Dntf-2r, a Young Drosophila Retroposed Gene With Specific Male Expression Under Positive Darwinian Selection

Esther Betrán and Manyuan Long[1]

Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637

Manuscript received September 20, 2002
Accepted for publication March 4, 2003

ABSTRACT

A direct approach to investigating new gene origination is to examine recently evolved genes. We report a new gene in the *Drosophila melanogaster* subgroup, *Drosophila nuclear transport factor-2-related* (*Dntf-2r*). Its sequence features and phylogenetic distribution indicate that *Dntf-2r* is a retroposed functional gene and originated in the common ancestor of *D. melanogaster, D. simulans*, *D. sechellia*, and *D. mauritiana*, within the past 3–12 million years (MY). *Dntf-2r* evolved more rapidly than the parental gene, under positive Darwinian selection as revealed by the McDonald-Kreitman test and other evolutionary analyses. Comparative expression analysis shows that *Dntf-2r* is male specific whereas the parental gene, *Dntf-2*, is widely expressed in *D. melanogaster*. In agreement with its new expression pattern, the *Dntf-2r* putative promoter sequence is similar to the late testis promoter of β*2-tubulin*. We discuss the possibility that the action of positive selection in *Dntf-2r* is related to its putative male-specific functions.

IT has been more than 3 decades since gene duplication was suggested to be a major source of evolutionary novelties (Ohno 1970). We are now able to examine this process in detail. Analyses of genome sequences have revealed high frequencies of gene duplicates in vertebrates, invertebrates, and plants (Arabidopsis Genome Initiative 2000; Blanc *et al.* 2000; Rubin *et al.* 2000; Ball and Cherry 2001; Li *et al.* 2001; Gu *et al.* 2002), which may suggest rapid evolution of novel functions. Our understanding of detailed molecular processes and underlying population evolutionary processes most often depends upon the direct observation of a young gene duplicate (Long *et al.* 1999; Long 2001; Betrán and Long 2002). Examples of genes recently arisen [<3 million years ago (MYA)] by several different mechanisms have been found in *Drosophila*: *jingwei* (*jgw*; Long and Langley 1993), *Adh-Finnegan* (Begun 1997), *Sdic* (Nurminsky *et al.* 1998, 2001), *exuperantia 1* X copy (Yi and Charlesworth 2000), and *Sphinx* (Wang *et al.* 2002; for a more comprehensive review see Long 2001). *Jgw* was created after a retroposition event that recruited duplicated exons of a neighboring gene. *Sdic* is the product of the fusion of two previously duplicated genes. *Exuperantia 1* X copy is a transposition of a whole gene by ectopic recombination. In these examples, positive selection has been shown to have played a crucial role in their evolution (Long and Langley 1993; Nurminsky *et al.* 1998; Yi and Charlesworth 2000; Nurminsky *et al.* 2001; Llopart *et al.* 2002).

Previous work (Long and Langley 1993) has revealed that retroposition can generate duplicates in Drosophila and leave clear evidence of the molecular process that generated the new gene copy from the parental copy. A newly derived retroposed gene can be identified by examining the hallmarks of retroposition (Li 1997): (1) one member of the pair is intronless in the coding region of similar sequence (new copy) while the other contains introns (parental copy); (2) the new copy contains a poly(A) tract; and (3) the new copy may still be flanked by short duplicate sequences. Properties 2 and 3 can be used to infer new retroposed genes if the parental genes do not contain introns. In a genome-wide analysis of retroduplicates, we inferred parental and derived copies by examining these fingerprints of retroposition process (Betrán *et al.* 2002). At the threshold of >70% protein sequence identity, we identified 24 pairs of retroposition events with various ages (Betrán *et al.* 2002). Here, we report the evolutionary analysis of one of these retroposed genes, *Nuclear transport factor-2-related* [*Dntf-2r* (*CG10174*)] and its parental gene, *Dntf-2* (*CG1740*). *Dntf-2*, which contains three introns, is located on the X chromosome, whereas *Dntf-2r*, which contains no introns, is likely a recent retroposed copy of *Dntf-2* and is located on the left arm of chromosome 2 in *D. melanogaster*. In addition, *Dntf-2r* contains a putative poly(A) tract at the end of the retroposed region (Figure 1). We determined the age of *Dntf-2r* by surveying its phylogenetic distribution using fluorescence *in situ* hybridization (FISH) experiments, genomic Southern blot, and PCR analyses. We found that *Dntf-2r* is present in only four species of Drosophila:

Sequence data from this article have been deposited with EMBL/GenBank Data Libraries under accession nos. AY150763–65, AY150768, AY150770–73, AY150775–78, AY150780–87, AY150789–90, AY150792–93, AY150796–97, and AY301039–AY301061.

[1]*Corresponding author:* Department of Ecology and Evolution, University of Chicago, 1101 E. 57th St., Chicago, IL 60637.
E-mail: mlong@midway.uchicago.edu

