DOI: 10.1534/genetics.105.041699

A Novel Chimeric Gene, *siren*, With Retroposed Promoter Sequence in the *Drosophila bipectinata* Complex

Masafumi Nozawa, Tadashi Aotsuka and Koichiro Tamura[1]

Department of Biological Sciences, Graduate School of Science, Tokyo Metropolitan University, Hachioji, Tokyo 192-0397, Japan

Manuscript received February 6, 2005
Accepted for publication August 9, 2005

ABSTRACT

Retrotransposons often produce a copy of host genes by their reverse transcriptase activity operating on host gene transcripts. Since transcripts normally do not contain promoter, a retroposed gene copy usually becomes a retropseudogene. However, in *Drosophila bipectinata* and a closely related species we found a new chimeric gene, whose promoter was likely produced by retroposition. This chimeric gene, named *siren*, consists of a tandem duplicate of *Adh* and a retroposed fragment of *CG11779* containing the promoter and a partial intron in addition to the first exon. We found that this unusual structure of a retroposed fragment was obtained by retroposition of *nanos*, which overlaps with *CG11779* on the complementary strand. The potential of retroposition to produce a copy of promoter and intron sequences in the context of gene overlapping was demonstrated.

GENE duplication and exon shuffling are the most important mechanisms for generation of new genes during genome evolution (reviewed by Brown 1999). Gene duplication produces two identical copies of an existing gene, which provides opportunities for one of them to accumulate mutations and acquire a new function and eventually become a new gene, while the other copy retains the original function supplied by the single-copy gene before duplication (Ohno 1970; Kimura and Ohta 1974). Because of the random nature of mutation, the protein-coding sequence of a duplicated gene is usually disrupted with time and the birth of a new functional gene has been thought to be rather rare event (Ohta and Kimura 1971). However, a recent study suggested that duplicated genes are almost as likely to acquire a new function as to be lost through acquisition of mutations that compromise the function of genes after genome duplications (Nadeau and Sankoff 1997). In general, many genes belonging to the same multigene family or gene superfamily are thought to result from gene duplications (see Graur and Li 2000 for review). Therefore, the cumulative contribution of gene duplication to genome evolution seems to be substantial. In addition to the accumulation of mutations, exon (or domain) shuffling can also produce a new gene by rearranging functional domains among duplicated and/or preexisting genes (Gilbert 1978). Many genes sharing common functional domains in different combinations (*e.g.*, Rubin *et al.* 2000; Li *et al.* 2001) indicate a substantial contribution of exon shuffling to the diversity of genes and gene functions in genomes. Long *et al.* (1995) estimated that at least 19% of exons in eukaryotic genes have experienced exon shuffling in their evolutionary histories.

In contrast to the observed indirect evidence of gene duplication and exon shuffling, there have been only a few opportunities to explore the processes and mechanisms of gene duplication and exon shuffling. Among them, recent studies suggest that retrotransposons and retroposons contribute to gene duplication and exon shuffling by means of reverse transcription of expressed genes and integration of the reverse transcripts into new genomic positions (*e.g.*, McCarrey and Thomas 1987; Noyce *et al.* 1997; Betrán and Long 2003). Since a retroposed gene copy does not contain the promoter sequence that resides in the untranscribed region, it will be inactive and eventually become a retropseudogene unless a promoter sequence is newly recruited (Long *et al.* 2003). For this reason, functional retrogenes often show a chimeric structure between a retroposed coding sequence and the part of the preexisting gene that supplied the promoter sequence (Long and Langley 1993; Long *et al.* 1999; Wang *et al.* 2002; Nisole *et al.* 2004; Sayah *et al.* 2004; Jones *et al.* 2005). Because the chance that a retroposed sequence will acquire a promoter sequence from a preexisting gene without damage to its function is expected to be very low, the potential of a retroposed sequence to become a new functional gene must be quite limited (see Graur and Li 2000). In other words, the mechanism for obtaining a

Sequence data from this article have been deposited with the DDBJ/EMBL/GenBank Data Libraries under accession nos. AB194414–AB194442.

[1]*Corresponding author:* Department of Biological Sciences, Graduate School of Science, Tokyo Metropolitan University, 1-1 Minami-ohsawa, Hachioji-shi, Tokyo 192-0397, Japan.
E-mail: ktamura@evolgen.biol.metro-u.ac.jp

promoter sequence is critical to the generation of functional retrogenes.

In this study, we report a novel chimeric gene consisting of a retroposed fragment of *CG11779* (annotated in the *Drosophila melanogaster* genome of Release 3.2) and a tandem duplicate of the alcohol dehydrogenase (*Adh*) gene in species of the *D. bipectinata* complex. We named it *siren* after Greek myths, in which the Siren is a chimera of a human and a bird. Interestingly, sequence homology and expression pattern suggested that the promoter sequence of *siren* was carried by the retroposed *CG11779* fragment. This unexpected structure was produced by retroposition of *nanos*, which overlaps with *CG11779* on the complementary strand. We show a potential of retroposition to produce a copy of promoter as well as protein-coding sequences, when genes are overlapping.

MATERIALS AND METHODS

Fly species and sequence data: For *D. parabipectinata* of the *D. bipectinata* complex, we determined, using molecular cloning techniques, a nucleotide sequence of a genomic region (17,499 bp) including the *Adh* gene and a new gene (*siren*) discovered in this study (DDBJ/EMBL/GenBank accession no. AB194414) and another sequence (17,580 bp) including *CG11779* and *nanos* that overlap each other on the opposite strands (accession no. AB194415). To clarify the gene structures of these genes, cDNA sequences for *siren*, *Adh*, *CG11779*, and *nanos* were determined by using RT-PCR (Powell *et al.* 1987) and 5′- and 3′-RACE (Frohman *et al.* 1988) techniques. Partial nucleotide sequences (~4.6–6.3 kb), including *siren*, were also determined for *D. bipectinata*, *D. malerkotliana* (*D. m. malerkotliana* and *D. m. pallens*), and *D. pseudoananassae* (*D. p. nigrens* and *D. p. pseudoananassae*) belonging to the *D. bipectinata* complex of the *D. ananassae* subgroup by using PCR techniques (accession nos. AB194416–AB194420). In addition, we determined partial nucleotide sequences of *Adh* and *CG11779/nanos* for all these species as well as for *D. ananassae*, *D. varians*, and *D. merina* of the *D. ananassae* subgroup for comparative sequence analyses (accession nos. AB194421–AB194436). The nucleotide sequences of *Adh* and *CG11779/nanos* regions for *D. melanogaster* (genome sequence of Release 3.2) were downloaded from GenBank.

