BMC Evolutionary Biology

Research article

Evolution and phylogeny of insect endogenous retroviruses
Christophe Terzian*, Alain Pélisson and Alain Bucheton

Address: IGH-CNRS, 141 rue de la Cardonille, Montpellier , France

E-mail: Christophe Terzian* - christophe.terzian@igh.cnrs.fr; Alain Pélisson - alain.pelisson@igh.cnrs.fr;
Alain Bucheton - alain.bucheton@igh.cnrs.fr
*Corresponding author

Abstract

Background: The genome of invertebrates is rich in retroelements which are structurally reminiscent of the retroviruses of vertebrates. Those containing three open reading frames (ORFs), including an *env*-like gene, may well be considered as endogenous retroviruses. Further support to this similarity has been provided by the ability of the *env*-like gene of *DmeGypV* (the *Gypsy* endogenous retrovirus of *Drosophila melanogaster*) to promote infection of Drosophila cells by a pseudotyped vertebrate retrovirus vector.

Results: To gain insights into their evolutionary story, a sample of thirteen insect endogenous retroviruses, which represents the largest sample analysed until now, was studied by computer-assisted comparison of the translated products of their *gag, pol* and *env* genes, as well as their LTR structural features. We found that the three phylogenetic trees based respectively on Gag, Pol and Env common motifs are congruent, which suggest a monophyletic origin for these elements.

Conclusions: We showed that most of the insect endogenous retroviruses belong to a major clade group which can be further divided into two main subgroups which also differ by the sequence of their primer binding sites (PBS). We propose to name **IERV-K** and **IERV-S** these two major subgroups of **I**nsect **E**ndogenous **R**etro **V**iruses (or **I**nsect **ER**ranti**V**irus, according to the ICTV nomenclature) which respectively use Lys and Ser tRNAs to prime reverse transcription.

Background

Retrovirus-like elements have been found in the genomes of most Eukaryotes. Their integrated/proviral forms consist of two long terminal repeats (LTRs) flanking an internal region which contains one to three major open reading frames (ORFs) coding for structural and enzymatic functions necessary for their replication cycle. Based on a phylogeny of their reverse transcriptase (RT) domains, the retrovirus-like elements were divided into two major groups: the *Ty1/Copia* and the *Ty3/Gypsy* families [1]. They differ by the order of enzymatic domains encoded in the *pol* gene: Integrase – Reverse Transcriptase in the case of the *Ty1/Copia* family, and Reverse Transcriptase – Integrase in the case of the *Ty3/Gypsy* family which is also the case for vertebrate retroviruses. Moreover, the *Ty3/Gypsy* family is more closely related to retroviruses than to the *Ty1/Copia* family. Recently, the International Committee on Taxonomy of Viruses (ICTV) has proposed to call these groups *Metaviridae* and *Pseudoviridae,* respectively [2]. Among the *Metaviridae* present in the genome of Invertebrates, a noticeable proportion contain an *env*-like third ORF which makes them strikingly similar to vertebrate endogenous retroviruses. *DmeGypV* (the Dro-

sophila *Gypsy* endogenous retrovirus) is the best studied of these elements, and its infectious properties due to its *env*-like gene have been demonstrated by feeding experiments [3,4] and by the observation that a moloney murine leukemia virus pseudotyped with the *DmeGypV env* gene was infectious for Drosophila cells [5].

A relationship between the envelope proteins of some of these insect endogenous retrovirus and the *Ld130* group of baculovirus envelope proteins was recently described [6]. Moreover, it was shown that members of this family contain a furin cleavage motif, a conserved motif downstream of this site, predicted coiled-coil domains, and a pattern of conserved cysteine residues [7]. Experimental data support these comparative analyses: it was recently reported that *Ld130* has the properties of a low-pH-dependant envelope fusion protein and may play a role in the infection cycle [8]. Moreover, IJkel et al. (2000) [9] have shown that *Se8*, the homologue of *Ld130* in *Spodoptera exigua,* is an envelope fusion protein, the R-X-K-R corresponding to the furin-like proprotein convertase cleavage site. Thirteen insect endogenous retroviral sequences are now available, providing an opportunity to analyse in detail their evolutionary relationships. The results presented here show that most of these sequences (including that of *DmeGypV*) cluster in a very homogeneous and specific group.

