Recurrent exon shuffling between distant P-element families

Danielle Nouaud, Hadi Quesneville, Dominique Anxolabéhère

▶ To cite this version:

Danielle Nouaud, Hadi Quesneville, Dominique Anxolabéhère. Recurrent exon shuffling between distant P-element families. Molecular Biology and Evolution, 2003, 20 (2), pp.190-199. 10.1093/molbev/msg036 . hal-02678803

HAL Id: hal-02678803
https://hal.inrae.fr/hal-02678803

Submitted on 31 May 2020

Recurrent Exon Shuffling Between Distant *P*-Element Families

Danielle Nouaud, Hadi Quesneville, and Dominique Anxolabéhère

Institut Jacques Monod, Dynamique du Génome et Evolution, CNRS–Universités PM Curie et D. Diderot, Paris, France

Two independent stationary *P*-related neogenes had been previously described in the *Drosophila obscura* species group and in the *Drosophila montium* species subgroup. In *Drosophila melanogaster*, *P*-transposable elements can encode an 87 kDa transposase and a 66 kDa repressor, but the *P*-neogenes have only conserved the capacity to encode a 66 kDa repressor-like protein specified by the first three exons. We have previously analyzed the genomic modifications associated with the transition of a *P*-element into the *montium P*-neogene, the coding capacity of which has been conserved for around 20 Myr (Nouaud, D., and D. Anxolabéhère. 1997. Mol. Biol. Evol. **14**:1132–1144). Here we show that the *P*-neogene of some species of the *montium* subgroup presents a new structure involving the capture of an additional exon from a very distant *P*-element subfamily. This additional exon is inserted either upstream or downstream of the first exon of the *P*-neogene. As a result of alternative splicing, these modified neogenes can produce, in addition to the repressor-like protein, a new protein which differs only by the NH2-terminal region. We hypothesize that this protein diversity within an organism results in a functional diversification due to the selective advantage associated with the domestication of the *P*-neogene in these species. Moreover, the autonomous *P*-element which provides the additional exons is still present in the genome. Its nucleotide sequence is more than 45% distant from the previously defined *P*-type element (M-type, O-type, T-type) and defines a new *P*-type element subfamily referred to as the K-type.

Introduction

The increase of transposons within a species is due primarily to their ability to replicate and not to a selective advantage for the host. These selfish and complex pieces of DNA harbor a powerful potential repertoire of new functional genetic abilities; consequently, if the host genome succeeds in domesticating such transposable elements (TEs) by taming their "anarchic behavior," these repetitive DNAs could be responsible for important evolutionary innovations (Miller et al. 1999; for a review see Kidwell and Lisch 2001). Transposable elements can be a factor in gene evolution not only by supplying *cis* regulatory domains to host genes but also by coding novel cellular functions. The transition from a genomic parasite to a stable integrated gene that is useful to the host has been described as "molecular domestication" (Miller et al. 1992). Recently, a scan of the draft human genome sequence has identified at least 47 genes probably derived from transposable elements (International Human Genome Sequencing Consortium 2001). It is notable that 43 of 47 of these TE-derived genes are composed of DNA transposon despite the fact that this class represents only a small proportion of the interspersed repeats in the human genome.

The two first examples of a molecular transition of a DNA transposon coding sequence into a stable integrated host gene were provided by studies on the Drosophila *P*-transposable element family. Indeed, stationary *P*-element–related neogenes have been discovered, one in a species belonging to the *obscura* species group (Paricio et al. 1991; Miller et al. 1992), and the other in a member of the *Drosophila montium* species subgroup (Nouaud and Anxolabéhère 1997). Both belong to the same *P*-subfamily: the T-type as defined by Hageman, Haring, and Pinsker (1996). Although the functional properties of the *P*-element-derived neogenes in their respective host are still unknown, this system provides the first example of multiple independent acquisition of the same type of TE-derived coding section in Drosophila evolution (Nouaud et al. 1999).

Autonomous *P*-transposable elements were initially discovered in *Drosophila melanogaster* (Rubin, Kidwell, and Bingham 1982) then in several other distant Drosophila lineages (Simonelig and Anxolabéhère 1991; Hagemann, Haring, and Pinsker 1996.). The molecular structure of the canonical *P*-element includes four exons (numbered 0 to 3) encoding two proteins completed by an alternative splicing of the primary transcript: an 87-kDa transposase (exons 0 to exon 3), and a 66-kDa protein (exons 0 to exon 2) which acts as a repressor of transposition (O'Hare and Rubin 1983; Rio, Laski, and Rubin 1986; Robertson and Engels 1989) (fig. 1*A*). The third intron is spliced exclusively in the germline and thus limits transposase synthesis to this tissue (Laski, Rio, and Rubin 1986).

In the stationary *P*-element–related neogenes, the terminal inverted repeats are missing or restricted to a skeleton and the coding region lacks the transposase-specific exon 3. In the *obscura* species group, the repeats are tandemly clustered (10 to 50 copies), but in the species of the *montium* subgroup they are present in a single copy. These *P*-neogenes probably derive from previously mobile *P*-elements which have undergone, in the course of the transposition at the neogene genomic location, structural modifications causing their immobilization and changing their *cis*-regulatory section. Although both the *obscura* and the *montium P*-derived neogenes are transcribed in adult flies into polyadenylated RNAs that encode *P*-repressor-like protein, the 5′ regulatory sections that drive the expression have different origins. In the case of the *obscura P*-neogene, a novel promoter region evolved from the TE-insertion of unrelated transposons (Miller et al. 1995, 2000). By contrast, the 5′ regulatory section of *montium P*-neogene might have evolved from an intergenic sequence flanking the *P*-element. Surprisingly, the

Key words: *P*-element, transposon, exon shuffling, molecular domestication, *Drosophila montium* subgroup.

E-mail: anxo@ccr.jussieu.fr.