The fly stocks of *D. bipectinata*, *D. malerkotliana* (*D. m. malerkotliana* and *D. m. pallens*), *D. merina*, *D. parabipectinata*, and *D. varians* were maintained at Tokyo Metropolitan University, and those of *D. ananassae* (stock no. 14024-0371.0), *D. p. nigrens* (stock no. 14024-0411.0), and *D. p. pseudoananassae* (stock no. 14024-0421.0) were originally obtained from the National Drosophila Species Stock Center at Bowling Green State University and maintained in Tokyo Metropolitan University.

DNA cloning, PCR, and sequencing: We constructed a genomic library for *D. parabipectinata* with λEMBL3 phage vector (Promega, Madison, WI) and obtained clones containing *Adh* and/or *siren*. The obtained clones were subcloned into plasmid vector pUC19 according to Henikoff (1984). The nucleotide sequences were determined by using BigDye Terminator v3.1 cycle sequencing kit and ABI Prism 377 DNA Sequencer (Applied Biosystems, Foster City, CA). We also constructed a SuperCos1-cosmid (Stratagene, La Jolla, CA) library to obtain *CG11779/nanos* clones and determined the nucleotide sequence by a shot-gun sequencing method in the DNA sequencing center (National Institute of Genetics). The standard PCR method was used to determine the genomic DNA sequences for *siren*, *Adh*, and *CG11779/nanos* for other species. The nucleotide sequences of these PCR products were determined by the direct sequencing method or via cloning in the cases of heterozygotes. For the latter cases, we followed Hanahan's (1985) procedures for cloning with pUC118 or pUC119 plasmid vector and determined the nucleotide sequences for multiple clones to correct PCR errors.

The template total RNA for RT-PCR and 5′- and 3′-RACE was extracted from 5–10 individuals by Boom *et al.*'s (1990) method with modifications for complete DNA removal. ThermoScript RT-PCR system (Invitrogen, San Diego) and 5′-RACE system ver.2.0 (Invitrogen) were used for RT-PCR and 3′-RACE and for 5′-RACE, respectively. The nucleotide sequences of these PCR products were determined by the direct sequencing method or via cloning as described above.

The RT-PCR was also used for detecting the presence or absence of transcripts at larval, pupal, and adult stages in male and female individuals to examine the expression pattern of *siren*, *Adh*, and *CG11779* genes. For the case of *Adh*, the template cDNA for 10 μl of PCR mixture was reverse transcribed from 10 ng of total RNA, whereas it was obtained from 100 ng of total RNA for *siren1*, *siren2*, and *CG11779*. Ribosomal protein L32 (*RpL32*) gene was used as a control to ensure the efficiency of RT-PCR. Since the *Adh* gene has two different promoters (Savakis *et al.* 1986), the corresponding types of transcripts were discriminated by specific primers. Detailed information for the cloning probes and PCR primers is described in the supplementary material at http://www.genetics.org/supplemental/.

Genetic linkage analyses: The relative chromosomal locations among *siren*, *Adh*, and *CG11779/nanos* were examined by genetic linkage analyses among *siren1* (upstream locus for tandem duplicates of *siren*), *Adh*, and *CG11779/nanos* loci using B133 and CJB162 strains of *D. bipectinata*, which are homozygous for different alleles from each other in all three loci. A single virgin female of CJB162 and a single male of B133 were crossed in a culture vial for 7 days and a single pair of their F_1 progeny was then crossed in the same manner. From each individual of the pair of F_1 progeny and 96 (48 female and 48 male) F_2 progenies, we extracted total DNA and amplified DNA fragments by PCR for each gene. The PCR products were digested by *Mbo*I for *siren1* and *CG11779/nanos* or by *Hae*II for *Adh* and agarose-gel electrophoresed to distinguish the alleles. Accession numbers for the nucleotide sequences of *siren1*, *Adh*, and *CG11779/nanos* are AB194437–AB194439 and AB194440–AB194442 for B133 and CJB162 strains of *D. bipectinata*, respectively.

Sequence analyses: To examine the homology between the genomic sequences obtained, we performed dot-plot analyses using the Nucleic Acid Dot Plots program (http://www.vivo.colostate.edu/molkit/). The parameter set of a 15-bp window size and a 1-bp mismatch limit was used. For detailed comparisons between the homologous parts identified, a mismatch limit of 3 bp was also used.

A molecular phylogenetic tree for *siren* and *Adh* was constructed by the minimum-evolution method (Rzhetsky and Nei 1992) with synonymous distances estimated by the modified Nei-Gojobori method (Zhang *et al.* 1998) with a transition/transversion ratio equal to 2 and a Jukes-Cantor correction (Jukes and Cantor 1969) for multiple hits. The 256 homologous codons between *siren* and *Adh* were used. The statistical confidence for each branch in the tree was evaluated by the bootstrap method (Felsenstein 1985) with 1000 replications. We used the MEGA3 program (Kumar *et al.* 2004) for these analyses.