Results

The sequences of thirteen insect endogenous retroviruses were included in this analysis, which represents the largest sample analysed until now. Their virus names, genome sequence accession numbers and assigned abbreviations are indicated in Table 1.

Table 1: Errantivirus species names, abbreviations and accession numbers

Species Name	Abbreviation	Accession number
Drosophila melanogaster Gypsy virus	*DmeGypV*	M12927
Drosophila melanogaster 17-6 virus	*Dwe176V*	X01472
Drosophila melanogaster 297 virus	*Dme297V*	X03431
Drosophila melanogaster ZAM virus	*DmeZamV*	AJ000387
Drosophila melanogaster Idefix virus	*DmeIdeV*	AJ009736
Drosophila melanogaster nomad virus	*DmeNomV*	AF039416
Drosophila melanogaster tirant virus	*DmeTirV*	X93507
Drosophila ananassae tom virus	*Dan Tom V*	Z24451
Drosophila subobscura Gypsy virus	*DsuGypV*	X72390
Drosophila virilis Gypsy virus	*DviGypV*	M38438
Drosophila virilis Tv1 virus	*DviTvlV*	AF056940
Trichoplusia ni TED virus	*TniTedV*	M32662
Ceratitis capitata yoyo virus	*CcaYoyV*	U60529

The primer binding sites (PBS) of insect endogenous retroviruses have unusual properties

The insect endogenous retroviruses can be divided into two main groups: (*DmeGypV DsuGypV DviGypV DmeNomV CcaYoyV*) using a tRNA^{Lys} and (*Dme176V Dme297V DanTomV DmeZamV DmeTirV DmeIdeV DviTv1V TniTedV*) using a tRNA^{Ser} binding sites. A striking feature of these element is that the last bp of the 5'LTR is always the first bp of the PBS. Hence, a mechanism should be involved in order to explain why the last nucleotide of the 5'LTR is not lost after a round of replication. Such a mechanism would imply the deletion of the 3' terminus nucleotide of the tRNA. It is also possible that the cleavage mediated by the RNase H encoded by these elements to remove the primer tRNA attached to the 5' end of the minus-strand template occurs in the CCA triplet between the terminal A and the adjacent C. The different products of the reverse transcription of these elements should be experimentally analysed in order to understand this unusual characteristic of errantiviruses.

Common motifs can be found in insect retroviral Gag and Env sequences

Gag sequences

None of the canonical domains described for Gag retroviral proteins are present in insect retroviral Gag sequences. Lack of a canonical cystein-rich nucleic acid-binding domain has also been reported in spumaviruses [10]. In the case of *DmeGypV, DviGypV and DsuGypV,* the C-terminal part of their Gag sequences contain an arginine-rich region which might act as an RNA binding motif [11] but might also play a role as a nuclear localization signal [12]. Three motifs can be identified in the N-terminal part of the sequences (Fig. 1). Moreover, these motifs are also present in the *burdock* element of the *Metaviridae* family, which does not have an *env*-like gene (see below).

