Comparative and Functional Studies of Drosophila Species Invasion by the *gypsy* Endogenous Retrovirus

Lucine Mejlumian, Alain Pélisson, Alain Bucheton and Christophe Terzian[1]

Institut de Génétique Humaine, 34396 Montpellier Cedex 5, France

Manuscript received May 14, 2001
Accepted for publication October 19, 2001

ABSTRACT

Gypsy is an endogenous retrovirus of *Drosophila melanogaster*. Phylogenetic studies suggest that occasional horizontal transfer events of *gypsy* occur between Drosophila species. *gypsy* possesses infective properties associated with the products of the *envelope* gene that might be at the origin of these interspecies transfers. We report here the existence of DNA sequences putatively encoding full-length Env proteins in the genomes of Drosophila species other than *D. melanogaster*, suggesting that potentially infective *gypsy* copies able to spread between sexually isolated species can occur. The ability of *gypsy* to invade the genome of a new species is conditioned by its capacity to be expressed in the naive genome. The genetic basis for the regulation of *gypsy* activity in *D. melanogaster* is now well known, and it has been assigned to an X-linked gene called *flamenco*. We established an experimental simulation of the invasion of the *D. melanogaster* genome by *gypsy* elements derived from other Drosophila species, which demonstrates that these non-*D. melanogaster gypsy* elements escape the repression exerted by the *D. melanogaster flamenco* gene.

RETROELEMENTS form a large and diverse class of mobile elements that can be found in all eukaryotes. The structural organization of retroelements reflects their common mechanism of replication: They propagate by reverse transcription of RNA intermediates and integrate their genetic information into the genome of a host cell. Their presence in all organisms suggests that they are very ancient. Several lines of evidence, such as their regulation and putative domestication by the host, indicate that the transposable elements and particularly retroelements (REs) can co-evolve with the host genome. Comparison of the phylogenies of retroelements and of their hosts provides a reliable background for the study of their evolutionary history. In particular, the mechanisms responsible for their distribution and their impact on the host genome can be explored.

Gypsy (*DmeGypV*) is an endogenous retrovirus of *Drosophila melanogaster*, showing a genomic structure remarkably similar to the proviral form of vertebrate retroviruses (Figure 1) and exhibiting infectious properties in particular conditions (Kim *et al.* 1994; Bucheton 1995). Sequences similar to *DmeGypV* are largely distributed among Drosophila species. Previous phylogenetic studies suggested that horizontal transfer events of *gypsy* elements between Drosophila species can occur occasionally (Mizrokhi and Mazo 1991; Alberola and de Frutos 1996; Terzian *et al.* 2000). However, the mechanisms by which horizontal transfers occur remain obscure. The cross-species horizontal transfer events are well documented for DNA transposons, such as *P* and *mariner* (Clark *et al.* 1994; Robertson 1997) or REs such as *copia* (Jordan *et al.* 1999). For these noninfectious elements, a vector-mediated transfer mechanism was proposed to explain transfer between individuals. By contrast, horizontal transfers of infectious elements like *gypsy* do not require any vector in principle. It has been shown in experimental conditions that *gypsy* can be horizontally transmitted between *D. melanogaster* individuals (Kim *et al.* 1994; Song *et al.* 1994) and from *D. melanogaster* cultured cells to *D. hydei* cultured cells (Syomin *et al.* 2001). The infectious properties of *gypsy* may result from the expression of the *envelope* (*env*) gene as illustrated in Drosophila cell culture using a Moloney murine leukemia virus-based retroviral vector pseudotyped with the *gypsy* envelope (Teysset *et al.* 1998).

However, the ability of *gypsy* to invade the genome of a new species is conditioned by its capacity to be expressed in this species. The genetic basis for the regulation of *gypsy* activity is now well known, and it has been assigned to an X-linked gene called *flamenco* (Prud'homme *et al.* 1995; Robert *et al.* 2001). The *flamenco* alleles belong to one of two categories: permissive, which allows the expression of *gypsy* in the somatic follicle cells surrounding the female germline, and restrictive, which represses this expression. The multiplication of vertically transmitted *gypsy* proviruses was shown to require the *env*-independent transfer of *gypsy* products from the permissive somatic cells to the female germline (Chalvet *et al.* 1999).

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. AJ308090–AJ308096.

[1] *Corresponding author:* IGH-CNRS, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France. E-mail: terzian@igh.cnrs.fr

We analyzed two regions of *gypsy*, known to be critical for the tissue-specific tropism of retroviruses: (i) the 5′ regulatory region containing both the 5′ long terminal repeat (LTR) and the untranslated leader region (ULR), which drives the expression of *gypsy* and is controlled by *flamenco* (Pelisson *et al.* 1994), and (ii) the ORF3 coding for the envelope protein. We report here the existence of DNA sequences putatively encoding full-length and functional Env proteins in the genome of the closely related species *D. simulans*, *D. erecta*, *D. orena*, *D. teissieri*, *D. yakuba*, and the more distant species *D. virilis* and *D. subobscura*. The presence of potentially functional *gypsy env* genes not only in the genome of *D. melanogaster* but also in the genomes of distantly related species supports the hypothesis that there are potentially infectious *gypsy* copies able to spread between sexually isolated species. An experimental simulation of the invasion of the *D. melanogaster* genome by *gypsy* elements derived from other Drosophila species demonstrates that these *gypsy* elements are likely to be resistant to the control by *flamenco*.

MATERIALS AND METHODS

Fly stocks: *w1118(R)* is the laboratory *white* reference stock described in Lindsley and Zimm (1992); it contains the restrictive *flam*$^{1118(R)}$ allele. The same mutation was introduced by recombination into two different stocks, *OR(P)* and *Rev(R)*, to give, respectively, the *wOR(P)* and *wRev(R)* stocks. (Chalvet *et al.* 1999 and our unpublished results). The *Rev(R)* stock is the one referred to as *RevI* (Desset *et al.* 1999).

All non-*melanogaster* Drosophila species were kindly provided by Françoise Lemeunier (Centre National de la Recherche Scientifique, Gif-sur-Yvette, France). The *OR(721)/FM3 D. melanogaster* strain (Pelisson *et al.* 1994) is known to contain multiple copies of the active *gypsy* element and was used as a positive control in protein truncation tests (PTT; see below).

PCR primers and conditions: The primers used for *env* amplification from the *melanogaster* subgroup species were designed from the *gypsy* sequence (M12927). The upstream primer was positioned at the beginning of ORF3 and permitted to start transcription at the underlined ATG: 5′ GGATCCTAATACGACTCACTATAGGAACAGACCACCATGTTCATACCCTTGGTAG 3′. The upstream primers used for *env* amplification from *D. subobscura* (5′ GGATCCTAATACGACTCACTATAGGAACAGACCACCATGTTTGTACTCACTTTACT 3′) and *D. virilis* (5′ GGATCCTAATACGACTCACTATAGGAACAGACCACCATGTTCGTCCATTTCAAATT 3′) were based on the *DsuGypV* and *DviGypV* published sequences (GenBank accession nos. X72390 and M38438, respectively) and contained the trinucleotide ATG (underlined) upstream in the *env* sequence in order to create a start codon, which is normally generated by RNA splicing (Pelisson *et al.* 1994; Alberola and de Frutos 1996). All upstream primers were modified at their 5′ end with the T7 promoter sequence to generate PCR products suitable for subsequent PTT analysis.

