Origin of *sphinx*, a young chimeric RNA gene in *Drosophila melanogaster*

Wen Wang*, Frédéric G. Brunet*, Eviatar Nevo†, and Manyuan Long*‡

*Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637; and †Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel

Contributed by Eviatar Nevo, February 4, 2002

Non-protein-coding RNA genes play an important role in various biological processes. How new RNA genes originated and whether this process is controlled by similar evolutionary mechanisms for the origin of protein-coding genes remains unclear. A young chimeric RNA gene that we term *sphinx (spx)* provides the first insight into the early stage of evolution of RNA genes. *spx* originated as an insertion of a retroposed sequence of the ATP synthase chain F gene at the cytological region 60DB since the divergence of *Drosophila melanogaster* from its sibling species 2–3 million years ago. This retrosequence, which is located at 102F on the fourth chromosome, recruited a nearby exon and intron, thereby evolving a chimeric gene structure. This molecular process suggests that the mechanism of exon shuffling, which can generate protein-coding genes, also plays a role in the origin of RNA genes. The subsequent evolutionary process of *spx* has been associated with a high nucleotide substitution rate, possibly driven by a continuous positive Darwinian selection for a novel function, as is shown in its sex- and development-specific alternative splicing. To test whether *spx* has adapted to different environments, we investigated its population genetic structure in the unique ''Evolution Canyon'' in Israel, revealing a similar haplotype structure in *spx*, and thus similar evolutionary forces operating on *spx* between environments.

RNA genes are of interest to evolutionary biologists for a number of reasons. In the RNA world, a hypothesis concerning the initial stage of life, RNA genes dominated all life activities (1). In extant genomes, there are many RNA genes with important roles in various biological processes (2, 3). In addition to ribosomal RNA (rRNA) and transfer RNA (tRNA), numerous other non-protein-coding RNAs (ncRNAs) have recently been identified (2–4). These ncRNAs fall in two categories: (*i*) small nonmessenger RNAs (snmRNAs) such as small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), BC1 RNA, and BC200 RNA (3–7); and (*ii*) mRNA-like ncRNAs (4, 5). It has been observed that some ncRNAs only exist in particular lineages or species, suggesting that nature must have independently created these new RNA genes during evolution. For example, Brosius and his coworkers proposed that neuron BC1 RNA gene in rodents was generated by retroposition of $\mathrm{tRNA}^{\mathrm{Ala}}$ just before the diversification of rodents (6, 7), and that primate BC200 RNA gene was only created after the emergence of Anthropoidea (8). The *Xist* gene, which is polyadenylated and plays a crucial role in inactivating X chromosome, is also only found in mammals (9, 10).

Thus, these observations give rise to an interesting question: how did new RNA genes originate and evolve throughout the history of gene evolution? Gilbert speculated that exon shuffling could serve to create new RNA genes in the RNA world (11). Indeed, there is ample evidence supporting a common role of exon shuffling in protein-coding genes (12, 13). However, whether the same mechanism operates on evolution of the extant RNA genes remains an open question. Does Darwinian selection that shapes new protein-coding genes (e.g., refs. 14 and 15) also drive the evolution of new RNA genes? A straightforward answer might be obtained by directly examining an RNA gene that originated recently, because the evolutionary characteristics related to the molecular processes of origin and subsequent evolutionary dynamics of new genes may be more easily observed in such a gene.

Phylogenetic comparison of genetic signals has proven to be an efficient method in identifying young protein-coding genes in *Drosophila* (e.g., refs. 14 and 15) and mammals (e.g., refs. 16 and 17), and is useful in investigating new gene evolution. We extended this approach to search for young genes in the *Drosophila melanogaster* subgroup, where we identified a young RNA gene that originated in *D. melanogaster* since it diverged from its sibling species about 2 million years ago. We name this gene *sphinx* (*spx*) to invoke an analogy to another chimera, the Sphinx from ancient Greek legend having a lion's body and a human head. In this paper, we report an analysis of the initial molecular mechanisms generating the gene structure and subsequent evolutionary process of the *spx* gene.

Materials and Methods

Fluorescent *in Situ* Hybridization (FISH) of Polytene Chromosomes. The Unigene library of *D. melanogaster* was obtained from Research Genetics (Huntsville, AL). The cDNA inserts were amplified and labeled by PCR using the vector primers (T7 and PM001). The probes were labeled with digoxigenin or biotin (purchased from Roche Molecular Biochemicals) and hybridized to polytene chromosomes of each member of the *D. melanogaster* subgroup according to ref. 18. By counting hybridization signals on the polytene chromosomes of these species, we were able to detect homologues that were translocated to different cytological loci by retroposition or other processes. With one of the many probes used, GH18886, we detected an extra signal on the fourth chromosome of *D. melanogaster*. By checking the genome sequence of this species, we found that this extra signal represents an intronless duplicate of the parental gene that transcribes GH18886. The parental gene was annotated as CG4692, coding for a product highly similar to ATP synthase chain F identified in other organisms (19).