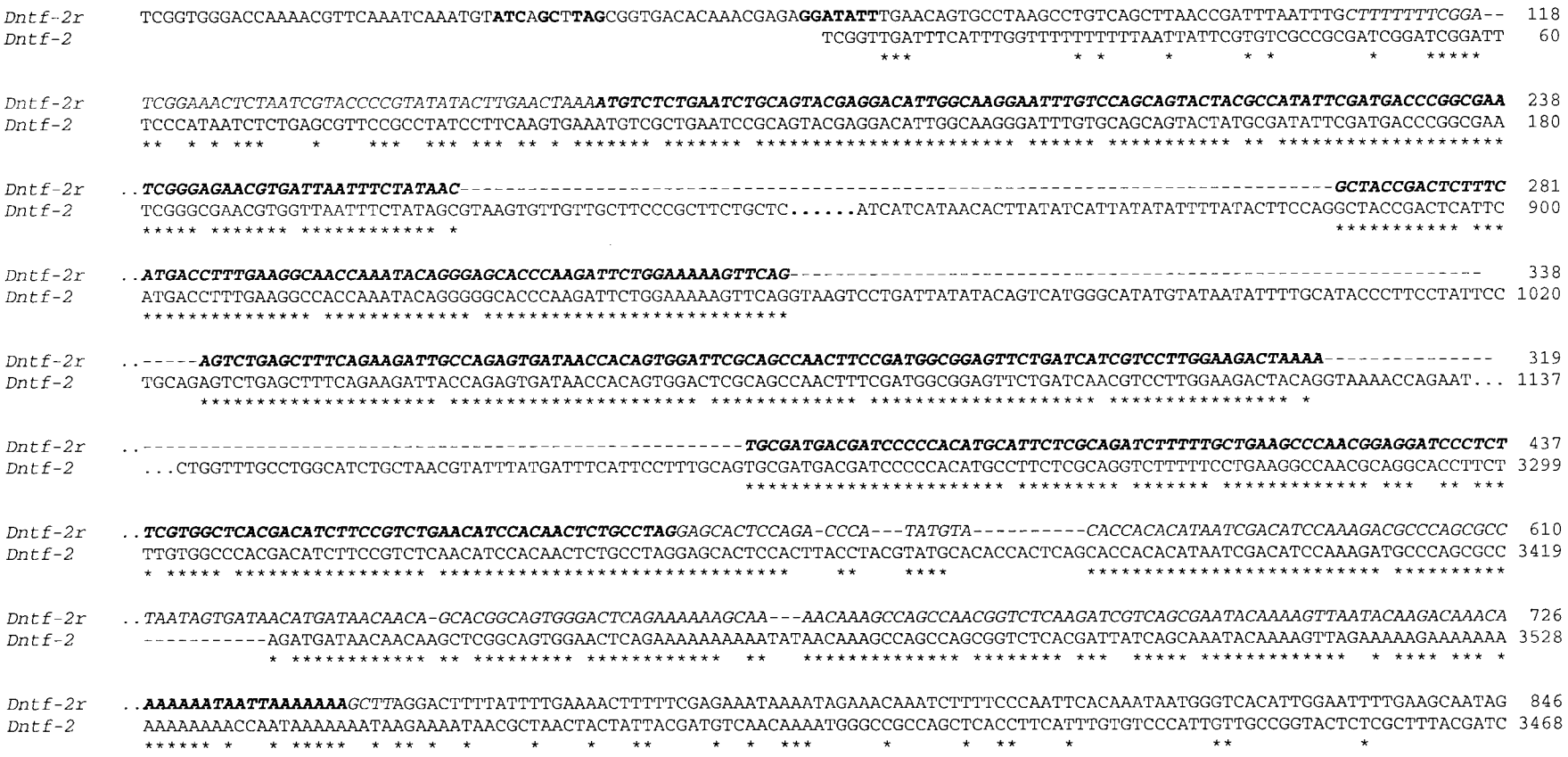

FIGURE 1.—Alignment of the *Dntf-2r* gene with its parental gene *Dntf-2* for *D. melanogaster*. The *Dntf2r* mRNA sequence in this species is shown in italics. The putative promoter, coding region, and poly(A) tract of *Dntf2r* are shown in boldface type.

D. melanogaster, D. simulans, D. sechellia, and *D. mauritiana*, suggesting that *Dntf-2r* originated recently, 3–12 MYA (Lachaise *et al.* 1988). Analysis of polymorphism and divergence for the new and parental genes reveals that *Dntf-2r* is a new functional gene whose protein has been subject to both purifying and positive Darwinian selection. Analyses of expression patterns in *D. melanogaster* indicate that *Dntf-2r* is male specific, in contrast to the wide expression of *Dntf-2*.

MATERIALS AND METHODS

Phylogenetic distribution of *Dntf-2r*: The phylogenetic distribution of *Dntf-2r* was determined by FISH to polytene chromosomes, Southern analysis, and polymerase chain reactions on the species of the *melanogaster* subgroup: *D. melanogaster, D. simulans, D. sechellia, D. mauritiana, D. yakuba, D. teissieri, D. erecta*, and *D. orena*. A probe of ~300 bp comprising the coding region of *Dntf-2r* was hybridized to polytene chromosomes of *D. melanogaster, D. simulans, D. yakuba*, and *D. erecta* following the Wang *et al.* (2000) protocol. The probe was synthesized by PCR from *Dntf-2r* and digoxygenin (DIG) labeled by random priming after gel purification of the specific band. For Southern analysis, 2 μg of genomic DNA from species of the *melanogaster* subgroup were digested and blotted to nylon membrane (Sambrook *et al.* 1989). Hybridizations and detection were carried out following the immunochemiluminescent protocol of Roche (Indianapolis). We did PCR with flanking and homologous primers for the *Dntf-2r* gene. Primer sequences were 5′ GCAGGGCGCATTGTTCAG 3′ and 5′ CATACGCCTGC CAATACGAGT 3′ to amplify *Dntf-2r* from its flanking region and 5′ TTGTCCAGCAGTACTACGCC 3′ and 5′ AGCCAC GAAGAGGGATCCTC 3′ to amplify *Dntf-2r* from its coding region.

DNA samples and sequencing: Single male fly genomic DNA was obtained using a Puregene kit. *Dntf-2r* and *Dntf-2* were amplified by PCR from this genomic DNA. *D. melanogaster* samples come from a worldwide distribution: 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 the United States; Rio from Brazil; and Rinanga, Bdx, Besançon, Prunay, and Capri from France. Other stocks used were *D. simulans* from Florida (provided by J. Coyne), *D. sechellia* (provided by J. Coyne), *D. mauritiana* (163.1, Lemeunier and Ashburner 1976), and *D. yakuba* (115, Lemeunier and Ashburner 1976). Oligoprimers 5′ ATCGGATCGGATTTC CCATAATCT 3′/5′ TGCCGAGCTTGTTGTTATCATCTG 3′ and 5′ CTGGCGGCCCATTTTGTTGACA 3′/5′ AGAAAAGT CGTCCCGAGCGAGGAA 3′ were used to amplify *Dntf-2* in *D. melanogaster* and *D. yakuba* (outgroup sequence; see below) and 5′ TGCAGGGCGCATTGTTCAG 3′ and 5′ CATACGCCT GCCAATACGAGT 3′ to amplify *Dntf-2r* in *D. melanogaster, D. simulans, D. sechellia*, and *D. mauritiana*, the four species where the gene is present. Haplotypes were obtained by direct sequencing for *Dntf-2* since it is on the X chromosome. Nine alleles of *D. melanogaster Dntf-2* were sequenced. For *Dntf-2r*, on the second chromosome, PCRs of individuals heterozygous at more than one site were cloned into a TOPO cloning vector (Invitrogen, San Diego) and one clone was sequenced to infer the haplotype. Only 1 allele randomly chosen in heterozygous individuals was analyzed, giving a total of 26 alleles. PCR products were sequenced directly after purification [QIAGEN (Valencia, CA) kit] on an ABI automated DNA sequencer (Applied Biosystems, Foster City, CA), using fluorescent DyeDeoxy terminator reagents.

Sequence analysis: Sequences were aligned by means of Clustal W (Thompson *et al.* 1994). Polymorphism and divergence patterns were studied in *Dntf-2* and *Dntf-2r* to determine the relative importance of natural selection and drift in the evolution of these genes.

Synonymous and nonsynonynous substitutions per site (K_S and K_A) were computed following Goldman and Yang (1994) and Yang (1998), using PAML 3.1 software (Yang 1997). This method accounts for transition/transversion bias (κ), for different base frequencies at different codon positions, and for the genetic code structure (Goldman and Yang 1994; Yang 1998). A model of a single rate for all sites was specified ($\omega = 0$; Yang 1997, 1998). For the analysis, a tree with *Dntf-2* of *D. yakuba* as outgroup (Figure 2A) was used. This tree was obtained by considering the age of the gene (see results) and the phylogenetic information (Ting *et al.* 2000). K_A/K_S ratio differences in different lineages were tested using the maximum-likelihood ratio test. Log likelihoods of different models were compared with a χ^2 distribution with as many degrees of freedom as the difference in number of variable parameters of the nested models (Yang 1998). Maximum-likelihood estimates of parameters for each branch (branch length and $\omega = K_A/K_S$) together with the estimate of κ can be used to calculate K_A and K_S per branch and construct nonsynonynous and synonymous trees.

π, the average number of nucleotide differences per site between two random sequences (Tajima 1989), and θ_W, Watterson's estimate of θ from the number of segregating sites (Watterson 1975), were calculated. Both values estimate the equilibrium neutral parameter $\theta = 4N_e\mu$ for autosomal loci and $\theta = 3N_e\mu$ for X-linked loci, where N_e is the effective population size and μ is the neutral mutation rate. The difference between π and θ_W (Tajima's D) reveals nonequilibrium conditions in the history of the sample. Tajima's D (Tajima 1989) was calculated and tested by 10,000 simulations using DNAsp 3.53 (Rozas and Rozas 1999). Fay and Wu's H test (Fay and Wu 2000) was also applied to the polymorphism data of *Dntf-2r*. The H statistic was used to measure the excess of derived variants at high frequency, a hallmark of recent positive selection (Fay and Wu 2000; Otto 2000). Fay and Wu's H test was computed at http://crimp.lbl.gov/htest.html and tested by 10,000 simulations. Recombination rate (R per gene) for all those simulations was estimated from the data using DNAsp 3.53 (Rozas and Rozas 1999).

