Hoppel, a P-like element without introns: a P-element ancestral structure or a retrotranscription derivative?

Daphné Reiss, Hadi Quesneville, Danielle Nouaud, Olivier Andrieu, Dominique Anxolabéhère

▶ **To cite this version:**

Daphné Reiss, Hadi Quesneville, Danielle Nouaud, Olivier Andrieu, Dominique Anxolabéhère. Hoppel, a P-like element without introns: a P-element ancestral structure or a retrotranscription derivative?. Molecular Biology and Evolution, 2003, 20 (6), pp.869-879. 10.1093/molbev/msg090 . hal-02676873

HAL Id: hal-02676873
https://hal.inrae.fr/hal-02676873

Submitted on 31 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Hoppel, a *P*-like Element Without Introns: a *P*-Element Ancestral Structure or a Retrotranscription Derivative?

Daphné Reiss, Hadi Quesneville, Danielle Nouaud, Olivier Andrieu, and Dominique Anxolabéhère

Institut Jacques Monod, UMR 7592 Dynamique du Génome et Evolution CNRS, Universités P. et M. Curie, D. Diderot, Paris, France

An in silico search for *P*-transposable-element–related sequences in the *Drosophila melanogaster* genome allowed us to detect sequences that are similar to *P*-element transposases. These sequences are located in the central region of 3.4-kb *Hoppel* elements, a class II transposon. Polymerase chain reaction (PCR) analysis of the insertional polymorphism revealed that these elements are mobile. The 3.4-kb elements are the longest copies of this family ever found. They contain an open reading frame that is long enough to encode a transposase, suggesting that the 3.4-kb elements are the full-length copies of the *Hoppel* family. Multiple alignments of several *P*-element transposases from different species and the *Hoppel*-element–encoded peptide showed that all of the *P*-element introns and the 5′ region of the transposase are absent from the *Hoppel* sequence. Sequence analysis combined with reverse transcriptase PCR analysis showed that the 3.4-kb *Hoppel* elements are intronless. *P* and *Hoppel* not only share similar amino acid sequences but also have terminal inverted repeats of the same length (31 bp), and their excision footprints present a similar structure, which suggests that their transposases are functionally very similar. Thus, we propose that the *Hoppel* element family be included in the *P*-element superfamily. Two evolutionary scenarios are discussed considering the presence /absence of introns within the *P*-element superfamily.

Introduction

The systematic sequencing of genomes has dramatically changed our view of their structure. Predictions about the ratios of coding and noncoding sequences were not always confirmed. One of the most unexpected findings was the very large number of repeated sequences or their derivatives. The importance of understanding the evolution of transposable elements and their impact on the evolution of the host genome is now widely recognized (Kidwell and Lisch 2001). The *P*-transposable element is a good model from which to study this subject, thanks to the large volume of data concerning its host species distribution and its mechanisms of transposition and regulation (Pinsker et al. 2001; Rio 2002).

The *P* element is a class II transposable element flanked by 31-bp terminal inverted repeats (TIRs) (O'Hare and Rubin 1983). It transposes via a cut-and-paste mechanism (Kaufman and Rio 1992). The canonical full-length *P* element is 2,907 bp in length and contains four exons (O'Hare and Rubin 1983). In germ line cells, all four exons are required and encode an 87-kDa transposase (Karess and Rubin 1984; Rio, Laski, and Rubin 1986). In somatic cells, the third intron is retained and produces a 66-kDa truncated transposase (Rio, Laski, and Rubin 1986), which acts as a repressor of transposition (Robertson and Engels 1989; Misra and Rio 1990).

The *P* element was first isolated in *Drosophila melanogaster* (Bingham, Kidwell, and Rubin 1982), but further investigations led to the discovery of *P* homologs in numerous Drosophila species (for review, see Pinsker et al. 2001) and even in other genera like *Scaptomyza* (Simonelig and Anxolabéhère 1991). Sequences homologous to the *P* element have also been detected in other Diptera, like *Musca domestica* (Lee, Clark, and Kidwell 1999) and *Lucilia cuprina* (Perkins and Howells 1992), and have recently been detected in humans (Hagemann and Pinsker 2001). The study of *P*-element distribution reveals several discontinuities suggesting the occurrence of horizontal transfers or of losses of the element (Pinsker et al. 2001). These discontinuities are well illustrated in the *melanogaster* subgroup. No *P* sequences have been detected in the *melanogaster* species subgroup, except in *D. melanogaster*, which acquired the *P* element by horizontal transfer from *D. willistoni* approximately 50 years ago (Daniels et al. 1990). The *D. melanogaster* laboratory strains collected before this time are devoid of *P* elements, as are closely related species (Kidwell 1983; Anxolabéhère, Kidwell, and Periquet 1988). However, the *melanogaster* subgroup belongs to the *Sophophora* subgenus, in which *P* sequences are widely distributed, suggesting that *P* sequences were present in the common ancestor of these species. Thus, the absence of *P* homologs in the *melanogaster* subgroup could result either from their strong divergence beyond recognition by the methods available until recently, or from their loss from these genomes (Clark, Kim, and Kidwell 1998). To distinguish between these two hypotheses, we carried out an in silico search for *P*-homologous sequences in the *D. melanogaster* genomic sequence, which is devoid of any recently acquired *P* elements. Our search led to the detection of a sequence with a significant level of sequence similarity to *P*-element proteins. Moreover, this sequence shared several structural characteristics with the *P* element (e.g., TIR size, footprint arrangement), providing evidence for a functional relationship with the *P*-element family. In fact, this sequence belongs to the *Hoppel*-element family, which is known to be composed exclusively of short copies (Kurenova et al. 1990). These short copies are deleted copies of the large element that we described here. Structural data from the coding region of this full-length *Hoppel* show that it is devoid of the canonical *P*-element

Key words: *Hoppel* element, *P* element, transposable element, introns, *Drosophila melanogaster*.

E-mail: reiss@ijm.jussieu.fr.

Mol. Biol. Evol. 20(6):869–879. 2003
DOI: 10.1093/molbev/msg090
© 2003 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

Table 1
Primers Used for Polymerase Chain Reactions

Primers	Positions	Composition
Flanking *Hoppel-1* and *Hoppel-2*[a]		
H1F5′	−332	5′-GAACGTTGCCCCTTGCGCTGCTTG-3′
H1F3′	+455	5′-CGAGGACGGCCGCAATCTTGAAGG-3′
H2F5′	−130	5′-GCCCTAGGAAACCGGGACGGAGGC-3′
H2F3′	+286	5′-TGACGAGGAGTCTGAAGAGGAGCCGG-3′
Internal to *Hoppel-1* and *Hoppel-2*[b]		
HIf1	892	5′-AACAGCATAAAAAAGTTACAGCCAGG-3′
HIf2	1236	5′-TGGGTTCAAAGTAAGGGGTG-3′
HIf3	1430	5′-ACTAGTTGACTTTGATGTCA-3′
HIf4	1966	5′-GCGTTCGAGCAAAAGGTGGTAACTGC-3′
HIr1	425	5′-TTATTAATTTTTCACACGCCGCAAGC-3′
HIr2	1370	5′-ACAGACTTGAAAAGATGCGG- 3′
HIr3	1708	5′-GGGGTTTTTGTCAGTTAATACTCTGC-3′
HIr4	2397	5′-GCCTTTAATTTCCCGAAGTCCGATTC-3′
Other		
DM actin5′		5′-GGTCGGCATGGGCCAGGAGGACT-3′
DM actin3′		5′-ACAGCTTCTCCTTGATGTCACGGAC-3′

[a] Positions are relative to the insertion site.
[b] Positions are *Hoppel-1* and *Hoppel-2* coordinates.

introns. To explain this feature, two alternative evolutionary scenarios will be discussed: (1) loss of introns as a result of the retrotranscription of a *P*-mRNA and (2) acquisition of introns.