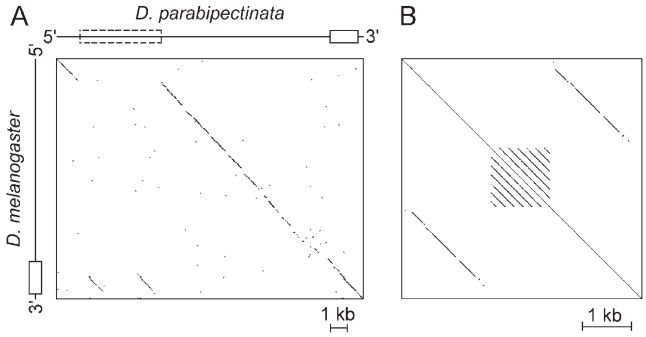

FIGURE 1.—Dot plot for *Adh* and its upstream region of *D. parabipectinata*. (A) Comparison of the genomic fragment including *Adh* and its upstream region between *D. parabipectinata* (horizontal axis) and *D. melanogaster* (vertical axis). The position of *Adh* is indicated by the open box and the 4.6-kb insertional sequence in the *D. parabipectinata* sequence is surrounded by the dashed line. (B) Comparison of the insertional sequence (position 1138–5957 in the *D. parabipectinata* genomic fragment) with itself. The parameter values used are a window size equal to 15 bp and a mismatch limit equal to 1 bp.

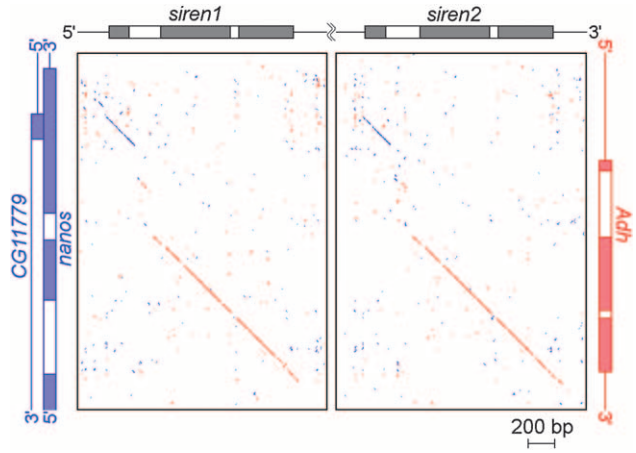

FIGURE 2.—Dot plots of *siren* and *CG11779/nanos* (blue dots) and *siren* and *Adh* (red dots) sequences of *D. parabipectinata*. The dot-plot parameters were a window size equal to 15 bp and a mismatch limit equal to 3 bp. (Top) The exon-intron structures deduced from cDNA sequences for *siren*, (left) for *CG11779/nanos*, and (right) for *Adh*, where solid and open boxes indicate exons and introns, respectively.

RESULTS

An insertional nucleotide sequence in the *D. parabipectinata* genome: We determined the nucleotide sequence of the *Adh* gene and its flanking region (17,499 bp) in the *D. parabipectinata* genome. This fragment was longer than its counterpart in the *D. melanogaster* genome sequence by 3943 bp. A dot plot of these homologous sequences indicates that the difference is largely attributed to a single insertional sequence (~4.6 kb) in the *D. parabipectinata* sequence, and >300 short (mostly <10 bp) indels are required to fill up the rest of the difference (~700 bp) (Figure 1A). Within this insertional sequence, two *Adh* homologous parts are clearly shown, suggesting that duplications involving the *Adh* sequence were responsible for the origin of the insertional sequence. The alternative possibility of a deletion in the *D. melanogaster* sequence is unlikely, since the insertional sequence does not exist in the sequence of an outgroup species (accession no. AADE01001152 for *D. pseudoobscura*). A dot plot of the insertional sequence with itself reveals the detailed structure, *i.e.*, a duplication of an ~1.5-kb segment including the *Adh* homologous sequence and a tandem direct repeat of a 180-bp unit length between the duplicates (Figure 1B). This structure was found in all other species of the *D. bipectinata* complex at the same location, while the number of repeat units varies among species (data not shown). The repeat sequence was not at all identified in the *D. melanogaster* whole-genome sequence (Release 3.2) by the standard BLAST search procedure (ALTSCHUL *et al.* 1990).

A new chimeric gene: Within the duplicated 1.5-kb segments, both GENSCAN (BURGE and KARLIN 1997) and Genie (KULP *et al.* 1996) gene-finding programs predicted the same open reading frames (ORFs). Using RT-PCR and 5′- and 3′-RACE techniques, we determined the complete cDNA sequences for both ORFs and confirmed the exon-intron structure of the genes corresponding to these two ORFs. These two genes were virtually two identical copies of the same gene. BLASTP (ALTSCHUL *et al.* 1990) showed that the second and third exons of this gene are homologous to the second and third exons of the *Adh* gene, respectively, at the translated amino-acid sequence level, as well as the nucleotide sequence level shown in Figure 1A. The remaining first exon showed the best homology to the first exon of *CG11779* in the *D. melanogaster* genome at the translated amino-acid sequence level. Therefore, this gene is a chimera consisting of the first exon of *CG11779* and the second and third exons of *Adh*. It should be noted that the encoded *siren* protein is also a chimera containing the 31-amino-acid N-terminal sequence of *CG11779* coupled onto the entire *Adh* protein sequence. As mentioned, we named it *siren* (*siren1* and *siren2* corresponding to the upstream and downstream loci, respectively) after Greek myths, following another chimeric gene, *sphinx*, found in *D. melanogaster* (WANG *et al.* 2002). In the *D. melanogaster* genome sequence, *CG11779* overlaps with another gene, *nanos*, on the complementary strand. We confirmed the same structure in the *D. parabipectinata* genome as well. A dot plot of the *siren* and *CG11779/nanos* sequences in the *D. parabipectinata* genome shows that the homology includes the 5′-untranscribed region, the first exon, and a part of the first intron of *CG11779*, which correspond to a part (mainly the 3′-untranslated region) of the third exon of *nanos* (Figure 2). The homology of *siren* to *Adh* starts after the *CG11779/nanos* homologous part and includes the second and third exons