Under neutrality, intraspecific variation is correlated with interspecific divergence (Kimura 1983). Deviations from this expectation can result from a number of causes including positive Darwinian selection (McDonald and Kreitman 1991; Nielsen 2001). We compared intraspecific variation with interspecific divergence at synonymous and replacement sites (McDonald and Kreitman 1991). DNAsp 3.53 software was used to carry out this comparison (Rozas and Rozas 1999).

Expression analysis: Tissues were homogenized and total RNA was prepared, as described by the QIAGEN protocol, from ~200 males and females, 15 virgin females, 15 gonadectomized males, 100 testes plus accessory glands, and 100 testes of *D. melanogaster*. Gonadectomized males (males from which we removed testes and accessory glands), testes plus accessory glands, and testes were obtained by dissecting mature males in saline solution. After dissection, tissues were preserved in RNA-later solution (Ambion, Austin, TX) at −20° after soaking them at 4° overnight until they were processed. mRNA was prepared from the total RNA of ~200 males and females following the QIAGEN protocol.

The full-length sequence of the *D. melanogaster Dntf-2r* transcript from testis was obtained by 5′ and 3′ rapid amplification

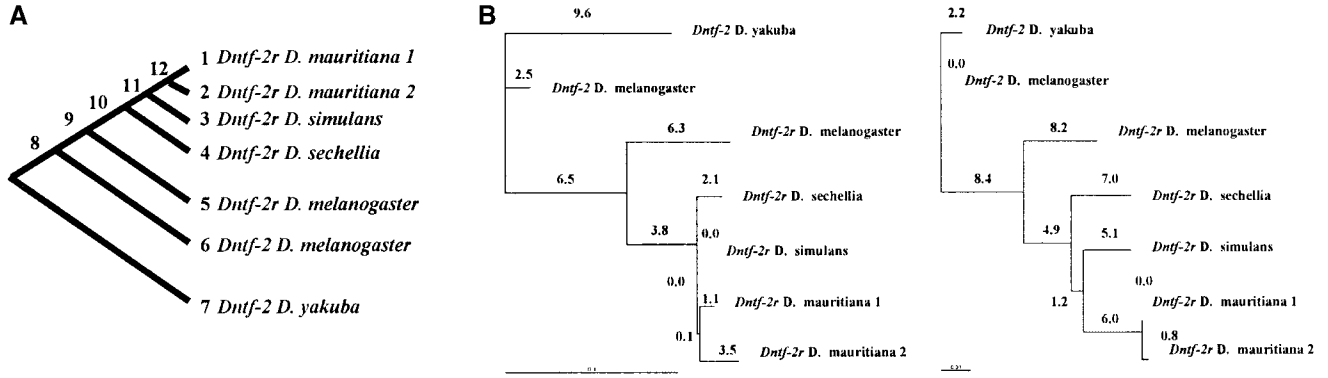

FIGURE 2.—(A) Gene and species genealogy used in the analyses of K_A/K_S. (B) K_S (left) and K_A (right) divergence tree for *Dntf-2r* and *Dntf-2* estimated under the "6 K_A/K_S ratio" model (Table 3). Estimated numbers of substitutions are shown in every branch.

of cDNA ends (RACE) experiments. Single-strand cDNA was synthesized from mRNA using Superscript (GIBCO-BRL, Gaithersburg, MD). Oligo(dT) was used to prime the synthesis of the 3′ end of the cDNA. Oligo(dT) and the specific primers 5′ TTGTCCAGCAGTACTACGCC 3′ and 5′ TCGTCCTTGGA AGACTAAAA 3′ were used to PCR amplify the 3′ end. The nested PCR product was subcloned and sequenced. Primer 5′ AGCCACGAAGAGGGATCCTC 3′ was used to synthesize the 5′ end of the *Dntf-2r* cDNA. This cDNA was tailed with dCTP by using terminal transferase (GIBCO-BRL 5′ race system). Oligo(dG) adaptors and the nested primers 5′ TTGGG CTTCAGCAAAAAGAT 3′ and 5′ GGGGATCGTCATCGCA TTT 3′ were used to PCR amplify the 5′ end of the cDNA (GIBCO-BRL 5′ race system).

RT-PCR was conducted on total RNA from virgin females, gonadectomized males, testes plus accessory glands, and testes for *Dntf-2r* and *Dntf-2*. Analysis of expression of intronless genes (such as *Dntf-2r*) is challenging because genomic contamination can produce a band of the same size as that expected from the cDNA. Therefore, we digested possible contaminating DNA from the total RNA (DNase I amplification grade; GIBCO-BRL) and ran controls including DNA-digested total RNA without retrotranscriptase. Single-strand complementary DNA (cDNA) was synthesized using Superscript and oligo(dT) (GIBCO-BRL). RT-PCR was carried out using specific primers 5′ TTGTCCAGCAGTACTACGCC 3′/5′ AGC CACGAAGAGGGATCCTC 3′ for *Dntf-2r* and 5′ TTGTGCAG CAGTACTATGCG 3′/5′ GGCCACAAAGAAGGTGCCTG 3′ for *Dntf-2*.

RESULTS

Structure of *Dntf-2r*: The complete *D. melanogaster* *Dntf-2r* transcript is given in Figure 1. The transcript consists of the retroposed regions and recruits seven additional nucleotides from its 5′ flanking region and four nucleotides from its 3′ flanking region. Unlike *jingwei* (LONG and LANGLEY 1993), *Dntf-2r* did not recruit any new coding region. Note that a nonconsensus polyadenylation signal must be used for this gene.

Phylogenetic distribution of *Dntf-2r*: We dated the appearance of *Dntf-2r* by establishing which species have the duplication, using several complementary techniques. Figure 3A shows polytene *in situ* hybridization in *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. erecta*. Positive hybridization of *Dntf-2r* probe in band 2L36F is shown in *D. melanogaster* and *D. simulans*. Two additional signals (not shown) were observed in the *D. melanogaster* and *D. simulans* genome corresponding to *Dntf-2* (X chromosome) and a lighter secondary signal (3R). Only two hybridization signals were observed in *D. yakuba* and *D. erecta*, in the X and 3R. Both signals are shown in *D. erecta* in addition to the lack of hybridization in 36F (2R in this species due to a pericentric inversion; see ASHBURNER 1989). Only *Dntf-2* (X chromosome) hybridization is shown in *D. yakuba*.

Southern blot analysis (Figure 3B) shows extra strong bands in *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia* corresponding to *Dntf-2r*. Figure 3, C and D, shows PCR with primers in the flanking and coding regions, respectively. A short product (lacking the *Dntf-2r* insertion) was obtained for *D. yakuba*, *D. teissieri*, and *D. erecta* (Figure 3C). The products from *D. yakuba*, *D. teissieri*, and *D. erecta* were sequenced. The sequence confirmed that this short fragment corresponds to the flanking region of *Dntf-2r* (Figure 4). In addition, primers in the coding region were unable to amplify *Dntf-2r* in *D. yakuba*, *D. teissieri*, and *D. erecta* (Figure 3D).