Southern Hybridization. Genomic DNAs of *D. melanogaster, Drosophila simulans, Drosophila mauritiana, Drosophila sechellia, Drosophila teissieri, Drosophila erecta*, and *Drosophila orena* were extracted by using the Puregene DNA isolation kit (Gentra Systems). *Hin*dIII-digested DNAs were transferred to a nylon membrane (Roche Molecular Biochemicals) by Southern blotting. The digoxigenin-labeled *sphinx* probe was hybridized to the membrane to confirm the copy numbers in different species detected in the FISH experiment.

Abbreviations: ncRNA, non-protein-coding RNA; NFS, north-facing slope; SFS, south-facing slope; RT, retrotranscription.

Data deposition: The DNA sequence of *spx* has been deposited in GenBank database (accession no. AF479655)

‡To whom reprint requests should be addressed. E-mail: mlong@midway.uchicago.edu.

www.pnas.org/cgi/doi/10.1073/pnas.072066399

Characterization of Transcripts of the Detected Gene. We used rapid amplification of cDNA ends (RACE) assay and retrotranscription (RT)-PCR to characterize the gene structure. Total RNA was extracted from adults of both sexes, adult females, males, 2-hour-old eggs, second- and third-instar larvae, or pupae, by using a Qiagen total RNA extraction kit. A series of forward and reverse gene-specific primers were designed based on the *Drosophila* genome sequences (19). Following the manufacturer's protocol, 5′ RACE was conducted by using the 5′ RACE System Version 2.0 (Life Technologies). As for the 3′ RACE, the adapter-linked oligo(dT) primer (Life Technologies) was used to synthesize the first strand of cDNA. Two forward primers were designed to amplify the 3′ end of the cDNA against the adapter primer (AUAP, Life Technologies).

Population Genetics Analyses. We surveyed *spx* variations in the "Evolution Canyon" (EC) populations to detect possible impact of the drastic changes in environment on the population genetic structure of *spx*. The EC, a microsite in Mount Carmel, Israel, is a microcosm of life's evolution mirroring regional and global evolutionary divergence patterns (20). The sharp interslope ecological contrast at EC was caused by differential solar radiation on the opposite slopes. The north-facing slope (NFS) is a relatively homogeneous, mesic, cool, live oak-forested biome. The south-facing-slope (SFS) obtains 6-fold more solar radiation than the NFS and is, therefore, xeric with an open park forest or savanna. SFS is warmer and drier with more fluctuating and unpredictable temperature than the NFS. Many organisms display remarkable qualitative and quantitative interslope biodiversity divergence (20–22). *D. melanogaster* diverges on the opposite slopes adaptively (23, 24) and displays incipient speciation (25, 26).

Previously we found that there exist two distinct haplotypes around the *spx* chromosomal region in natural populations, which suggested that a mechanism of balancing selection might be responsible for this pattern (27). By sequencing 7 more SFS isofemale lines, plus the 11 NFS and 2 SFS isofemale lines reported in ref. 27, we analyzed the local population structure and divergence of *spx* at EC. We then assayed the frequency of the two haplotypes by genotyping 29 more NFS and 35 more SFS lines to thoroughly evaluate the environmental effect on the haplotype structure of the *spx* locus. In addition, to reveal the evolutionary processes in the origin of *spx*, we also amplified and sequenced two ancestral sequences; the flanking region of *spx* in *D. sechellia* and a partial sequence of the parental gene, the ATP synthase chain F gene, in *D. simulans*.

Results

FISH and Southern Hybridization. We identified a cDNA probe, GH18886, which displays a signal in a FISH experiment only on the fourth chromosome of *D. melanogaster* (Fig. 1*a*). The parental signal that exists in all of the species under investigation is located at 60D8, 2R, of polytene chromosomes of *D. melanogaster*. Genomic Southern hybridization with the same digoxigenin-labeled probe confirmed that only the *D. melanogaster* genome contains an additional hybridization signal (Fig. 1*b*).

Gene Structure and Expression Patterns. We retrieved the cDNA sequence of GH18886 from Flybase (http://flybase.bio.indiana.edu), and sequenced the clone used for the FISH to confirm that the clone was indeed GH18886. By searching the Genome Annotation Database of *Drosophila* (Gadfly; ref. 19), we found that this transcript matches a gene annotated as CG4692. This gene possesses two introns and its mRNA transcript is 569 nt long, coding for a product homologous to the ATP synthase chain F of *Caenorhabditis elegans*, *Mus musculus*, and *Homo sapiens* (19). A BLAST search against the *Drosophila* genome matches the second sequence located at 102F8 of the fourth chromosome, which is consistent with our FISH data. However, the homologous sequence at 102F8 contains no intron. We also observed a stretch of four A's after the original polyadenylation site of the parental gene (Fig. 2). Therefore, the duplicate on the fourth chromosome appears to be a processed sequence created by retroposition.