Env sequences

The insect retroviral Env sequences share structural properties with vertebrate retroviral Env proteins: a signal peptide, N-glycosylation sites and a transmembrane domain [13]. It has been shown experimentally that the *env* gene of *DmeGypV, DmeZamV, DanTomV* and *DmeNomV* is expressed from a spliced mRNA [13–16]. Using the MEME program, we have identified six collinear motifs in 12 out of the 13 insect retroviral Env sequences (Fig. 2). Considering the high variability of viral envelope proteins generally described, this strongly suggests that these Env sequences form a monophyletic group. Moreover, the motif II, previously described [6,7], is common to the thirteen Env sequences. The R-X-K-R sequence present at the beginning of this motif was previously described as a common motif present in some in-
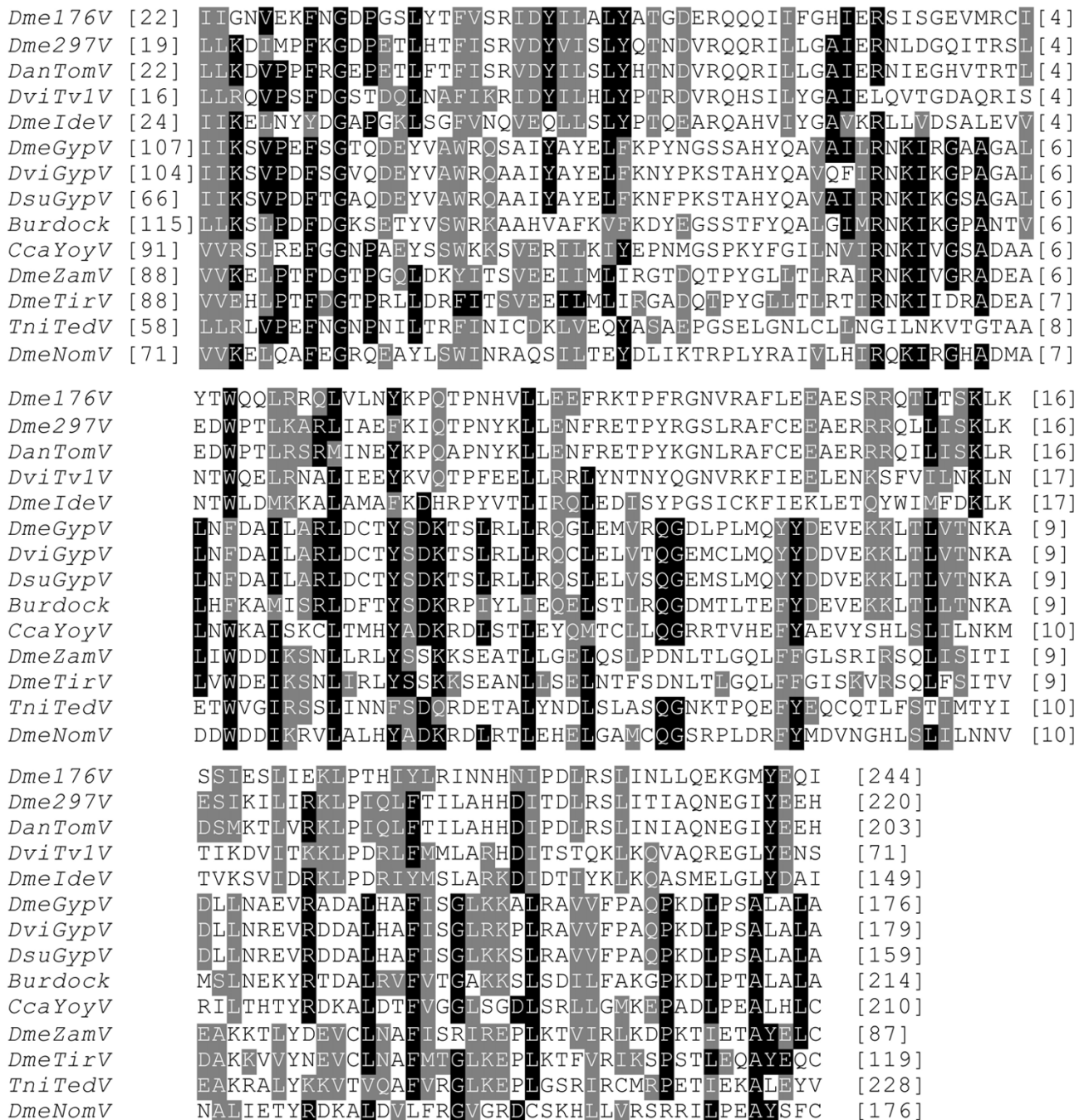
```
Dme176V  [22]  IIGNVEKFNGDPGSLYTFVSRIDYILALYATGDERQQQIIFGHIERSISGEVMRCI [4]
Dme297V  [19]  LLKDIMPFKGDPETLHTFISRVDYVISLYQTNDVRQQRILLGAIERNLDGQITRSL [4]
DanTomV  [22]  LLKDVPPFRGEPETLFTFISRVDYILSLYHTNDVRQQRILLGAIERNIEGHVTRTL [4]
DviTv1V  [16]  LLRQVPSFDGSTDQLNAFIKRIDYILHLYPTRDVRQHSILYGAIELQVTGDAQRIS [4]
DmeIdeV  [24]  IIKELNYYDGAPGKLSGFVNQVEQLLSLYPTQEARQAHVIYGAVKRLLVDSALEVV [4]
DmeGypV  [107] IIKSVPEFSGTQDEYVAWRQSAIYAYELFKPYNGSSAHYQAVAILRNKIRGAAGAL [6]
DviGypV  [104] IIKSVPDFSGVQDEYVAWRQAAIYAYELFKNYPKSTAHYQAVQFIRNKIKGPAGAL [6]
DsuGypV  [66]  IIKSVPDFTGAQDEYVAWRQAAIYAYELFKNFPKSTAHYQAVAILRNKIKGSAGAL [6]
Burdock  [115] LLKSLPDFDGKSETYVSWRKAAHVAFKVFKDYEGSSTFYQALGIMRNKIKGPANTV [6]
CcaYoyV  [91]  VVRSLREFGGNPAEYSSWKKSVERILKIYEPNMGSPKYFGILNVIRNKIVGSADAA [6]
DmeZamV  [88]  VVKELPTFDGTPGQLDKYITSVEEIIMLIRGTDQTPYGLLTLRAIRNKIVGRADEA [6]
DmeTirV  [88]  VVEHLPTFDGTPRLLDRFITSVEEIILMLIRGADQTPYGLLTLRTIRNKIIDRADEA [7]
TniTedV  [58]  LLRLVPEFNGNPNILTRFINICDKLVEQYASAEPGSELGNLCLINGILNKVTGTAA [8]