These data established that the distribution of *Dntf-2r* is limited to the four species in the *D. melanogaster* clade: *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana*. Therefore, the *Dntf-2r* gene is between 3 and 12 million years old (*i.e.*, the time length from the common ancestor of all *D. melanogaster* subgroup species to the four-species clade; see LACHAISE *et al.* 1988).

Sequence analysis: Sequence variants in the coding region for *Dntf-2* and *Dntf-2r* in related species are shown in Table 1, and Table 2 shows variants for the noncoding region of *Dntf-2* in *D. melanogaster*.

Divergence analyses were carried out using the consensus sequence for *D. melanogaster* and two alleles (haplotypes 1 and 2) for *D. mauritiana* (Table 2). Log-likelihood values and maximum-likelihood estimates of the

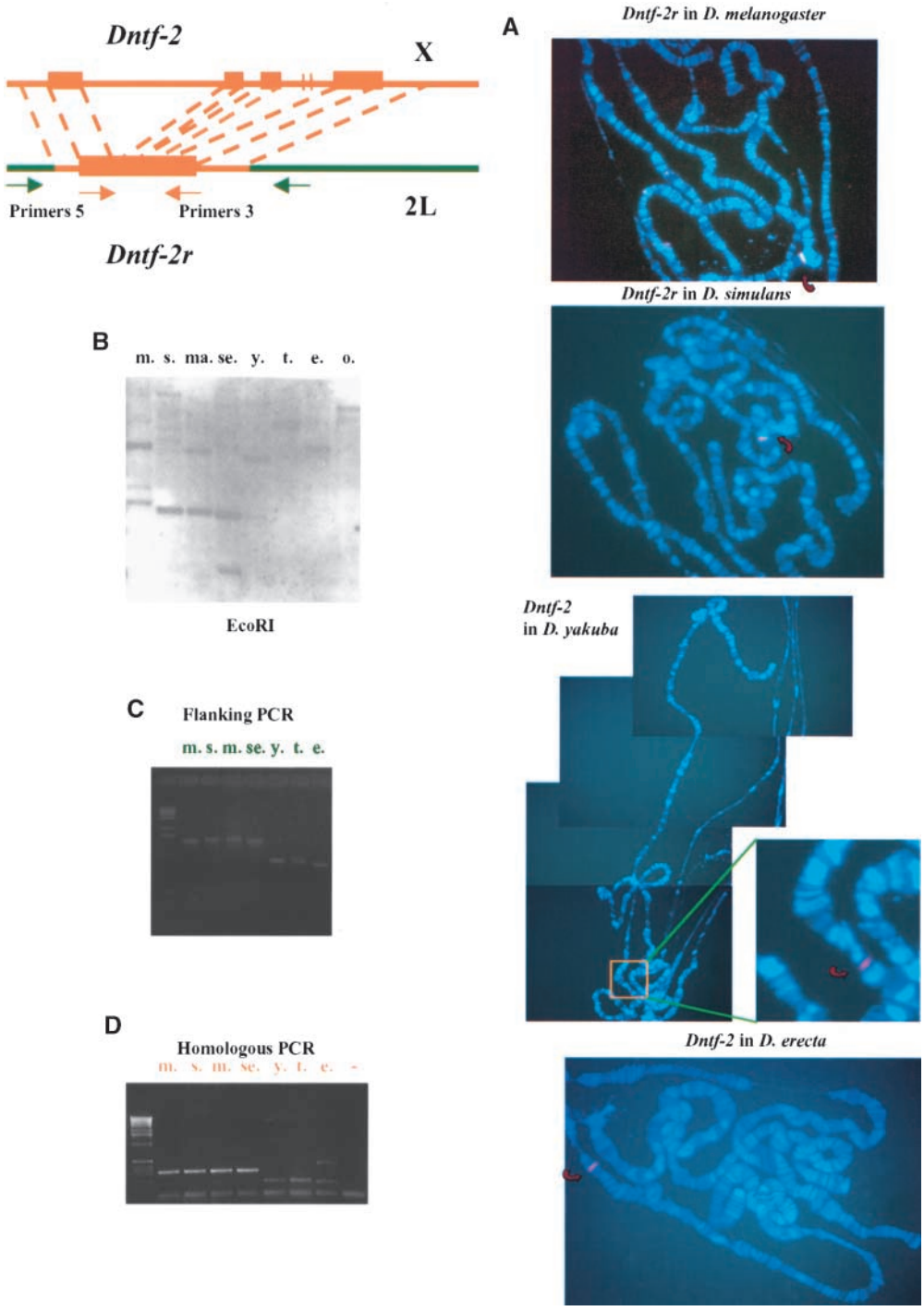

FIGURE 3.—(A) Polytene *in situ* hybridization with a probe of *Dntf-2r* in *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. erecta*. (B) Southern blot hybridized with a probe of *Dntf-2r*. Species name is shown at top of each lane. *Eco*RI digest is shown. (C) *Dntf-2r* was amplified with flanking region primers. One band of the same size was obtained in *D. melanogaster*, *D. sechellia*, *D. simulans*, and *D. mauritiana* (Table 1). A shorter band was obtained for *D. yakuba*, *D. teissieri*, and *D. erecta*. This band was sequenced in *D. yakuba*, *D. teissieri*, and *D. erecta*. Sequences aligned well with the flanking sequence of *Dntf-2r* but the gene insert (new gene) was missing (Figure 4). (D) *Dntf-2r* was amplified with primers from the coding region. One band of the same size was obtained in *D. melanogaster*, *D. sechellia*, *D. simulans*, and *D. mauritiana*. No band of the expected size was obtained for *D. yakuba*, *D. teissieri*, and *D. erecta*.

K_A/K_S ratio for each branch of the tree for *Dntf-2r* and *Dntf-2* sequences (Figure 2A) under several models are given in Table 3. A free-ratio model (B) was first applied to the data (YANG 1998). This model with 23 parameters differs significantly from the one-ratio model (A) with 13 parameters ($\ln L_B = -935.41$, $\ln L_A = -948.38$; $X^2_{(10)} = 2(\Delta \ln L) = 25.94$; $P = 0.0038$). Thus, we conclude that $\omega(K_A/K_S)$ differs among different branches of the tree. However, model B does not differ significantly from model C, the six-ratio model ($X^2_{(5)} = 0.12$; $P > 0.05$). So, the six-ratio model is the simplest model that still contains all the information from the free model. Figure 2B shows the estimated numbers of synonymous and replacement substitutions per branch under model C. Now that we know that there are differences in K_A/K_S ratios along the tree, we want to answer two questions. Is DNTF-2R evolving faster than DNTF-2? If so, is positive Darwinian selection acting in any of the *Dntf-2r* lineages? We compared different models to answer these questions.

Model A *vs.* C [$X^2_{(5)} = 25.82$; $P = 0.0001$], A *vs.* D [$X^2_{(1)} = 11.6$; $P = 0.0007$], and A *vs.* F [$X^2_{(2)} = 19.56$;

FIGURE 4.—Alignment of the flanking region of *Dntf-2r* of *D. melanogaster* with the short product (lacking the *Dntf-2r* insertion) obtained from *D. yakuba*, *D. teissieri*, and *D. erecta*. The first and last codons of *Dntf-2r* are shown in boldface type.

