogenase	**0.1767**	0.67448
8	*CG11825*	2R_47A	—	*CG17734*	3R_86D	LP	Unknown	**0.1691**	0.73976
9	*CG12334*	3R_90C	AT	*CG1534*	**X_9E**	GM/LP/LD	Unknown	**0.1405**	0.73999
10	*Act42A*	2R_42A	AT	*Act79B*	3L_79B	GH/LP	Actin	**0.0400**	0.74529
11	*Vha36*	2R_52A	AT/GM/GH	*CG8310*	**X_3A**	—	Transporter	**0.1583**	0.80580
12	*Trxr2*	3L_79E	AT	*Trxr1*	**X_7D**	AT/GM/LD	Glutathion reductase	**0.2102**	0.89926
13	*CG7768*	3L_70D	AT	*Cyp33*	2R_54C	LD	Chaperone	**0.2940**	0.90764
14	*Ef1α48E*	2R_48D	LD/HL/SD	*Ef1α100E*	3R_100E	HL/LP	Translation Ef.	**0.0654**	0.91664
15	*CG7235*	2L_25F	AT/GH	*Hsp60*	**X_10A**	AT/GM/LD	Chaperone	**0.1362**	0.92378
16	*Pros28.1A*	3R_92F	AT/LP	*Pros28.1*	**X_14B**	LD	Endopeptidase	**0.1637**	0.95103
17	*CanB*	X_4F	at/gdm	*CanB2*	2R_43E	SD	Protein phosphatase	**0.0177**	1.03241
18	*CG9819*	X_14F	LP	*CanA1*	3R_100B	GH	Protein phosphatase	**0.1007**	1.16483
19	*CG9873*	2R_59C	at	*CG9091*	**X_13B**	GM/SD/LP	Ribosomal protein	**0.1693**	1.22348
20	*Sep5*	2R_43F	LD	*Sep2*	3R_92E	GM/LD/SD	Cytoskeletal protein	**0.1647**	1.23927
21	*CG13340*	2R_50C	AT	*CG8040*	3L_67D	AT/GH/LP	Peptidase	**0.1256**	1.31892
22	*Cdlc2*	2L_22A	AT/GH/LP	*ctp*	**X_4C**	AT/GM/LP	Dynein light chain	**0.0030**	1.43721
23	*CG4706*	3R_86D	AT	*Acon*	2L_39B	GM/LD/HL	Aconitase	**0.0789**	1.52027
24	*CG8602*	3L_65F	GH/LD/SD	*CG12194*	2L_25B	—	Sugar transporter	**0.1099**	1.56227

This is a subsample of young retroposed copies whose parental gene lies in a different chromosome (see text for details). The K_A/K_S ratio is in bold when it is significantly smaller than 0.5. We have checked expression experimentally for some genes in adult males and females (a), gonadectomized males (gdm), and testis (at) (see Fig. 1 for details) and used information from Berkeley Drosophila Genome Project (BDGP) EST libraries for the other genes. Tissues in this latter case are named following BDGP nonmenclature: LD (embryos), LP (larvae and early pupae), HL and GH (both adult heads), SD (Schneider L2 cells), AT (adult testis), and GM (ovaries). When the gene is expressed in more than three tissues, only the three in which the gene is most highly expressed or more relevant for discussion, i.e., AT and GM, have been listed. The lowercase letters indicate the data from our expression experiments, and the uppercase letters the data from the BDGP EST library.

and *CG13732*. The parental genes have diverse functions, consistent with results from the human genome (Gonçalves et al. 2000).