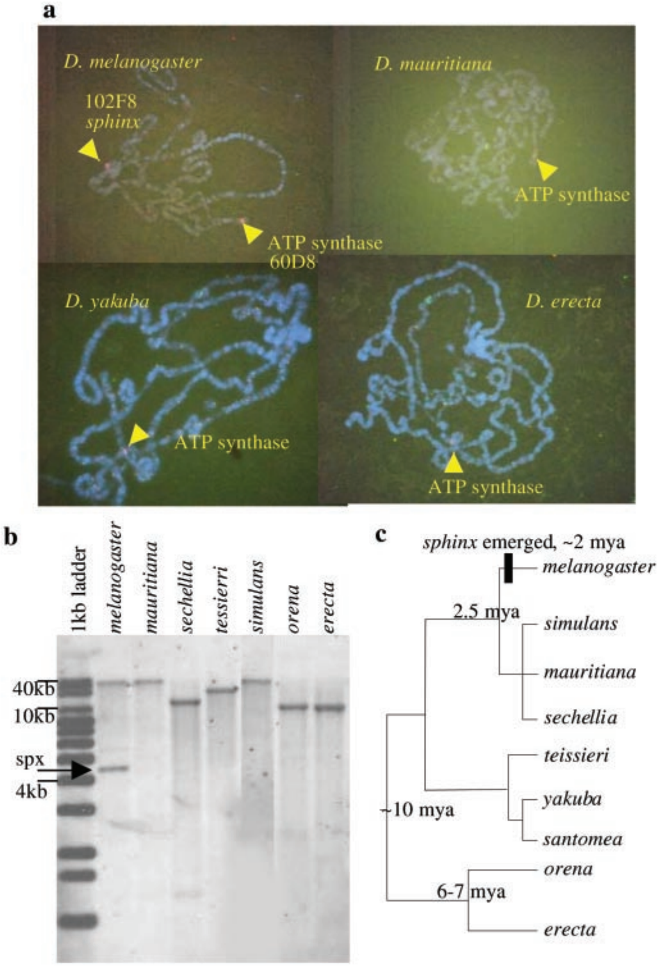

Fig. 1. (*a*) FISH results showing that there are two red signals (arrows) in *D. melanogaster*. (*b*) Southern hybridization results also show that *D. melanogaster* has two copies homologous to the probe, which is the cDNA of ATP synthase chain F gene, but the other species have only one. (*c*) Phylogenetic tree of *D. melanogaster* subgroup (36). Divergence time of some nodes and emergence time of *sphinx* is indicated.

Numerous molecular changes and reorganizations have occurred at this locus. The start codon for translation of ATP synthase was lost by substitution, a substitution at the 18th codon position would have led to a premature stop codon, and three deletions are scattered in various positions (Figs. 2 and 3). A comparison of this region to the orthologous genomic region we obtained in *D. sechellia* defined an insertion of 751 bp as a consequence of retroposition. That this insertion was indeed a result of retrotransposition is also supported by a short direct repeat (TTCG) that we identified at both boundaries of the insertion (Fig. 2), a further hallmark of retroposition. A sequence of 472 bp in the insertion is similar to the ATP synthase gene. The remaining 5′ sequence of the insertion is similar to the terminal inverted repeat of the *S* elements that are *Tc1*-like DNA transposons (28).

A gene, CG11091, is predicted by Gadfly (19) containing the retroposed ATP synthase sequence and a small first exon.