DmeNomV  [71]  VVKELQAFEGRQEAYLSWINRAQSILTEYDLIKTRPLYRAIVLHIRQKIRGHADMA [7]

Dme176V   YTWQQLRRQLVLNYKPQTPNHVLLEEFRKTPFRGNVRAFLEEAESRRQTLTSKLK  [16]
Dme297V   EDWPTLKARLIAEFKIQTPNYKLLENFRETPYRGSLRAFCEEAERRRQLLISKLK  [16]
DanTomV   EDWPTLRSRMINEYKPQAPNYKLLENFRETPYKGNLRAFCEEAERRRQILISKLR  [16]
DviTv1V   NTWQELRNALIEEYKVQTPFEELLRRLYNTNYQGNVRKFIEELENKSFVILNKLN  [17]
DmeIdeV   NTWLDMKKALAMAFKDHRPYVTLIRQLEDISYPGSICKFIEKLETQYWIMFDKLK  [17]
DmeGypV   LNFDAILARLDCTYSDKTSLRLLRQGLEMVRQGDLPLMQYYDEVEKKLTLVTNKA  [9]
DviGypV   LNFDAILARLDCTYSDKTSLRLLRQCLELVTQGEMCLMQYYDDVEKKLTLVTNKA  [9]
DsuGypV   LNFDAILARLDCTYSDKTSLRLLRQSLELVSQGEMSLMQYYDDVEKKLTLVTNKA  [9]
Burdock   LHFKAMISRLDFTYSDKRPIYLIEQELSTLRQGDMTLTEFYDEVEKKLTLLTNKA  [9]
CcaYoyV   LNWKAISKCLTMHYADKRDLSTLEYQMTCLLQGRRTVHEFYAEVYSHLSLILNKM  [10]
DmeZamV   LIWDDIKSNLLRLYSSKKSEATLLGELQSLPDNLTLGQLFFGLSRIRSQLISITI  [9]
DmeTirV   LVWDEIKSNLIRLYSSKKSEANLLSELNTFSDNLTLGQLFFGISKVRSQLFSITV  [9]
TniTedV   ETWVGIRSSLINNFSDQRDETALYNDLSLASQGNKTPQEFYEQCQTLFSTIMTYI  [10]
DmeNomV   DDWDDIKRVLALHYADKRDLRTLEHELGAMCQGSRPLDRFYMDVNGHLSLILNNV  [10]

Dme176V   SSIESLIEKLPTHIYLRINNHNIPDLRSLINLLQEKGMYEQI  [244]
Dme297V   ESIKILIRKLPIQLFTILAHHDITDLRSLITIAQNEGIYEEH  [220]
DanTomV   DSMKTLVRKLPIQLFTILAHHDIPDLRSLINIAQNEGIYEEH  [203]
DviTv1V   TIKDVITKKLPDRLFMMLARHDITSTQKLKQVAQREGLYENS  [71]
DmeIdeV   TVKSVIDRKLPDRIYMSLARKDIDTIYKLKQASMELGLYDAI  [149]
DmeGypV   DLLNAEVRADALHAFISGLKKALRAVVFPAQPKDLPSALALA  [176]
DviGypV   DLLNREVRDDALHAFISGLRKPLRAVVFPAQPKDLPSALALA  [179]
DsuGypV   DLLNREVRDDALHAFISGLKKSLRAVVFPAQPKDLPSALALA  [159]
Burdock   MSLNEKYRTDALRVFVTGAKKSLSDILFAKGPKDLPTALALA  [214]
CcaYoyV   RILTHTYRDKALDTFVGGLSGDLSRLLGMKEPADLPEALHLC  [210]
DmeZamV   EAKKTLYDEVCLNAFISRIREPLKTVIRLKDPKTIETAYELC  [87]
DmeTirV   DAKKVVYNEVCLNAFMTGLKEPLKTFVRIKSPSTLEQAYEQC  [119]
TniTedV   EAKRALYKKVTVQAFVRGLKEPLGSRIRCMRPETIEKALEYV  [228]
DmeNomV   NALIETYRDKALDVLFRGVGRDCSKHLLVRSRRILPEAYSFC  [176]
```