Several lines of evidence indicate that these newly derived genes are functional. First, many of them are known genes with identified bona fide proteins (Table 1). Second, we examined functional constraints on these new genes by comparative analysis of the rates of nonsynonymous substitutions per site (K_A) and synonymous substitutions per site (K_S) between the members of each gene pair. In general, a K_A/K_S ratio that is significantly lower than unity is considered to indicate functional constraint. However, the expected K_A/K_S ratio for divergence between a functionless new retrogene duplicate and a functional parental gene should be smaller than unity but higher than 0.5, dependent upon the selective constraint on the parental gene (Li 1997). In a conservative test, we considered K_A/K_S significantly lower than 0.5 to indicate functional constraint on both genes. We found that the K_A/K_S ratios of 20 of the 24 gene pairs are significantly lower than 0.5 (Table 1); the ratios of four genes are not significantly lower than 0.5.

We surveyed nucleotide polymorphism in these four genes by sequencing 12 to 36 alleles for each gene, which suggested strong selective constraints (Table 2). First, in these genes, nonsynonymous polymorphism is significantly lower than synonymous polymorphism ($\chi^2 = 21.25$, $P < 0.00001$).

Table 2. Polymorphism Analysis of the Retroposed Copies of Genes With Lower K_S (see Table 1)

New gene	L	N	S	M_S	M_N	π_S	θ_S	π_N	θ_N
CG12628	456 bp	33	5	2	3	0.0047	0.0042	0.0030	0.0022
CG12324	390 bp	16	8	7	1	0.0182	0.0232	0.0004	0.0010
CG10174	390 bp	36	7	5	2	0.0125	0.0134	0.0015	0.0016
CG13732	630 bp	12	8	4	5	0.0098	0.0103	0.0019	0.0033

L, length of the gene; N, number of alleles sequenced; S, segregating sites; M, number of mutations; π, average nucleotide pairwise differences, and θ, estimator of $4N_e\mu$, where N_c and μ are effective population size and neutral mutation rate, respectively. The subscripts N and S refer to nonsynonymous sites and silent sites, respectively. Stop codon position or codons with deletions for *CG12628* were excluded from the analysis. Values were calculated using DNAsp software (Rozas and Rozas 1999).
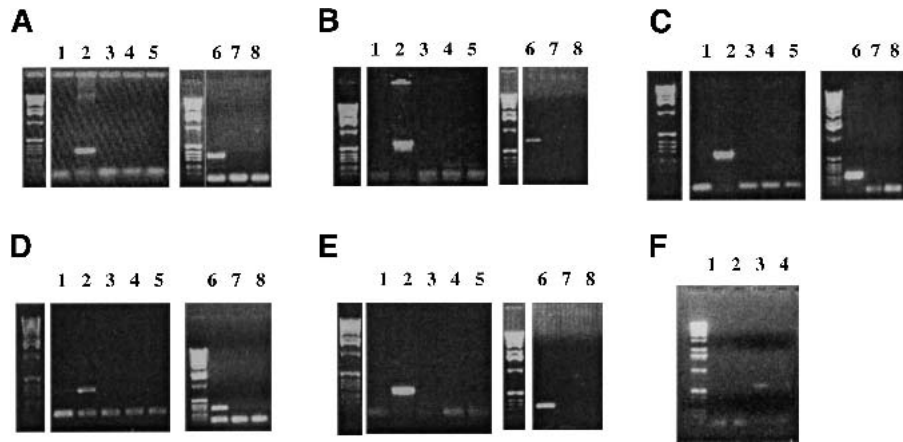
Figure 1 RT-PCR for several genes. (*A*) *CG10174*, (*B*) *CG13732*, (*C*) *CG17856*, (*D*) *CanB*, and (*E*) *CG9873*. Lane *1* corresponds to gonadectomized male cDNA, lane *2* is testis + accessory glands cDNA; lanes *3* and *4* are the negative controls after DNA digestion for the experiments of lanes *1* and *2*, respectively, and lane *5* is the negative control of the PCR. Lane *6* is the PCR experiment using testis cDNA; lane *7* is the negative control after DNA digestion, and lane *8* is the negative control of the PCR. (*F*) Lane *1* is *CG15645* RT-PCR using cDNA from polyA selected RNA from a mixed sample of males and females; lane *2* is the PCR from this mRNA without being reverse-transcribed from the mixed sample; lanes *3* and *4* are the nested PCR experiments using the PCR products of lanes *1* and *2* as templates. The DNA marker, as shown here, is a 1-kb DNA ladder (Gibco).