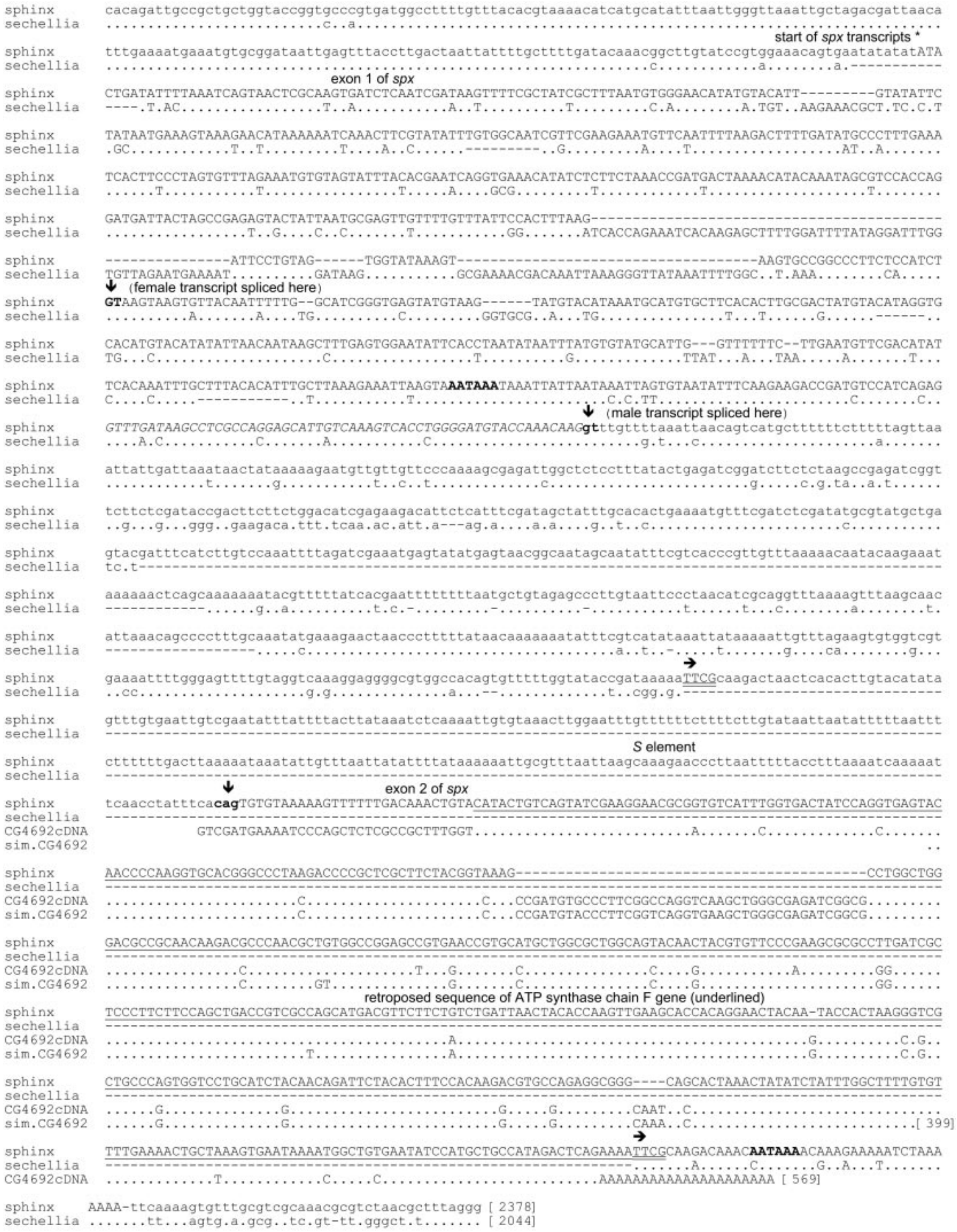
sphinx
sechellia
start of spx transcripts *
exon 1 of spx
(female transcript spliced here)
(male transcript spliced here)
S element
exon 2 of spx
CG4692cDNA
sim.CG4692
retroposed sequence of ATP synthase chain F gene (underlined)
[399]
[569]
[2378]
[2044]

Fig. 2. Alignment of the *sphinx* locus sequence of *D. melanogaster* with the sequences of the correspondent region of *D. sechellia*, CG4692 cDNA of *D. melanogaster*, and deduced partial cDNA of *D. simulans.* Asterisk indicates the start base of *spx* transcripts, dashes indicate deletions, and dots show the identical bases. The two exon sequences of the male-specific transcript (*sphinx-m*) are in capital letters. The two alternative adenylation signals (AATAAA) are in bold. All of the splicing sites (gt or cag) are indicated by boldface and vertical arrows. The retroposed sequence is flanked by the short direct repeats of TTCG, which are double underlined and indicated by the horizontal arrows. The region homologous to the ATP synthase gene is underlined. The remaining retroposed sequence is homologous to the terminal inverted repeat of *S* element, and therefore, the recipient slicing site was provided by the *S* element sequence.

According to the annotation, this gene would be transcribed from the antisense strand, and hence its conceptual amino acid sequence has no significant similarity to the ATP synthase chain F. No other similar proteins were found when we searched this sequence against the protein databases using the BLAST. Furthermore, our RACE experiments showed that the predicted transcript of CG11091 from the antisense strand was undetectable.

On the other hand, we did detect transcripts from the sense strand defined by the sequence of the retroposed ATP synthase chain F gene. Remarkably, the 5′ RACE experiment revealed an additional exon and intron recruited by the retroposed sequence. This finding demonstrates the manifestation of a new chimeric gene, likely with a novel function as implied by its fused coding structure and a new regulatory sequence. The major problem of how a newly arisen retrosequence becomes activated is solved by recruiting previously existing regulatory sequences and associ-