Mol. Biol. Evol. 20(2):190–199. 2003
DOI: 10.1093/molbev/msg036

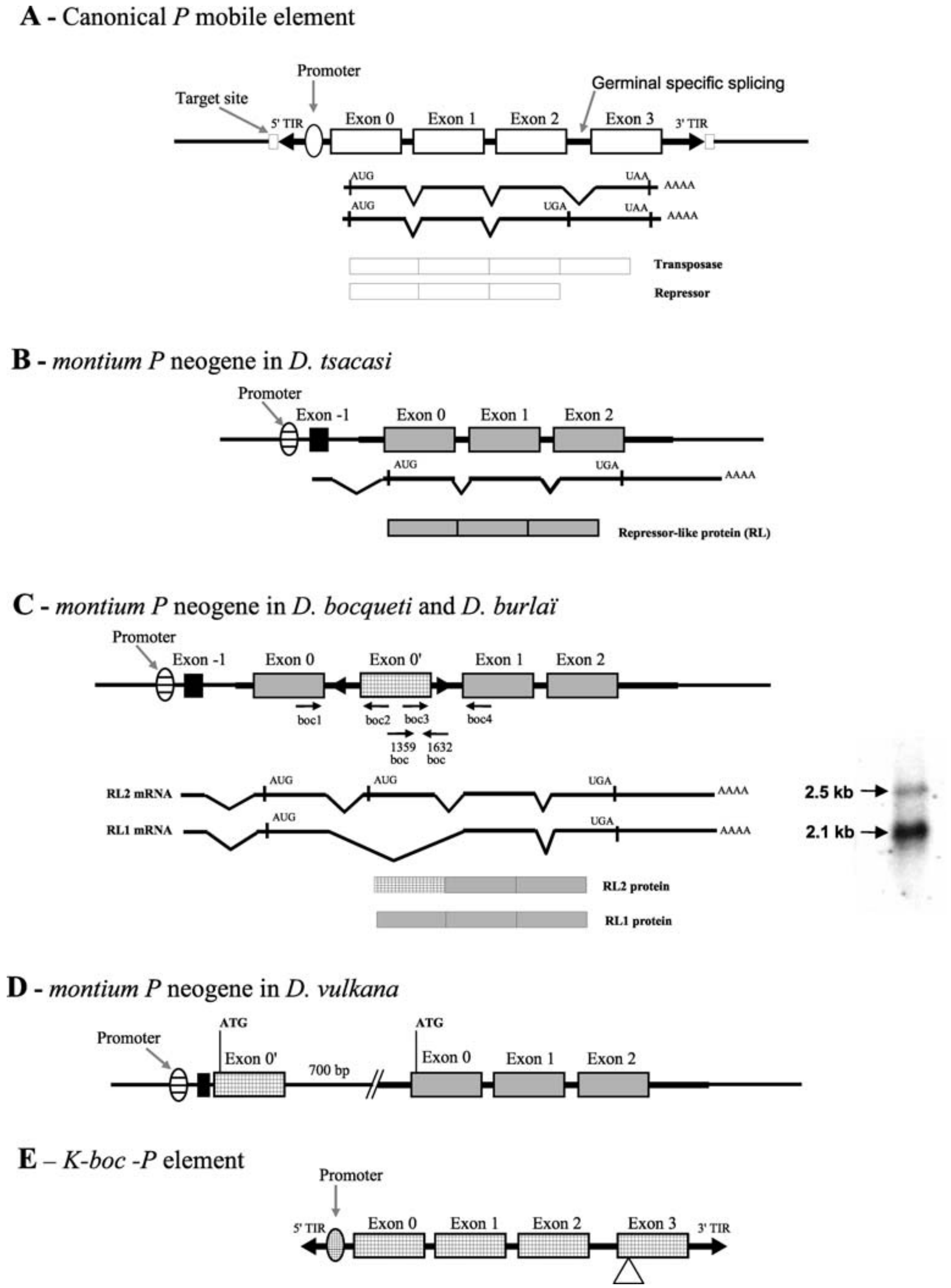

FIG. 1.—Schematic representations of the organization of the *P*-autonomous elements and of the *montium P*-neogenes. The gene structures are represented with their products where these have been identified. (*A*) Autonomous *P*-element as described in *D. melanogaster* (Rio, Laski, and Rubin 1986). (*B*) Standard *P*-neogene as described in *D. tsacasi* (Nouaud et al. 1999). (*C*) Rearranged *montium P*-neogene as described in *D. bocqueti*. On the right, Northern blot on adult transcripts hybridized with a riboprobe spanning exons 1 and 2. (*D*) Rearranged *montium P*-neogene as described in *D. vulkana*. (*E*) New type of autonomous *P*-element as cloned from *D. bocqueti*. Boxes correspond to exons and their corresponding regions of the proteins (open boxes specific to the canonical *P*-element; gray boxes specific to standard *montium P*-neogene; angled bar boxes specific to the K-type *P*-element). The additional intron present in exon 3 of K-type *P*-element is shown as a triangular insertion.

montium P-neogene contains a new exon (exon-1) and a new intron (intron-1) upstream of the original *P*-sequence insertion site (fig. 1*B*) (Nouaud et al. 1999). Thus, the two *P*-repressor-like neogenes have recruited flanking genomic sequences as new regulatory regions that may result in different expression patterns leading to distinct novel functions of the proteins.

In the present article we describe two independent events of exon shuffling that have taken place within the *montium P*-neogene. They have resulted in the capture of an additional exon from a very distant *P*-element subfamily described here for the first time. Each novel structure of the *P*-neogenes encodes two putative proteins sharing the same COOH-terminal region but strongly divergent for their NH2-terminal part. We hypothesize that this protein diversity within an organism results in functional diversification.

Materials and Methods

Fly Stock Sources

Stock flies were obtained from the CNRS Laboratoire Populations, Génétique et Evolution, Gif-sur-Yvette, France.

DNA Hybridization Analysis and Cloning

Genomic DNA was digested with restriction enzymes according to the manufacturers' instructions. Restriction fragments were separated by electrophoresis in agarose gels, and then transferred onto a nitrocellulose membrane (Schleicher and Schuell) according to standard protocols (Maniatis, Fritsch, and Sambrook 1982). The probes used were synthesized by polymerase chain reaction (PCR), either from the cloned neogene *P-boc* (Nouaud et al. 1999) for the probe specific to exon 0′ or from the cloned *K-boc-P* (present work) as a template for the probe specific to exon 3. The primers 1359 (5′-TGTGGGAAAAATCCTTAGAATGC–3′) and 1632 (5′CTAGATGATAGTTGTTGCA 3′) yield an amplified fragment of 293 bp specific to exon 0′ of the *P-boc* neogene (see *Results* and fig. 1*C*). The primers 1938 (5′-CATTCACATTTTTCGCAGCC-3′) and Reverse primer belonging to the polylinker of the TA-cloning vector at the *K-boc-P* 3′ yield an amplified fragment of 1.1 kb specific to the region of exon 3. Probes were labeled with ^{32}P, with the random primed kit (Amersham). Prehybridization and hybridization conditions were 6× SSC, 5× Denhardt, 0.5% SDS, and 150 μg/ml of salmon sperm at 65°C, and washing was done twice at 65°C in 2× SSC and 0.1% SDS.

RNA Isolation and Northern Analysis

Total RNA was isolated from adults using RNAzol reagent (Bioprobe system). Poly(A)$^+$ RNA was purified through an oligo(T) column and separated by electrophoresis in 1.3% agarose formaldehyde gel and transferred onto a nitrocellulose membrane.

RT-PCR Experiments

Reverse transcription (RT) of total RNA and subsequent PCR were carried out with the OneStep RT-PCR kit (Qiagen) according to the supplier's recommendations. The primers used to detect transcripts of the *P-boc* neogene are shown in figure 1*C*. From the mRNAs, the primers boc1 (5′-GCATTTTGATGCGTCCCAGTGG-3′) and boc2 (5′-GTCTTGGCAGGGCGTTTGGC-3′) were expected to amplify a product of 437 bp; the primers boc3 (5′-GACACACATTTCAAAGCATCGG-3′) and boc4 (5′-ACTGCTCGAGCTGCTGACGC-3′) were expected to amplify a product of 248 bp; and the primers boc1 and boc4 gave a product of 261 bp. The amplified products were cloned into the pCR2 vector from the Topo TA-cloning kit (Invitrogen) and introduced into *Escherichia coli* INV α F′ competent cells. Plasmid DNA was prepared for sequencing with the QIAprep kit (Qiagen). Automatic sequencing was done with the ABI Prism BigDye Terminator Cycle Sequencing Ready Reaction (Applied Biosystems).