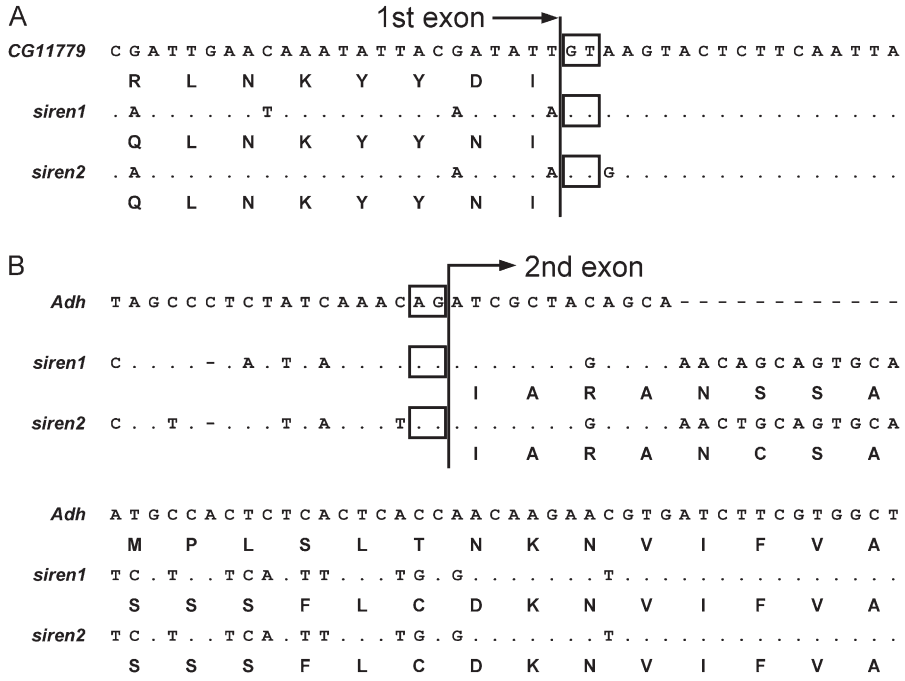

FIGURE 3.—Nucleotide and amino-acid sequence alignments for the exon-intron boundaries of *siren* in *D. parabipectinata*. (A) The boundary between the first exon and the first intron of *siren1* and *siren2* with its homologous part of *CG11779*. (B) The boundary between the first intron and the second exon of *siren1* and *siren2* with its homologous part of *Adh*. Dashes represent alignment gaps and each dot represents the nucleotide identical with that in the topmost sequence. The donor (A) and acceptor (B) sites of the first intron of *siren* and their counterparts in *CG11779* and *Adh*, respectively, are surrounded by open boxes.

(Figure 2). Although the boundaries of these distinct parts are rather ambiguous due to substantial gaps (>600 bp in total) and base substitutions (Figure 2), it is clear that the introns of *CG11779* and *Adh* are joined to be the first intron of *siren*, in which the donor and accepter sites are provided by *CG11779* and *Adh*, respectively (Figure 3). The amino-acid sequence at the beginning of the second exon of *siren* is considerably modified by a 12-bp indel and base substitutions compared to the original second exon of *Adh* (Figure 3).

Chromosomal location of *siren*, *Adh*, and *CG11779/nanos*: The relative chromosomal locations among *siren1*, *Adh*, and *CG11779/nanos* were examined by the genetic linkage analyses (Table 1). Since the F_1 progeny were heterozygous for all three loci irrespective of their gender, these loci were identified to be autosomal. As expected by the physical distance between *siren1* and *Adh* at the nucleotide sequence level (~13 kb), these two loci showed a perfect linkage. On the other hand, the distribution of the observed numbers of F_2 genotypes gave an excellent goodness of fit ($\chi^2_{\text{d.f.}=8} = 0.32$, $P > 0.99$) with the expected values under independent assortment of *siren1* (or *Adh*) and *CG11779/nanos* without linkage. The results show that *CG11779/nanos* belongs to a different linkage group from *siren* and *Adh*; *i.e.*, *CG11779/nanos* is located on a different chromosome or at least in a chromosomal region distant from *siren* and *Adh*.

Expression patterns of *siren*, *Adh*, and *CG11779*: The transcription patterns of *siren1*, *siren2*, *Adh*, and *CG11779* for *D. parabipectinata* were examined by RT-PCR at larval, pupal, and adult stages and in female and male individuals separately (Figure 4). The transcription of *siren1* and *siren2* was detected only at pupal and adult stages in

TABLE 1

The observed and expected numbers of individuals for each genotype in F_2 progeny of B133 and CJB162 strains of *D. bipectinata*

	No. of individuals								
Genotype:[a]	$a^1a^1b^1b^1$	$a^1a^1b^1b^2$	$a^1a^1b^2b^2$	$a^1a^2b^1b^1$	$a^1a^2b^1b^2$	$a^1a^2b^2b^2$	$a^2a^2b^1b^1$	$a^2a^2b^1b^2$	$a^2a^2b^2b^2$
a, siren1; *b, Adh*									
Observed	24	0	0	0	51	0	0	0	21
Expected[b]	24.0	0.0	0.0	0.0	48.0	0.0	0.0	0.0	24.0
a, siren1 (or *Adh*); *b, CG11779/nanos*									
Observed	7	11	6	14	24	13	4	9	8
Expected[c]	6.0	12.0	6.0	12.0	24.0	12.0	6.0	12.0	6.0

[a] The genotypes of B133 and CJB162 were $a^1a^1b^1b^1$ and $a^2a^2b^2b^2$, respectively.
[b] The expected numbers of F_2 individuals under complete linkage between two loci.
[c] The expected numbers of F_2 individuals under random association between two loci.