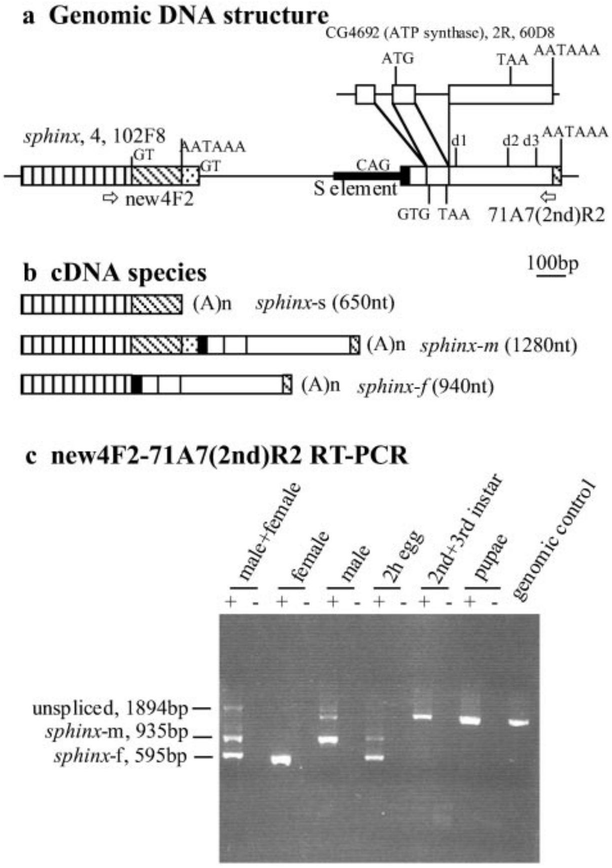

Fig. 3. (*a*) Gene structure of *sphinx* and its parental gene, ATP synthase chain F. Blank blocks represent the exon of the ATP synthase gene. Black blocks represent the terminal inverted repeat sequence of *S* element, a small part (27 bp) of which is recruited into the second exon of *sphinx*. Other stripped blocks are exon sequences that are endogenous in the fourth chromosome. Some important changes in the retroposed ATP synthase sequence are indicated, including the change of start codon into GTG, introduction of a stop codon, and three deletions ($d_1 = 42$ bp, $d_2 = 1$ bp, $d_3 = 4$ bp). The positions of the two primers used for RT-PCR are indicated by arrows. Splicing sites are marked by GT or CAG. Polyadenylation sites are marked by AATAAA. (*b*) mRNA species resulted from alternative adenylation and splicing. *Sphinx-m* is detected in males and eggs, *sphinx-f* is in females and eggs, *sphinx-s* is in both females and males, and the unspliced one is in males, larvae, and pupae. (*c*) RT-PCR results. ''−'' indicates the RT-PCR-negative controls in which everything is the same as the positive (+) except omitting reverse transcriptase. The last lane is a positive control using genomic DNA as the template. The primer locations on the gene are shown in *a*.

ated exon(s) and intron(s), thus displaying creative molecular tinkering.

Moreover, the 3′ RACE and RT-PCR analyses showed that the transcripts of the *spx* gene are alternatively spliced in sex- and development-specific fashions (Fig. 3). The male-specific transcript (*spx-m*) is 1,280 nt, and the female-specific transcript (*sphinx-f*) is 940 nt long. The major mRNA species in larvae and pupae is the unspliced primary RNA transcript (*spx-p*), but the female-specific transcript can also be weakly detected in larvae (Fig. 3*c*). In males, we also detected two other minor mRNA species: an unspliced primary RNA and a spliced transcript longer than *spx-m* (Fig. 3*c*). When primers located in the recruited first exon were used, our 3′ RACE experiments also identified a short transcript (*spx-s*) with a lower level of abundance in adults, which uses an alternative adenylation signal in the recruited exon of the *spx* gene (Figs. 2 and 3). This finding suggests that the recruited exon itself was likely derived from a preexisting independent gene whose regulatory system confers the *spx* gene with a controlled expression.

Evolutionary Analyses. We sequenced 500 bp of the ATP synthase chain F gene of *D. simulans*, including 399 bp of the coding region. We used this sequence as the outgroup and were able to infer the substitution events in the gene lineages of *spx*, ATP synthase gene in *D. melanogaster* and *D. simulans*. Excluding the three deletions, we observed as many as 18 fixed substitutions in the *spx* lineage, but only 2 synonymous substitutions in ATP synthase gene of *D. melanogaster*. Because there are only 87 synonymous sites out of the total alignable 352 sites in the *D. melanogaster* ATP synthase gene, 2 synonymous changes out of 87 sites are equivalent to 8.1 changes in a comparable region of *spx*, if we assume *spx* is nonfunctional. This is a conservative assumption that also assumes that all replacement sites in the ATP synthase gene are not changeable in evolution. Otherwise, the two changes in *D. melanogaster* ATP synthase gene would have represented more neutral sites. Thus, these two changes would have been equivalent to fewer changes (<8.1) in the comparable region of the *spx* gene. Under this assumption, therefore, the substitution rate of the *spx* gene is significantly higher than the rate expected for a nonfunctional (neutral) sequence ($P < 0.05$ for a null hypothesis of equal evolutionary rate; $\chi^2 = 3.85$ with df = 1). Thus, the accelerated rate of *spx* cannot be interpreted with the relaxation of selective constraint on *spx*.