Table 1
***P*-Sequences Used in This Study**

Sequences	Species	Code	GenBank Accession Number
P-tsa[a]	*Drosophila tsacasi*	Dtsa	AF0160036
P-boc[b]	*Drosophila bocqueti*	Dboc	AF169142
P-vul[b]	*Drosophila vulkana*	Dvulk	AY116625
P-bu[b]	*Drosophila burlai*	Dbur	AY1166252
K-boc-P[b]	*Drosophila bocqueti*	Kboc	AY116624
*P*π*25.1*[c]	*Drosophila melanogaster*	Dmel	PPI251
O-type[d]	*Drosophila bifasciata*	DbifO	X71634
M-type[e]	*Drosophila bifasciata*	DbifM	X61795
P-helvet[f]	*Drosophila hervetica*	Dhelvet	AF313771
T-type[g]	*Drosophila ambigua*	Damb	AF012414
PS18[h]	*Scaptomyza pallida*	Spal18	M63342
P-musca[i]	*Musca domestica*	Musca	AF183396
P-luci[j]	*Lucilia cuprina*	Luci	M89990

[a] Nouaud and Anxolabéhère 1997.

[b] This study.

[c] O'Hare and Rubin 1983.

[d] Hagemann, Miller, and Pinsker 1992.

[e] Hagemann, Miller, and Pinsker 1994.

[f] Haring, Hagemann, and Pinsker 2000.

[g] Hagemann, Haring, and Pinsker 1996.

[h] Simonelig and Anxolabéhère, 1991.

[i] Lee and Kidwell 1999.

[j] Perkins and Howells 1992.

Sequence Analysis

Sequences used in this study are listed, together with their accession numbers, in table 1. Nucleotide and amino acid alignments of autonomous *P*-elements were made using PILEUP program (Genetics Computer Group 1991) with the default options and then optimized by hand. Pairwise distance matrices were inferred using the Kimura correction methods. Phylogenetic analysis was performed by the Neighbor-Joining method following the procedure indicated in the text.

Results

Exon Insertion Events in the *montium* Stationary *P*-Neogene

In a previous study we cloned and totally or partially sequenced 12 of 18 *montium P*-neogenes. In seven species (*D. bicornuta*, *D. davidi*, *D. jambulina*, *D. nikananu*, *D. seguyi*, *D. serrata*, *D. tsacasi*), the size of the *P*-neogene is consistent with the size expected from a *P*-neogene similar to that described in *D. tsacasi* (fig. 1*B*) (Nouaud et al. 1999). In the five other species (*D. bakoue*, *D. bocqueti*, *D. burlai*, *D. malagassya*, *D. vulcana*), the size of the *P*-neogenes is greater than expected, suggesting the presence of DNA insertions. The *P*-neogenes of *D. bocqueti* (*P-boc*) and *D. vulkana* (*P-vul*) have been entirely sequenced (accession numbers AF169142 and AY116625).

```
Dtsa  CCTACCTATCGATGTTTATCG...................CCCGCCACTAATGCCTTGTGCCATCATTAGTTTA..TGTTCAGCTGGAAA......AATAATCATCAACTATTAGGACTTT
Dboc  ---------------------...................------t-c-g-t--c---------g--------....------c-----aataat---------a--------c------
Dbur  ---------------------...................------t---g-t--c---------g--------....------c-----aataat---------a---------------
Dvul  ---------------------.............cccccc---c--c---gc------------c----c-c-gc------.t-ca--..........--ca-------g---a------
Dbak  ---------------------.............cccccc---c--c---gc------------c----c-c-gc------.t-ca--..........--ca-------g---a------
Dmala --------------c------acgcgcccccccccccccccc---c--caagg----------t----c------c-gc----------t-..........--ca------g---a------
```

FIG. 2.—Conservation of the promoter region: nucleotide sequence alignments of the 5′ region of the *P*-neogenes with or without exon duplication. Gray box: promoter region of the *P*-neogene in *D. tsacasi*; the arrow indicates the start of transcription (Nouaud et al. 1999). Names of sequences are defined in table 1. The identical nucleotides are represented by dashes. The nonconserved nucleotides are in lower case.

Insertion of a New Coding Exon Downstream of Exon 0 of the P-Neogene of Drosophila bocqueti

A comparison of the structures of the *D. tsacasi* and *D. bocqueti P*-neogenes (fig. 1*B* and *C*) shows that an immobilized and internal deleted *P*-element is inserted inside the intron (0, 1) separating exon 0 and exon 1 in the *D. bocqueti P*-neogene. This *P*-sequence insertion is 556 bp long (accession number AF169142 from nucleotides 1049 to 1604). It is flanked by a direct 8 bp duplication corresponding to the duplication of the target site, with one mismatch. The 31 bp of the 3′ terminal inverted repeat (TIR) are 87% identical to the sequence of the *D. melanogaster P*-mobile element TIR. The first 13 bp of the 5′ TIR are missing. This internal insertion retains an intact open reading frame (ORF) corresponding to exon 0 of the canonical *P*-element. Hereafter, this insertion will be called *InsPboc* and its exon, exon 0′. The identity between exon 0′ and the first coding exon (exon 0) of the *P-boc* neogene is 54.4% and 43.3% at the nucleotide and amino-acid levels, respectively. Northern blot analysis was performed on adult poly(A)$^+$ RNA with a riboprobe obtained from the subcloned region of exons 1 and 2 of the *P-tsa* neogene. The probe was synthesized using T7 RNA polymerase and labeled with [^{32}P]UTP. As shown in figure 1*C*, a 2.5-kb transcript and a 2.1-kb transcript were detected. The difference between the sizes of the two transcripts corresponds to that expected if alternative splicing occurs, joining either exon 0 to exon 0′ and exon 0′ to exon 1, or exon 0 to exon 1. The complete RNA processing results in two mRNAs: one including exons –1, 0, 0′, 1 and 2 (2.5 kb) and the second including exons –1, 0, 1 and 2 (2.1 kb) (fig. 1*C*). As the probe used for the Northern blot covers the same part of the two transcripts, the difference in intensity between them probably results from quantitative differences in the adults. This alternative splicing was confirmed by RT-PCR. Transcripts were extracted from adults and the cDNA was synthesized as described in *Materials and Methods*. The primers designed for the cDNA amplification are shown in figure 1*C*. The sequences of the amplified products confirm that the alternative splice uses the donor and acceptor splicing sites corresponding to those in the canonical *P*-transposable element (Laski et al. 1986).

The sequence of the 2.1-kb transcript has the coding capacity for a protein 574 amino acids long. Hereafter this protein will be called repressor-like 1 (RL1). The 2.5-kb transcript could also be translated from the conventional start of translation present in exon 0 or in exon 0′. The translation initiated from exon 0 ceases at the beginning of exon 0′ because of the presence of a stop codon (the splicing between exon 0 and exon 0′ does not conserve the phase in exon 0′). In contrast, the translation initiated from the conventional AUG of exon 0′ leads to a protein of 570 AA, which will hereafter be called repressor-like 2 protein (RL2).