The three downstream primers were positioned into the 3′ LTR of *DmeGypV* (5′ GGCGATAGCGATTTGATTGTAAA 3′) and *DviGypV* (5′ TTATCTTGGGTAAGTTGCGTTGA 3′), and at the 3′ end of the *DsuGypV env* gene (5′ TCAGAATGATGACCGTCC 3′) to amplify complete *env* sequences.

Reaction volumes were 50 μl and contained 50 ng template DNA and 2.5 units Taq DNA polymerase (Promega, Madison, WI). Reaction parameters were as follows: 3 cycles of 93° for 45 sec, 45° for 45 sec, 72° for 3 min; 25 cycles of 93° for 30 sec, 61° (56° for *D. virilis*) for 30 sec, 72° for 2 min, followed by 72° for 10 min.

A primer pair was designed (upper 5′ TCGCATGCCAACTACATT 3′, lower 5′ AGGCTCGTCTTCTCCTTA 3′) according to regions conserved between *DmeGypV*, *DsuGypV*, and *DviGypV* Env sequences to amplify and sequence the *env* gene derived from *gypsy* elements of *D. simulans*.

PTT: A T7-coupled reticulocyte lysate system (Promega) was used for the PTT analysis according to the protocol recommended by the manufacturer with minor modifications. The reactions were carried out in a volume of 20 μl, and half of the recommended volume of [^{35}S]methionine was used. Purified full-length *env* gene PCR products as well as cloned *env* PCR product were transcribed and translated in a TNT T7-coupled reticulocyte lysate system at 30° for 1 hr. The translation products were separated by discontinuous SDS-PAGE through a 15% separating gel with Tris-glycine buffer. The signals were detected by autoradiography.

Cloning of the *gypsy* regulatory elements from various Drosophila species: pDsugyp and pDvigyp were constructed from PCR-amplified *gypsy* LTR-ULR regions derived from *D. virilis* and *D. subobscura*, fused transcriptionally to the *lacZ* reporter gene, and subcloned into the transformation vector pW7. Primers were designed according to the *D. subobscura* and *D. virilis gypsy* published sequences.

The *gypsy* LTR-ULR from *D. subobscura* was isolated by PCR amplification using 50 ng genomic DNA as template and 2.5 units *Pfu* DNA polymerase in the following conditions: 93°, 30 sec; 57°, 45 sec; 72°, 2 min for one cycle; annealing at 61°, 30 sec for 25 cycles; and 72°, 10 min as final extension. The upstream primer was 5′ GGAATTCCAGTTAAGAACTAAGTACATAAGTTATTCCC 3′ and the downstream primer was 5′ CGGGATCCCGCATTGGCTTATGGTTGGCAC 3′.

The *DviGypV* LTR-ULR was obtained by PCR amplification using as template 5 ng of cloned full-length *DviGypV* (supplied by M. Evgen’ev) and 2.5 units *Pfu* DNA polymerase in the following conditions: 93°, 30 sec; 55°, 30 sec; 72°, 1 min 30 sec for 25 cycles. The upstream primer was 5′ GGAATTCCAGTTAACAACTAAGCATAAATATATTGCCC 3′, and the downstream primer was 5′ CGGGATCCCGCTCATTGGCTTATGGTTGGC 3′. The underlined nucleotides correspond to the *Eco*RI- and *Bam*HI-cleavable extensions used in the cloning strategy.

***P*-element-mediated germline transformation and genetic crosses:** All transgenes were recovered as colored-eyed w^+ individuals after *P*-mediated transformation (Spradling 1986) into the *wOR(P)* permissive stock that is devoid of functional *gypsy* element. Transgenic males were backcrossed to *w1118(R)/w1118(R)* and *wRev(R)/wRev(R)* restrictive females, on one hand, and to *wOR(P)/wOR(P)* females, on the other hand, to keep the transgenes in restrictive and permissive backgrounds, respectively. The same kind of crosses were performed during this experiment with the pgyp68.1 transgenic line, which contains the *DmeGypV* LTR-ULR (Figure 1) fused transcriptionally to *lacZ* (named pgyp in this article; Pelisson *et al.* 1994).

Histochemistry and quantitative measurements of β-galactosidase activity: Histochemical staining for β-galactosidase and quantitative measurements of β-galactosidase activity in the ovaries of the pgyp, pDsugyp, and pDvigyp transgenic flies were done according to Pelisson *et al.* (1994).

Sequencing: PCR products of interest were cloned into the pPCR-Script Amp SK(+) vector (Stratagene, La Jolla, CA) and both strands were sequenced using vector-specific T7 and T3 primers.

Nucleotide and amino acid sequence analyses: Multiple alignments of LTR-ULR nucleotide sequences were performed with

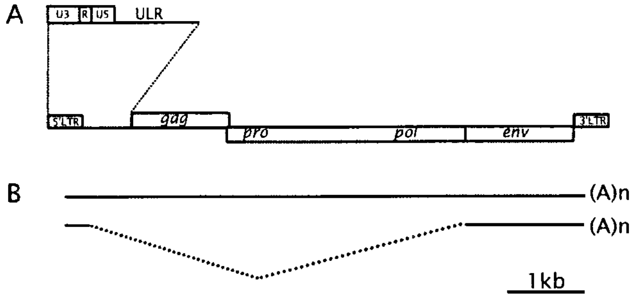

Figure 1.—Structure of *DmeGypV*. (A) Organization of *DmeGypV* with the location of 5′ LTR and 3′ LTR, the ULR, and the genes corresponding to the three open reading frames. (B) Schematic drawing of the unspliced genomic and spliced subgenomic RNAs.

the DIALIGN2 program (Morgenstern 1999). The MEME program was used in addition to visually adjust the initial alignments (Bailey and Elkan 1995). Multiple alignments of amino acid sequences were performed with CLUSTALX (Thompson *et al.* 1997). Alignments were displayed using Bio-Edit (Hall 1999). Jukes-Cantor (JC) nucleotide distances were estimated using Distances from the University of Wisconsin Genetics Computer Group package (GCG, version 10.0). Diverge (GCG v10.0) was used to estimate the numbers of synonymous (K_s) and nonsynonymous (K_a) substitutions per site between two sequences coding for proteins. PlotSimilarity (GCG v10.0) was used to plot the average similarity between two aligned sequences at each position in the alignment, using a window of comparison of 50 nucleotides. The window of comparison was moved along all sequences, one position at a time, and the average similarity over the entire window was plotted at the middle position of the window. The average similarity across the entire alignment is plotted as a dotted line.

RESULTS

Sequences homologous to *gypsy* from *D. melanogaster* (*DmeGypV*) are widely distributed among Drosophila species. Complete *gypsy* elements have been previously cloned and sequenced from *D. subobscura* (*DsuGypV*) and *D. virilis* (*DviGypV*) species (Mizrokhi and Mazo 1991; Alberola and de Frutos 1996). Previous sequence comparisons suggested that the genetic organization of *gypsy* is similar in these three species and that only *D. melanogaster* contains elements possessing a complete functional *env* gene (Alberola and de Frutos 1996). However, as the genomes of these species contain many defective copies of *gypsy* in addition to complete elements, the complete and potentially functional proviruses might have escaped these analyses. For this reason, we searched in several Drosophila species, using a PCR-based approach, for the presence of not-yet-described *env* genes putatively coding full-length proteins.