To examine the population structure of *spx*, we analyzed the sequences of the *sphinx* locus (2,624 bp long) for 11 NFS and 9 SFS strains. The segregating sites are listed in Table 1 to show the haplotype structure. Insertions and deletions were found only in nonexon regions. Both haplotypes were found in the two populations. With more strains genotyped, Table 2 further reveals that NFS and SFS consist of a similar haplotype structure. Thus, the population structure does not significantly differ from the observation of Wang *et al.* (27) in the haplotype distribution in a worldwide sample.

Discussion

Origin of *Sphinx*. One advantage of studying young genes is that the features acquired during their origin (e.g., the complex molecular tinkering process creating new gene structures) are well preserved, and thus the early steps in their origination become directly observable. As a young gene that originated within the last 2–3 million years, the sources for all different portions of *spx* can be clearly traced. It clearly recruited an exon/intron unit with its regulatory sequences from a functionally unknown sequence. Its second exon features all of the hallmarks of a retrosequence derived from the gene encoding the ATP synthase chain F: intron loss and partial poly(A)-stretch as present in the template RNA of the parental gene, and the flanking direct repeat resulting from the retroposition event (Figs. 2 and 3).

Although the role of retroposition is well defined in the origin of this gene, it should be pointed out that this is an unusual retroposition process. An independent DNA transposon, *S* element, moved together in the process with the ATP synthase chain F gene. A consequence of this process leaves a partial *S* fragment attached to the ATP synthase element-derived region in *spx*. There are several hypothetical scenarios for the origin of this complex structure. The first hypothesis is that the retroposed sequence of ATP synthase gene might have been inserted first into the *S* element located in the current position of the chromosome. Then the chimeric gene structure evolved by using the sequence of degenerated *S* element as the recipient site for splicing of the newly created intron between the recruited exon and ATP synthase chain F derived exon. The second hypothesis is that the retrosequence might have landed first in the *S*

Table 1. Segregating sites in the 11 NFS and 9 SFS sequences

Stocks	Code in	Collected	Segregating sites													
	ref. 27	year				11	1	1	1	2	2	2	2	2	2	2222222222222
			2	3	4	11	2	5	7	0	0	3	3	3	5	5555555555555
			5	4	5	22	5	5	1	1	3	2	3	7	0	1122222222223
			4	3	1	12	0	2	8	7	2	1	9	1	2	8901234567890
NFS6.45	Is1.1	1995	T	A	G	AT	C	T	G	T	T	C	A	T	G	GGCGGATTGTGGG
NFS7.6	Is1.2	1995	.	.	.	..	.	.	.	.	.	.	.	.	.	....
NFSN4	Is1.3	1997	.	.	.	..	.	.	.	.	.	.	.	.	.	....
NFSN34	Is1.4	1997	.	.	.	..	.	.	.	.	.	.	.	.	.	....
NFS6.3	Is1.5	2000	.	.	.	..	.	.	.	.	.	.	.	.	.	....
NFS6.8	Is1.6	2000	.	.	.	..	.	.	.	.	.	.	.	.	.	....
NFS97line1	Is1.7	1997	.	.	.	..	.	.	.	.	.	.	.	.	.	....
NFS97line2	Is1.8	1997	.	.	.	..	.	.	.	.	.	.	.	.	.	....
NFS7.13	Is1.9	1995	A	T	T	--	G	A	T	.	G	A	-	A	.	TGTGTGT------
NFS6.28	Is1.10	1995	A	T	T	--	G	A	T	.	G	A	-	A	.	TGTGTGT------
NFS97line5	Is1.11	1997	A	T	T	--	G	A	T	A	G	A	-	A	A	TGTGTGT------
SFS1.10	south1.10	1995	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFSA9	southA9	1997	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFS1.51		1995	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFS2.31		1995	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFSA27		1997	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFS.2		2000	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFS.3		2000	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFS.8		2000	.	.	.	..	.	.	.	.	.	.	.	.	.	....
SFS2.39		1995	A	T	T	--	G	A	T	.	G	A	–	A	A	TGTGTGT------
			exon1(male)			intron					exon2		3′ flanking			

Positions correspond to the sequence we obtained. The exon and intron regions are indicated.

element, located in another portion of the genome, before the *S* element carrying the retrosequence jumped into the current position and degenerated in the *S* element structure. The third hypothesis is that the portion of the *S* element, which was located upstream of *spx*, might have been cotranscribed with the ATP synthase chain F gene and retroposed. The observation that the short repeats flank both *S* element fragment and the ATP synthase derived portion of *spx* is consistent with the third hypothesis.