A similar structure is found in *D. burlai*. (accession number AY116626), a sibling species of the *bocqueti* complex of species (Lemeunier et al. 1986). In this species, the *P*-neogene contains an insertion of 501 bp, inserted at the same site as in *D. bocqueti*, indicating that the primary insertion event took place in a common ancestor of the two species. This insertion, hereafter called *InsPbur*, present TIRs which have the same characteristics as *InsPboc*, except for a 7-bp insertion inside the 3′ TIR. Thus, it cannot be *trans*-mobilized. *InsPbur* presents an ORF with 93 amino acids showing 92.5% identity with exon 0′ of *InsPboc* The identities between exon 0′ for *InsPbur* and exon 0 of the *P-bur* neogene are 51.5% and 42.2% at the nucleotide and amino-acid levels, respectively. Moreover, the sequence analysis shows the conservation of the same splice sites experimentally determined in *P-boc* neogene. Consequently, the *P-bur* neogene would provide two proteins with 96.5% and 95.3% identity with the corresponding RL1 and RL2 proteins, respectively, of the *P-boc* neogene.

Another Example of Exon Shuffling: Insertion of a New Exon Upstream of Exon 0 of the D. vulcana *P-Neogene*

A comparison of the structure of the *D. tsacasi P*-neogene with that of *D. vulkana* (fig. 1*B* and *D*) shows that an internal deleted *P*-element is inserted inside exon –1 of the *D. vulcana P*-neogene. This insertion, hereafter called *InsPvul*, is 350 bp long and has conserved an intact ORF corresponding to exon 0′ described above. A skeletal *P*-element 5′ TIR can still be identified in the sequence upstream of this ORF, but no significant identity with a 3′ TIR is detectable in the downstream region. The nucleotide comparison between the *InsPvul* coding sequence and exon 0 of the *P-vul* neogene shows an identity of 51.1%. The structural similarity between *InsPboc* and *InsPvul* and their high nucleotide sequence identity (83.9%) make it possible to deduce the putative transcripts of the *P-vul* neogene from the splicing sites experimentally identified for the *P-boc* neogene (see *Discussion*).

The *P*-neogenes of *D. bakoue* and *D. malagassya* have been partially sequenced; upstream of exon 0, they present the same insertion as the *P-vul* neogene, located at the same target site (data not shown). These two species belong to the same complex of species as *D. vulkana* (the *bakoue* complex of species, Lemeunier et al. 1986). This indicates that this insertion event occurred in their common ancestor. The additions of exons into the *P*-neogenes described above are not accompanied by any other structural modifications. It is remarkable that, as shown in figure 2, the sequence upstream of exon −1 is highly conserved when compared to the promoter region in the *P*-neogene of *D. tsacasi* (Nouaud et al. 1999).

Identification of the Exon 0′ Master Copy

The nucleotide divergences between the insertions *InsPboc* or *InsPvul* and the numerous *P*-sequences registered in the data banks are all greater than 35%, implying that they do not belong to a previously described *P*-element subfamily (Clark and Kidwell 1997; Pinsker et al. 2001). Moreover, each of them could result from the insertion of a complete *P*-element, followed by large deletions, leaving the region (including the full coding region of the first exon) inserted. Because of their identity (83.9%), these insertions should derive from a same *P*-element subfamily. These results support the hypothesis that the genome of the species *D. bocqueti* and *D. vulcana* and their related species harbor an active *P*-element family which is at the origin of exons 0′ identified in several *montium P*-neogenes.

Southern blot experiments were performed with genomic DNA from six species belonging to the *montium* subgroup (*D. bocqueti*, *D. burlaï*, *D. kikkawai*, *D. nikananu*, *D. tsacasi*, and *D. vulkana*). DNA samples were digested with *Pst I* endonuclease, and after electrophoresis the restriction fragments were bi-transferred onto a nitrocellulose membrane. One filter was hybridized with the exon 0′–specific fragment amplified with the primers 1359 and 1632 from the clone containing the *P-boc* neogene as a template (see *Materials and Methods*). A number of hybridization signals are present in *D. bocqueti*, as well as in other species (fig. 3*A*), showing that the inserts *InsPboc* and *InsPvul* belong to a repeated dispersed *P*-element family. In an attempt to isolate *P*-elements at the origin of exon 0′, a long-range PCR amplification was performed on *D. bocqueti* DNA as a template with a primer (5′CATAATGG-AATAACTATAAGGTGG3′) corresponding to the first 24 bp of the 3′ TIR sequence of *Insboc*. Full-length and deleted *P*-elements have been cloned by the TA-cloning method (Invitrogen) from PCR products. Some have been sequenced. The sequence of a complete *P*-element (accession number AY116624), described in figure 4, has the coding capacity of an autonomous *P*-element. This element is called the *K-bok-P*-element (Kenya-*bocqueti P*-element, for the *D. bocqueti* strain originated in Kenya). Six other *K-boc* sequences are partially sequenced. The divergence between them is less than 5%. They are available by request. The *K-boc-P*-element is 3300 bp long and its termini are formed by 31 bp inverted repeats. The difference in length between *K-boc-P* and the canonical *P*-element (fig. 1*A*) results from two features: (1) the intron between exon 0 and exon 1 is unusually long in *K-boc-P* (264 bp as opposed to only about 50 bp in the other *P*-elements), and (2) exon 3 is interrupted by an additional 172bp intron. However, the *K-boc-P*-element shares a number of structural features with the autonomous *P*-element from other Drosophila species (*D. melanogaster*, *D. bifasciata*, *S. pallida*). Subterminal inverted repeats (SIRs) of 10 bp (positions 33–42 and 3259–3268) and 11 bp with one mismatch (positions 127–137 and 3161–3171) are found in the 5′ and 3′ noncoding regions. These locations correspond to those of SIRs in the *P*-elements of the other species, thus implying a functional equivalence. Moreover, exon 1, like the *D. melanogaster* and *Scaptomyza pallida P*-elements (Simonelig and Anxolabéhère 1991), presents inverted repeats of 17 bp separated by 29 bp (positions 942–958 and 988–1004). The consensus 5′- and 3′-splice sites of the exons are conserved and the additional intron inside the exon retains the coding capacity of the *K-boc-P*-element. The putative protein is 721 amino acids long and has a molecular weight of 83 kDa (fig. 4). It is remarkable that Cys, His, Arg, Lys, and Trp are over-represented in the first 70 amino-acids of the N-terminal section (35.7 % compared to 17.5% in the rest of the protein). Moreover, the CCHC putative metal-binding site present in the canonical *P*-element (Miller et al. 1995; Lee, Mul, and Rio 1996; Miller et al. 1999) can be recognized at the same position in the *K-bok-P* protein. These results suggest that the features of DNA-binding domains are present in the N-terminal sections of the putative transposase of the *K-boc-P*-element. Furthermore, by comparison with the *D. melanogaster P*-element, other functionally important sections are also conserved: the three leucine-zipper motifs are found at the same locations as is the helix-turn-helix motif, which shows only four mismatches out of 19 residues (fig. 4).