Figure 1
Multiple alignment of the three conserved motifs found in Gag sequences. The numbers in brackets indicate the amino acids between motifs not used in the alignment. The alignment is shaded using to a 50% consensus with gray and black shading indicating similar and identical residues respectively.

sect and vertebrate retro viral Env sequences [14,17]. It is the consensus cleavage site recognized by a cellular endopeptidase that cleaves the precursor envelope protein into the surface (SU) and transmembrane (TM) polypeptides [18].

We first addressed the specificity of the motif II. For this purpose, we scanned for the R-x(2)-R-X(5,6)-[GE]-x(5)-[LV]-x-G-x(2)-D-x(2)-D pattern in TrEMBL using the ScanProsite program and got eight hits. Six of them are indeed insect retroviral Env sequences. The other ones
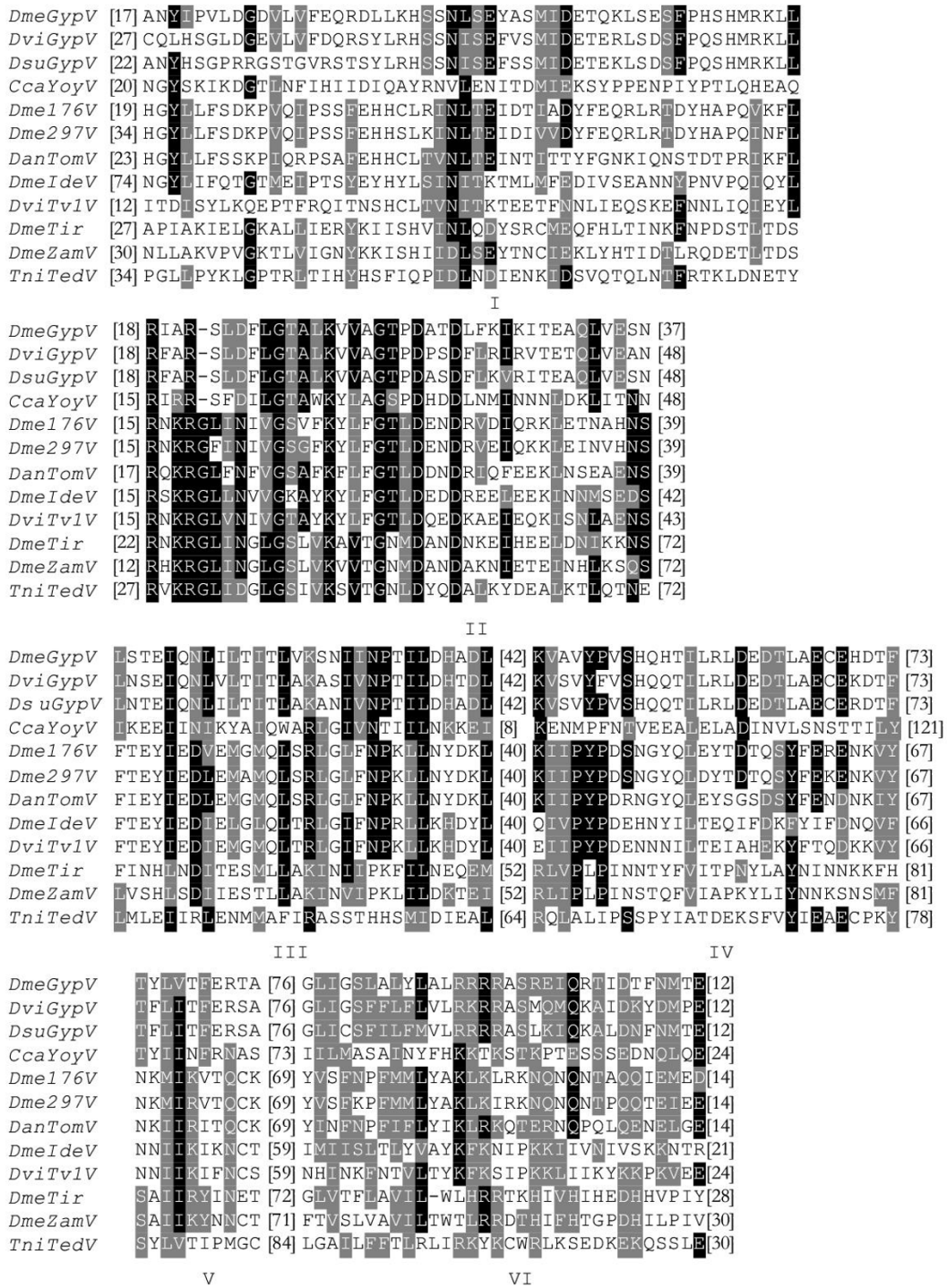
Figure 2
Multiple alignment of the six motifs common to all Env sequences, except for *DmeNomV* (see Material and Methods). The numbers in brackets indicate the amino acids between motifs not used in the alignment. The alignment is shaded using to a 50% consensus with gray and black shading indicating similar and identical residues respectively.