Second, variation in these genes does not significantly differ from the values for average functional genes in *Drosophila* (π_s = 0.0135, π_{total} = 0.0040), whereas one could predict that functionless DNA should have higher variation (Powell 1997). Finally, none of the alleles, with the exception of some alleles of *CG12628*, contain a frameshift mutation and/or premature stop codon. Although *CG12628* shows a premature stop codon or one base pair deletion in some alleles, a large proportion (60.61%) of alleles maintain an intact reading frame. Furthermore, nonsynonymous polymorphism is lower than synonymous polymorphism in both the normal alleles and the truncated alleles in which a shorter predicted open reading frame (ORF) remains. Thus, the functional role for this retrogene cannot be ruled out. These polymorphism data together with K_A/K_S values significantly lower than 0.5 in the rest of the genes suggest that almost all new retrogenes identified are subject to strong functional constraints. Furthermore, in RT-PCR experiments and BDGP EST libraries (Fig. 1, Table 1), we observed that most new retrogenes are expressed in one or more of the investigated tissues, further suggesting that these genes are functional. Population genetic analyses of the gene sequences with newly evolved expression patterns suggest that some of these new genes may have evolved functions that did not exist previously (E. Betrán and M. Long, unpubl.).

Examination of the physical positions of these newly evolved functional genes revealed an unexpected pattern. We observed that 12 pairs (50%) originated from parental genes located on the X chromosome despite its low gene number (17% of the genes in the genome), whereas we found only 12 from autosomes, 3 to X and 9 to autosomes (Tables 1, 3). This pattern is significantly different from the expected (*P* = 0.0084; Table 3). If every gene in the genome is retroposed with equal probability, a sample of 24 parental genes should include only 5.6 (23.3%) from the X chromosome and 18.4 (76.7%) from autosomes (see Methods). Therefore, there is an excess of new genes retroposed from the X-linked pa-

rental genes to autosome; correspondingly, there is a deficiency of retroposed genes originated from autosomes (Table 3).

Although this result suggests that many new genes originated from the X chromosome, it is unclear whether or not this observation is limited to the identified new genes in the group defined by 70% amino acid identity. Thus, we extended a similar analysis (see Methods) to the new retrogenes of 50% or higher identity at the amino acid level with their parental genes and observed a similar phenomenon. Of 159 putative interchromosomal retroposition events, 63 (40%) originated from X-linked genes, indicating a highly significant excess of X-linked origination events over the 23.3% expected under the assumption of random retroposition (*P* < 0.0001, χ^2 = 23.81, *df* = 1). Therefore, the pattern that we observed is not limited to a certain subset of genes.

We had ignored retroposed copies from the X chromosome that inserted elsewhere in the same chromosome in all previous analyses, to ensure that we were not looking at tandem duplicates or at ancient tandem duplicates now separated by paracentric inversions within the same chromosome (Powell 1997). However, we examined the frequency of retroposition among different sections within the X chromosome. In the retrogenes with 50% or higher amino acid identity with parental genes, we found that of 67 putatively retroposed copies from the X chromosome, only four inserted into different X chromosomal sections. The expected value of within-X transpositions is 10.1, which is significantly higher than the observed value (*P* = 0.039, χ^2 = 4.33, *df* = 1).

Four possible explanations could account for the observed pattern: (1) nonrandom generation of retrogenes by a disproportionate number of X-linked genes that express in the germline cells; (2) negative selection against insertions in the X chromosome; (3) different recombination rates (or possibly deletion rates) between the autosomes and the X chromosome; and (4) positive Darwinian selection favoring retrogenes generated from the X chromosome to the autosomes.