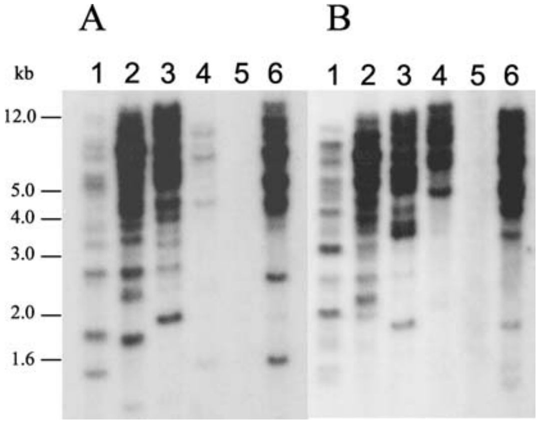

FIG. 3.—Southern blot analysis of genomic DNA from six species of the *montium* subgroup. All DNA samples were digested by PstI and probed (*A*) with the PCR amplified fragment of 293 bp by primers 1359 and 1632 from the *P-boc* clone (see fig. 1*C*) and (*B*) with the PCR amplified fragment by the primers 1938 and reverse primer (see *Materials and Methods*) from the *K-boc-P* clone. Each lane contained 10 μg of DNA and the gel was bi-transferred onto the nitrocellulose filters A and B. 1: *D. tsacasi*, 2: *D. bocqueti*, 3: *D. burlai*, 4: *D. vulkana*, 5: *D. kikkawai*, 6: *D. nikananu*.

The second filter from the bi-transferred DNA samples described above was hybridized with a PCR product synthesized from exon 3 specific to the transposase of the cloned *K-boc-P*-element. As shown in figure 3*B*, a number of hybridization signals are detected in *D.*

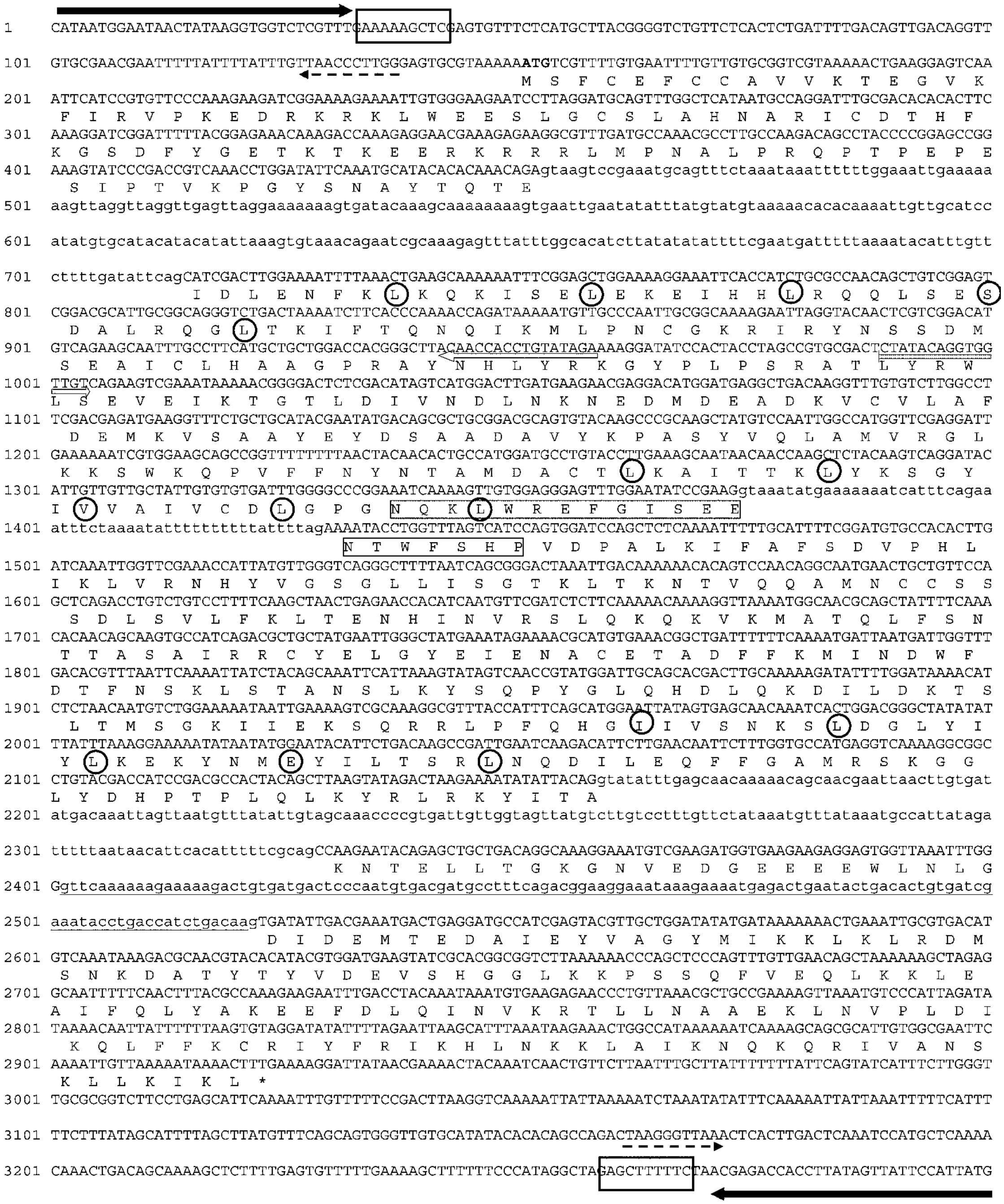

FIG. 4.—Nucleotide and derived amino acid sequences for the *K-boc-P*-element of *D. bocqueti* (accession number AY116624). The terminal inverted repeats are identified with solid arrows. The 10-bp transposase binding sites that have been defined at both ends of the *D. melanogaster P*-element (Kaufman, Doll, and Rio 1989) are boxed. Internal inverted repeats are underlined with broken and shaded arrows. The intronic sequences are indicated by lower case. The extra intron within exon 3 is underlined. The amino acid residues constituting the three leucine zippers are circled. The helix-turn-helix motif is boxed.

bocqueti, *D. burlai*, *D. nikananu*, *D. tsacasi*, and *D. vulkana* (but not in *D. kikkawai*), indicating the presence of numerous *P*-elements containing the exon 3 specific to the transposase coding sequence.

To define the relation between the *K-boc-P*-element and the major *P*-element subfamilies as they have been previously characterized in *D. ambigua* (T-type), *D. bifasciata* (M-type and O-type), *D. helvetica* (M-type), *D. melanogaster* (M-type), and *Scaptomyza pallida* (M-type) (for review, see Hagemann, Miller, and Pinsker 1996), the nucleotide and amino acid alignments of these elements together with the *K-boc-P*-element were performed using the Pileup program of the GCG package (Madison, Wis.) and improved manually. The pairwise distances are shown in table 2. The *K-boc-P*-element is very distant from all other *P*-elements

Table 2
Nucleotide and Amino Acid Distances Between Seven Autonomous *P*-Elements

	M-type				O-type	T-type	K-type
	Dhelvet	*DbifM*	*Spa18*	*Dmel*	*DbifO*	*Damb*	*K-boc-P*
Dhelvet	—	6.72	9.68	24.20	37.07	44.58	47.39
DbifM	9.44	—	7.80	22.88	36.52	41.83	46.45
Spa18	11.16	8.60	—	24.33	37.31	44.89	46.38
Dmel	28.07	26.16	28.15	—	38.71	45.59	48.70
DbifO	39.48	36.18	41.03	42.36	—	48.93	54.72
Damb[a]	45.35	42.34	44.69	43.08	46.55	—	58.81
K-boc-P	52.15	49.61	50.15	48.85	52.67	58.55	—

NOTE.—Values are calculated using Kimura's (1980, 1983) method. Numbers above the diagonal are nucleotide distances. Numbers below the diagonal are the amino acid distances.

[a] Deduced from a degenerate T-type *P*-element.