The coding capacities for Env proteins are conserved in *gypsy* elements from various Drosophila species: The PTT, a mutation-detection method generally used to scan for premature termination mutations (Den Dunnen and Van Ommen 1999), was chosen to analyze the coding capacity of the *gypsy env* gene present in the genomes of different Drosophila species. *gypsy env* PCR amplifications were carried out with genomic DNAs extracted from six species of the *melanogaster* subgroup and the more distantly related *D. virilis* and *D. subobscura* species. In addition to blank reactions, several other controls were performed to guard against cross-contamination of genomic DNA. In this series of controls, primers amplifying the internal spacer regions of the multicopy Drosophila rDNA genes were used as previously described (Terzian *et al.* 2000). Each sample resulted in PCR products of a unique size distinct from that of *D. melanogaster* (data not shown), indicating that no cross-species contamination occurred.

A major 1.8-kb product and a few minor bands were observed in genomic DNA amplifications from the species of the *melanogaster* subgroup except *D. simulans* for which no product was obtained. The expected *D. melanogaster* PCR product size is 1.8 kb, corresponding to the 1.4-kb complete *env* gene plus 0.4 kb of the 3′ LTR. The specificity of the amplified fragments was checked by Southern blot hybridization at high stringency, using a radioactively marked internal fragment of *DmeGypV env* as a probe, revealing only the 1.8-kb fragment out of several PCR products obtained with the species of the *melanogaster* subgroup (data not shown). The 1.8-kb major PCR product from each species was subsequently gel purified and used for PTT analyses.

The 1.4- and 1.8-kb fragments observed with *D. subobscura* and *D. virilis* did not hybridize to the *D. melanogaster* probe in the Southern blot experiments, presumably because of the divergence between *DmeGypV*, *DsuGypV*, and *DviGypV*. In these two cases, we gel purified the 1.4- and 1.8-kb PCR products from *DsuGypV* and *DviGypV*, respectively.

The major protein products obtained in the PTT assay with *D. teissieri*, *D. yakuba*, *D. erecta*, and *D. orena* have molecular weights corresponding to the expected 54-kD Env protein of *D. melanogaster* (Figure 2). In the case of *D. virilis* and *D. subobscura*, we obtained PTT products of ∼54 and 53.5 kD, respectively (Figure 2). These results show that the genomes of these species contain at least one copy of *gypsy* putatively encoding a complete envelope protein. One should note that protein products with smaller sizes are also present in this assay. They might correspond to the previously described truncated Env putative proteins of 52.8 and 42.7 kD for *D. virilis* and *D. subobscura*, respectively (Alberola and de Frutos 1996).

The PCR products of *D. virilis*, *D. subobscura*, *D. erecta*, and *D. teissieri* used in the PTT assays were cloned into the pCRScript cloning vector, and two positive clones for each species were randomly selected and sequenced. Although it was concluded from the published sequence of DsuGypV that *D. subobscura* does not contain *gypsy* elements encoding functional Env products (Alberola and de Frutos 1996), the sequences of the *env* gene

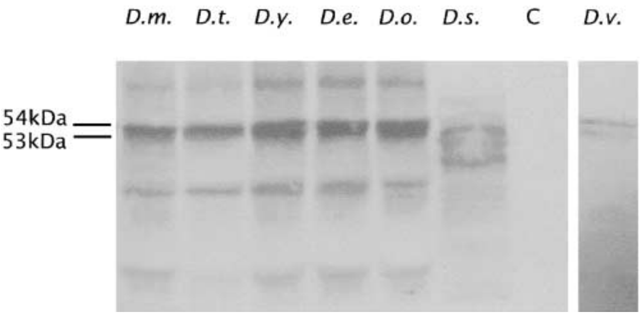

Figure 2.—*gypsy* Env proteins encoded by different Drosophila species as revealed by PTT. *D.m.*, *D. melanogaster*; *D.t.*, *D. teissieri*; *D.y.*, *D. yakuba*; *D.e.*, *D. erecta*; *D.o.*, *D. orena*; *D.s.*, *D. subobscura*; *D.v.*, *D. virilis*. C, control with no DNA added at the beginning of the reaction. Except for *D.v.* where a cloned *env* PCR product was transcribed and translated, proteins were directly *in vitro* transcribed and translated from purified env PCR products.

that we obtained indicate that this species does contain such elements. This confirms the results of the PTT assays. The full-length coding capacity of the *env* sequences we obtained is restored due to the lack of a single T nucleotide present at position 5545 in the X72390 sequence eliminating a stop codon, and the addition of an A nucleotide at position 6639.

Similarly, the PCR products of *D. virilis* encode a complete Env protein. The deletion in our *D. virilis env* sequences of a C nucleotide present at position 5848 in the M38438 sequence eliminates a downstream stop codon present in the previously published sequence (Mizrokhi and Mazo 1991) and restores the N-terminal amino acid sequence highly similar to that of *DmeGypV*.

One of the two *gypsy env* sequences obtained from the *D. teissieri* genome encodes a full-length Env protein, whereas the other contains a deletion of 19 nucleotides that generates a frameshift mutation resulting in a stop codon.

Both *gypsy env* sequences of *D. erecta* show high levels of similarity, and none has a complete protein-coding capacity. Comparison with the *DmeGypV env* sequence reveals multiple point mutations and several significant deletions that create stop codons.

The *env*-based phylogeny is congruent with the *int*-based phylogeny of *gypsy*: We performed a phylogenetic analysis of *gypsy* based on the *env* gene sequences. A previous phylogenetic analysis based on the *integrase* gene showed that the distribution of *gypsy* among the eight *melanogaster* subgroup species does not follow the phylogeny of the host species, defining two main lineages: GypA and GypB (Terzian *et al.* 2000). The GypA lineage is restricted to *D. simulans*, *D. sechellia*, and *D. mauritiana*, whereas the presence of the GypB lineage is limited to *D. melanogaster*, *D. teissieri*, *D. yakuba*, *D. erecta*, and *D. orena*.

Since we could not amplify a *gypsy* homologous *env* gene from the *D. simulans* genome using the PTT assay primers, we carried out a PCR using a couple of primers conserved in *DmeGypV*, *DsuGypV*, and *DviGypV* Env sequences. A single fragment was obtained. It was cloned in pCRScript, and two independent clones were sequenced. The results indicate that *D. simulans* contains *gypsy* elements putatively encoding full-length Env proteins. As the percentages of nucleotide identity between both sequences from the same species were very high (>98%), only one sequence from each species was used to construct a multiple alignment with CLUSTALX (Figure 3). This alignment shows that the putative cellular endopeptidase cleavage site, the two putative N-linked glycosylation sites, the six cysteine residues, and the transmembrane domain described in Pelisson *et al.* (1994) were found at exactly the same positions in all sequences. Moreover, the rates of K_s are much higher than the rates of K_a ($5 \leq K_s/K_a \leq 9$), suggesting that the *env* sequences have evolved under purifying selection. Altogether, these data indicate the existence of a selective pressure for the conservation of the Env function in the *gypsy* elements from various Drosophila species.