We found similar proportions of X-linked and autosomal genes expressed in germline cells in the Berkley EST libraries of ovary and adult testis (E. Betrán, K. Thornton, and M. Long, unpubl.), ruling out the first possible explanation that a disproportionate number of genes that express in the germline are X-linked resulting in the larger number of X-originated retrogenes. Alternatively, if insertions are slightly deleterious because of possible disruption of the regulation of gene activity, there will be stronger selection against X-linked than autosomal insertions because of male hemizygosity for the X (Charlesworth et al. 1987). This selection would reduce the number of insertions surviving in the X chromosome by a small proportion, e.g., lower than 2%, under the assumptions that the selection intensity is an order of magnitude lower than the inverse of effective population size and that the fitness effects of insertions are recessive (see Methods). This can

Table 3. Analysis of the Pattern of Retroposition

Direction of the gene formation event	Expectation		Observed No.	Excess (%)
	%	No.		
X→A	23.3	5.6	12	114
A→X	20.3	4.9	3	−39
A→A	56.4	13.5	9	−33
	$X^2 = 9.55$, df = 2, P = 0.0084			

X, X chromosome; A, autosome; Excess = $[O − E)/E] \times 100$; E, expected; O, observed.

account only for a negligible part of the deficiency of new gene insertions in the X chromosome. Therefore, the negative selection from this hypothetical process cannot explain the excess of retroposition from X-linked parent genes.

The ectopic exchange model predicts that insertion elements will be more abundant in regions of low recombination because they are less likely to be deleted by unequal recombination (Langley et al. 1988). Hence, under this model, different recombination rates of the autosomes and the X chromosome would be likely to be associated with different deletion rates, thus yielding different rates of new retrogenes between the X and the autosomes, as we observed. However, there is no evidence for different recombination rates between autosomes and the X chromosome. Recombination rates per base pair in these chromosomes are similar (Ashburner 1989), and the product between the population size and the time spent in females (recombining sex) is the same for X chromosomes

$$\left(\frac{3}{4} \cdot \frac{2}{3} = \frac{1}{2}\right)$$

and autosomes

$$\left(1 \cdot \frac{1}{2} = \frac{1}{2}\right).$$

The fourth hypothesis, positive selection, seems more parsimonious to interpret the excess of retroposition from X to autosomes. X inactivation during early spermatogenesis could produce a selective advantage for the retroposed genes with novel functions that escape X linkage and become expressed in testis, as previously suggested (Lifschytz and Lindsley 1972; McCarrey 1994). X inactivation early in spermatogenesis is well documented in *Drosophila*, mouse, and human (Lifschytz and Lindsley 1972; Richler et al. 1992). Thus, a mutant with a newly retroposed gene on autosomes will have some advantage over an X-linked form, because the mutant can carry out a new function putatively required in male germline cells after the X chromosome becomes inactivated. This hypothesis assumes that retroposition occurs from genes on all chromosomes with the same probability but natural selection favors the ones that avoid X-linkage by moving to an autosome and developing expression in testis.

The hypothesis of selective advantage by avoiding X linkage predicts that most of the new retrogenes that evolved from X-linked parent genes would be expressed in the male germline, nonexclusively. The new genes can also develop or retain additional functions in other tissues (McCarrey 1994). Data in Table 1 and Figure 1 confirm this prediction, showing that 10 of the 11 genes retroposed from the X chromosome,

for which expression information is available, are expressed in adult male testis. Such a high percentage (91%) of retrogenes expressed in the testis is unlikely to be a random pattern, considering that transcripts of only ~10% of the ~13,600 genes of the *Drosophila* genome have been detected in testis (Andrews et al. 2000), and it is in agreement with the prediction of the hypothesis of positive selection. Nevertheless, it is also possible that the expression pattern of a new copy could be a by-product of the region into which it fortuitously inserted (Bownes 1990; Pasyukova et al. 1997). However, these explanations predict such elements to be nonfunctional pseudogenes, against our observations above and the fact that these retrogenes have been kept, according to our phylogenetic data (see Methods), far longer than the half-life of pseudogenes in *Drosophila* (Watterson 1983; Petrov et al. 2000).