(>0.45): this new full-length *P*-element belongs to a so far unidentified *P*-subfamily. We define this subfamily as the K-type.

A Neighbor-Joining analysis performed on the putative proteins of these *P*-sequences and two additional *P*-sequences from more distant species, *Lucilia cupina* (*Calliphoridae*) (Perkins and Howells 1992) and *Musca domestica* (*Muscidae*) (Lee, Clark, and Kidwell 1999), produces a dendrogram in which the *K-boc-P*-element groups with the elements from the *Drosophilidae* (fig. 5).

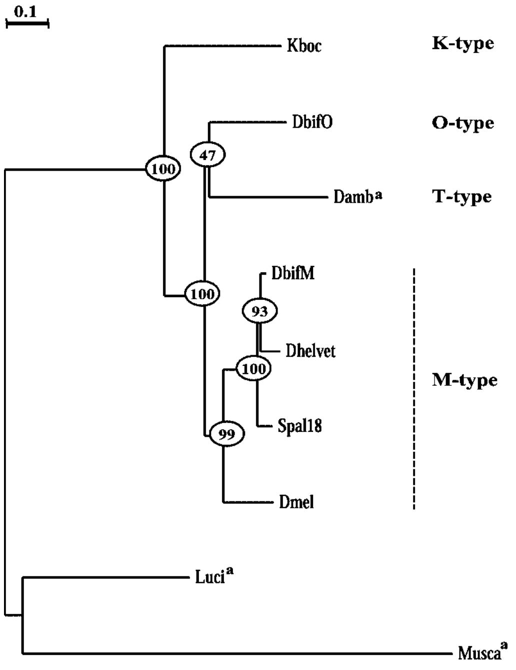

FIG. 5.—Relationship between the *K-boc-P*-element and other *P*-elements based on their amino acid sequences. Multiple alignments were performed by ClustalW (Thompson, Higgins, and Gibson 1994). The dendrogram was created by Neighbor-Joining analysis of a matrix for 497 informative sites using the Phylo_win program (Galtier, Gouy, and Gautier 1996). The scale bar indicates genetic distances in units of residue substitution per site. The bootstrap values are given for each node (they correspond to the percent of 1000 replications). *P*-sequences are noted by code name (see table 1). [a]Amino acid sequence deduced from the degenerate *P*-sequence.

Clark and Kidwell (1997) have performed an extensive phylogenetic analysis of *P*-sequence with 40 species in the *Drosophilidae* using a partial *P*-sequence (449 bp from exon 2). This analysis provided a cladogram in which 16 clades are well supported. To define the position of the *K-boc* element relative to these *P*-element subfamilies, a Neighbor-Joining analysis was performed using this partial internal sequence. Only one or two *P*-sequences representative of each clade defined by Clark and Kidwell's work were included in the analysis. In the new cladogram (fig. 6) the *K-boc-P*-element does not group inside any previously identified clades, confirming that the *K-boc-P*-element does not belong to one of the subfamilies already described.

The position and coding capacity of exons 0′ suggest that the rearranged *P*-neogenes are under host level selection. Direct evidence is provided by a test for selection at the sequence level. The pairwise comparisons of the substitution rates between the exon 0 of the *K-boc* full-length *P*-element and the exon 0′ of the *P*-neogenes in *D. bocqueti*, *D. burlai*, and *D. vulkana*, are presented in table 3 (not enough sequence data were available for the neogenes of *D. malagassia* and *D. bakoue*). All significant results ($P < 0.05$) are due to $d_N/d_S < 1$; that is, they showed evidence of conservative selection. These results are in accordance with those of Witherspoon (1999), obtained using partial sequences of the *P*-neogenes of *D. davidi*, *D. tsacasi*, and *D. kikkawai*. As very few changes occur between exon 0′ of RL2bur and exon 0′ of RL2boc, the test has less power than in the other comparisons giving a nonsignificant statistic.

Discussion

Exon Shuffling Increases the Diversity of Proteins Encoded by *P*-Sequences

The stationary *P*-neogene of the *montium* subgroup has been detected in the 18 species from the *montium* subgroup in which it has been sought, but not in species belonging to any other related subgroup (Nouaud et al. 1999). This result strongly suggests that the domestication event took place in the ancestor of the *montium* subgroup more than 20 MYA. The primordial *P*-element insertion event at the origin of the *P*-neogene was accompanied by a large 3′ terminal deletion covering the last exon and by the capture of a new promoter from the 5′ flanking region, associated with a short noncoding exon and a new intron (Nouaud et al. 1999). Activation by the captured promoter might have resulted in a different expression pattern, thus initiating a novel function for the protein encoded by exons 0–2. This protein, as discussed earlier, has been named repressor-like protein, or RL, since in the canonical *P*-element the corresponding protein acts as a repressor of *P*-transposition. Moreover, through successive speciation events, the orthologous *P*-neogenes have undergone several independent modifications, such as the exon acquisition reported in the present work. The *P-boc* and *P-vul* neogenes have retained a new exon 0 (exon 0′) respectively downstream and upstream of the exon 0 of their neogene. In both cases, the new exon originates from the new K-type *P*-element subfamily.

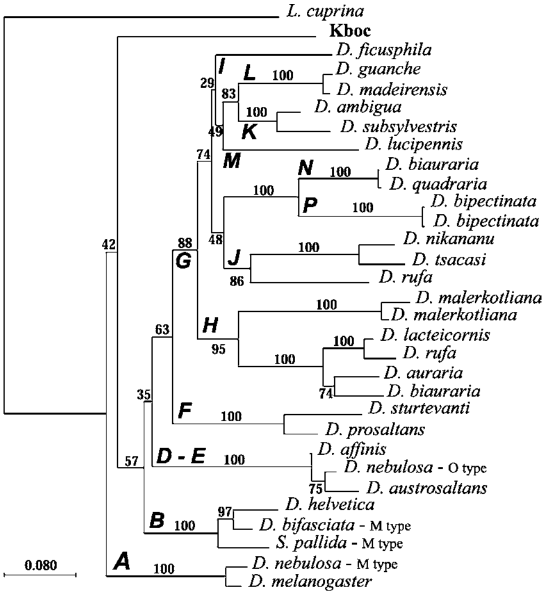

FIG. 6.—Relationship between the *K-boc-P*-element and *P*-nucleotide sequences from the family *Drosophilidae*. Each *P*-sequence is named by the species from which it derives (note that some genomes retain more than one *P*-subfamily). Multiple alignments were performed by ClustalW (Thompson, Higgins, and Gibson 1994). The dendrogram was created by Neighbor-Joining analysis of a matrix for 385 informative sites using the Phylo_win program (Galtier, Gouy, and Gautier 1996). The scale bar indicates genetic distances in units of residue substitution per site. Numbers show the bootstrap percentages (1000 replications). Letters refer to the clades as defined by Clark and Kidwell (1997).

Northern blot and RT-PCR analyses show the presence of two transcripts for the *P-boc* neogene. Nucleotide comparisons between the *P-boc* and the *P-vul* neogenes reveal a high conservation of the acceptor and donor sites at the boundaries of exon 0, exon 0′, and exon 1. On the basis of the sequence of the *P-vul* neogene, we can deduce the existence of transcripts encoding two putative proteins analogous to the RL1 and RL2 proteins of the *P-boc* neogene initiated from the *P*-canonical start codon present in exon 0 and exon 0′, respectively (fig 1*D*). It must be emphasized that three out of four splice sites used to join exons 0, 0′, and 1 are similar to the functional splice sites of the *P*-element. Conversely, the last splice site is a cryptic acceptor site upstream of exon 0′ (*P-boc*) and exon 0 (*P-vul*). In both cases, the splicing of exons 0′ and 1 specific to the RL2 protein has probably been functional from the outset. In each species, the two proteins differ only in their NH2 terminal region.