TABLE 1

DNA sequence variation in the coding region of *Dntf-2r* and *Dntf-2*

	1 684	223 573	345 857	566 905	677 604	788 956	999 247	111 001 570	111 111 467	111 233 059	111 456 036	112 880 392	222 002 365	222 333 479	222 456 490	223 770 790	333 000 135	333 112 391	333 223 480	333 334 590	333 444 356	333 455 847	33 67 92	No.
Dntf-2																								
D. yakuba	GTC	GCC	GGT	CGT	CGC	AGC	TGA	CGC	AAC	TGC	AAC	TCA	CGC	TCT	GAA	CGC	GCA	GGC	GGC	CCA	TCT	CCC	CG	1
D. melanogaster	...	...	...	...	...	...	...	T..	C..	AA.	.G.	...	...	AT.	...	..T	..T	.T.	...	...	...	T..	TC	9
Dntf-2r																								
D. sechellia	...	.AA	.CC	.C.	G.T	..A	C..	T.T	C..	.A.	..T	CA.	GTT	AT.	.TT	AAT	T.T	ATG	.CG	GAT	...	G..	.C	1
D. simulans	.C.	AAA	.CC	.C.	...	..A	..C	T..	C..	.A.	G..	C..	ATT	AT.	.TC	AAT	T.T	ATG	.CG	GAT	..A	T..	TC	1
D. mauritiana																								
Haplotype 1	...	.AA	.CC	ACC	...	T.A	..C	T..	CTA	.A.	...	C..	ATT	AT.	ATT	AAT	T.T	ATG	.CG	GAT	.T.	T..	TC	1
Haplotype 2	...	.AA	.CC	ACC	.T.	T.A	..C	T..	CTA	.A.	...	C..	ATT	AT.	ATT	AAT	T.T	ATG	.C.	GAT	...	..T	TC	1
D. melanogaster																								
Haplotype 1	T.T	...	ACC	.C.	...	..A	.A.	TA.	C.T	.AA	...	..G	..T	ATC	..T	AAT	.AT	ATG	.C.	GAT	C..	.T.	T.	9
Haplotype 2	T.T	...	ACC	.C.	...	..A	.A.	TA.	C..	.AA	...	..G	..T	ATC	..T	AAT	.AT	ATG	.C.	GAT	C..	.T.	T.	9
Haplotype 3	T.T	...	ACC	.C.	...	..A	.A.	TA.	C.T	.AA	...	..G	..T	AT.	..T	AAT	..T	ATG	AC.	GAT	C..	.T.	T.	2
Haplotype 4	T.T	...	ACC	.C.	...	..A	.A.	TA.	C.T	.AA	...	..G	..T	ATC	..T	AAT	..T	ATG	.C.	GAT	C..	.T.	T.	1
Haplotype 5	T.T	...	ACC	.C.	...	.AA	.A.	TA.	C..	.AA	...	..G	..T	ATC	..T	AAT	.AT	ATG	.C.	GAT	C..	.T.	T.	1
Haplotype 6	T.T	...	ACC	.C.	...	..A	.A.	TA.	C..	.AA	...	..G	..T	AT.	..T	AAT	..T	ATG	.C.	GAT	C..	.T.	T.	1
Haplotype 7	T.T	...	ACC	.C.	...	..A	.A.	TA.	C..	.AA	...	.TG	..T	ATC	..T	AAT	.AT	ATG	.C.	GAT	C..	.T.	T.	1
Haplotype 8	T.T	...	ACC	.C.	...	.AA	.A.	TA.	C..	.AA	...	..G	..T	AT.	..T	AAT	..T	ATG	.C.	GAT	C..	.T.	T.	1
Haplotype 9	T.T	...	ACC	.C.	...	..A	.A.	TA.	C.T	.AA	...	..G	..T	ATC	..T	AAT	.AT	ATG	.C.	GAT	C..	.TT	T.	1
Consensus	T.T	...	ACC	.C.	...	..A	.A.	TA.	C.T	.AA	...	..G	..T	ATC	..T	AAT	.AT	ATG	.C.	GAT	C..	.T.	T.	26

Samples are described in the text. The coding region of *Dntf-2r* and *Dntf-2* is 390 bp. Variation in noncoding regions of *Dntf-2* is given in Table 2. No., the number of times the particular haplotype has been observed.

TABLE 2

DNA sequence variation in the *Dntf-2* region

	1111111111111222222222222222
	2233556666666777777777777111344445688880144444456666777
	6834130113589147777778888147334451025991315668840049 24
D. melanogaster	0690415276070593456789012343451780447838549169684566369
253.27	CCAGAGATTTAGTAGTCATCATAACAGCACCAAGCTCACGTAGT-AGTGTTAGCC
Bdx	..--.....-....--....
yep25	..--.....AG.....-....--....
Besançon	T.....A.....TCA.A..G..-....--.C..
Rinanga	.T--..T.-A.AACT-----.AGTGT-A..A..ACGACATAC.TTG.TT
yep18	.T--..T.-A.AACT-----.AG.GT-A.....ACATAC.T-GC.T
Rio	.T--..T.-A.AACT-----.AG.GT-ATCA.A.CG..-....--..TT
Capri	.T--..T.-A.AACT-----.AG.GT-ATCA.A.CG..-....--....
yep3	.T--GA.G-.TAACT-----.AG.GT-ATC.C.....-...T--....

Samples are described in the text. Part of the third intron has not been sequenced (~400 bp). Sequence length was 2954 bp. Positions of the four exons (coding regions) are 30–137, 813–884, 953–1051, and 2764–2874. Deletions are shown with dashes.

$P = 0.00006$] tests reveal that DNTF-2 evolved much more slowly than DNTF-2R ($K_A/K_S = 0.0499$ *vs.* $K_A/K_S = 0.5405$ on average, respectively). Model C does not differ from model F [$X^2_{(3)} = 6.26$; $P > 0.05$], showing that K_A/K_S for *Dntf-2* is on average very small: ~0.0502. Thus, *Dntf-2* evolved under strong purifying selection, suggesting high functional constraint. The significantly accelerated evolution of DNTF-2R (A *vs.* D and A *vs.* F) can be the result of two different phenomena: relaxation of selection or positive Darwinian selection for novel function after the duplication. Relaxation of selection would occur if the protein product translated from *Dntf-2r* is under less constraint than the protein from *Dntf-2*. However, if amino acid substitutions in a lineage occur faster than the neutral rate, K_A/K_S ratios will exceed 1, revealing the action of positive selection.