Here we observed that new functional retrogenes, mostly with newly evolved testis expression, tend to avoid X-linkage by moving to an autosome. Consistently, it was observed that, in *Drosophila*, autosomal mutations for male sterility have mostly late spermatogenesis effects (Castrillon et al. 1993) and, in the nematode *C. elegans*, X-linked sperm-enriched and germline-intrinsic genes are scarce (Reinke et al. 2000). This pattern reveals a possible role of Darwinian selection for the retroposed new genes that escape from the spermatogenesis X inactivation, although there may be additional mechanisms contributing to the retroposition process, for example, the hypothetical sexual antagonism that genetic variants are advantageous for one sex but disadvantageous for the other sex (Rice 1984; C.-I. Wu, pers. comm.). The pattern also supports the view that genomic location matters for gene function (Hurst and Randerson 1999). Genes that escape X-linkage by retroposing to an autosome and are expressed in the male germline have been found in mammals (Dahl et al. 1990; McCarrey 1994), although a comparable general pattern has not been detected in the human genome (Venter et al. 2001). If this pattern exists in the human genome, it could be obscured by the enormous number of degenerating retroposed copies in this genome (Gonçalves et al. 2000). A large number of X-linked genes expressed in spermatogonia have been reported in the mouse (Wang et al. 2001). Our finding is not necessarily contradictory to this interesting observation. These mouse genes, observed from the early stage (mitotic cells) of spermatogenesis, are expressed prior to X inactivation. When we analyzed locations of the known mammalian genes that are expressed exclusively during male meiosis (Eddy and O'Brien 1998), we found that all 26 genes are located on autosomes and none are on the X chromosome (E. Betrán and M. Long, unpubl.). This result, revealing a different pattern from that of Wang et al. (2001) in a different spermatogenesis stage, suggests that the mammalian late spermatogenesis was likely subject to selection as we observed in *Drosophila*.

METHODS

Genome Analysis of Retroposed Copies of Genes

Sequence data (Adams et al. 2000) were obtained from the BDGP Web site (www.fruitfly.org). The database of real and predicted amino acid sequences of Release 2 was first purged of peptides resulting from alternative transcription, retaining only the longest peptide sequence. Paralogous pairs were identified from the fasta33_t program (Pearson 1990) alignments of this entire database with a criterion of at least 70% amino acid identity or ≥50% amino acid identity in a mini-

mum overlap of 35 amino acids in the region of local alignment (Thornton and Long 2002).

The coding regions of the pairs with 70% amino acid identity were aligned with the corresponding genomic region and inspected for retroposition features: (1) one pair member was intronless in the region of sequence similarity whereas the other had introns; (2) one of them had a poly-A tail when both copies were intronless; and/or (3) one copy was flanked by short repeats. All three hallmarks of retroposition can be found in a retrogene, sometimes two, sometimes only one. Only pairs that were on different chromosomes were considered. The retroposition features plus the fact that all pairs are in different chromosomes ensure that we are not looking at tandem duplicates or at tandem duplicates that are separated by paracentric or pericentric inversions (Powell 1997); they are instead retroposed copies of genes. In the case of families (more than two homologs), the parental gene was considered to be the one with the smaller K_S. Pairs with homology to mobile elements were discarded.

In the case of paralogous pairs with amino acid identity ≥50%, we obtained the numbers of exons for each gene in each paralogous pair from the BDGP annotation. We only included gene pairs where one member is predicted to contain introns (parental gene) and the member has no predicted introns (new gene) that locate in different chromosomes, that is, the duplication arose by a retroposition event. Tandem duplicated members of gene families would look like many events but, for our purpose, they were considered a single retroposition event.