Materials and Methods

Fly Stocks: *Drosophila melanogaster* Strains

Strains collected since 1993: Agadir, Kenya, Tanna, Valencia, Costa Rica, Cuba, Florida, Kilimanjaro, Guadeloupe. Old laboratory strains: *ISO1*, w^{1118}, Canton, Harwich, Gruta, Tautavel 67, ICA, Guillin, Hikon. Species obtained from the Laboratoire Population Génétique et Evolution (Gif-sur-Yvette, France): *D. simulans*, *D. yakuba*, *D. teissieri*, *D. santomea*, *D. mauritiana*, *D. sechellia* (*melanogaster* subgroup), *D. ficusphila* (*ficusphila* subgroup), *D. mimetica* (*mimetica* subgroup), *D. ananassae* (*ananassae* subgroup), *D. elegans* (*elegans* subgroup), *D. tsacasi* (*montium* subgroup), *D. guanche* (*obscura* group), *Scaptomyza pallida* (*scaptomyza* genus).

DNA Amplification, Cloning, and Sequencing

Total genomic DNA was extracted from 50 flies from each strain, as described by Junakovic, Caneva, and Balario (1984). For single fly extractions, the same protocol was followed. The sequences of the oligonucleotide primers used are given in table 1. Amplification was performed with 1.25 units of AmpliTaq DNA polymerase (Perkin) in a 25 μl volume, with 50 ng of genomic DNA and buffer adjusted to 1.5 mM of $MgCl_2$. When screening species, the concentration of $MgCl_2$ was adjusted to 2.5 mM. The reaction conditions were: 92°C for 5 min and 30 cycles of 92°C for 1 min, hybridization temperature (depending on the primer) for 30 s and 72°C for 1 min 30 s, followed by 10 min at 72°C. For the single fly amplifications with three primers (one forward and two reverse), the concentration of the forward primer was double that of the reverse primers. Polymerase chain reaction (PCR) products were separated in a 1% agarose gel or in a 1.5% agarose gel for single fly PCR products. Long PCR was performed with the Expand Long PCR system (Boehringer Mannheim), according to the supplier's recommendations.

Polymerase chain reaction products were cloned into the pCR2 vector by use of the TOPO TA-Cloning kit (Invitrogen). Plasmid DNA was prepared for sequencing by use of the QIAprep kit (Qiagen). Automatic sequencing was performed using the ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction (Applied Biosystems).

RNA Isolation and Reverse Transcriptase Amplification

Total RNA was isolated from the ovaries of 25 *ISO1* flies by use of the RNeasy Mini Kit (Qiagen), which includes a DNase step. cDNA was obtained with the Omniscript RT (reverse transcriptase) Kit (Qiagen), by use of the oligo-dT primer and DNA amplification followed at the conditions described above.

Sequence Analysis

The GenBank accession numbers of all the sequences used in this study are shown in table 2. *D. melanogaster* genomic sequences were analyzed at the Berkeley Drosophila Genome Project (BDGP) Web site (http://www.fruitfly.org). Blast searches were performed using the WU-Blast package (http://www.fruitfly.org/blast/; W. Gish 1996, 2002; http://www.blast.wustl.edu) with the default parameter values. Putative promoters were identified by use of the Neural Network Promoter Prediction tool (http://www.fruitfly.org/seq_tools/promoter.html). Polyadenylation signals were identified by use of the Nucleotide Sequence Analysis tool in the GCG (1990) program (http://genomic.sanger.ac.uk/gf.gf.html).

We used hidden Markov models (HMM) to analyze the gene-coding structure. The algorithms used are described in Durbin et al. (1998). Hidden Markov models

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

Table 2
Sequences Used in This Study and Their Accession Numbers

Name of *P* Sequences	Species	Code	GenBank Accession Number
P-tsa	*Drosophila tsacasi*	Dtsa	AF016036
K-boc-P	*Drosophila bocqueti*	Kboc	AY116624
Pπ25.1	*Drosophila melanogaster*	Dmel	X06779
O-type	*Drosophila bifasciata*	DbifO	X71634
M-type	*Drosophila bifasciata*	DbifM	X61795
PS18	*Scaptomyza pallida*	Spal18	M63342
PS2	*Scaptomyza pallida*	Spal2	M63341
P-luci	*Lucilia cuprina*	Luci	M89990
P-musca	*Musca domestica*	Musca	AF183396
P-madeir	*Drosophila madeirensis*	Dmad	X79804
P-gua	*Drosophila guanche*	Dgua	X61720
Phsa	*Homo sapiens*	Homo	AK026973
D. helvet	*D. helvetica*	Dhel	AF313771
Hoppel-1	*Drosophila melanogaster*	Hoppel-1	AY138841
Hoppel-2	*Drosophila melanogaster*	Hoppel-2	AF540061
Δ5′-Hoppel	*Drosophila melanogaster*	Δ5′-Hoppel	AF533772

can combine several Markov chains (called state of the HMM). Each Markov chain has its own set of parameters, reflecting the base composition for that state. Given a path in the HMM, which is a sequence of states, the joint probability of the sequence S and the path π is calculated as follows:

$$P(S,\pi) = a_{0,\pi_1} * \prod P_{MC}(S_i \mid \theta_{\pi i}) a_{\pi_i,\pi_{i+1}}.$$

$P_{MC}(S_i \mid \theta)$ is the probability of nucleotide S_i using the Markov chain parameters θ. $a_{\pi_i,\pi_{i+1}}$ reflects the probability of switching from state π_i to state π_{i+1} (transition probability). Thus, the parameters of an HMM with n states are n^2 transition probabilities and n sets of Markov chain parameters (also called *emission* probabilities).

We used an HMM with five states: coding exons in phases 1, 2, and 3, intron, and ''terminal'' sequence. A transition departing from a coding exon state can go to the next phase coding exon state (most probable), or the intronic state; it can also stay in the same state or go to the preceding coding state, to account for frameshifts. The terminal state represents the noncoding, non-intronic parts at the beginning and end of the sequences. A path in the HMM starts and ends in this state. For each state, the emission probabilities are simple second-order Markov chains.

The models were trained using the Baum-Welsch algorithm, an EM (expectation maximization) algorithm (Durbin et al. 1998) with labeled sequences of *P* elements. The starting point for the iterative Baum-Welsch algorithm was an HMM with random emission probabilities but fixed transition probabilities. Training estimates these emission probabilities and refines the transition probabilities, but the overall structure of the HMM is maintained. The training was repeated several times with different random starting points, and the estimate with the best likelihood was kept.

Given a nucleotide sequence S, we calculated the most probable state sequence using the Viterbi algorithm (Durbin et al. 1998). We also calculated the posterior probability of states, that is the probability that nucleotide i lies in a state k of the HMM: $P(\pi_i = k \mid S)$. By calculating this for each nucleotide of S, we were able to plot the probability of state k along the sequence.

Multiple alignments were obtained with the PILEUP program of the GCG package (Madison, Wis.) using the default options.

Phylogenetic Analysis

Aligned sequences were analyzed by the Neighbor-Joining method in the PHYLO_WIN program (Galtier, Gouy, and Gautier 1996). PAM distance and global gap removal options were chosen for Neighbor-Joining analysis. Five hundred bootstrap replicates were performed.

Results

Identification of Sequences Homologous to the *P* Element in the *ISO1 D. melanogaster* Line

We used TBlastN to screen the whole Berkeley Drosophila Genome Group (BDGP) Drosophila sequence dataset for sequences homologous to the *P* element in the *D. melanogaster* genome (Altschul et al. 1990, 1997) using the following *P*-element proteins as a query: Spal2, Luci, DbifM, DbifO, Dmel, Dmad, Dgua, Dtsa. All of the *P*-element proteins identified the same nucleic sequence in the *D. melanogaster* genome (accession number AC005453, of unknown chromosomal localization) with amino acid similarity values ranging from 45% to 47% and significant P values (1.8×10^{-29} to 1.3×10^{-18}). The result of the TBlastN search showed that the length of the nucleotide-matching region varies from 1,170 bp to 1,550 bp depending on the P protein used for comparison. The region of the AC005453 fragment associated with the *P*-element proteins will be called the ''*P*-like sequence'' hereinafter. The match with P transposases expands from the middle of exon 1 to the beginning of exon 3.