We estimated the JC nucleotide distances between the aligned *env* sequences and compared them graphically with the *int* JC distances (Figure 4). The data demonstrate that there is a strong correlation between *env* and *int* distances ($r = 0.87$), indicating that the *env* and *int* phylogenetic trees are congruent. Moreover, the *int*:*env* distance ratios are very close to 1 (Figure 4), suggesting that *int* and *env* evolve at the same rate in the same *gypsy* genome, most likely because of the lack of recombination between *int* and *env*. These results provide significant information supporting the phylogeny of *gypsy* that we previously established (Terzian *et al.* 2000). Since the *int/env*-based phylogenies are different from those of the species, we propose that horizontal transmission of *gypsy* occurred between Drosophila species.

The expression of *gypsy* in *D. subobscura* and *D. virilis* is not repressed by the *flamenco* gene in *D. melanogaster*: We developed an experimental model of the invasion of the genome of *D. melanogaster* by a *gypsy* element from another Drosophila species. A series of experiments was carried out to study the ability of alien *gypsy* regulatory sequences to drive the expression of the bacterial *lacZ* reporter gene in *D. melanogaster*.

It has been previously shown that *gypsy* in *D. melanogaster* is controlled by the *flamenco* host gene (Prud'homme *et al.* 1995) and that the 5′ LTR and ULR of *DmeGypV* contain *cis*-acting sequences responsible for the expression pattern of *gypsy* in the function of *flamenco* permissive and restrictive alleles (Pelisson *et al.* 1994). The LTR-ULR sequences of *gypsy* elements from *D. subobscura* and *D. virilis* were transcriptionally fused to *lacZ* and inserted into the pW7 transformation vector. The resulting constructs, pDsugyp and pDvigyp, were introduced into the genome of the *wOR(P)* by *P*-mediated transformation and the resulting transgenic, as well as the pgyp transgenic, individuals were crossed to *w1118(R)/*

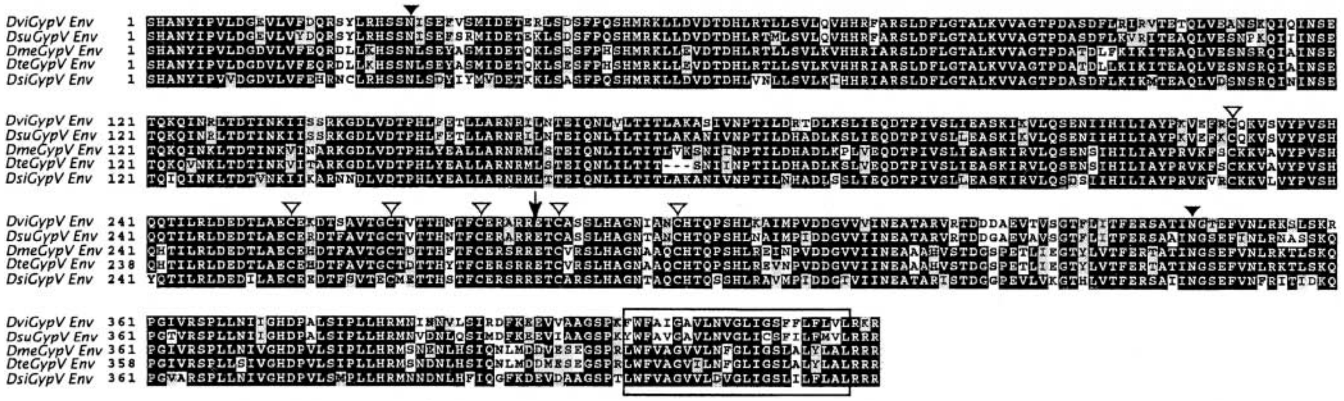
DviGypV Env 1 SHANYIPVLDGEVLVFDQRSYLRHSSNISEFVSMIDETERLSDSFFPQSHMRKLLDVDTDHLRTLLSVLQVHHRFARSLDFLGTALKVVAGTPDASDFLRIRVTETQLVEANSKQIQINSE
DsuGypV Env 1 SHANYIPVLDGEVLVYDQRSYLRHSSNISEFSRMIDETEKLSDSFFPQSHMRKLLDVDTDHLRTMLSVLQVHHRFARSLDFLGTALKVVAGTPDASDFLKVRITEAQLVESNPKQIIINSE
DmeGypV Env 1 SHANYIPVLDGDVLVFEQRDLLKHSSNLSEYASMIDETQKLSESFPHSHMRKLLEVDTDHLRTLLSVLKVHHRIARSLDFLGTALKVVAGTPDATDLFKIKITEAQLVESNSRQIAINSE
DteGypV Env 1 SHANYIPVLDGDVLVFEQRDLLKHSSNLSEYASMIDETQKLSESFPHSHMRKLLEVDTDHLRTLLSVLKVHHRIARSLDFLGTALKVVAGTPDATDLLKIKITEAQLVESNSRQIAINSE
DsiGypV Env 1 SHANYIPVVDGDVLVFEHRNCLRHSSNLSDYIYMVDETKKLSASFPQSHMRKLLDVDTDHLVNLLSVLKIHHRIARSLDFLGTALKVVAGTPDASDFLKIKMTEAQLVDSNSRQININSE
DviGypV Env 121 TQKQINRLTDTINKIISSRKGDLVDTPHLFETLLARNRILNTEIQNLVLTITLAKASIVNPTILDRTDLKSLIEQDTPIVSLIEASKIKVLQSENIIHILIAYPKVEFRCQKVSVYPVSH
DsuGypV Env 121 TQKQINRLTDTINKIISSRKGDLVDTPHLFETLLARNRILNTEIQNLILTITLAKANIVNPTILDHADLKSLIEQDTPIVSLLEASKIKVLQSENIIHILIAYPKVEFKCQKVSVYPVSH
DmeGypV Env 121 TQKQINKLTDTINKVINARKGDLVDTPHLYEALLARNRMLSTEIQNLILTITLVKSNIINPTILDHADLKPLVEQDTPIVSLIEASKIRVLQSENSIHILIAYPRVKFSCKKVAVYPVSH
DteGypV Env 121 TQKQVNKLTDTINKVITARKGDLVDTPHLYEALLARNRMLSTEIQNLILTIT---SNIINPTILDHADLKSLVEQDTPIVSLIEASKIRVLQSENSIHILIAYPRVKFSCKKVAVYPVSH
DsiGypV Env 121 TQIQINKLTDTVNKIIKARNNDLVDTPHLYEALLARNRMLTTEIQNLILTITLAKANIVNPTILNHADLSSLIEQDTPIVSLLEASKIRVLQSDSIIHILIAYPRVKVRCKKVLVYPVSH
DviGypV Env 241 QQTILRLDEDTLAECEKDTSAVTGCTVTTHNTFCERARRETCASSLHAGNIANCHTQPSHLKAIMPVDDGVVVINEATARVRTDDDAEVTVSGTFLITFERSATINGTEFVNLRKSLSKR
DsuGypV Env 241 QQTILRLDEDTLAECERDTFAVTGCTVTTHNTFCERARRETCASSLHAGNTANCHTQPSHLNAIMPIDDGVVIINEATARVRTDDGAEVAVSGTFLITFERSAAINGSEFINLRNASSKQ
DmeGypV Env 241 QHTILRLDEDTLAECEHDTFAVTGCTDTTHFTFCERSRRETCVRSLHAGNAAQCHTQPSHLREINPVDDGVVIINEAAAHVSTDGSPETLIEGTYLVTFERTATINGSEFVNLRKTLSKQ
DteGypV Env 238 QHTILRLDEDTLAECEHDTFAVTGCTDTTHYTFCERSRRETCVRSLHAGNAAQCHTQPSHLREVNPVDDGVVIINEAAAHVSTDGSPETLIEGTYLVTFERTATINGSEFVNLRKTLSKQ
DsiGypV Env 241 YQTILRLDEDILAECEKDTFSVTECMETTHSTFCERSRRETCARSLHAGNTAQCHTQSSHLRAVMPTDDGTVIINEATARISTDGGPEVLVKGTHLVTFERSAIINGSEFVNFRITIDKQ
DviGypV Env 361 PGIVRSPLLNIIGHDPALSIPLLHRMNINNVLSIRDFKKEVVAAGSPKFWFAIGAVLNVGLIGSFFLFLVLRKR
DsuGypV Env 361 PGTVRSPLLNIIGHDPALSIPLLHRMNVDNLQSIMDFKEEVIAAGSPKYWFAVGAVLNVGLICSFILFMVLRRR
DmeGypV Env 361 PGIVRSPLLNIVGHDPVLSIPLLHRMSNENLHSIQNLMDDVESEGSPRLWFVAGVVLNFGLIGSLALYLALRRR
DteGypV Env 358 PGIVRSPLLSIVGHDPVLSIPLLHRMSNDNLHSIQNLMDDMESEGSPRLWFVAGVILNFGLIGSLALYLALRRR
DsiGypV Env 361 PGVARSPLLNIVGHDPVLSMPLLHRMNNDNLHFIQGFKDEVDAAGSPTLWFVAGVVLDVGLIGSLILFLALRRR