K_A and K_S estimation and K_A/K_S ratio test

K_A and K_S were estimated in the region of sequence similarity using K-estimator software (Comeron 1999). We used a likelihood ratio test to determine whether K_A/K_S between pairs of duplicates was smaller than 0.5. The Codeml program of PAML 3.1 (Yang 1998) was run twice for every gene pair; first fixing $\omega = 0.5$ and second estimating omega. The log likelihood value of the 0.5 model (l_0) was compared to the free model (l_1). We considered the ratio significantly smaller than 0.5 if the free model was significantly more likely than the 0.5 model. Significance at the 5% level was tested by comparing twice the log likelihood difference, $2\Delta l = 2(l_1 - l_0)$, to a χ^2 distribution with one degree of freedom (Yang 1998).

Expected Number of Retropositions

Considering the number of genes per chromosome and the size (euchromatin) of the chromosome as the source and target of insertion, respectively, the fact that X-linked genes are dosage-compensated, and assuming independent generation and landing on a chromosome site and equal numbers of males and females in the population, we calculated the expected frequency (P_{KL}) (i.e., $P_{X \to A}$, $P_{A \to X}$, and $P_{A \to A}$, where "$\to$" indicates the direction of retroposition, from the parental gene to the new gene [A→A includes $A_2 \to A_3$ and $A_3 \to A_2$]).

$$P_{KL} = \frac{\sum N_i L_j f_{ij}}{\sum \sum N_i L_j f_{ij}},$$

where N_i and L_j are the proportions of gene number at the source chromosome i and the euchromatic size of the targeted chromosome, respectively, and f_{ij} is the frequency of occurrence of this type of retroposition to a given chromosome in the population. According to genome data (Adams et al. 2000) and the existence of males and females in the population, i, j: X, 2 and 3, N_i: 0.17, 0.38, 0.45; L_j: 0.19, 0.36, 0.44 (chromosome 4 ignored for its minuscule size); and f_{ij}: 0.75 for j = X and 1 for j = 2 or 3; reflecting the relative population sizes of the X chromosome and autosomes. When i = j, the expectation within chromosomes is calculated. The expected percent-

age of interchromosomal retroposition events that originate from the X chromosome to autosomes is 23.3% (see Table 3 for the other expected values). The expected percentage of copies originated from X chromosome that become inserted in the X chromosome is 15%.

Relative Fixation Rates of X Chromosome and Autosomes

The difference of relative fixation rates between X chromosome (K_X) and autosome (K_A) for a slightly deleterious mutation model with selection in one or both sexes and dosage compensation is given by $K_A/K_X = 1 + 1/3N_e s(h - 1/2)$ (Charlesworth et al. 1987); where h is the dominance coefficient, N_e the effective population size, and s the selection coefficient. When considering reasonable magnitudes of these parameters, e.g., $N_e S = -0.1$ and h = 0, we have $K_x = 0.98K_A$, indicating that X-linked genes would evolve at slightly slower rates than autosomal genes.

Population Genetic Analysis and Worldwide Samples

Genes were PCR-amplified from single *Drosophila* individuals from a worldwide sample of *D. melanogaster*. *D. melanogaster* strains used were: OK17, HG84, and Z(s)56 from Africa; yep3, yep18, yep25, Cof3, BLI5, cal4, y10, and y2 from Australia; 253.4, 253.27, 253.30, and 253.38 from Taiwan; Closs3, Closs10, Closs16, Closs19, and Seattle from USA; Rio from Brazil; Rinanga, Bdx, Besançon, Prunay, and Capri from France.

Primers used to amplify genes for sequencing were: 5′ATTCCGGATTGCAAGTATGAGC3′ / 5′GAACCCAAGATCC GGATTTATTTT3′ for *CG12628*; 5′GCTGCCAACTCGCTTC ATAA3′ / 5′AACGTAGGAAATGTTGAAGCTG3′ for *CG12324*; 5′TGCAGGGCGCATTGTTCAG3′ / 5′CATACGCCTGCCAA TACGAGT3′ for *CG10174*; and 5′TTACGCAATTCAAT GGCACCT3′ / 5′GAGAAGCAGCAGCGGGAGAT3′ for *CG13732*. Sequence was obtained for both strands and haplotypes determined directly or by subcloning and sequencing individual clones. Sequences were aligned and revised by eye considering the information from the literature (Adams et al. 2000).