Interspecific comparison shows that the similarity is 87.1% between the two RL1 proteins and 85.8% between the two RL2 proteins. These values strongly suggest that the exon shuffling has been associated with a selective advantage for the host. Evidence that RL2 proteins are under host level selection is given by an excess of synonymous versus nonsynonymous substitutions in their coding sequence. These d_N/d_S ratios, being lower than 1, suggest that the protein region encoded by exon 0′ of the neogenes has conserved functional characteristics similar to those of functional *K-boc-P*-elements. Moreover, the pairwise comparisons between the *P*-neogenes exon 0′ are in accordance with a host selective pressure.

Table 3
Evidence of Selection in Comparison of the RL2 Products of the *P*-Neogene and the *K-boc-P*-Element (first coding exon)

	d_N/d_S	*P*-Value
Exon 0′ RL2bur–Exon 0′ RL2boc	1.13	>0.05
Exon 0 K-boc–Exon 0′ RL2boc	0.28	<0.01
Exon 0 K-boc–Exon 0′ RL2bur	0.30	<0.05
Exon 0′ RL2vul–Exon 0′ RL2boc	0.27	<0.01
Exon 0′ RL2vul–Exon 0′ RL2bur	0.27	<0.01
Exon 0′ RL2vul–Exon 0 K-boc	0.20	<0.001

NOTE.—Nucleic sequences alignment was performed using ClustalW and the tree was edited to cluster the *P*-neogenes. Pairwise comparisons of the ratio of per site rates of nonsynonymous and synonymous changes (d_N/d_S) was performed using the program CODEML (PAML, version 3.12; Yang 1997). Likelihood ratio tests were calculated to infer statistical significance of the d_N/d_S ratios.

Preliminary experiments showed that the RL1 and RL2 proteins from *P-boc* are produced in vivo in *D. melanogaster* transgenic flies (unpublished data), but the functions of these two proteins are still unknown. Their common region (exons 1 and 2) contains the same three leucine zipper motifs and the coiled-coiled domains characteristic of the P protein. However, pairwise comparisons between RL1 and RL2 restricted to the region corresponding to the first exon show similarities of 57.6%, 53.8%, and 53.7% in *D. bocqueti*, *D. burlai*, and *D. vulkana*, respectively. Thus, in each species, the neogene products strongly differ. Given that the DNA-binding domain is conserved in the N-terminal region of the two proteins, this amino-acid divergence could indicate a diversification of the DNA-binding specificity of the proteins, which in turn could correspond to a functional differentiation.

Recurrent Exonic Insertion Inside the *montium* *P*-neogenes

Surprisingly, the first exon of the *K-boc-P*-elements has been captured twice by the *P*-neogene in the *montium* subgroup, once downstream of exon 0 (in the common ancestor of *D. bocqueti* and *D. burlai*), and once upstream of exon 0 (in the ancestor of *D. vulcana*, *D. bakoue*, and *D. malagassya*) (fig. 1*CD*). These two independent events could be due to the tendency of the *K-boc-P*-element to insert inside the *P*-neogene and to selective advantage associated with the production of the chimeric protein RL2. As observed in *D. melanogaster*, the canonical *P*-element tends to insert in the 5′ end regions of genes (Bellen et al. 1989). If the *K-boc-P*-element has the same property, two insertions inside the 5′ region of the *P*-neogene might have occurred independently in the *montium* subgroup, and the elements could subsequently have undergone internal deletions leaving an intact exon 0. However, a more parsimonious scenario would be that of an insertion of a *K-boc-P*-element into the intron separating exon 0 and exon 1 of the *P*-neogene in the common ancestor of these five species, followed by an internal deletion event. This in turn would have been followed by a local transposition just upstream of exon 0 in the ancestral species at the origin of the clade including *D. vulkana*, *D. bakoue*, and *D. malagassya*. Another scenario can be proposed, *mutatis mutandis*, but with the primary insertion upstream of exon 0. These scenarios are supported by two properties of the *D. melanogaster P*-element: the homing phenomenon and local transpositions. *P*-element transposition occurs by a nonreplicative "cut-and-paste" mechanism beginning with an excision of the element and followed at the donor site by a double-strand gap repair according to a process similar to gene conversion (Engels et al. 1990; Kaufman and Rio 1992). The appearance of double *P*-elements has been explained by a homing phenomenon: the *P*-element transposase, which has an affinity for *P*-elements, may remain attached to the excised element and sometime helps to target it to another *P*-element elsewhere in the genome. The insertion target will thus frequently be the copy of the excised element present on the sister chromatid or on the homologous chromosome (Delattre, Anxolabéhère, and Coen 1995). This could explain why a significant fraction of *P*-element transpositions are local and often lead to *P*-element insertions within or near a second *P*-element (Eggleston 1990; Daniels and Chovnick 1993; Tower et al. 1993; Zhang and Spradling 1993; Dorer and Henikoff 1994; Golic 1994; Delattre, Anxolabéhère, and Coen 1995). These local transpositions can represent up to 80% of transposition events, depending on the insertion site (Golic 1994). This process has been proposed to be at the origin of nested rearranged double *P*-elements. In this study the first event inserted a *K-boc-P*-element directly into the *P*-neogene, which was then followed by a local transposition event.

So far, the *K-boc-P*-element has been detected only in species belonging to the *montium* subgroup. It occurs in species in which the *P*-neogene presents exonic duplications (e.g., *D. bocqueti*), as well as in species in which the neogene does not have such a duplication (i.e., *D. tsacasi*). It should be noted that the *K-boc-P*-element has not been found in the *obscura* group, in which another type of *P*-element domestication took place (Paricio et al. 1991; Miller et al. 1992). As shown by the Neighbor-Joining analysis (fig. 5), the *K-boc-P*-family is very distant from all the other *P*-families, and it is not possible to speculate on the origin of this new *P*-subfamily. It is present in *D. tsacasi*, *D. bocqueti*, *D. burlai*, *D. vulkana*, and *D. nikananu*, but it has not been detected in *D. kikkawai*, or in *D. davidi* and *D. serrata* (data not shown). For the moment, we cannot speculate on the origin of the *K-boc-P*-element, nor do we know whether the patchy distribution inside the *montium* subgroup results from horizontal transfer events.

The molecular domestication of *P*-coding sequences described here, and the two similar events previously described in the *montium* subgroup and in the *obscura* group, demonstrate the creative force of a transposable element as an evolutionary motor that can restructure the genome and lead to the acquisition of novel proteins

Acknowledgments

We thank S. Ronsseray and W. Miller for insightful discussions and comments on the manuscript. This work was supported by the Centre National de Recherche Scientifique (CNRS), the Universities P. and M. Curie and D. Diderot (Institut Jacques Monod, UMR 7592, Dynamique du Génome et Evolution), and the GDR-CNRS 2157 "Evolution des éléments transposables du génome aux populations."