FIGURE 3.—Alignment of partial amino acid of Env sequences from *DviGypV*, *DsuGypV*, *DmeGypV*, *DteGypV*, and *DsiGypV*. Identical residues have a black background, and similar residues are shaded. The first residue of the partial *DmeGypV* Env sequence corresponds to position 21 of the 483-amino-acid full-length Env sequence, whereas the last residue corresponds to position 454. Open arrowhead, cysteine residue; solid arrowhead, N-linked putative glycosylation site; arrow, putative endopeptidase cleavage site; box, transmembrane domain.

w1118(R) and *wRev(R)/wRev(R)* restrictive females or to *wOR(P)/wOR(P)* permissive females in order to keep the transgenes in restrictive and permissive backgrounds (see MATERIALS AND METHODS). Several transgenic lines were obtained for each construct and at least three independent transformed lines were analyzed to rule out possible chromosomal position effects of the insertion sites of the transgenes. All transgenic lines bearing the same construct gave identical results.

lacZ expression was detected in follicle cells of vitellogenic stages of oogenesis in the ovaries of *flamenco* permissive pgyp transgenic flies, as previously shown by PELISSON *et al.* (1994). Both squamous (nurse-cell-associated) and columnar (oocyte-associated) follicle cells were stained in permissive ovaries. The highest β-galactosidase activity was observed in the centripetal follicle cells. *lacZ* expression was strongly repressed in *flamenco* restrictive females. This is in agreement with the previous results reported by PELISSON *et al.* (1994).

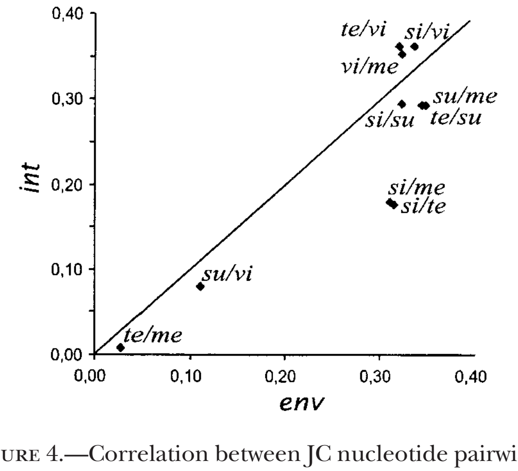

FIGURE 4.—Correlation between JC nucleotide pairwise distances among *env* and *int* sequences. me, *DmeGypV*; te, *DteGypV*; si, *DsiGypV*; su, *DsuGypV*; vi, *DviGypV*. The line marks the 1:1 ratio that would be expected if *int* and *env* genes had the same divergence rate.

Histochemical staining of dissected ovaries of permissive transgenic females homozygous for the pDsugyp and pDvigyp constructs show that reporter gene expression remains restricted to the somatic follicular epithelial cells surrounding egg chambers (Figure 5). However, some significant differences were observed: First, the β-galactosidase activity promoted by the LTR-ULR regions of *DsuGypV* and *DviGypV* is higher than that of the LTR-ULR of *DmeGypV*; second, the pDsugyp and pDvigyp transgenes are expressed at all stages of oogenesis; third, the expression of pDsugyp and pDvigyp is not much repressed by restrictive alleles of *flamenco*, namely, *w1118(R)* and *wRev(R)*, in contrast to pgyp, which is completely repressed in the presence of these alleles. Quantitative measurements of β-galactosidase activity in the ovaries of transgenic flies confirmed the histochemical observations. High levels of β-galactosidase activity were observed in the ovaries of permissive and restrictive transgenic flies for pDsugyp and pDvigyp while pgyp was expressed only in permissive females (Figure 6). A significant repression was observed for pDsugyp in the repressive *wRev(R)/wRev(R)* genetic background. However, the level of β-galactosidase activity in this case is still much higher than that observed with pgyp in the same background. Hence, *DviGypV* and *DsuGypV* are both able to express themselves at a high level in *D. melanogaster* ovaries, whichever *flamenco* alleles are present in the females.

The extent of divergence between the gypsy LTR-ULR sequences varies along the nucleotide sequence: In an attempt to understand the evolutionary forces that shaped the regulatory sequences of *gypsy* leading to the expression profiles that we observed, we sequenced the LTR-ULR from pDsugyp and pDvigyp and compared these sequences with the sequence of the

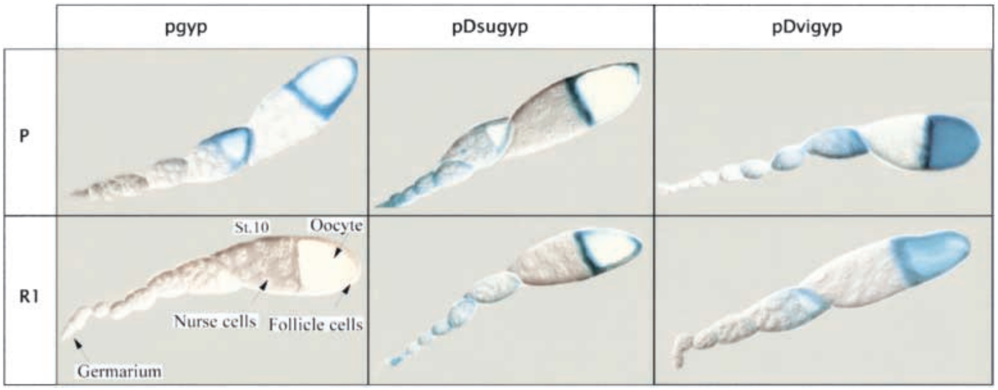

Figure 5.—*lacZ* expression pattern of pgyp (*D. melanogaster*), pDsugyp, and pDvigyp homozygous transgenic flies in permissive [P: *wOR(P)/wOR(P)*] or restrictive [R1: *w1118(R)/w1118(R)*] ovaries. Age of females was 2–3 days.