Phylogenetic Inference

Chromosomes with standard arrangement of *D. melanogaster* (CS), *D. simulans* (Florida), *D. yakuba* (115) or *D. teissieri* (128.2), and *D. erecta* (154.1), representing different lineages in the *D. melanogaster* subgroup of species (Lemeunier and Ashburner 1976; Powell 1997) were hybridized with fluorescent probes (Wang et al. 2000) of the retroposed copy of the pair in most cases. Presence or absence of this copy was investigated using *D. melanogaster* maps cut and pasted to reconstruct the other species maps. All retroposed genes except the first four genes in Table 1 are older than the estimated age of the *D. melanogaster* subgroup (data not shown), 15 My (Powell 1997).

Expression Analysis

Using RT-PCR experiments (Wang et al. 2000), transcription was addressed for several genes. Analysis of expression of intronless genes is challenging because genomic contamination can produce a band the same size as that expected from the cDNA. To ensure that we were getting product from the cDNA, we obtained poly-A selected RNA or, alternatively, we obtained total RNA and digested the possible DNA contaminant by RNAse-free DNAse treatment (Gibco) and ran controls including mRNA without being reverse-transcribed. Primer sequences were: 5′TTGTCCAGCAGTACTACGCC3′ / 5′TTGGGCTTCAGCAAAAAGAT3′ for *CG10174*; 5′AGAAGT TGCTCGAGCAGAGC3′ / 5′CTCCGAGGCAGTTACATCCA3′ for *CG13732*; 5′TGTCTGGATTCAACCAATAC3′ / 5′GCTCTT

CGCGCTCCTTTTGC3′ for *CG17856*; 5′ACTCGGGTGCGC TGAGCATA3′ / 5′CCTTGTCCGCAAAGCAAATG3′ for *CG4209*; 5′TGACCAAGGGAACCACTAGT3′ / 5′TCTTAGCG GCACCTCCTTCA3′ for *CG9873*; and 5′ATGGAATTCAAT TACCTTGCT3′ / 5′CTTGCAACTTCTGCTGTAGG3′ for *CG15645*.

REFERENCES

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

Andrews, J., Bouffard, G.G., Cheadle, C., Lu, J., Becker, K.G., and Oliver, B. 2000. Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. *Genome Res.* **10:** 2030–2043.

Ashburner, M. 1989. *Drosophila: A laboratory handbook*. Cold Spring Harbor Laboratory Press, New York.

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

Bownes, M. 1990. Preferential insertion of P elements into genes expressed in the germ-line of *Drosophila melanogaster*. *Mol. Gen. Genet.* **222:** 457–460.

Brosius, J. 1991. Retroposons—Seeds of evolution. *Science* **251:** 753.

Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238:** 115–134.

Castrillon, D.H., Gonczy, P., Alexander, S., Rawson, R., Eberhart, C.G., Viswanathan, S., DiNardo, S., and Wasserman, S.A. 1993. Toward a molecular genetic analysis of spermatogenesis in Drosophila melanogaster: Characterization of male-sterile mutants generated by single P element mutagenesis. *Genetics* **135(2):** 489–505.

Charlesworth, B., Coyne, J.A., and Barton, N.H. 1987. The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* **130:** 113–146.

Comeron, J.M. 1999. K-Estimator: Calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* **15:** 763–764.

Dahl, H-H.M., Brown, R. M., Hutchison, W.M., Maragos, C., and Brown, G.K. 1990. A testis-specific form of the human pyruvate dehydrogenase E1 α subunit is coded for by an intronless gene on chromosome 4. *Genomics* **8:** 225–232.

Eddy, E.M. and O'Brien, D.A. 1998. Gene expression during mammalian meiosis. *Curr. Top. Dev. Biol.* **37:** 141–199.