Relaxation of selective constraints and positive selection in *Dntf-2r* can be investigated by considering additional models. Model F is significantly more likely than both model D [$X^2_{(1)} = 7.96$; $P = 0.0048$] and model E [$X^2_{(2)} = 12.48$; $P = 0.00195$], and model G is more likely than model E [$X^2_{(1)} = 10.92$; $P = 0.00095$], revealing that K_A/K_S ratios are significantly <1 in some *Dntf-2r* lineages (12-1, 12-2, 9-10, 9-5, and 8-9). Thus, we see

TABLE 3

Log-likelihood values and parameters estimated under different maximum-likelihood models

Branch	A. One $\hat{\omega}$	B. $\hat{\omega}$ free	C. Six $\hat{\omega}$	D. Two $\hat{\omega}$	E. One $\hat{\omega}$	F. Three $\hat{\omega}$	G. Two $\hat{\omega}$
12-1	0.3702	0.0001	0.0001	0.5405	1.0000	0.3316	0.3286
12-2	0.3702	0.0984	0.0714	0.5405	1.0000	0.3316	0.3286
11-12	0.3702	13.3651	13.3440	0.5405	1.0000	2.2573	1.0000
11-3	0.3702	∞	∞	0.5405	1.0000	2.2573	1.0000
10-11	0.3702	∞	∞	0.5405	1.0000	2.2573	1.0000
10-4	0.3702	1.0285	1.0283	0.5405	1.0000	2.2573	1.0000
9-10	0.3702	0.4319	0.4024	0.5405	1.0000	0.3316	0.3286
9-5	0.3702	0.4076	0.4024	0.5405	1.0000	0.3316	0.3286
8-9	0.3702	0.3812	0.4024	0.5405	1.0000	0.3316	0.3286
8-6	0.3702	0.0001	0.0001	0.0499	0.0508	0.0502	0.0497
8-7	0.3702	0.0625	0.0714	0.0499	0.0508	0.0502	0.0497
p	13	23	18	14	13	15	14
l	−948.38	−935.41	−935.47	−942.58	−944.84	−938.60	−939.38
$\hat{\kappa}$	1.60492	1.62703	1.62606	1.61576	1.81635	1.61931	1.56083

See Figure 2 for branch definition. *p* is the number of parameters in the model, *l* is the log likelihood of the model, $\hat{\kappa}$ is the estimated transition/transversion ratio, and $\hat{\omega}$ is the estimated K_A/K_S ratio for the branch in a given model (Yang 1998). See text for the comparisons of the likelihood of the models. Model C (six-ratio model) was chosen as the simpler model that retains all information from the free model. Models D and E allow only two ratios, one for a parental gene and one for a new locus. Model E sets the new locus to be a pseudogene. Model F allows a different rate for the fast-evolving and slow-evolving branches of the new locus. Model G sets the new locus to be a pseudogene in the fast-evolving branches.

TABLE 4

Polymorphism analysis of *Dntf-2r* and *Dntf-2*

Gene	*L* (bp)	*N*	*S*	π_T	θ_T	π_R	θ_R	π_S	θ_S
Dntf-2r	390	26	7	0.0040	0.0047	0.0014	0.0017	0.0128	0.0146
Dntf-2	2954 (390 coding)	9	37	0.0053	0.0046	0.0000	0.0000	0.0000	0.0000

L, length of sequenced part of the gene; *N*, number of alleles sequenced; *S*, segregating sites (only point mutations); π, average nucleotide pairwise differences and θ, Watterson's (1975) estimator of $4N\mu$ for autosomal genes or $3N\mu$ for X-linked genes; subscripts T, R, and S mean all sites, replacement sites, and silent sites in the coding regions, respectively.

clear effects of purifying selection in these branches ($K_A/K_S \sim 0.33$), indicating functional constraint for this gene. However, K_A/K_S ratios could be larger than or equal to one in segments 11-12, 11-3, 10-11, and 10-4 because model F is not a significant improvement over model G [$X^2_{(1)} = 1.56$; $P > 0.05$]—which does not support the action of positive selection. Significance of the other likelihood-ratio tests remains after correcting for multiple comparisons (Bonferroni correction, $P < 0.005$; Sokal and Rohlf 1995).

We have shown that the K_A/K_S ratio that maximizes the likelihood is ~0.33 in some *Dntf-2r* lineages and ≥1.0000 in others. However, these values do not discriminate between the alternatives of relaxation of selection or positive selection on *Dntf-2r*. This is because the boundary $K_A/K_S > 1$ sets a high threshold for testing positive selection (Kreitman and Akashi 1995; Wyckoff *et al.* 2000). Only a part of the protein is likely to be susceptible to advantageous mutations while the remainder remains subject to purifying selection.

Polymorphism data (Tables 1 and 2) were analyzed next. Levels of variation at synonymous and nonsynonymous sites and sites in noncoding regions were calculated (Table 4). The *Dntf-2* sequence is variable only in noncoding regions, confirming the action of strong purifying selection on its coding region. Tajima's *D* for *Dntf-2* was 0.7416 ($P > 0.10$); *i.e.*, the frequency spectrum shows no deviation from neutrality.

We detected variation in the coding region for *Dntf-2r* in *D. melanogaster* (Tables 1 and 4). Tajima's *D* for *Dntf-2r* was −0.45276 ($P > 0.10$) and Fay and Wu's *H* was −1.8338 ($P = 0.0752$; assuming the value of back mutation of 0.10 and recombination rate estimated for the data, 0.0344 per base, and using *Dntf-2* of *D. melanogaster* as ancestral sequence). Although the frequency spectrum of *Dntf-2r* variation in the coding region shows no deviation from neutrality, the negative Fay and Wu's *H* is at a marginal level of significance, suggesting that the three derived sites (117, 239, and 303) are at high frequency. Thus, these derived alleles may have been driven by the effects of positive Darwinian selection. While these tests can detect selection, they have power to detect it only for a short time (~$0.5N$ generations in the favorable case of no recombination; Simonsen *et al.* 1995; Fay and Wu 2000). This is equivalent to only ~0.05 MY if we consider 10^6 as the population size and 10 generations per year for Drosophila.

The McDonald-Kreitman test (McDonald and Kreitman 1991), which considers both within- and between-species variation, was next performed for the *Dntf-2r* data. Table 5 shows the results for three comparisons with *D. melanogaster*. The comparisons (*D. melanogaster vs. D. mauritiana*, *D. melanogaster vs. D. simulans*, and *D. melanogaster vs. D. sechellia*) are not independent (see Figure 2). However, if we consider the comparison in which we have polymorphism data for both species (*D. melanogaster vs. D. mauritiana*), we see the most significant pattern ($P = 0.0072$). The significance remains after Bonferroni correction for multiple comparisons ($P < 0.017$; Sokal and Rohlf 1995). We also made a comparison pooling all the independent information we have in the four species: divergence, mapped in every independent lineage from Figure 2B ($R/S = 32/12$), and polymorphism from Table 5 ($R/S = 3/8$). This test

TABLE 5

***Dntf-2r* McDonald-Kreitman test**

	Substitutions			Polymorphic sites	
	D.m-D.si	*D.m-D.se*	*D.m-D.ma*	*D.m*	*D.m + D.ma*
Silent	7	8	6	5	8
Replacement	18	18	18	2	3

$G_{\text{value }(D.m\text{-}D.si)} = 4.317$, $P = 0.0377$; $G_{\text{value }(D.m\text{-}D.se)} = 3.779$, $P = 0.0519$; and $G_{\text{value }(D.m\text{-}D.ma)} = 7.228$, $P = 0.0072$ for the comparison between *D. melanogaster* (*D.m*) *vs. D. simulans* (*D.si*), *D. sechellia* (*D.se*), and *D. mauritiana* (*D.ma*), respectively.

shows a highly significant excess of amino acid replacements ($G = 7.648$; $P = 0.0057$). The significantly higher ratios of replacement to silent substitutions compared to the ratios of replacement to silent polymorphic sites are consistent with positive selection in the *Dntf-2r* lineages.

***Dntf-2r* expression and promoter analysis:** RT-PCR results for *Dntf-2r* and *Dntf-2* in different tissues of *D. melanogaster* are shown in Figure 5. We differentially amplified the two genes with specific primers. We observed that while *Dntf-2* is expressed in all tissues studied, *Dntf-2r* is expressed only in testes.