Functionality of *Sphinx*. At first glance, the premature stop codon and deletions in the retrosequence region seem to indicate a processed pseudogene. However, there are several independent lines of evidence against this possibility, supporting the hypothesis that this is a newly evolved functional gene. (*i*) Expression data show that the retroposed ATP gene is transcribed by recruitment of a proximal exon/intron unit including the regulatory sequence, thus yielding a new chimeric gene, *spx*. This is an important mechanism that retrosequences evolve into new functional sequences by acquiring a new regulatory system and a new functional domain (14, 18, 29–32). One example involving this mechanism is the *jingwei* gene in *Drosophila* that was created by the retroposition of a processed *Adh* gene sequence into the third intron of a duplicate of *yellow emperor* gene (14, 18). In the case of *spx*, whether the recruited regulatory sequence and exon/intron unit are parts of a preexisting gene will be subject to further studies by comparison with other species. (*ii*) The *spx* gene has developed and maintained well-defined splicing sites (Figs. 2 and 3), which is usually not expected for a pseudogene (33, 34). (*iii*) We found sex- and development-specific alternatively splicing (Fig. 3), suggesting various biological functions associated with male and female individuals in various developmental stages and the expression is under tight control. (*vi*) Population genetic data revealed insertions and deletions only in the noncoding regions, suggesting some likely functional constraints in the "coding" region. (*v*) Finally, the *spx* gene has a substitution rate significantly higher than a nonfunctional sequence (e.g., synonymous sites and pseudogene, see for example ref. 35), likely suggesting a positive selection for novel functions. Accelerated change driven by positive selection is an expected phenomenon for young functional genes as a consequence of modifying the raw material (i.e., the retrosequence) for new functions (e.g., refs. 14 and 15). All these features support the hypothesis that *sphinx* is a newly evolved functional gene. In fact, pseudogenes are very rare in *Drosophila* (36), and some presumed recently created *Drosophila* pseudogenes have been shown to be functional genes (14, 37, 38).

Table 2. Similar haplotype structures in NFS and SFS populations

Population	Major haplotype	Minor haplotype	Heterozygotes
NFS	22	7	11
SFS	27	7	10

$\chi^2 = 0.752$; df = 2, $P > 0.7$.

Several lines of evidence suggest *spx* is likely a ncRNA gene. (*i*) It appears that *spx* does not contain any ORFs with significant coding potential. There are three ORFs larger than 40 aa in the exon sequences. The first one is 46 aa long and located in the recruited exon, but this ORF does not exist in the *D. sechellia* sequence. The second ORF is 45 aa long and spans the boundaries of the two exons of the male-specific mRNA. Therefore, this ORF does not exist in *spx-f* and *spx-s*. The third ORF is 83 aa in the second exon; however, it leads into the polyadenylation signal by using its TAA as the hypothetical stop codon, which, to our knowledge, has not been observed yet in a protein-coding gene. Perhaps the observed rapid elimination of the coding ability in the retroposed sequence is driven by the RNAlization that changes a protein coding sequence into a RNA gene. (*ii*) Even if the ORF length and coding potential are necessary, but not sufficient conditions for discriminating a ncRNA gene from a protein coding gene, the comparative analysis proposed by Eddy (3) provides further evidence for the noncoding property of *spx*. Our polymorphism data in Table 1 show that all three polymorphic sites, which are in the first and third ORFs, are at nonsynonymous sites, reflecting no protein-coding constraint as required by functional proteins (39). (*iii*) All these small conceptual peptides do not match with any known functional peptides in databases. These observations together provide strong evidence for the non-protein-coding property of *spx* as an RNA gene.

Origin of RNA Genes. How a new RNA gene originated has been explored with interesting insights. Brosius and coworkers have identified two young RNA genes, BC1 and BC200 (5, 40), which were created by retroposition tens of millions of years ago (6, 7). These genes used adjacent genomic regions as raw materials for new regulatory sequences during the exaptation process for new functions. In our investigation, *spx*, which is much younger (around 2 million years old) than these genes, provides an opportunity to further characterize the early stage of new genes in evolution. The details in the molecular tinkering are so clear that it is observable that the retroposition of the ATP synthase chain F gene was even associated with a degenerated DNA transposon. Subsequent events involved the recruitment of the regulatory sequence together with a novel exon. All fine features in sequence evolution were observed, including rapid changes of gene structures and a high rate of base substitutions, which were likely positively selected for a novel RNA gene function. Therefore, a similar exon-shuffling molecular mechanism, which created *jingwei*, can also be used to generate new RNA genes although the reincarnated consequences are different: JINGWEI became a protein with chimeric peptide domains (14), whereas *spx* was evolved into a noncoding RNA gene with both sex- and development-specific expression patterns.

The new functions assumed by *spx* may be fundamental in the physiology of *D. melanogaster*, because the dimorphism pattern of polymorphisms was found to be similar across different environments, as the two local populations (NFS and SFS), and worldwide samples (populations from all continents) revealed (27). The dimorphisms in all locations provide strong evidence for the role of balancing selection in the evolution of the chromosome fragment that contains *spx* although it is unclear whether or not the *spx* is the target of selection.

We thank the members of M. Long laboratory for critical discussions, W. Gilbert, J. Brosius, J.-L. Nahon, L. Patthy, and J. J. Emerson for their comments on the manuscript. This work was supported by two grants from the National Sciences Foundation, and by David and Lucile Packard Fellowships for Science and Engineering (to M.L.). Support was also provided by Israel Science Foundation (ISF Grant no. 157/98-17.3) and the Ancell–Teicher Research Foundation for Molecular Genetics and Evolution (to E.N.).