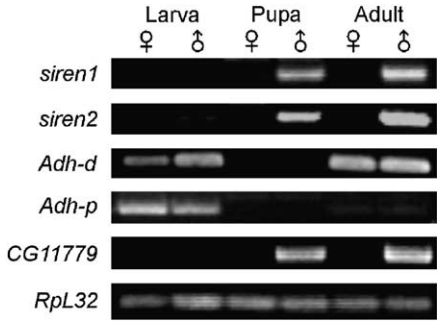

FIGURE 4.—Agarose-gel electrophoresis for RT-PCR products, showing expression patterns of *siren1*, *siren2*, *Adh*, *CG11779*, and *RpL32* (control) genes at larval, pupa, or adult developmental stage in female or male individuals of *D. parabipectinata*. Transcripts from the distal (*Adh-d*) and proximal (*Adh-p*) promoters of the *Adh* gene were separately examined.

male individuals. This pattern was virtually identical with that of *CG11779* but distinct from the sex-independent and pupa-depressed pattern of both types of *Adh* transcripts. In addition, the amount of transcripts for *Adh* was substantially larger than that for *siren* and *CG11779* (data not shown). To equalize the production of RT-PCR for the sake of experimental convenience, we used 10 times more total RNA for *siren* and *CG11779* than for *Adh*. These results indicate that the promoter sequence of *siren* originated from *CG11779*.

Molecular evolution of *siren*: A phylogenetic tree for *siren1*, *siren2*, and *Adh* sequences was constructed to infer the evolutionary relationships among these genes (Figure 5). It is clear that a duplication of *Adh* occurred after the divergence of the *D. bipectinata* complex from other species of the *D. ananassae* subgroup to generate the main part of *siren* (solid circle in Figure 5). An insertion of the *CG11779/nanos*-derived sequence must have

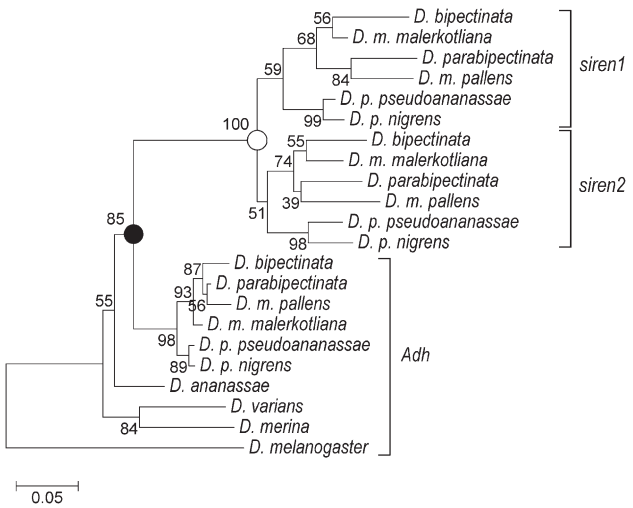

FIGURE 5.—Minimum-evolution tree for *siren* and *Adh* sequences. The solid and open circles indicate the duplications of *Adh* and ancestral *siren*, respectively.

TABLE 2

The average numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions per site (±standard error) and the d_N/d_S ratios for *siren1*, *siren2*, and *Adh* among species of the *D. bipectinata* complex

Gene	d_N	d_S	d_N/d_S
Adh	0.003 ± 0.002	0.038 ± 0.009	0.084*
siren1[a]	0.005 ± 0.003	0.121 ± 0.016	0.046*
siren2[a]	0.009 ± 0.003	0.122 ± 0.016	0.075*

* P ($d_N/d_S = 1$) $\ll$ 0.001.
[a] Only for the *Adh* homologous part (256 codons).

occurred during the same evolutionary period, since the inserted sequence was found in the species of the *D. bipectinata* complex but not in their outgroup species. However, given that the average synonymous distance between *siren* and *Adh* (0.25 ± 0.03) was very close to the distance between *siren* and *CG11779* (0.23 ± 0.10), whether the duplication of *Adh* occurred before or after the insertion of the *CG11779/nanos* sequence is unclear. Rearrangements of the *Adh* and *CG11779/nanos*-derived parts were thought to occur subsequently to produce ancestral single-copy *siren*, and a tandem duplication of the ancestral *siren* occurred finally to produce *siren1* and *siren2* before the speciation within the *D. bipectinata* complex (open circle in Figure 5).

Rates of synonymous (d_S) and nonsynonymous (d_N) substitution were computed to examine functional constraints on *siren1* and *siren2* among species of the *D. bipectinata* complex (Table 2). The average d_N/d_S ratios are significantly lower than unity for *siren1* and *siren2* as well as for *Adh*, suggesting that both copies of *siren* are under functional constraints at the amino-acid sequence level. However, the d_S values for *siren1* and *siren2* are higher than that for *Adh*. This is because the synonymous rate for *Adh* is reduced by a selective constraint for high expression efficiency (SHIELDS *et al.* 1988). The higher synonymous rate for *siren* suggests that the selective constraint on *siren* is much weaker than that on *Adh*.

DISCUSSION

The *siren* gene includes two distinct components, *i.e.*, almost a whole part of *Adh* and a part of *CG11779/nanos*. Since *siren* is located close to *Adh* (Figure 1A and Table 1), it is quite natural to assume that the *Adh* homologous part was created by a simple tandem duplication due to unequal crossing over or unequal sister-chromatid exchange (BROWN 1999 for review) similar with other *Adh* duplicated genes found around *Adh* in Drosophila (*e.g.*, FISHER and MANIATIS 1985; SCHAEFFER and AQUADRO 1987).

On the other hand, given that *CG11779/nanos* is located on a distant chromosomal region from *siren* as indicated by the genetic linkage analysis (Table 1), the *CG11779/nanos* homologous part is hardly explained by tandem duplication. Retroposition seems to be the best candidate to explain small-scale gene duplications into such distant genomic location (see Graur and Li 2000 for review). The essence of retroposition is the replication of parasitic retrotransposons and retroposons and involves reverse transcription of their own transcripts and integration of the reverse transcripts into the host genome (Moran *et al.* 1996). However, the reverse transcriptase that catalyzes this reaction occasionally operates on non-self transcripts, resulting in duplications of an expressed host gene into a genomic location irrelevant to the location of the parent gene (*e.g.*, McCarrey and Thomas 1987; Noyce *et al.* 1997; Esnault *et al.* 2000; Wei *et al.* 2001; Betrán and Long 2003). Actually, Ohshima *et al.* (2003) reported that 87% of retropseudogenes were located on chromosomes different from their parental genes in the human genome. The *CG11779/nanos* homologous part of *siren* fulfills this condition.