The β*2-tubulin* gene has a late testis-specific promoter (Michiels *et al.* 1989). This promoter is known to exhibit only two elements: β*2-tubulin* upstream element 1 (β*2UE1*; 14 bp), which is essential for spermatocyte-specific expression, and a quantitative element (7 bp; Michiels *et al.* 1989). Interestingly, we identified a region of sequence similarity with these late testis-specific promoter elements of β*2-tubulin* at −42 bp from the transcription initiation site of *Dntf-2r* in *D. melanogaster*. *Dntf-2r* putative promoter elements, upstream element (ATCAGC-TTAGCGGT −62) and quantitative element (GGATATT −42), have a 57% nucleotide identity with the β*2UE1* (**ATC-GC**AG**TAG**TCTA) and a 100% identity with the β*2-tubulin* quantitative element (**GGAT ATT**) of *D. hydei*, respectively.

Examination of the flanking region of the insertion site in *D. teissieri, D. yakuba*, and *D. erecta* (outgroups lacking the insertion; Figure 4) reveals similarity to the 5′ putative promoter sequence of *Dntf-2r* in *D. melanogaster*. The GGATATT putative quantitative element is present in these outgroup sequences as well as three nucleotides TAG of the putative *Dntf-2r* upstream element (see Figure 4). This would favor the hypothesis that *Dntf-2r* developed a new promoter with late testis expression after retroposition by acquiring only a few modifications to the preexisting 5′ sequence.

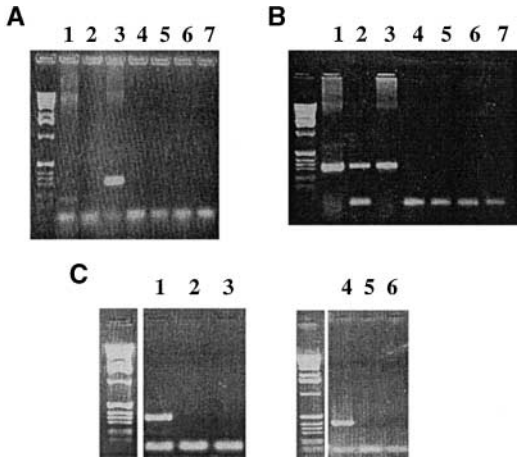

Figure 5.—RT-PCR for (A) *Dntf-2r* and (B) *Dntf-2* in *D. melanogaster*. Lane 1, PCR from female cDNA; lane 2, gonadectomized male cDNA; lane 3, testes plus accessory gland cDNA; lanes 4–6, the respective negative controls after DNA digestion; and lane 7, negative control of the PCR. (C) RT-PCR from testes cDNA for *Dntf-2r* and *Dntf-2*. Lane 1, testes cDNA; lane 2, negative control after DNA digestion; lane 3, the negative control of the PCR for *Dntf-2r*; lane 4, testes cDNA; lane 5, negative control after DNA digestion; and lane 6, negative control of the PCR for *Dntf-2*.

DISCUSSION

We investigated evolution of a recently originated gene and its parental copy in *D. melanogaster*, *Dntf-2r* (*CG10174*) and *Dntf-2* (*CG1740*). Sequence comparison revealed that the new gene was generated in a retroposition event. Recent work on the parental copy *Dntf-2* in *D. melanogaster* revealed that this gene, playing a role in the nuclear transport of proteins with nuclear localization signals, is essential for the antimicrobial immune response (Bhattacharya and Steward 2002). This important role is in agreement with the strong sequence constraint on *Dntf-2* that we observed: there was no variation in the coding region within *D. melanogaster* and the K_A/K_S ratio was ∼0.05 between *D. melanogaster* and *D. yakuba*.

On the other hand, there was no information on the function of *Dntf-2r*. Our sequence analyses of divergence and polymorphism for this gene, as well as our expression evidence, indicate that this gene may produce a functional protein. First, we observed that polymorphism is higher for synonymous than for replacement sites: $\pi_R/\pi_S = 0.11$ (Table 4), revealing the action of purifying selection. Second, K_A/K_S ratios for substitutions in *Dntf-2r* are on average significantly lower than unity (∼0.5), which is not consistent with the hypothesis that the gene is a pseudogene in many of the species. The K_A/K_S ratio of ∼0.5 for *Dntf-2r* is higher than that for *Dntf-2*. However, the McDonald-Kreitman test revealed a significant excess of amino acid substitutions, suggesting that the accelerated protein sequence evolution is likely a consequence of the action of positive Darwinian selection. Consistent with this interpretation, the Fay-Wu test, with an H statistic of marginal significance, gives a strong hint of an excess of high-frequency variants that would be coupled with the fixation of beneficial mutations. Thus, both purifying selection and adaptive evolution detected in these analyses argue that *Dntf-2r* encodes a protein, possibly with an evolving novel function. *Dntf-2r* provides new evidence for the role of positive Darwinian selection in the origin of new genes.

Whether or not a retroposed sequence recruits a new promoter is a critical step to its future fate. If a retroposed sequence integrates in a genomic region devoid of expression potential, it would be doomed to evolve into a pseudogene (Jeffs and Ashburner 1991). However, *Dntf-2r* has developed a tissue-specific expression pattern in *D. melanogaster* while *Dntf-2* is widely expressed in this species (this work and Bhattacharya and Steward 2002). Consistent with these results, we observe that

the 5′ flanking region of *Dntf-2r* in *D. melanogaster* is similar to the β*2-tubulin* promoter elements: β*2UE1*, which is essential for spermatocyte-specific expression, and the quantitative element. Although *Dntf-2r* tissue expression remains to be analyzed in *D. simulans*, *D. mauritiana*, and *D. sechellia*, a similar pattern of expression may be possible, considering that the two putative promoter elements are 100% conserved in these species (data not shown). The *Dntf-2r* retroposed sequence, as expected from its mRNA origin, did not contain the promoter sequence of the parental gene. It instead recruited a novel 5′ regulatory sequence from the insertion site, needing few mutations to become a promoter and leading to a testis-specific pattern of expression. The examination of the *D. teissieri, D. yakuba*, and *D. erecta* orthologous regions of the insertion site for *Dntf-2r* reveals an element similar to the putative promoter sequence of *Dntf-2r* in *D. melanogaster* with only a few substitutions. However, it is unclear if this previously existing sequence is a functional promoter for some unknown gene in the region or is just a random genomic sequence that happens to be similar to a promoter sequence. Given its high similarity to the promoter region of β*2-tubulin* and its similar expression site (testis), it would be tempting to take the first possibility as a working hypothesis in further research. The promoter capabilities of the 5′ preexisting sequence and the 5′ region of *Dntf-2r* in *D. melanogaster* should be further investigated to test this hypothesis. In conclusion, the *Dntf-2r* gene is a chimera: the regulatory sequences and protein-coding regions originated from different sources.

An accelerated rate of evolution has been widely observed in some reproduction-related genes, probably due to competition among sperm from different males, female choice, and/or intersexual genomic conflict (Eberhard 1985; Metz and Palumbi 1996; Ting *et al.* 1998; Tsaur *et al.* 1998; Aguadé 1999; Wyckoff *et al.* 2000; Swanson *et al.* 2001a,b). We can speculate that the detected positive Darwinian selection on *Dntf-2r* may be related to its newly evolved male-specific function(s). On the other hand, *Dntf-2* is also expressed in the germline. The strong purifying selection on this parental gene could be a consequence of its expression in other tissues, a possibly different timing in germline expression, or a different function.