Gilbert, W. 1978. Why genes in pieces? *Nature* **217:** 501.

Gonçalves, I., Duret, L., and Mouchiroud, D. 2000. Nature and structure of human genes that generate retropseudogenes. *Genome Res.* **10:** 672–678.

Hurst, L.D. and Randerson, J.P. 1999. An eXceptional chromosome. *Trends Genet.* **15:** 383–385.

Lander, E.S., Linton, L.M., Birren B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N., and Charlesworth, B. 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genet. Res.* **52:** 223–235.

Lemeunier, F. and Ashburner, M. 1976. Relationships in the melanogaster species subgroup of the genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based upon polytene banding sequences. *Proc. R. Soc. Lond. B.* **193:** 257–294.

Li, W-H. 1997. *Molecular evolution*. Sinauer Associates Sunderland, MA.

Lifschytz, E. and Lindsley, D.L. 1972. The role of X-chromosome inactivation during spermatogenesis. *Proc. Nat. Acad. Sci.* **69:** 182–186.

Long, M. 2001. Evolution of novel genes. *Curr. Opin. Genet. Dev.* **11:** 673–680.

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

Long, M., Wang, W., and Zhang, J. 1999. Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in Drosophila. *Gene* **238:** 135–141.

McCarrey, J.R. 1994. Evolution of tissue-specific gene expression in mammals. How a new phosphoglycerate kinase was formed and refined. *Bioscience* **44:** 20–27.

McCarrey, J.R. and Thomas, K. 1987. Human testis-specific PGK gene lacks introns and possesses characteristics of a processed gene. *Nature* **326:** 501–505.

Nurminsky, D.I., Nurminskaya, M.V., Aguilar, D.D., and Hartl, D.L. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* **396:** 572–575.

Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, New York, NY.

Pasyukova, E., Nuzhdin, S., Li, W., and Flavell, A.J. 1997. Germ line transposition of the copia retrotransposon in Drosophila melanogaster is restricted to males by tissue-specific control of copia RNA levels. *Mol. Gen. Genet.* **255:** 115–124.

Pearson, W.R. 1990. Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* **183:** 63–98.

Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., and Shaw, K.L. 2000. Evidence for DNA loss as a determinant of genome size. *Science* **287:** 1060–1062.

Powell, J.R. 1997. *Progress and prospects in evolutionary biology: The Drosophila model*, p. 355, Oxford University Press, New York, NY.

Reinke, V., Smith, H.E., Nance, J., Wang, J., Van Doren, C., Begley, R., Jones, S.J.M., Davis, E.B., Scherer, S., Ward, S., et al. 2000. A global profile of germline gene expression in C. elegans. *Mol. Cell* **6:** 605–616.

Rice, W.R. 1984. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38:** 735–742.

Richler, C., Soreq, H., and Wahrman, J. 1992. X inactivation in mammalian testis is correlated with inactive X-specific transcription. *Nat. Genet.* **2:** 192–195.

Rozas, J. and Rozas, R. 1999. DnaSP version 3.52: An integrated program for molecular population genetics and molecular evolution. *Bioinformatics* **15:** 174–175.

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

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wang, P.J., McCarrey, J.R., Yang, F. and Page, D.C. 2001. An abundance of X-linked genes expressed in spermatogonia. *Nat. Genet.* **27:** 422–426.

Wang, W., Zhang, J., Alvarez, C., Llopart, A., and Long, M. 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.

Wang, W., Brunet, F.G., Nevo, E., and Long, M. 2002. Origin of *sphinx*, a young chimeric RNA gene in Drosophila melanogaster. *Proc. Nat. Acad. Sci.* **99:** 4448–4453

Watterson, G.A. 1983. On the time for gene silencing at duplicate loci. *Genetics* **105:** 745–766.

Yang, Z. 1998. Likelihood ratio test for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15:** 568–573.

WEB SITE REFERENCES

www.fruitfly.org; BDGP Web site.