The in silico analysis of the regions upstream and downstream of the *P*-like sequence revealed the presence of TIRs that are characteristic of class II transposable elements. These TIRs were 31 bp long and were flanked by 7-bp target site duplications (TSDs). The size of the hypothetical transposon is 3,408 bp. Another copy of the same length and sharing 99.5% identity was detected in the chromosomal site 38A (accession number AE003664).

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

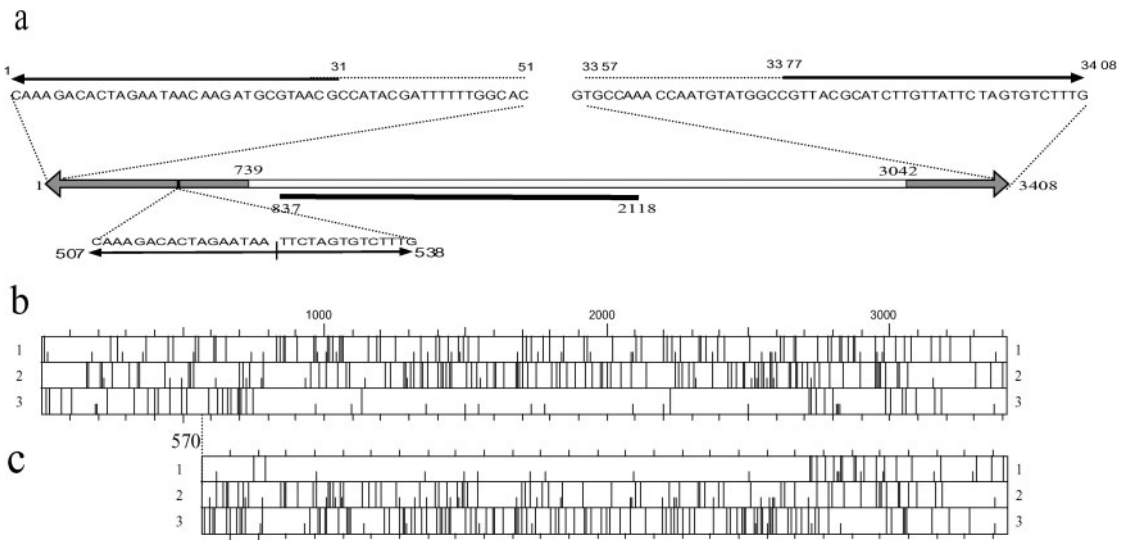

FIG. 1.—Schematic structure and predicted ORFs of *Hoppel-1*, *Hoppel-2*, and Δ5′-*Hoppel* elements. *a*, Structure of the 3.4-kb *Hoppel* elements. Gray areas indicate the regions shared with the internal deleted *Hoppel* element (1100 bp). The single black arrows indicate the perfect 31-bp TIRs, and the dotted line extends to 51 bp, giving imperfect TIRs. The double black arrow corresponds to an internal footprint formed by the first 17 bp of the 5′ TIR and the first 14 bp of the 3′ TIR. The black line indicates the *P*-like region identified by the TBlastN search with the P proteins. *b*, Three-frame ORF map deduced from *Hoppel-1* as printed by DNA Strider. Each horizontal open bar corresponds to a reading frame. The long vertical bars represent the stop codons, and the short ones represent the methionine residues. Only frames in orientation with the P protein as deduced by the alignment are shown. *c*, 5′-*Hoppel* ORFs in scale with those of *Hoppel-1*.

A BlastN search was carried out to detect other genomic copies of this transposon. At least 100 copies carrying deletions in their terminal or central region were found throughout the genome. A particular class of these dispersed and deleted sequences has been previously described and is called the *1360* or *Hoppel*-element family (Kholodilov et al. 1988; Kurenova et al. 1990). The *Hoppel* element is about 1,100 bp long and is entirely encompassed within the 3.4-kb element under investigation (fig. 1*a*). The only elements of the *Hoppel* family that have been described to date are internally deleted elements that do not contain large enough open reading frames (ORFs) to encode a transposase and are devoid of the *P*-like sequence. We propose that the 3.4-kb elements be included in the *Hoppel* family. It is noteworthy that the 3.4-kb sequences can also be found in Repbase Update (Jurka 2000) with the *protoP* identifier.

To investigate the molecular structure of these 3.4-kb *Hoppel* elements, without introducing into our study potential assembly mistakes found in the repeated region of the BDGP sequences, a long PCR amplification of the two complete copies was done in the *ISO1* strain with primers corresponding to the flanking regions of the two insertion sites (table 1). Two PCR products of the expected sizes were cloned and their sequences were compared to those of the corresponding BDGP sequences. One copy, hereinafter called *Hoppel-1*, was 99.5% identical to the corresponding AC005453 sequence, and the other, to be called *Hoppel-2*, was 98.3% identical with the AE003664 sequence.

Sequence Analysis of the 3.4-kb *Hoppel* Transposable Elements: *Hoppel-1* and *Hoppel-2*

Sequence analysis of *Hoppel-1* and *Hoppel-2* elements has shown that TIRs can be extended up to 51 bp long. These 51-bp TIRs are composed of the 31-bp perfect TIRs cited previously and 20 additional base pairs that are not perfectly repeated: only 15 bp of the 20 are identical (fig. 1*a*). In addition, a 31-bp footprint of *Hoppel* element (see below) is present within the element at positions 507–538. It consists of the first 17 bp of the 5′ *Hoppel* TIR and the first 14 bp of the 3′ TIR.

To determine the capability of the 3.4-kb *Hoppel* elements to encode a transposase, ORFs were sought. The result for *Hoppel-1* is shown in figure 1*b*. Three consecutive large ORFs, separated by a single stop codon, were found in the central region of the transposon (frame +3). These ORFs correspond to the peptide that was similar to the P proteins found by TBlastN. *Hoppel-2* presents the same ORFs, except for a frame shift in the first ORF at position 1019 (data not shown).

To locate the position of *P*-element introns in the *Hoppel* sequence, the following proteins were aligned with the region of the *Hoppel-1* element that contains the three ORFs including the two stop codons: DbifM, Dhel, Spal2, Dmel, and DbifO. The alignment was performed by the PILEUP program, and the results are shown in figure 2. The similarity between *P* and *Hoppel* extends from the beginning of the first *Hoppel* ORF, which matches the middle of the P-transposases exon 1, to the end of the third *Hoppel* ORF, which matches the end of the third exon of the P transposases (fig. 2). The *Hoppel* matching region cannot be expanded upstream in any of the reading frames. Remarkably, the amino acid sequence of the *Hoppel* element does not present any discontinuity of the similarity with the splicing regions of exons 1–2 and 2–3 of the P transposases. Moreover, the two stop codons are not located in these regions. Thus, the *P*-element introns are not present in the *Hoppel* sequence.