Literature Cited

Bellen, H. J., C. H. O'Kane, C. Wilson, U. Grossnilklaus, R. K. Pearson, and W. J. Gehring. 1989. *P*-element-mediated enhancer detection: a versatile method to study development in Drosophila. Genes Dev. **3**:1288–1300.

Clark, J. B., and M. G. Kidwell. 1997. Phylogenetic perspective on *P* transposable element evolution. Proc. Natl. Acad. Sci. USA **94**:11428–11433.

Daniels, S. B., and A. Chovnick. 1993. *P* element transposition in *Drosophila melanogaster*: an analysis of sister-chromatid pairs and the formation of intragenic secondary insertions during meiosis. Genetics **133**:623–636.

Delattre, H., D. Anxolabéhère, and D. Coen. 1995. Prevalence of localized rearrangements *vs.* transpositions among events induced by Drosophila *P* element transposase on a *P* transgene. Genetics **141**:1407–1424.

Dorer, D. R., and S. Henikoff. 1994. Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. Cell **77**:993–1002.

Eggleston, W. B. 1990. *P* element transposition and excision in Drosophila: interactions between elements. Ph.D. Thesis, Univeristy of Wisconsin, Madison.

Engels, W. R., D. M. Johnson-Schlitz, W. B. Eggleston, and J. Swed. 1990. High-frequency *P* elements loss in *Drosophila* is homologue dependent. Cell **62**:515–525.

Galtier, N., M. Gouy, and C. Gautier. 1996. SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. Comput. Applic. Biosci. **12**:543–548.

Genetics Computer Group. 1991. Wisconsin sequence analysis package. Version X. Genetics Computer Group, Madison, Wis.

Golic, K. G. 1994. Local transposition of P element in *Drosophila melanogaster* and recombination between duplicated elements using a site-specific recombinase. Genetics **137**:551–563.

Hagemann, S., E. Haring, and W. Pinsker. 1996. A new *P* element subfamily from *Drosophila tristis*, *D. ambigua* and *D. obscura*. Genome **39**:978–985.

Hagemann, S., W. J. Miller, and W. Pinsker. 1992. Identification of a complete *P* element in the genome of *D. bifasciata*. Nucleic Acids Res. **20**:409–413.

———. 1994. Two distinct *P* element subfamilies in the genome of *D. bifasciata*. Mol. Gen. Genet. **244**:168–175.

Haring, E., S. Hagemann, and W. Pinsker. 2000. Ancient and recent horizontal invasions of drosophilids by *P* elements. J. Mol. Evol. **51**:577–586.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

Kaufman, P. D., and D. C. Rio. 1992. *P* element transposition *in vitro* proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. Cell **69**:27–39.

Kaufman, P. D., R. F. Doll, and D. C. Rio. 1989. *Drosophila P* element transposase recognizes internal *P* element DNA sequences. Cell **59**:359–371.

Kidwell, M. G., and D. R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution **55**:1–24.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

———. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Laski, F. A., D. C. Rio, and G. M. Rubin. 1986. Tissue specificity of Drosophila *P* element transposition is regulated at the level of mRNA splicing. Cell **44**:7–19.

Lee, C. C., Y. M. Mul, and D. C. Rio. 1996. The *Drosophila P*-element *KP* repressor protein dimerizes and interacts with multiple sites on the *P*-element DNA. Mol. Cell. Biol. **16**:5616–5622.

Lee, S. H., J. B. Clark, and M. G. Kidwell. 1999. A *P* element-homologous sequence in the house fly *Musca domestica*. Insect Mol. Biol. **8**:491–500.

Lemeunier, F., J. R. David, L. Tsacas, and M. Ashburner. 1986. The *melanogaster* species group. Pp. 147–256 *in* M. Ashburner, H. L. Carson, and J. M. Thompson, Jr., eds. The genetics and biology of *Drosophila*. Academic Press, New York.

Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Miller, W. J., S. Hagemann, E. Reiter, and W. Pinsker. 1992. *P* element homologous sequences are tandemly repeated in the genome of *D. guanche*. Proc. Natl. Acad. Sci. USA **89**: 4018–4022.

Miller, W. J., J. F. McDonald, D. Nouaud, and D. Anxolabéhère. 1999. Molecular domestication—more than a sporadic episode in evolution. Genetica **107**:197–207.

Miller, W. J., A. Nagel, J. Bachmann, and L. Bachmann 2000. Evolutionary dynamics of the SGM transposon family in the *Drosophila obscura* species group. Mol. Biol. Evol. **17**: 1597–1609.

Miller, W. J., N. Paricio, S. Hagemann, M. J. Martinez-Sebastian, W. Pinsker, and R. DeFrutos, 1995. Structure and expression of the clustered *P* element homologous in *Drosophila subobscura* and *D. guanche*. Gene **156**:167–174.

Nouaud, D., and D. Anxolabéhère. 1997. *P* element domestication: a stationary truncated *P* element may encode a 66-kDa repressor-like protein in the *Drosophila montium* species subgroup. Mol. Biol. Evol. **14**:1132–1144.

Nouaud, D., B. Boëda, L. Levy, and D. Anxolabéhère. 1999. A *P* element has induced intron formation in *Drosophila*. Mol. Biol. Evol. **16**:1503–1510.

O'Hare, K., and G. M. Rubin. 1983. Structure of transposable elements and the site of insertion and excision in the *Drosophila melanogaster* genome. Cell **34**:25–35.

Paricio, N. M., M. Perez-Alonso, M. J. Martinez-Sebastian, and R. De Frutos. 1991. *P* sequences of *D. subobscura* lack exon 3 and may encode a 66kd repressor-like protein. Nucleic Acids Res. **19**:6713–6718.

Perkins, H. D., and A. J. Howells. 1992. Genomic sequences with homology to the *P* element of *Drosophila melanogaster* occur in the blowfly *Lucilia cuprina*. Proc. Natl. Acad. Sci. USA **89**:10753–10757.

Pinsker, W., E. Haring, S. Hagemann, and W. J. Miller. 2001. The evolutionary life history of *P* transposons: from horizontal invaders to domesticated neogenes. Chromosoma **110**:148–158.

Rio, D. C., F. A. Laski, and G. M. Rubin. 1986. Identification and immunochemical analysis of biologically active Drosophila *P* element transposase. Cell **44**:21–32.

Robertson, H. M., and W. R. Engels. 1989. Modified *P* elements that mimic the *P* cytotype in *Drosophila melanogaster*. Genetics **123**:815–824.

Rubin, G. M., M. G. Kidwell, and P. M. Bingham. 1982. The molecular basis of P-M dysgenesis: the nature of induced mutations. Cell **29**:987–994.

Simonelig, M., and D. Anxolabéhère. 1991. A *P* element of *Scaptomyza pallida* is active in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **88**:6102–6106.

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

Tower, J., G. H. Karpen, N. Craig, and A. C. Spradling. 1993. Preferential transposition of *Drosophila P*-elements to nearby chromosomal sites. Genetics **188**:347–349.

Witherspoon, D. J. 1999. Selective constraints on *P*-element evolution. Mol. Biol. Evol. **16**:472–478.

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

Zhang, P., and A. C. Spradling. 1993. Efficient and dispersed local *P* element transposition from Drosophila females. Genetics **188**:361–375.

Thomas Eickbush, Associate Editor

Accepted October 1, 2002