LTR-ULR of *DmeGypV* previously published (M12927). A multiple sequence alignment was generated using the DIALIGN segment-to-segment approach, which is particularly appropriate for detecting local similarities, and some editing was done manually according to the MEME results (Figure 7).

Many differences are observed between the *gypsy* LTR-ULR sequences. They are not equally distributed along the LTR-ULR regions. The divergences among these sequences result from point mutations and large insertions and deletions (indels): In particular, the *DsuGypV* LTR region contains several insertions compared to the other two sequences.

We tried to characterize the differences in the nucleotide sequences that might be responsible for the differences in the patterns of expression observed previously among pgyp, pDsugyp, and pDvigyp in transgenic flies. Because pDsugyp and pDvigyp transgenic flies share a similar pattern of expression different from that of pgyp, we determined whether they also share similarities in the regions that diverge from *DmeGypV*. This was done by plotting *DviGypV vs. DsuGypV*, *DviGypV vs. DmeGypV*, and *DmeGypV vs. DsuGypV* divergences along the sequence (Figure 8) using PlotSimilarity. Distance values were calculated for a window of 50 nucleotides after removal of all gaps from the alignments. The plots showed clearly that the ULR region allows us to discriminate *DmeGypV* from *DviGypV* and *DsuGypV*. We found one domain where the extent of divergence between *DviGypV* and *DsuGypV* is minimal whereas the extent of divergence between *DviGypV* and *DmeGypV* or *DmeGypV* and *DsuGypV* is maximal (Figure 8). This 30-bp sequence lies within the ULR region, upstream of the first Su(Hw)-binding domain (Figure 7). It is also worthy of note that the similarity scores between the *gypsy* LTR-ULR sequences change suddenly at the U3/R boundary (Figure 8). The highest level of similarity observed in the U3 region concerns *DmeGypV* and *DviGypV*, whereas the highest similarity score observed in the ULR region concerns *DsuGypV* and *DviGypV*. This difference in the degree of similarity between the U3 and ULR regions suggests that *DviGypV* results from recombination events between *DmeGypV* and *DsuGypV*.

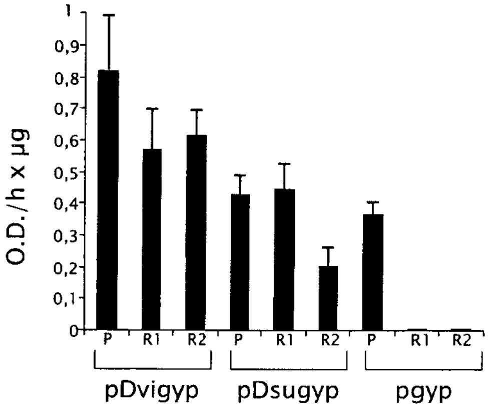

Figure 6.—β-Galactosidase activity in ovaries of transgenic females homozygous for pDvigyp, pDsugyp, and pgyp, respectively. Except for pgyp, each value is the average of β-gal activity from three independent transgenic lines and the standard deviation is given. Concerning pgyp, each value is the average of β-gal activity from three independent protein extracts from the same transgenic line. Data are expressed as 1 OD per hour and per microgram of total protein. P, *wOR(P)/wOR(P)*; R1, *w1118(R)/w1118(R)*; R2, *wRev(R)/wRev(R)*.

DISCUSSION

The Envelope proteins can be a major determinant of horizontal *gypsy* transfers: An increasing amount of experimental data indicates that many classes of transposable elements have been transferred horizontally between species (Flavell 1999). We previously reported that the *int* sequences of *gypsy* elements from *D. subobscura* and *D. virilis*, on one hand, and *D. melanogaster* and *D. teissieri*, on the other hand, are too close together to have diverged from a vertically inherited ancestor (Terzian *et al.* 2000). The phylogenetic analysis that we report here shows that the *env*-based and *int*-based phylogenies are congruent and that these two genes evolved at the same rate. These results also suggest that these two genes have evolved as parts of the same *gypsy* genomes during their horizontal and vertical transfers. Altogether, these data support our hypothesis that hori-