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

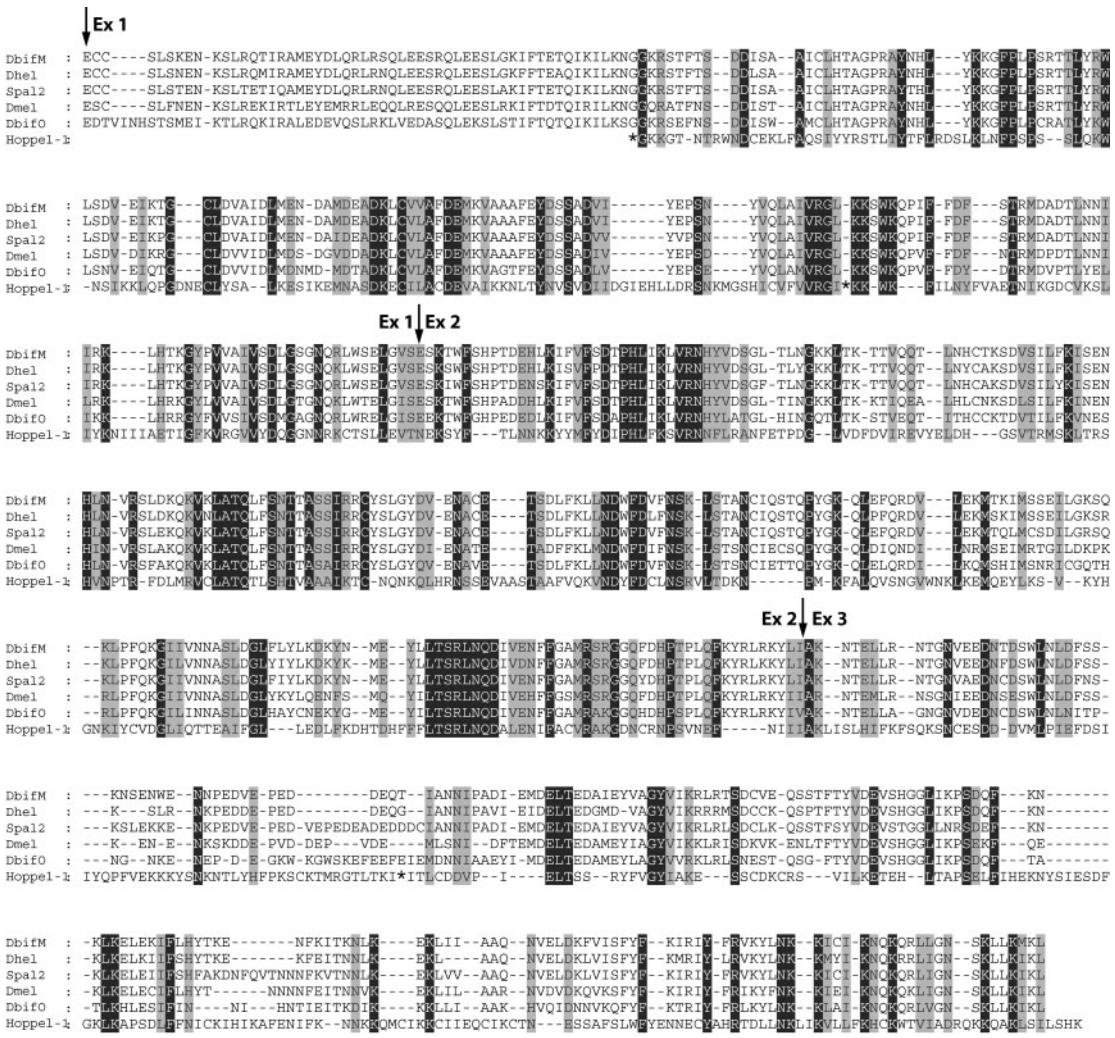

FIG. 2.—Pileup multiple alignment of P proteins with the peptide encoded by the three ORFs of *Hoppel-1*. The sequences are named by their code names (see table 2). Black boxes indicate 100% amino acid identity, and gray boxes indicate 100% amino acid similarity. Gaps are shown by dashes. Arrows show the positions of *P*-element introns. The region encoded by exon 0 of the *P* elements is not shown. An asterisk marks the stop codons in the Hoppel-1 sequence.

The presence of two stop codons in *Hoppel-1* and *Hoppel-2* means either that these elements have recently lost their coding capacity or that the stop codons are located within intronic sequences. To determine which of these possibilities is so, the coding and noncoding regions were sought, by means of an HMM analysis, on the basis of their nucleotide composition. The following *P* sequences were used as a learning set: Spal2, Spal18, DbifM, DbifO, Dmel, Luci, Dhel, and Kboc. The predictions are shown in figure 3*A*. The graphs indicate the posterior probability of each nucleotide to belong to exonic or intronic *P* sequences, or to non-exonic and non-intronic sequences. A large coding region was predicted from nucleotides 617 to 2432, which covers virtually all three ORFs. However, only one frame shift was detected by the Viterbi algorithm; it is located at the beginning of the coding region (nucleotide 780). The upstream region contains numerous stop codons in the three frames, suggesting that this is no longer a coding region. Nevertheless, according to the HMM analysis, the nucleotide composition of this region is characteristic of *P*-coding regions. With regard to the 3′ extremity, the HMM analysis located the end of the coding region at nucleotide 2432, even though the stop codon is located at nucleotide 2715. This is probably due to the low level of nucleotide diversity in the region between nucleotides 2432 and 2715, which is similar to microsatellite sequences and is consequently recognized as being noncoding by the HMM algorithm. This is also the case for *P*-element sequences: at the end of their last exon, the nucleotide composition is similar to that of microsatellites and was not detected as a coding sequence by the HMM method (data not shown). The region between nucleotides 1100 and 1350 was ambiguously identified by HMM (exonic versus intronic). The stop codon at position 1134 is included in this region, whereas the stop codon at position 2223 is located in an exonic region, according to HMM.

To clarify the exonic versus intronic status of these two regions, RT-PCR analysis was performed on mRNA from *ISO1* line ovaries (fig. 3). Both primer pairs amplified

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

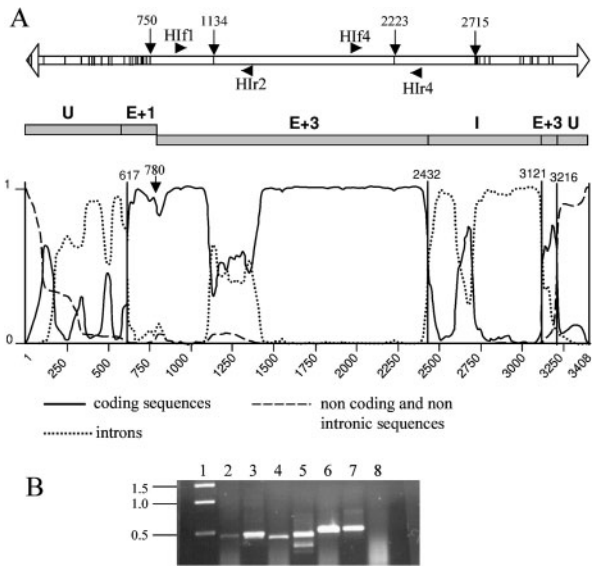

FIG. 3.—Search for introns in the genomic *Hoppel* elements of the *ISO1* line: HMM analysis of the *Hoppel-1* element and RT-PCR analysis. *A*, HMM profile of *Hoppel-1*. Open double arrow indicates the *Hoppel-1* element. Bars correspond to the stop codons of the +3 frame. The black arrowheads indicate the primers used for RT-PCR. The graph is in scale with the open arrow. X-axis: nucleotide positions of *Hoppel-1*. Y-axis: posterior probabilities of the state (see *Materials and Methods*). The gray boxes represent the most probable HMM state path as predicted by the Viterbi algorithm, with a frame shift at position 780. E: exonic sequence; I: intronic sequence; U: noncoding and non-intronic sequence. *B*, RT-PCR and PCR experiments on *ISO1* ovary transcripts. Size marker: lane 1; RT-PCRs: lanes 2, 4, and 6; PCRs: lanes 3, 5, and 7. Lanes 2 and 3: primers HIf1 and HIr2. Lanes 4 and 5: primers HIf4 and HIr4. Lanes 6 and 7: actin gene–specific primers used as RT-PCR positive control. Lane 8: negative control PCR using actin-specific primers and an RNA extract that was not subjected to reverse transcription. The absence of any amplification product in this sample shows that the RNA extracts are DNA-free. All RT-PCRs were performed on the same RNA extract.

fragments of the same size with RNA and DNA templates, suggesting that these regions do not contain introns. The amplified fragments were cloned and sequenced. The nucleotide sequences differed by 1% to 5% from the corresponding regions of *Hoppel-1* and *Hoppel-2*. Thus, *Hoppel* elements are transcribed, but the transcripts analyzed here are not generated by the *Hoppel-1* or *Hoppel-2* copies. Interestingly, the translation deduced from the 892–1370 region of the transcript did not contain the stop codon present in the corresponding region of *Hoppel-1* and *Hoppel-2*. This suggests that the *ISO1* genome contains an intact coding *Hoppel* sequence.