It is known that, in Drosophila, X inactivation occurs early in spermatogenesis (Lifschytz and Lindsley 1972). This implies that *Dntf-2* (on the X chromosome) is inactivated at an early stage in spermatogenesis. *Dntf-2r* on an autosome and expressed in male germline stages could carry out newly evolved function(s) in spermatogenesis. Although this is a plausible interpretation, an alternative scenario ought to be discussed: *Dntf-2r* maintains the functions of its X-linked parental copy, *Dntf-2*, in male germline cells after X inactivation. In this scenario, the signature of positive selection observed in the *Dntf-2r* sequence must be explained by its being subject to a new set of selective pressures in a specific testis tissue despite maintaining the same function. However, this explanation would encounter a difficulty. X inactivation in Drosophila evolved before *Dntf-2r* originated, since there is evidence for X inactivation in *D. pseudoobscura* (Lifschytz and Lindsley 1972), which diverged from the *melanogaster* subgroup ~40 MYA (Powell 1997). Lifschytz and Lindsley (1972) observed that in the spermatocytes of *D. pseudoobscura*, the arm of the X homologous to the *D. melanogaster* X chromosome is heteropycnotic, whereas the other arm homologous to the right arm of the *D. melanogaster* second chromosome does not show heteropycnosis. Thus, there would be a long period (at least 30 MY) from the emergence of the X inactivation to the origin of *Dntf-2r*, during which the putative *Dntf-2* function would be silenced in late male germline stages. Silencing of a previously existing function for such a long evolutionary period might not be tolerable. Thus, it is more parsimonious to assume that the *Dntf-2r* does not maintain the same function as its parental copy. The observed accelerated substitution in the protein encoded by *Dntf-2r* is more likely a consequence of positive selection for novel function.

We thank J. Coyne, P. Gibert, F. Lemeunier, and M.-L. Wu for providing Drosophila strains used in this work; Janice Spofford for critically reading the manuscript; and members of the Long lab, especially Kevin Thornton, for valuable discussions. We also thank two anonymous reviewers for their comments and corrections that improved the manuscript. This work was supported by grants from National Science Foundation (Career Award) and Packard Fellowship in Science and Engineering to M.L.

LITERATURE CITED

Aguadé, M., 1999 Positive selection drives the evolution of the Acp29AB accessory gland protein in Drosophila. Genetics **152:** 543–551.

Arabidopsis Genome Initiative, 2000 Analysis of the genome sequence of the flowering plant Arabidopsis *thaliana*. Nature **408:** 796–815.

Ashburner, M., 1989 *Drosophila: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Ball, C. A., and J. M. Cherry, 2001 Genome comparisons highlight similarity and diversity within the eukaryotic kingdoms. Curr. Opin. Chem. Biol. **5:** 86–89.

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

Betrán, E., and M. Long, 2002 Expansion of genome coding regions by acquisition of new genes. Genetica **115** (1): 65–80.

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

Bhattacharya, A., and R. Steward, 2002 The Drosophila homolog of NTF-2, the nuclear transport factor-2, is essential for immune response. EMBO Rep. **3** (4): 378–383.

Blanc, G., A. Barakat, R. Guyot, R. Cooke and M. Delseny, 2000 Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell **12:** 1093–1101.

Eberhard, W. G., 1985 *Sexual Selection and Animal Genitalia*. Harvard University Press, Cambridge, MA.

Fay, J. C., and C.-I Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics **155:** 1405–1413.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Gu, Z., A. Cavalcanti, F-C. Chen, P. Bouman and W.-H. Li, 2002 Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. Mol. Biol. Evol. **19:** 256–262.

Jeffs, P., and M. Ashburner, 1991 Processed pseudogenes in *Drosophila*. Proc. R. Soc. Lond. Ser. B **244:** 151–159.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge, UK.

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26:** 403–422.

Lachaise, D., M.-L. Cariou, J. R. David, F. Lemeunier, L. Tsacas *et al.*, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–225.

Lemeunier, F., and M. Ashburner, 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. Ser. B **193:** 257–294.

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

Li, W. H., Z. Gu, H. Wang and A. Nekrutenko, 2001 Evolutionary analyses of the human genome. Nature **409:** 847–849.

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

Llopart, A., J. M. Comeron, F. Brunet, D. Lachaise and M. Long, 2002 Intron presence/absence polymorphism in Drosophila driven by positive Darwinian selection. Proc. Natl. Acad. Sci. USA **99** (12): 8121–8126.

Long, M., 2001 Evolution of novel genes. Curr. Opin. Genet. Dev. **11** (6): 673–680.

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

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

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

Metz, E. C., and S. R. Palumbi, 1996 Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein binding. Mol. Biol. Evol. **13:** 397–406.

Michiels, F., A. Gasch, B. Kaltschmidt and R. Renkawitz-Pohl, 1989 A 14 bp promoter element directs the testis specificity of the Drosophila beta 2 tubulin gene. EMBO J. **8:** 1559–1565.

Nielsen, R., 2001 Statistical tests of selective neutrality in the age of genomics. Heredity **86:** 641–647.

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

Nurminsky, D., D. De Aguilar, C. D. Bustamante and D. L. Hartl, 2001 Chromosomal effects of rapid gene evolution in Drosophila melanogaster. Science **291:** 128–130.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer, Berlin.

Otto, S. P., 2000 Detecting the form of selection from DNA sequence data. Trends Genet. **16:** 526–529.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology—The Drosophila Model*. Oxford University Press, New York.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. G. Miklos, C. R. Nelson *et al.*, 2000 Comparative genomics of the eukaryotes. Science **287:** 2204–2215.

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Sokal, R. R., and F. J. Rohlf, 1995 *Biometry*. Freeman, San Francisco.

Swanson, W. J., C. F. Aquadro and V. D. Vacquier, 2001a Polymorphism in abalone fertilization proteins is consistent with the neutral evolution of the egg's receptor for lysin (VERL) and positive Darwinian selection of sperm lysin. Mol. Biol. Evol. **18:** 376–383.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner and C. F. Aquadro, 2001b Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc. Natl. Acad. Sci. USA **98:** 7375–7379.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Ting, C.-T., S.-C. Tsaur, M. L. Wu and C.-I Wu, 1998 A rapidly evolving homeobox at the site of a hybrid sterility gene. Science **282:** 1501–1504.

Ting, C.-T., S.-C. Tsaur and C.-I Wu, 2000 The phylogeny of closely related species as revealed by the genealogy of a speciation gene, *Odysseus*. Proc. Natl. Acad. Sci. USA **97:** 5313–5316.

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

Tsaur, S. C., C. T. Ting and C.-I Wu, 1998 Positive selection driving the evolution of a gene of male reproduction, Acp26Aa, of Drosophila: II. Divergence versus polymorphism. Mol. Biol. Evol. **15:** 1040–1046.

Wang, W., J. Zhang, C. Alvarez, A. Llopart and M. Long, 2000 The origin of the Jingwei gene and the complex modular structure of its parental gene, yellow emperor, in Drosophila melanogaster. Mol. Biol. Evol. **17:** 1294–1301.

Wang, W., F. G. Brunet, E. Nero and M. Long, 2002 Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **99** (7): 4448–4453.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wyckoff, G. J., W. Wang and C.-I Wu, 2000 Rapid evolution of male reproductive genes in the descent of man. Nature **403:** 304–309.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13:** 555–556.

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

Yi, S., and B. Charlesworth, 2000 A selective sweep associated with a recent gene transposition in *Drosophila miranda*. Genetics **156:** 1753–1763.

Communicating editor: M. A. F. Noor