1. Gesteland R. F., Cech, T. R. & Atkins, J. F. (1999) *The RNA World* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
2. Eddy, S. R. (1999) *Curr. Opin. Genet. Dev.* **9,** 695–699.
3. Eddy, S. R. (2001) *Nat. Rev. Genet.* **2,** 919–929.
4. Erdmann, V. A., Szymanski, M., Hochberg, A., de Groot, N. & Barciszewski, J. (2000) *Nucleic Acid Res.* **28,** 197–200.
5. Huttenhofer, A., Kiefmann, M., Meier-Ewert, S., O'Brien, J., Lehrach, H., Bachellerie, J. P. & Brosius, J. (2001) *EMBO J.* **20,** 2943–2953.
6. DeChiara, T. M. & Brosius, J. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2624–2628.
7. Martignetti, J. A. & Brosius, J. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 9698–9702.
8. Skryabin, B. V., Kremerskothen, J., Vassilacopoulou, D., Disotell, T. R., Kapitonov, V., Jurka, J. & Brosius, J. (1998) *J. Mol. Evol.* **47,** 677–685.
9. Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J. & Willard, H. F. (1992) *Cell* **71,** 527–542.
10. Avner, P. & Heard, E. (2001) *Nat. Rev. Genet.* **2,** 59–67.
11. Gilbert, W. (1987) *Cold Spring Harb. Synp. Quant. Biol.* **52,** 901–905.
12. Patthy, L. (1997) *Gene* **238,** 103–114.
13. Long, M., Rosenberg, C. & Gilbert, W. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 12495–12499.
14. Long, M. & Langley, C. H. (1993) *Science* **260,** 91–95.
15. Nurminsky, D. I., Nurminskaya, M. V., De Aguiar, D. & Hartl, D. L. (1998) *Nature (London)* **396,** 572–575.
16. Thomson, T. M., Lozano, J. J., Loukili, N., Carrio, R., Serras, F., Cormand, B., Valeri, M., Diaz, V. M., Abril, J., Burset, M., *et al.* (2000) *Genome Res.* **10,** 1655–1657.
17. Courseaux, A. & Nahon, J.-L. (2001) *Science* **291,** 1293–1297.
18. Wang, W., Zhang, J., Alvarez, C., Llopart A. & Long M. (2000) *Mol. Biol. Evol.* **17,** 1294–1301.
19. Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000) *Science* **287,** 2185–2195.
20. Nevo, E. (1995) *Proc. R. Soc. London B* **262,** 149–155.
21. Nevo, E. (1997) *Theor. Popul. Biol.* **52,** 231–243.
22. Nevo, E. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 6233–6240.
23. Nevo, E., Rashkovetsky, E., Pavlicek, T. & Korol, A. (1998) *Heredity* **80,** 9–16.
24. Harry, M., Rashkovetsky, E., Pavlicek, T., Baker, S., Derzhavets, E., Capy, P., Cariou, M.-L., Lachaise, D., Asada, N. & Nevo, E. (1999) *Biologia, Bratislava* **54,** 683–703.
25. Korol, A., Rashkovetsky, E., Konstantin, I., Michalak, P., Ronin, Y. & Nevo, E. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 12637–12642.
26. Michalak, P., Minkov, I., Helin, A., Lerman, D. N., Bettencourt, B. R., Feder, M., Korol, A. B. & Nevo E. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 13195–13200.
27. Wang, W., Thornton, K., Berry, A. & Long, M. (2002) *Science* **295,** 134–137.
28. Merriman, P. J., Grimes, C. D., Ambroziak, J., Hackett, D. A., Skinner, P. & Simmons, M. J. (1995) *Genetics* **141,** 1425–1438.
29. Long, M., Wang, W. & Zhang, J. (1999) *Gene* **238,** 135–141.
30. Brosius, J. (1991) *Science* **251,** 753.
31. Brosius, J. (1999) *Gene* **238,** 115–134.
32. Makalowski, W. (2000) *Gene* **259,** 61–67.
33. Stephen, R. M. & Schneider, T. D. (1992) *J. Mol. Biol.* **228,** 1124–1136.
34. Long, M. & Deutsch, M. (1999) *Mol. Biol. Evol.* **16,** 1528–1534.
35. Pritchard, J. K. & Schaeffer, S. W. (1997) *Genetics* **147,** 199–208.
36. Powell, J. R. (1997) *Progress and Prospects In Evolutionary Biology: The Drosophila Model* (Oxford Univ. Press, New York).
37. Sullivan, D. T., Starmer, W. T., Curtiss, S. W., Menotti-Raymond, M. & Yum, J. (1994) *Mol. Biol. Evol.* **11,** 443–458.
38. Begun, D. (1997) *Genetics* **145,** 375–382.
39. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, U. K.).
40. Tiedge, H., Chen, W. & Brosius, J. (1993) *J. Neurosci.* **13,** 2382–2390.