Another TBlastN analysis was carried out with the peptide generated from the three ORFs of *Hoppel-1* (including the two stop codons). One of the output sequences presented the same coding capacity as the query (91% similarity), but without the two stop codons. Thus, it gives rise to an ORF of 1,746 nucleotides (fig. 1*c*). The *Hoppel* element corresponding to this sequence (AC010916.8 location unknown) is truncated in its 5′ extremity up to the nucleotide 570. It will be called Δ5′-*Hoppel*. This sequence was not considered in the first TBlastN analysis because it is not a full-length copy.

Taken together, the sequence analyses for *Hoppel-1*, *Hoppel-2* elements and Δ5′-*Hoppel*, suggest a coding structure for the *Hoppel* element (fig. 4). The putative protein is 582 amino acids long and has an estimated molecular weight of 67.3 kDa. It is 41% similar to the *Scaptomyza pallida P* transposase.

In the *ISO1* line, the *Hoppel* transposase could be encoded by the Δ5′-*Hoppel* element or provided by hypothetical *Hoppel* elements located in an unsequenced genomic region (i.e., a heterochromatic region). Alternatively, *ISO1* could have lost functional *Hoppel* elements by genetic drift, as it is an inbred line.

Mobility of Hoppel-1 *and* Hoppel-2

To check the mobility of *Hoppel-1* and *2*, the insertional site polymorphism was investigated in 18 strains of *D. melanogaster* (listed in *Materials and Methods*). This polymorphism was investigated by PCR using primers based on flanking and internal sequences (table 1). DNA from 50 individuals from each strain was analyzed, corresponding to 100 insertion sites for each element. A first PCR was carried out with a primer pair corresponding to the flanking regions of *Hoppel-1* or *Hoppel*-2 (H1F 5′–H1F 3′ for *Hoppel-1* and H2F 5′–H2F 3′ for *Hoppel-2*). A second PCR was performed using the 5′ flanking primer coupled with the HIr1 reverse primer internal to the element. If there is an insert, no amplification product should be obtained with the first PCR, because the expected band is too large to be amplified in the conditions used. However, a band should be obtained with the second PCR and the internal primer. On the contrary, if there is no insertion, an amplification product should be obtained only with the first PCR, corresponding to the target devoid of the *Hoppel* element. In all tested strains, the specific *Hoppel-1* primers amplified a product only in the second PCR. Thus, a *Hoppel-1* insertion is present in all of the tested strains. Furthermore, no insertional polymorphism is present in any of them. Fourteen of the 18 strains tested contained the *Hoppel-2* insertion. Amplification products were seen for the two types of PCR in 10 cases, a finding that provides evidence for insertional polymorphism (table 3).

To confirm the insertional polymorphism of the *Hoppel-2* element, PCR experiments were performed on single fly DNA from a strain that displayed a polymorphism in the mass analysis (table 3) on a recently collected strain from Cuba (1997). For each DNA sample, three primers were used in the same reaction (H2F5′, H2F3′, and HIr1; see table 1). The simultaneous use of the three primers makes it possible to test whether an individual is homozygous or hemizygous for the *Hoppel-2* element in a single amplification reaction. In total, 70 individuals were analyzed; figure 5*A* shows the profiles of 20 of them. The expected sizes were 555 bp in the presence of *Hoppel-2* and 416 bp in the absence of *Hoppel-2*. Individuals presenting only the 555-bp band are homozygous for *Hoppel-2* (lanes 5 and 20), and individuals presenting only the 416-bp band are homozygous for the absence of *Hoppel-2* in this site (lanes 3, 4, 7, 8, 11, 13–17). Hemizygous individuals present the two expected bands (lanes 1, 2, 6). Three other genotypes were detected. These genotypes presented additional bands of about 450

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

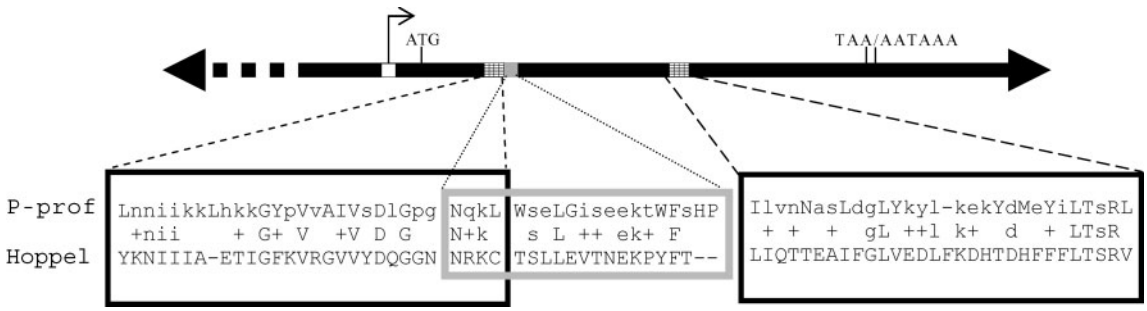

FIG. 4.—Structure of the *Hoppel*-master element. The solid lines define the region present in the Δ5′-*Hoppel*; the dashed lines, its deleted 5′ extremity. The broken dashed arrow indicates the transcription initiation site at position 841, and the open box indicates the promoter at positions 801–851. The LZ and HTH motifs are indicated by the angled bar boxes and the gray box, respectively, with their corresponding alignments between P profiles and Hoppel protein. LZ2 is located at positions 1180–1293, LZ3 is at positions 1819–1931, and HTH, which overlaps the LZ2 motif, is located at positions 1254–1337. The methionine codon (ATG) is located at position 969, the stop codon (TAA) at position 2715 and the poly-adenylation signal (AATAAA) at position 2770.

bp (lanes 9, 10, 12, 18, and 19). The DNA fragments corresponding to these two supplementary bands were cloned and sequenced. Sequence analysis showed that footprints generated by *Hoppel*-element excision accounted for the intermediate sizes in both cases. One was composed of 47 bp formed by the 17 most external nucleotides of the 5′ and 3′ TIRs, separated by a ''filler,'' here composed of 13 bp, formed by overlapping 9-bp inverted repeats (fig. 5*B*a). The other footprint was composed of the first 17 bp of the 5′ TIR and by only the first 14 bp of the 3′ TIR (fig. 5*B*b). These two footprints are flanked by the same 7-bp TSD (TTTTTAC) as *Hoppel-2*. The presence of two footprints provides evidence for two distinct excision events.

We also sought *Hoppel*-footprints in other sites in the whole *D. melanogaster* genome in BDGP. This led to the detection of 11 footprints in distinct locations (fig. 5*B*c). Four of them were 34 bp long (17 bp of 5′ and 3′ TIRs) with perfect 7-bp TSDs. These footprints may have resulted from the excision of either the full-length or the deleted copies, as Coelho et al. (1998) have described an insertional polymorphism for *Hoppel*-deleted copies.

Similarities Between Hoppel *and* P-*Element Features*

Full-length *Hoppel* elements were found in the *D. melanogaster* genome because their amino acid sequences were similar to those of *P* elements. To confirm that the *Hoppel* element is a member of the *P*-element family, a profile HMM was derived from a multiple alignment of *P*-element transposases (HMMER package [Eddy 1998]): Dmel, DbifM, DbifO, Dhel, Spal2, Spal18, Luci, and Kboc. The *P*-element profile was then aligned with the Hoppel protein deduced from Δ5′-*Hoppel* (data not shown). The E value resulting from the alignment was highly significant (2.2×10^{-4}), strongly suggesting that the Hoppel protein is related to the P transposases. The profile HMM of the three leucine zipper motifs (LZ1, LZ2, and LZ3, according to their relative positions on the *P* element) and the helix-turn-helix motif (HTH) present on the P transposases were established with the same set of P amino acid sequences. Five amino acids on each side of these motifs, as described in Nouaud and Anxolabéhère (1997), were included in the profile construction. Three of the four profiles could be identified on the Hoppel protein (fig. 4) with E values of 7.9×10^{-3}, 1.3×10^{-4}, and 4.8×10^{-3}, respectively, for the LZ2, LZ3, and HTH profiles. The positions of the motifs in the Hoppel protein corresponded to the positions in the P proteins. It is noteworthy that the LZ2 motif overlaps with the HTH domain as in the P proteins, and that LZ1 was not found in the Hoppel protein. The latter finding was expected, because the region including this motif does not exist in this protein.