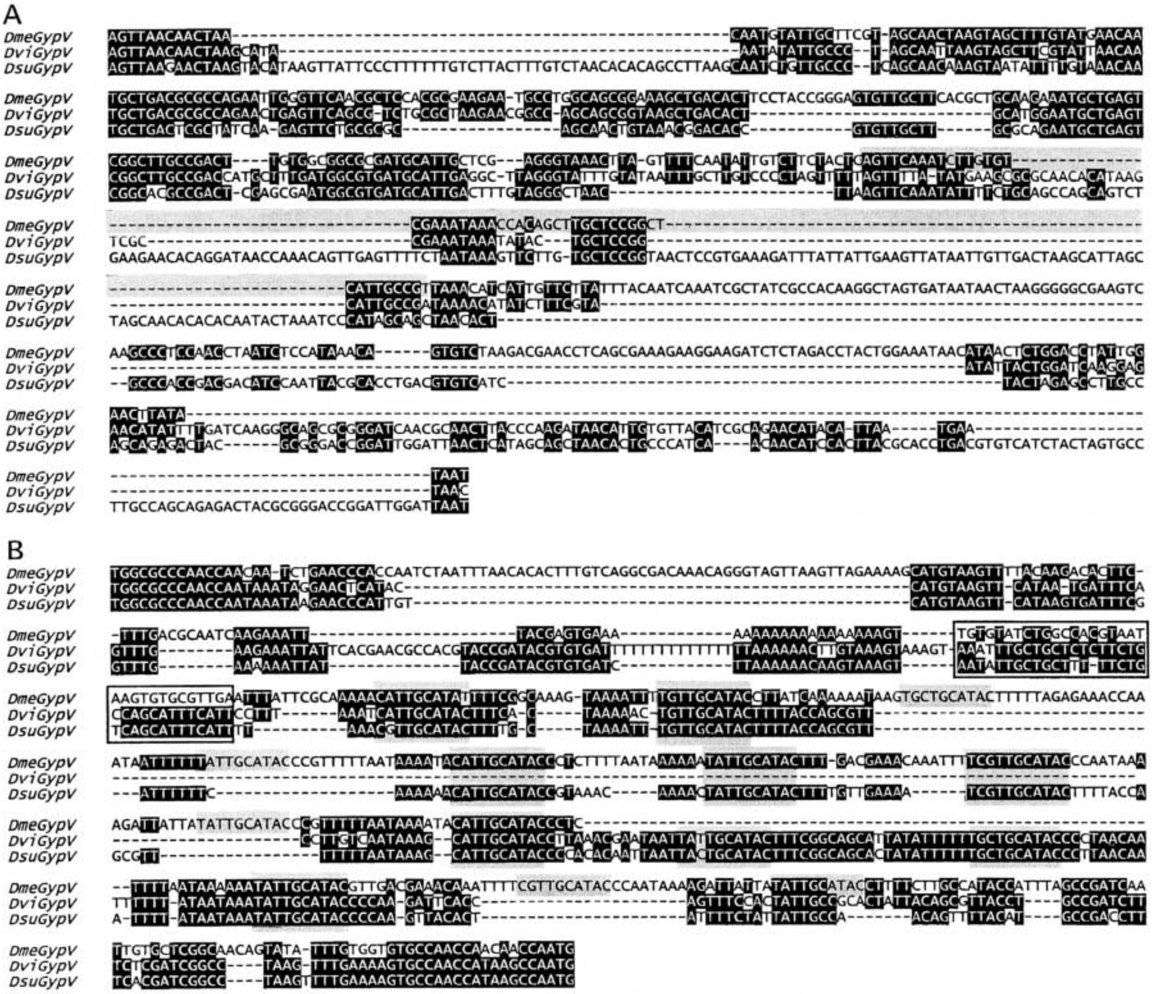

FIGURE 7.—Alignment of the LTR (A) and ULR (B) sequences of *DmeGypV*, *DviGypV*, and *DsuGypV*. (A) The R region, which lies between the U3 and U5 regions, is shaded. (B) The binding sites for Su(Hw) are shaded (PARKHURST *et al.* 1988), and the region that discriminates *DmeGypV* from *DviGypV* and *DsuGypV* (see Figure 8) is boxed.

zontal transfers have played a crucial role in the evolutionary history of these *gypsy* genomes (TERZIAN *et al.* 2000).

The mechanisms by which horizontal transfers of noninfectious elements could occur remain obscure and probably involve diverse vectors (HOUCK *et al.* 1991). Polytropic retroviruses are known to be capable of infecting new host species by horizontal transfer due to their infectious properties provided by the envelope protein. The *gypsy* endogenous retrovirus of *D. melanogaster* encodes an Env protein responsible for its infectious properties (KIM *et al.* 1994; SONG *et al.* 1994; TEYSSET *et al.* 1998). Therefore, the *gypsy* endogenous retrovirus can potentially jump from one individual to another without the need of a vector, increasing the probability of transfer compared to noninfectious retrotransposons. It has been proposed that horizontal transfer is a strategy that transposons use to escape the host-mediated mechanisms repressing transposition (FLAVELL 1999; JORDAN *et al.* 1999). Since the mechanisms repressing *gypsy* activity, such as the repression of *DmeGypV* by the *flamenco* gene, appear to be very efficient, the maintenance of *gypsy* in Drosophila species might be ensured by horizontal transmissions. Our present results show that the ability to encode full-length Env proteins has been conserved in different species, supporting the hypothesis of the existence of selective forces that favor horizontal transfers between individuals from the same species or from different species. Moreover, the excess of synonymous to nonsynonymous substitutions indicates that purifying selection is acting on Env function. Preservation of Env functionality could then represent an advantage for *gypsy* maintenance in Drosophila species. The *gypsy* Env protein thus might be a major component of the mechanisms involved in the evolution of this endogenous retrovirus.

We cannot exclude that the *gypsy env* gene was instead co-opted by the host genome for some cellular functions and that the *env* genes described in this work are not part of full-length active *gypsy* copies. For instance, recent results suggest that the HERV-W Env protein might have a physiological role during pregnancy and placenta

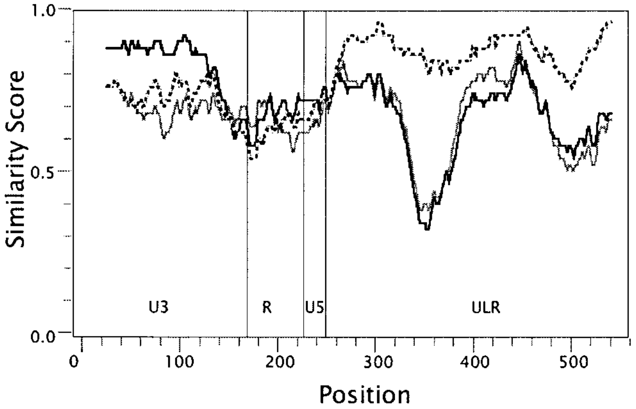

FIGURE 8.—Sequence similarity between *DmeGypV* and *DviGypV* (solid line), *DmeGypV* and *DsuGypV* (shaded line), and *DviGypV* and *DsuGypV* (dotted line) using PlotSimilarity (see MATERIALS AND METHODS). The *x*-axis indicates the nucleotide positions along the alignment (gaps were removed from the alignment). The U3, R, U5, and ULR regions are shown.

formation whereas the HERV-W *gag* and *pol* genes are defective (BLOND *et al.* 2000). However, the fact that the *gypsy int* and *env* genes have the same evolutionary history of a complex pattern of horizontal and vertical transfers and are both under positive selective pressures, in addition to the fact that *int* sequences were issued from complete *gypsy* copies (TERZIAN *et al.* 2000), strongly suggest that both genes are members of full-length *gypsy* elements.

***DsuGypV* and *DviGypV* might invade *flamenco* restrictive *D. melanogaster* populations:** *Gypsy* proviruses of the present-day populations of *D. melanogaster* (*DmeGypV*) are efficiently repressed by the restrictive alleles of the *flamenco* gene. Although the molecular mechanism of this repression is yet unknown, it appears that this regulation acts primarily on the accumulation of the transcripts of *DmeGypV* in the follicle cells of the ovaries (PELISSON *et al.* 1994). We show here that, in contrast to *DmeGypV*, *DsuGypV* and *DviGypV* are not repressed by the restrictive *flamenco* alleles and that the difference in this behavior lies in the LTR-ULR regulatory sequences of *gypsy*. Unless *flamenco* can downregulate them at a post-transcriptional level, *DsuGypV* and *DviGypV* should therefore be able to transpose in *D. melanogaster*. We should note that there is no experimental evidence that *flamenco* controls *DmeGypV* post-transcriptionally. The LTRs of *DmeGypV*, *DsuGypV*, and *DviGypV* all contain *cis*-acting sequences driving the expression of follicle cells. The follicle cells are assumed to be the somatic source of the soma-toward-germline transfer responsible for the *DmeGypV* amplification (CHALVET *et al.* 1999). This suggests that the various *gypsy* elements likely share a common mechanism of transposition.

A pairwise detailed comparison of the three aligned sequences revealed a stretch of 30 nucleotides that is shared by *DviGypV* and *DsuGypV* and is strongly divergent in *DmeGypV*. This sequence was previously described as part of a negative regulatory domain of *DmeGypV* transcription under the conditions of transient expression in Drosophila cultured cells (MAZO *et al.* 1989). We do not know if this is a coincidental correlation or if this divergence of sequence is responsible for the differences in *lacZ* expression driven by pgyp, pDsugyp, and pDvigyp in *D. melanogaster* transgenic restrictive females. Indeed, we cannot exclude the hypothesis that indels present in the LTR-ULR of *DmeGypV* and absent in the two other sequences are the major factors responsible for the differences of expression of the elements in the three species.