Undergoing transcription, reverse transcription, and integration, the resultant retrogenes automatically acquire the following peculiarities: (1) absence of intron, (2) addition of a poly(A) tract, and (3) addition of flanking short direct repeats (Tokunaga *et al.* 1985). Absence of intron seems to be consistently observable after long evolutionary time, whereas the latter two characteristics are often masked by base substitutions, insertions, and deletions that superimpose on them with time. For instance, of 24 retrogenes identified in the *D. melanogaster* genome by the definition of a different chromosomal location from the parent gene and the absence of intron, only the youngest gene retains the direct repeats and the four youngest genes retain a degenerated poly(A) tract (Betrán *et al.* 2002).

In the case of *siren*, none of these characteristics were recognized for retroposition of *CG11779*. Especially, it is quite unlikely that the nonexon sequences (the 5′-untranscribed region and the first intron) were included in the retroposed sequence. Alternatively, retroposition of *nanos* can produce this sequence, since the entire *CG11779* homologous part of *siren* is included in the third exon of *nanos* on the complementary strand (Figure 6). This is a plausible scenario in the context of the expression pattern: *nanos* mRNA is localized in germline cells (*e.g.*, Forbes and Lehmann 1998), where the activity of retrotransposons is generally enhanced (Zhao and Bownes 1998; Kogan *et al.* 2003). Positive associations between the levels of transcription and transposition frequencies in male germline cells has been demonstrated for the *copia* retrotransposon in *D. melanogaster* (Pasyukova *et al.* 1997). If this is the case, we can expect absence of intron and presence of a poly(A) tract with respect to the gene structure of *nanos*

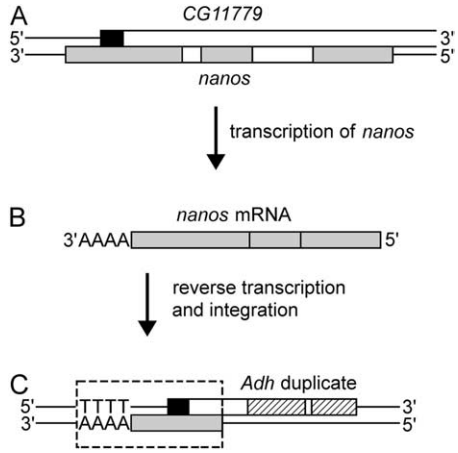

Figure 6.—Schematic of the generation of *siren*. The exons of *CG11779*, *nanos*, and *Adh* (including their homologous parts in *siren*) are represented by solid, shaded, and hatched boxes, respectively, whereas introns are represented by open boxes. (A) *CG11779* and *nanos* overlap each other on the complementary strands. (B) When a mature *nanos* mRNA is reverse transcribed and integrated into the genome, (C) the resultant insert sequence (surrounded by the dashed line) is expected to contain the *CG11779* promoter sequence and an upstream poly(T) tract corresponding to the poly(A) tract of the *nanos* mRNA.

rather than that of *CG11779*. Unfortunately, it is impossible to examine the former condition, since only the single exon of *nanos* is involved in the current structure of *siren*. In the *D. bipectinata* sequence, corresponding to the poly(A) tract of *nanos*, a 9-bp poly(T) tract was identified on the complementary strand at the position exactly matching with the polyadenylation site of *nanos* pre-mRNA (Figures 6C and 7). This gives clear evidence that a *nanos* mRNA was retroposed to be a part of *siren*. Although a similar poly(T) tract was not identified in other species, it is likely that such sequences have been already blurred by base substitution, insertion, and deletion as the date of the hypothesized retroposition is estimated to be old enough; the average d_S value between *siren* and *CG11779* is 0.23, which seems to be at the borderline for identifying the footprint of poly(A) tract (Betrán *et al.* 2002). Although the length of the current *CG11779/nanos* homologous part in *siren* is

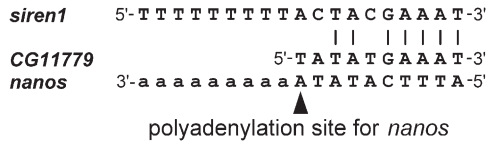

Figure 7.—Sequence alignment of the upstream region of *siren1* and the corresponding region of *CG11779/nanos* for *D. bipectitnata*. The polyadenylation site (arrowhead) of the *nanos* transcript was deduced from the cDNA sequence, at which the poly(A) tract (lowercase characters) is expected to be appended. Identical nucleotides between *siren* and *CG11779/nanos* sequences are indicated by vertical lines.

much shorter than the expected full length of the *nanos* transcript, it is known that the majority of retrotransposons and retropseudogenes are 5′ truncated (Maestre *et al.* 1995; Pavlíček *et al.* 2002; Farley *et al.* 2004). Alternatively, it is also plausible that the rest of the expected *nanos*-derived sequence was lost by deletions and/or destroyed by the tandem duplication of *Adh* if it occurred later. A number of gaps identified around the boundary between the *CG11779/nanos* and *Adh* homologous parts (Figure 2) indicate that many deletions and insertions have actually occurred during the evolution of *siren*.

Retropositions of exon-intron structure of a gene via reverse transcription of its overlapping gene have been also reported by Courseaux and Nahon (2001) and Ejima and Yang (2003) in primate genomes. However, a remarkable aspect found in this study is that the retroposition of *nanos* provided the promoter sequence of *CG11779* for *siren*. This is strongly supported by the fact that the expression pattern of *siren* is the same as that of *CG11779* but distinct from the alternative candidate, *Adh* (Figure 4). In the current point of view, a retroposed gene copy will normally become a retropseudogene due to a lack of promoter and it can survive as a functional gene only when it recruits a new promoter sequence. In fact, retrosequence-derived genes were often found as a chimera with a preexisting gene that supplied the promoter (Brosius 2003; Long *et al.* 2003 for reviews). *Jingwei* (Long and Langley 1993; Long *et al.* 1999; Wang *et al.* 2000) and *TRIMCyp* (Nisole *et al.* 2004; Sayah *et al.* 2004) are the best paradigms of such chimeric genes. Although *siren* is also a chimeric gene, the promoter sequence is provided by the retroposed sequence rather than by the preexisting gene. This is contrary to the current viewpoint and reveals an unexpected potential of retroposition to generate new genes, when genes are overlapping.