Three other features imply that the *Hoppel* element is closely related to the *P*-element family. First, the *Hoppel* element has perfect 31-bp TIRs like the *P* element (O'Hare and Rubin 1983); however, they do not share significant similarity with any *P*-element TIRs. Second, the *Hoppel* footprints were the same size as those of the *P* element: 17 bp of the 5′ extremity and 17 bp of the 3′ extremity (Takasu-Ishikawa, Yoshihara, and Hotta 1992; Staveley et al. 1995). Third, the TSD is 7 bp long for the *Hoppel* element and 8 bp long for the *P* element. These three characteristics of the DNA sequences are functionally associated with the transposase for type II transposable elements.

The DNA and protein data described above confirm that the *Hoppel* element is closely related to the *P* family, although its structure in a single exon is different from the canonical *P* structure, which is formed by four exons.

To define the relationship between *Hoppel* and other members of the *P*-element family, a phylogenetic analysis

Table 3
Screening of *D. melanogaster* Strains for the Presence of *Hoppel-2* and Its Insertional Polymorphism

Strains	Fragment Amplified with H2F5′ and H2F3′	Fragment Amplified with H2F5′ and HIr1	Insertion
Agadir, Cuba, Guadeloupe, ICA, Kenya, Kilimanjaro, Madrid, Tanna, Valence, w^{1118}	+	+	Polymorph
Harwich, Hikon, ISO1, Tautavel 67	−	+	Yes
CantonS, Florida, Gruta, Guillin	+	−	No

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

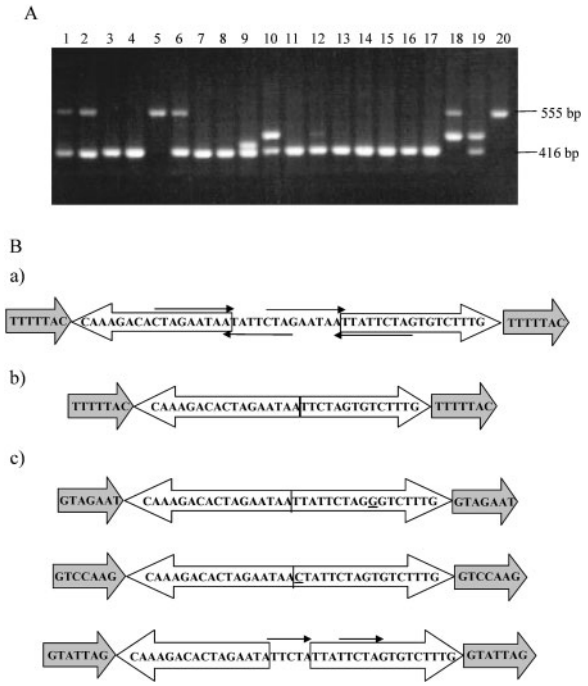

FIG. 5.—*Hoppel-2* element mobility. *A*, Single fly PCR. Each lane corresponds to one individual. The 555-bp band corresponds to the presence of a *Hoppel-2* insert and the 416-bp band corresponds to the absence of this insert. The intermediate bands correspond to two different *Hoppel*-footprints. *B*, *Hoppel* footprints. White arrows indicate the TIR sequences of the footprints; gray arrows indicate the TSDs, the nucleotides between two TIRs constitute the filler. (a) A 47-bp long *Hoppel-2* footprint, the thin arrows correspond to the 9-bp overlapping inverted repeats; (b) a 31-bp-long *Hoppel-2* footprint, (c) three of the 11 *Hoppel* footprints found in the *ISO1* genome with the BlastN search. Underlined nucleotides are not in accordance with the corresponding nucleotides in the inverted repeats. The thin arrows indicate 5-bp direct repeats. The accession numbers of the scaffolds in which the footprints are located are AE003209, AE002853, and AC005437. The third footprint is located at chromosomal site 47C1-C7.

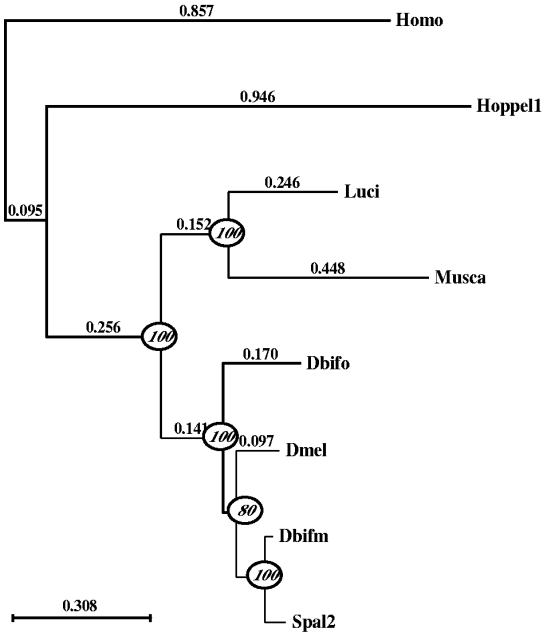

FIG. 6.—Relationships estimated by the Neighbor-Joining method based on the amino acid sequences of the *P* transposases and the peptide encoded by the three ORFs of *Hoppel-1*. The species are noted by code name (see table 2). The bootstrap values are given for each node (they correspond to the percentage of 500 bootstrap replications). Numbers on the branches represent percentage divergence.

was performed using the Neighbor-Joining method. For this study, distantly related P transposases were chosen: Dmel, DbifM, DbifO, Spal2, Luci, Musca, Homo, and the Hoppel protein. The phylogenetic tree resulting from the Neighbor-Joining method is shown in figure 6. *Hoppel* is located out of the cluster formed by the other *P* elements. Thus, given their phylogenetic distance and their structural difference (1 versus 4 exons), we propose that the *Hoppel* element be included in the *P*-element superfamily.

Presence of the Hoppel *Element in Other Species*

The *Hoppel* element has also been detected in other species of the *Drosophila melanogaster* subgroup (Coelho et al. 1998). However, the work that demonstrated *Hoppel* was performed using the internal deleted *Hoppel-1360* element. Here, we sought homologous regions of the full-length *Hoppel* element in 12 Drosophila species (see *Materials and Methods*). Polymerase chain reaction experiments were performed using several sets of primers (HIf2–HIr3; HIf3–HIr3; HIf4–HIr4), chosen according to the coding region of the *Hoppel-1* element under low stringency conditions (data not shown). Amplified fragments were obtained from all species except *D. ficusphila*, *D. tsacasi*, and *D. guanche*. Two of the 430-bp amplified products from *D. lutescens* and *Scaptomyza pallida* were cloned and sequenced. The respective sequences present 82.1% and 86.2% identity with the *D. melanogaster Hoppel-1* element, meaning that the *Hoppel* element is widely represented in the *Drosophilidae*.

Discussion

Mobility of *Hoppel* Elements

Coelho et al. (1998) described an insertional polymorphism of *Hoppel*-deleted copies in *D. melanogaster* strains. We confirm these data by showing that *Hoppel*-element footprints resulting from recent excisions are present in the genome. Until our study, only deleted copies (about 1.1 kb in length) of the *Hoppel* family had been described. We characterized 3.4-kb copies, presenting ORFs that are large enough to encode a transposase. However, the two complete 3.4-kb copies, *Hoppel-1* and *Hoppel-2*, present two stop codons interrupting their coding capability. Introns eliminating these two stop codons were not found by HMM or RT-PCR analysis. Therefore, these copies are not able to encode the functional transposase required for *Hoppel* mobility. Conversely, the $\Delta 5'$*-Hoppel* element has intact coding capacity; thus it could provide the *Hoppel* transposase. However, we cannot exclude the possibility that another full-length copy encoding transposase is present in the unsequenced regions of the *ISO1* genome. A last obvious possibility is that *ISO1* lost the functional *Hoppel* elements as a consequence of genetic drift.