These data raise the question of the status of a naive genome and its potential invasion by retroelements. One simple hypothesis is that a naive genome is *a priori* permissive for a retroelement because acquisition of mechanisms of resistance have a fitness cost. Our results suggest that an "alien" *gypsy* element could propagate into a naive host genome. Indeed, *flamenco* seems to be a highly specific host defense system that could not control *gypsy* elements from other Drosophila species. It is probably one of the less "costly" mechanisms for *D. melanogaster*, but its inefficiency in the case of invasion by *DsuGypV* or *DviGypV* would make it necessary for the host to select new mechanisms to avoid the deleterious effect of *gypsy* activity. Several examples (reviewed in BOEKE and STOYE 1997) indicate that the host/retroelement "arms race" results in highly specific interactions. We do not yet know the *flamenco* gene at the molecular level, but it will be very useful to determine the evolutionary history of this gene in Drosophila species in order to understand the relationships between endogenous retroviruses and their hosts.

We thank Dr. Michael Evgen'ev for generously providing the *DviGypV* clone. This work was supported by grants from the Association pour la Recherche sur le Cancer and the Fondation pour la Recherche Médicale (FRM). L.M. was the recipient of fellowships from the French Government and the FRM.

LITERATURE CITED

ALBEROLA, T. M., and R. DE FRUTOS, 1996 Molecular structure of a gypsy element of Drosophila subobscura (gypsyDs) constituting a degenerate form of insect retroviruses. Nucleic Acids Res. **24:** 914–923.

BAILEY, T. L., and C. ELKAN, 1995 The value of prior knowledge in discovering motifs with MEME. Proc. Int. Conf. Intell. Syst. Mol. Biol. **3:** 21–29.

BLOND, J. L., D. LAVILLETTE, V. CHEYNET, O. BOUTON, G. ORIOL *et al.*, 2000 An envelope glycoprotein of the human endogenous retrovirus HERV-W is expressed in the human placenta and fuses cells expressing the type D mammalian retrovirus receptor. J. Virol. **74:** 3321–3329.

BOEKE, J. D., and J. P. STOYE, 1997 Retrotransposons, endogenous retroviruses, and the evolution of retroelements, pp. 343–436 in *Retroviruses*, edited by J. M. COFFIN, S. H. HUGHES and H. E. VARMUS. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

BUCHETON, A., 1995 The relationship between the flamenco gene and gypsy in Drosophila: how to tame a retrovirus. Trends Genet. **11:** 349–353.

Chalvet, F., L. Teysset, C. Terzian, N. Prud'homme, P. Santamaria *et al.*, 1999 Proviral amplification of the Gypsy endogenous retrovirus of Drosophila melanogaster involves env-independent invasion of the female germline. EMBO J. **18:** 2659–2669.

Clark, J. B., W. P. Maddison and M. G. Kidwell, 1994 Phylogenetic analysis supports horizontal transfer of P transposable elements. Mol. Biol. Evol. **11:** 40–50.

Den Dunnen, J. T., and G. J. Van Ommen, 1999 The protein truncation test: a review. Hum. Mutat. **14:** 95–102.

Desset, S., C. Conte, P. Dimitri, V. Calco, B. Dastugue *et al.*, 1999 Mobilization of two retroelements, ZAM and Idefix, in a novel unstable line of Drosophila melanogaster. Mol. Biol. Evol. **16:** 54–66.

Flavell, A. J., 1999 Long terminal repeat retrotransposons jump between species. Proc. Natl. Acad. Sci. USA **96:** 12211–12212.

Hall, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. **41:** 95–98.

Houck, M. A., J. B. Clark, K. R. Peterson and M. G. Kidwell, 1991 Possible horizontal transfer of Drosophila genes by the mite Proctolaelaps regalis. Science **253:** 1125–1128.

Jordan, I. K., L. V. Matyunina and J. F. McDonald, 1999 Evidence for the recent horizontal transfer of long terminal repeat retrotransposon. Proc. Natl. Acad. Sci. USA **96:** 12621–12625.

Kim, A., C. Terzian, P. Santamaria, A. Pelisson, N. Prud'homme *et al.*, 1994 Retroviruses in invertebrates: The gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. Proc. Natl. Acad. Sci. USA **91:** 1285–1289.

Lindsley, D. L., and G. G. Zimm, 1992 *The Genome of Drosophila melanogaster*. Academic Press, London.

Mazo, A. M., L. J. Mizrokhi, A. A. Karavanov, Y. A. Sedkov, A. A. Krichevskaja *et al.*, 1989 Suppression in Drosophila: su(Hw) and su(f) gene products interact with a region of gypsy (mdg4) regulating its transcriptional activity. EMBO J. **8:** 903–911.

Mizrokhi, L. J., and A. M. Mazo, 1991 Cloning and analysis of the mobile element gypsy from D. virilis. Nucleic Acids Res. **19:** 913–916.

Morgenstern, B., 1999 DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics **15:** 211–218.

Parkhurst, S. M., D. A. Harrison, M. P. Remington, C. Spana, R. L. Kelley *et al.*, 1988 The Drosophila su(Hw) gene, which controls the phenotypic effect of the gypsy transposable element, encodes a putative DNA-binding protein. Genes Dev. **2:** 1205–1215.

Pelisson, A., S. U. Song, N. Prud'homme, P. A. Smith, A. Bucheton *et al.*, 1994 gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. EMBO J. **13:** 4401–4411.

Prud'homme, N., M. Gans, M. Masson, C. Terzian and A. Bucheton, 1995 Flamenco, a gene controlling the gypsy retrovirus of *Drosophila melanogaster*. Genetics **139:** 697–711.

Robert, V., N. Prud'homme, A. Kim, A. Bucheton and A. Pelisson, 2001 Characterization of the flamenco region of the *Drosophila melanogaster* genome. Genetics **158:** 701–713.

Robertson, H. M., 1997 Multiple Mariner transposons in flatworms and hydras are related to those of insects. J. Hered. **88:** 195–201.

Song, S. U., T. Gerasimova, M. Kurkulos, J. D. Boeke and V. G. Corces, 1994 An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. Genes Dev. **8:** 2046–2057.

Spradling, A. C., 1986 P element-mediated transformation, pp. 175–197 in *Drosophila: A Practical Approach*, edited by D. B. Roberts. IRL Press, Oxford.

Syomin, B. V., L. I. Fedorova, S. A. Surkov and Y. V. Ilyin, 2001 The endogenous Drosophila melanogaster retrovirus gypsy can propagate in Drosophila hydei cells. Mol. Gen. Genet. **264:** 588–594.

Terzian, C., C. Ferraz, J. Demaille and A. Bucheton, 2000 Evolution of the gypsy endogenous retrovirus in the Drosophila melanogaster subgroup. Mol. Biol. Evol. **17:** 908–914.

Teysset, L., J. C. Burns, H. Shike, B. L. Sullivan, A. Bucheton *et al.*, 1998 A Moloney murine leukemia virus-based retroviral vector pseudotyped by the insect retroviral gypsy envelope can infect Drosophila cells. J. Virol. **72:** 853–856.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin and D. G. Higgins, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.

Communicating editor: M. Veuille