It has been shown recently that a certain fraction of genes has overlapping genes in various organisms. In mammals, many sense-antisense transcript (SAT) pairs produced from the same locus have been found, *e.g.*, 5880 (22%) of 26,741 transcripts in humans (Chen *et al.* 2004). For 65% of them, however, either sense or antisense transcript is nuclear and poly(A) minus, which strongly restricts retroposition. Among the rest, only 30% have the configuration of head-to-head overlapping (Veeramachaneni *et al.* 2004), where a promoter sequence is involved in the overlapping part. Consequently, ~600 genes remain as potential targets of the promoter retroposition. In mouse, SATs were found in 4962 (15%) of 33,360 transcripts (Kiyosawa *et al.* 2003), and 1886 of them were identified to form a coding/coding pair (Kiyosawa *et al.* 2005). Since the frequency of head-to-head overlapping is 37% (Veeramachaneni *et al.* 2004), 700 genes can be the candidate. Analyzing the *D. melanogaster* genome sequence data (Release 3.2), we obtained that 1863 (13.8%) of the 13,472 annotated protein-coding genes have at least one overlapping protein-coding gene on the complementary strand, which is consistent with Misra *et al.* (2002). After the deduction for the non-head-to-head overlapping fraction, 222 genes remain as the candidate. A similar extent of gene overlapping is also observed in plants. In *Arabidopsis thaliana*, 2680 (8.9%) of 29,993 transcripts are generated from overlapping protein-coding genes, of which 370 genes are in a head-to-head overlapping configuration (Wang *et al.* 2005). In rice, 1060 (5.2%) protein-coding genes have an overlapping protein-coding gene (Osato *et al.* 2003). When antisense transcripts are retroposed, the chance of the resultant sequences surviving as a new gene is possibly higher than in the case of the usual retroposition of sense transcripts. This is because intron sequences facilitate exon shuffling between retroposed and genomic preexisting genes without disturbing their protein-coding frames and because the promoter is essential for transcription—which is the case in *siren*.

The identification of retrogenes commonly relies on two major definitions, *i.e.*, (1) a distant chromosomal location from the parent gene and (2) the absence of an intron (Betrán *et al.* 2002). As *siren* does not meet the latter criterion, any gene in a genome that fulfills only the former criterion can also be a candidate of retrogene when its parent gene has an overlapping gene. Consequently, the significance of retroposition as a mechanism of new gene generation may be greater than appreciated previously. In this study, we have reported a new chimeric gene named *siren* and shown a further potential of retroposition to generate new genes.

We greatly appreciate the reading of the nucleotide sequences of cosmid clones by Toshiro Aigaki, Yuji Kohara, and Tadasu Shin-I. We are also grateful to Yoshihiro Kawahara, Shigeyuki Koshikawa, and Kazumi Yumita for supporting our experiments. We also thank Shigeru Saito, Claire T. Saito, Ryoko D. Segawa, and Norikazu Kitamura for valuable comments on the manuscript. This work was supported by grants from the Japan Society for the Promotion of Science to T.A. (12375002 and 15255006) and K.T.

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Betrán, E., and M. Long, 2003 *Dntf-2r*, a young Drosophila retroposed gene with specific male expression under positive Darwinian selection. Genetics **164:** 977–988.

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

Boom, R., C. J. A. Sol, M. M. M. Salimans, C. L. Jansen, P. M. E. Wertheim-van Dillen *et al.*, 1990 Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. **28:** 495–503.

Brosius, J., 2003 The contribution of RNAs and retroposition to evolutionary novelties. Genetica **118:** 99–116.

Brown, T. A., 1999 *Genomes*, pp. 367–389. BIOS Scientific, Oxford.

Burge, C., and S. Karlin, 1997 Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268:** 78–94.

Chen, J., M. Sun, W. J. Kent, X. Huang, H. Xie *et al.*, 2004 Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Res. **32:** 4812–4820.

Courseaux, A., and J-L. Nahon, 2001 Birth of two chimeric genes in the *Hominidae* lineage. Science **291:** 1293–1297.

Ejima, Y., and L. Yang, 2003 *Trans* mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. Hum. Mol. Genet. **12:** 1321–1328.

Esnault, C., J. Maestre and T. Heidmann, 2000 Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. **24:** 363–367.

Farley, A. H., E. T. L. Prak and H. H. Kazazian, Jr., 2004 More active human L1 retrotransposons produce longer insertions. Nucleic Acids Res. **32:** 502–510.

Felsenstein, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783–791.

Fisher, J. A., and T. Maniatis, 1985 Structure and transcription of the *Drosophila mulleri* alcohol dehydrogenase genes. Nucleic Acids Res. **13:** 6899–6917.

Forbes, A., and R. Lehmann, 1998 Nanos and Pumilio have critical roles in the development and function of *Drosophila* germline stem cells. Development **125:** 679–690.

Frohman, M. A., M. K. Dush and G. R. Martin, 1988 Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proc. Natl. Acad. Sci. USA **85:** 8998–9002.

Gilbert, W., 1978 Why genes in peaces? Nature **271:** 501.

Graur, D., and W-H. Li, 2000 *Fundamentals of Molecular Evolution*, Ed. 2, pp. 249–366. Sinauer Associates, Sunderland, MA.

Hanahan, D., 1985 Techniques for transformation of *E. coli*, pp. 109–135 in *DNA Cloning, Vol. I: A Practical Approach*, edited by D. M. Glover. IRL Press, Oxford.

Henikoff, S., 1984 Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene **28:** 351–359.