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

The Full-Length *Hoppel* Element Is a Member of the *P*-Element Superfamily

We present structural and functional evidence that the *Hoppel* element is related to the *P*-transposable elements and should be included in the *P* superfamily. Multiple alignments of P proteins and the Hoppel-1 peptide provided by three consecutive ORFs (fig. 2) show that the region of similarity extends from the middle of exon 1 to the end of exon 3 of the P proteins. Pairwise comparisons of P proteins with the Hoppel protein displayed similarity values of about 40%. Moreover, two LZ motifs and a HTH motif, both characteristic of P transposases, are present on the Hoppel protein. In addition, TBlastN searches in Repbase Update (all organisms section merged) using the Hoppel protein as a query, detected significant similarity only with P transposases.

Regarding their working modalities, the *P* and *Hoppel* elements share two main characteristics. First, the perfect TIRs of the two elements are the same length, 31 bp. However, the *Hoppel* TIRs can be expanded to 51-bp imperfect TIRs. This can be explained by an extension of TIRs in the *Hoppel* lineage conferring a transposition efficiency advantage. Indeed, TIRs are important *cis*-acting DNA sequences that are necessary for efficient transposition. The two fixation sites of the *P* transpsosase (a 20-bp region recognized by the transposase) are located 16 bp from the 5′ TIR and 4 bp from the 3′ TIR (Kaufman, Doll, and Rio 1989). These two sites have a consensus sequence of 10 bp that are in inverted orientation with respect to the ends of the *P* element. In the *Hoppel* element, the two closely located inverted structures (TIRs and transposase fixation sites) could have been assembled, forming a 51-bp imperfect inverted sequence.

The second characteristic shared by the two elements is the structure of their footprints, despite the absence of significant similarity between their TIRs. The excision of the *P* element is mediated by 17-nucleotide staggered cleavages at its ends, which are without precedent for all known transposase and restriction endonuclease cleavage sites determined to date (Beall and Rio 1997). These cleavages result in the *P*-element footprint characteristics. Most of the footprints resulting from germ-line excision are formed by 16 ± 1 bp of the 5′ and 16 ± 1 bp of the 3′ ends (Takasu-Ishikawa, Yoshihara, and Hotta 1992; Staveley et al. 1995). Several *Hoppel* footprints detected on the whole genome sequence correspond to similar structures like the perfect 34-bp-long *P* footprints (17 bp of 5′ and 17 bp of 3′ TIRs) (e.g., fig. 5). Moreover, *P*-element footprints often present short sequences separating the two *P*-element extremities called ''fillers.'' The filler usually corresponds to part of the 5′ or the 3′ TIRs. Filler sequences are probably generated by replication-slippage-replication during the breakpoint repair of the donor site (Kurkulos et al. 1994). Several *Hoppel* footprints also present additional nucleotides (fig. 5). Thus, the excision of the *Hoppel* elements leaves footprints with the same structural characteristics as those of *P*. The 17-bp staggering of cleavage sites is specific to the DNA binding site and the endonuclease site of the transposase (Beall and Rio 1997). Thus, the similarity of their footprints suggests that the *Hoppel* and the *P* transposases were derived from a common origin.

Moreover, the TSDs are 8 bp and 7 bp long for the *P* and *Hoppel* elements, respectively. This TSD size is a characteristic of the endonuclease activity of the transposase that causes a staggered cleavage at the target site. This observation alone has no significance concerning the relationship of the two elements, but together with the common features shared by their proteins, it reinforces the hypothesis that these elements share common functional modalities.

The Presence/Absence of Introns in the *P* Superfamily Results from Intron Loss or Gain Events?

Despite having homologous coding sequences, the structures of the *P* and *Hoppel* elements differ in two main ways. First, the coding region corresponding to exon 0 and the first half of exon 1 of the *P* element is totally absent from the Hoppel protein. Thus, if the amino acid multiple alignments are read from the COOH terminal to the NH2 end, there is an abrupt breakpoint of the similarity in the middle of exon 1. Second, the *Hoppel* coding region does not present the introns that interrupt the *P*-coding sequence. An appealing hypothesis explain these two differences with a single event : the *Hoppel* coding sequence arose from the retrotranscription of a *P*-element processed mRNA. Retropseudogenes and retrogenes can be recognized by the following characteristics: the lack of introns found in their functional counterparts, the presence of a poly-A tail at their 3′ end and the presence of flanking TSDs (Vanin 1985). In addition, the interruption of the retrotranscription before the 5′ end of the mRNA can lead to the 5′ truncated cDNA observed in numerous cases. Thus, the coding region of the *Hoppel* element was probably derived from a 5′ truncated *P*-element mRNA. However, there is no evidence for the presence of a poly-A tail in the downstream region of the *Hoppel* coding sequence or of the TSDs in its vicinity. This is expected if the retrotranscription event occurred a very long time ago. Another explanation for the lack of poly-A tail in the retrotranscripts is that the reverse transcription was initiated within a 3′ A-rich region upstream of the poly-A tail (Kleene et al. 1998). Remarkably, the 3′ regions of all *P* elements are A-rich, making this process possible.

Under this hypothesis we propose the following retrotranscription scenario: a *P*-element mRNA is retrotranscribed in the germ line; this retrotranscription is initiated inside a deleted *P* element from the same family as the one that provided the *P*-processed mRNA. This could result from a *P* element homing process (Delattre, Anxolabéhère, and Coen 1995; Delattre, Tatout, and Coen 2000). In fact, the *P*-element transposase, which has affinity for the *P* sequence, may remain attached to the *P*-mRNA and target it to a deleted *P* element inside of which the reverse transcription takes place. Finally, the retrotranscript is interrupted in the middle of exon 1. Thus, the resulting sequence is formed by the partial reverse transcription of the coding regions of a *P* element nested inside the 5′ and 3′ *cis*-acting sequences of the host *P*-deleted element. Although common in vertebrates

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

(Vanin 1985; Wilde 1986), processed pseudogenes are rare in Drosophila for protein-coding genes (Jeffs and Ashburner 1991). However, flies do possess retroelements that generate reverse-transcriptase enzymes and sequences derived from the reverse transcription of RNA (Finnegan 1989). Very few retrogenes have been described in the Drosophila genome (Neufeld, Carthew, and Rubin 1991), and if this hypothesis is true, the *Hoppel* element would be the first example of a retrotranscribed class II transposable element. However, the *Hoppel* element is not the only member of the *P* superfamily presenting these characteristics. A *P* sequence without the canonical *P* introns, exon 0, and TIRs was recently described in the human genome and called *Phsa* (Hagemann and Pinsker 2001). This sequence has two introns located inside the region corresponding to the *P*-element exon 1, but it does not present the two canonical *P* introns separating exons 1–2 and 2–3. According to the loss-of-intron hypothesis, either *Hoppel* and *P* elements form a monophyletic group and two independent retrotranscription events are required in the *Hoppel* and *Phsa* lineage, or *Hoppel* elements form a monophyletic group with *Phsa*, and a single retrotranscription event is required before their divergence, followed by a horizontal transfer of the *Hoppel* element from deuterostomians to protostomians.

An alternative hypothesis is that the absence of introns corresponds to the ancestral structure of *P* elements. In this case, the *P*-element canonical introns were inserted into the *P* lineage after the divergence of *P* and *Hoppel* elements. As described previously, the amino acids flanking the two *P*-element introns are conserved in the Hoppel protein, which implies that in the case of intron gain, splicing does not change the protein encoded by the *P* element compared to the intronless *Hoppel* sequence. Indeed, several examples of such intron insertions have been described in the literature. A first example is the insertion of 16 introns in the triosephosphate isomerase gene from different species (Kwiatowski et al. 1995; Logsdon et al. 1995). Multiple alignments of the proteins encoded by these genes clearly show that the amino acids in the vicinity of the inserted introns are perfectly conserved. Within the *P* superfamily, two intron-insertion events have taken place without the addition or deletion of any amino acids; the first, in the exon 1 region of the *Phsa* sequence as mentioned above, and the second in exon 2 of a *P* element carried by the common ancestor of *L. cuprina* and *M. domestica* (Perkins and Howels 1992; Lee, Clark, and Kidwell 1999). Therefore, intron gain events without any changes in the transposase sequence are observed within the *P*-element lineage. According to this hypothesis, the lack of exon 0 in the *Hoppel* element sequence could result from the strong divergence of its 5′ region from *P* sequences. Indeed, the 5′ coding region of *P* elements is weakly conserved between species. Obviously, another possibility is the gain of the exon 0 in the *P* lineage. Three examples of new exons have been described in the domesticated *P* sequences of the *montium* subgroup: the first is a noncoding exon located upstream of exon 0, and the other two are additional exons 0 inserted either upstream of the canonical exon 0 or between the canonical exons 0 and 1 (Nouaud et al. 1999; Nouaud, Quesneville, and Anxolabéhère 2003). Additional data concerning the presence or absence of canonical *P* introns within the *P* superfamily are required to define the most parsimonious scenario.

The most closely related sequence to the *P* element in the *D. melanogaster* genome is the *Hoppel* element. However, the presence of the *Hoppel* element does not explain the absence of *P*-family sequences in the *melanogaster* subgroup. Indeed, the rate of divergence between them and the presence of *Hoppel* elements in species that do not belong to the *melanogaster* subgroup suggests that their split occurred before the radiation of this taxon. It is noteworthy that, beside the *Hoppel* sequence, no other kind of *P* sequence has been detected in the *ISO1* genome, suggesting that this absence resulted from genetic drift or a deletion process as described in Petrov and Hartl (1998).

Acknowledgments

We thank E. Bonnivard and D. Higuet for supplying numerous natural populations of *D. melanogaster*, and Isabelle Gonçalves for helping us with the phylogenetic analysis. We also thank S. Ronsseray and S. Charlat for helpful comments on the manuscript. This work was supported by the Centre National de la Recherche Scientifique (CNRS) and the Universités 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

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.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Anxolabéhère, D., M. G. Kidwell, and G. Periquet. 1988. Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of *Drosophila melanogaster* by mobile *P* elements. Mol. Biol. Evol. **5**:252–269.

Beall, E. L., and D. C. Rio. 1997. Drosophila *P*-element transposase is a novel site-specific endonuclease. Genes Dev. **11**: 2137–2151.

Bingham, P. M., M. G. Kidwell, and G. M. Rubin. 1982. The molecular basis of P-M hybrid dysgenesis: the role of the *P* element, a *P*-strain–specific transposon family. Cell **29**:995–1004.

Clark, J. B., P. C. Kim, and M. G. Kidwell. 1998. Molecular evolution of *P* transposable elements in the genus Drosophila. III. The *melanogaster* species group. Mol. Biol. Evol. **15**: 746–755.

Coelho, P., J. Queiroz-Machado, D. Hartl, and S. Ce. 1998. Pattern of chromosomal localisation of the *Hoppel* transposable element family in the *Drosophila melanogaster* subgroup. Chromosome Res. **6**:385–395.

Daniels, S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick. 1990. Evidence for horizontal transmission of the *P* transposable element between Drosophila species. Genetics **124**:339–355.

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013

Delattre, M., 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.

Delattre, M., C. Tatout, and D. Coen. 2000. *P*-element transposition in *Drosophila melanogaster*: influence of size and arrangement in pairs. Mol. Gen. Genet. **263**:445–454.

Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. 1998. Biological sequence analysis. Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge.

Eddy, S. R. 1998. Profile hidden Markov models. Bioinformatics **14**:755–763.

Finnegan, D. 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. **5**:103–107.

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

GCG Sequence Analysis Software Package. 1990. Madison, Wis.

Hagemann, S., and W. Pinsker. 2001. Drosophila *P* transposons in the human genome? Mol. Biol. Evol. **18**:1979–1982.

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

Junakovic, N., R. Caneva, and P. Balario. 1984. Genome distribution of *copia*-like elements in laboratory stocks of *Drosophila melanogaster*. Chromosome **90**:378–382.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. Trends Genet. **16**:418–420.

Karess, R. E., and G. M. Rubin. 1984. Analysis of *P* transposable element functions in Drosophila. Cell **38**:135–146.

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.

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.

Kholodilov, N. G., V. N. Bolshakov, V. M. Blinov, V. V. Solovyov, and I. F. Zhimulev. 1988. Intercalary heterochromatin in Drosophila. III. Homology between DNA sequences from the Y chromosome, bases of polytene chromosome limbs, and chromosome 4 of *D. melanogaster*. Chromosoma **97**:247–253.

Kidwell, M. G. 1983. Evolution of hybrid dysgenesis determinants in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **80**:1655–1659.

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

Kleene, K. C., E. Mulligan, D. Steiger, K. Donohue, and M. A. Mastrangelo. 1998. The mouse gene encoding the testis-specific isoform of Poly(A) binding protein (Pabp2) is an expressed retroposon: intimations that gene expression in spermatogenic cells facilitates the creation of new genes. J. Mol. Evol. **47**:275–281.

Kurenova, E. V., B. A. Leibovich, I. A. Bass, D. V. Bebikhov, M. N. Pavlova, and O. N. Danilevskaia. 1990. [*Hoppel*-family of mobile elements of *Drosophila melanogaster*, flanked by short inverted repeats and having preferential localization in the heterochromatin regions of the genome]. Genetika **26**:1701–1712.

Kurkulos, M., J. M. Weinberg, D. Roy, and S. M. Mount. 1994. *P* element-mediated in vivo deletion analysis of *white-apricot*: deletions between direct repeats are strongly favored. Genetics **136**:1001–1011.

Kwiatowski, J., J. Krawczyk, M. Kornacki, K. Bailey, and F. J. Ayala. 1995. Evidence against the exon theory of genes derived from the triose-phosphate isomerase gene. Proc. Natl. Acad. Sci. USA **92**:8503–8506.

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.

Logsdon, J. M., M. G. Tyshenko, C. Dixon, J. D. Jafari, V. K. Walker, and J. D. Palmer. 1995. Seven newly discovered intron positions in the triose-phosphate isomerase gene: evidence for the introns-late theory. Proc. Natl. Acad. Sci. USA **92**:8507–8511.

Misra, S., and D. C. Rio. 1990. Cytotype control of Drosophila *P* element transposition: the 66 kd protein is a repressor of transposase activity. Cell **62**:269–284.

Neufeld, T., R. Carthew, and G. Rubin. 1991. Evolution of gene position: chromosomal arrangement and sequence comparison of the *Drosophila melanogaster* and *Drosophila virilis sina* and *Rh4* genes. Proc. Natl. Acad. Sci. USA **88**:10203–10207.

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

Nouaud, D., H. Quesneville, and D. Anxolabéhère. 2003. Recurrent exon shuffling between distant *P*-element families. Mol. Biol. Evol. **20**:190–199.

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

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.

Petrov, D. A., and D. L. Hartl. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Biol. Evol. **15**:293–302.

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. 2002. *P* Transposable Elements in *Drosophila melanogaster*. Pp 484–518 *in* Mobile DNA II. ASM Press, Washington, D.C.

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.

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.

Staveley, B. E., T. R. Heslip, R. B. Hodgetts, and J. B. Bell. 1995. Protected *P*-element termini suggest a role for inverted-repeat-binding protein in transposase-induced gap repair in *Drosophila melanogaster*. Genetics **139**:1321–1329.

Takasu-Ishikawa, E., M. Yoshihara, and Y. Hotta. 1992. Extra sequences found at *P* element excision sites in *Drosophila melanogaster*. Mol. Gen. Genet. **232**:17–23.

Vanin, E. F. 1985. Processed pseudogenes: characteristics and evolution. Annu. Rev. Genet. **19**:253–272.

Wilde, C. D. 1986. Pseudogenes. CRC Crit. Rev. Biochem. **19**: 323–352.

Pierre Capy, Associate Editor

Accepted January 13, 2003

Downloaded from http://mbe.oxfordjournals.org/ at INRA Institut National de la Recherche Agronomique on June 20, 2013