Jones, C. D., A. W. Custer and D. J. Begun, 2005 Origin and evolution of a chimeric fusion gene in *Drosophila subobscura*, *D. madeirensis* and *D. guanche*. Genetics **170:** 207–219.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–132 in *Mammalian Protein Metabolism III*, edited by H. N. Munro. Academic Press, New York.

Kimura, M., and T. Ohta, 1974 On some principles governing molecular evolution. Proc. Natl. Acad. Sci. USA **71:** 2848–2852.

Kiyosawa, H., I. Yamanaka, N. Osato, S. Kondo, Riken Ger Group *et al.*, 2003 Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. Genome Res. **13:** 1324–1334.

Kiyosawa, H., N. Mise, S. Iwase, Y. Hayashizaki and K. Abe, 2005 Disclosing hidden transcripts: mouse natural sense-antisense transcripts tend to be poly(A) negative and nuclear localized. Genome Res. **15:** 463–474.

Kogan, G. L., A. V. Tulin, A. A. Aravin, Y. A. Abramov, A. I. Kalmykova *et al.*, 2003 The GATE retrotransposon in Drosophila melanogaster: mobility in heterochromatin and aspects of its expression in germline tissues. Mol. Gen. Genomics **269:** 234–242.

Kulp, D., D. Haussler, M. G. Reese and F. H. Eeckman, 1996 A generalized hidden Markov model for the recognition of human genes in DNA. Proc. Int. Conf. Intell. Syst. Mol. Biol. **4:** 134–142.

Kumar, S., K. Tamura and M. Nei, 2004 MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief. Bioinformatics **5:** 150–163.

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

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., C. Rosenberg and W. Gilbert, 1995 Intron phase correlations and the evolution of the intron/exon structure of genes. Proc. Natl. Acad. Sci. USA **92:** 12495–12499.

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.

Long, M., E. Betrán, K. Thornton and W. Wang, 2003 The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. **4:** 865–875.

Maestre, J., T. Tchénio, O. Dhellin and T. Heidmann, 1995 mRNA retroposition in human cells: processed pseudogene formation. EMBO J. **14:** 6333–6338.

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

Misra, S., M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell *et al.*, 2002 Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. Genome Biol. **3:** 0083.

Moran, J. V., S. E. Holmes, T. P. Naas, R. J. DeBerardinis, J. D. Boeke *et al.*, 1996 High frequency retrotransposition in cultured mammalian cells. Cell **87:** 917–927.

Nadeau, J. H., and D. Sankoff, 1997 Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics **147:** 1259–1266.

Nisole, S., C. Lynch, J. P. Stone and M. W. Yap, 2004 A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. Proc. Natl. Acad. Sci. USA **101:** 13324–13328.

Noyce, L., J. Conaty and A. A. Piper, 1997 Identification of a novel tissue-specific processed *HPRT* gene and comparison with X-linked gene transcription in the Australian marsupial *Macropus robustus*. Gene **186:** 87–95.

Ohno, S., 1970 *Evolution by Gene Duplication*, pp. 71–82. Springer-Verlag, Berlin.

Ohshima, K., M. Hattori, T. Yada, T. Gojobori, Y. Sakaki *et al.*, 2003 Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. Genome Biol. **4:** R74.

Ohta, T., and M. Kimura, 1971 Functional organization of genetic material as a product of molecular evolution. Nature **233:** 118–119.

Osato, N., H. Yamada, K. Satoh, H. Oota, M. Yamamoto *et al.*, 2003 Antisense transcripts with rice full-length cDNAs. Genome Biol. **5:** R5.

Pasyukova, E., S. Nuzhdin, W. Li and A. J. Flavell, 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.

Pavlíček, A., J. Pačes, R. Zíka and J. Hejnar, 2002 Length distribution of long interspersed nuclear elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection. Gene **300:** 189–194.

Powell, L. M., S. C. Wallis, R. J. Pease, Y. H. Edwards, T. J. Knott *et al.*, 1987 A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell **50:** 831–840.

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.

Rzhetsky, A., and M. Nei, 1992 A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. **9:** 945–967.

Savakis, C., M. Ashburner and J. H. Willis, 1986 The expression of the gene coding for alcohol dehydrogenase during the development of *Drosophila melanogaster*. Dev. Biol. **114:** 194–207.

Sayah, D. M., E. Sokolskaja, L. Berthoux and J. Luban, 2004 Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. Nature **430:** 569–573.

Schaeffer, S. W., and C. F. Aquadro, 1987 Nucleotide sequence of the *Adh* gene region of *Drosophila pseudoobscura*: evolutionary change and evidence for and ancient gene duplication. Genetics **117:** 61–73.

Shields, D. C., P. M. Sharp, D. G. Higgins and F. Wright, 1988 "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

Tokunaga, K., K. Yoda and S. Sakiyama, 1985 Structure of a processed gene of mouse cytoplasmic γ-actin transposed into a BAM5 sequence: insertion has created 13 base-pair direct repeats. Nucleic Acid Res. **13:** 3031–3042.

Veeramachaneni, V., W. Makałowski, M. Galdzicki, R. Sood and I. Makałowska, 2004 Mammalian overlapping genes: the comparative perspective. Genome Res. **14:** 280–286.

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

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

Wang, X-J., T. Gaasterland and N-H. Chua, 2005 Genome-wide prediction and identification of *cis*-natural antisense transcripts in *Arabidopsis thaliana*. Genome Biol. **6:** R30.

Wei, W., N. Gilbert, S. L. Ooi, J. F. Lawler, E. M. Ostertag *et al.*, 2001 Human L1 retrotransposition: *cis* preference versus *trans* complementation. Mol. Cell. Biol. **21:** 1429–1439.

Zhang, J., H. F. Rosenberg and M. Nei, 1998 Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95:** 3708–3713.

Zhao, D., and M. Bownes, 1998 The RNA product of the *Doc* retrotransposon is localized on the *Drosophila* oocyte cytoskeleton. Mol. Gen. Genet. **257:** 497–504.

Communicating editor: T. H. Eickbush