corresponds to *Ld130* and *Se8* previously described in Malik *et al.*[6] and Rohrmann and Karplus [7]. Because all sequences are not present in TrEMBL, we also performed a PHI-BLAST on *Dme176V* Env (default parameters) using this same pattern sequence as the seed. We detected all the expected Env sequences at significant level. Hence, the R-x(2)-R-X(5,6)-[GE]-x(5)-[LV]-x-G-x(2)-D-x(2)-D pattern can be used as a highly specific probe for the *in silico* detection of insect endogenous retroviral envelope protein sequences in databanks.

Phylogeny based on the Reverse Transcriptase (RT) and RNaseH domains
A phylogeny based on alignment of the RT and RNaseH domains using the characteristic motifs previously defined by Malik & Eickbush [19] is shown in Fig. 3B. In order to illustrate the relationships among the insects retroelements from the *Ty3/Gypsy* family, we added five insect *Metaviridae* sequences which do not have an env-like gene. This phylogenetic analysis indicates that the insect endogenous retroviruses form a homogeneous group, whereas the other elements are attached to the tree by longer branchs. One discrepancy emerged: the *burdock* element clusters with the insect endogenous retroviruses.

Gag and Env based phylogenies
Two unrooted phylogenetic trees were generated based on the Gag motifs I to III (Fig. 3A) and the Env motif II (Fig. 3C) domains previously identified. Both trees are in good agreement with the RT/RnaseH-based phylogeny (Fig. 3A). The insect retroviruses are then mainly divided into two major groups: (*DmeGypV DsuGypV DviGypV DmeNomV CcaYoyV*) and (*Dme176V Dme297V DanTomV DmeZamV DmeTir DmeIdeV DviTvIV DmeB104 TniTedV*). Moreover, this division is perfectly correlated with the form of tRNA used to prime the reverse transcription.

DmB104V, a member of the BEL clade, encodes an envelope containing the motif II
The PHI-BLAST search has revealed a match with the truncated Env-like amino-acid sequence (CAA88414.1) of *DmeB104*T(Z48503, also known as *roo*), a member of the BEL clade which is widespread in metazoans [6]. Using TBLASTN, we found multiple copies of *DmB104V env*-like genes in the Drosophila genome database. Most of these sequences display premature stop codons due to a couple of insertions/deletions (indels), but a few complete env genes are present (e.g. AC008209, 112295–114043). However, it should be noted that the *DmB104V* Env full-length sequence does not share the five other common motifs found previously in errantiviral *env* translated sequences. Moreover, the *DmeB104V* sequence is closer to the Ld130 protein family than to the

errantiviral Env sequences as revealed by the score obtained from a PSI-BLAST analysis (data not shown). This result does not agree with Frame *et al.'s* claim [20] that *DmeB104V* captured the envelope coding region from a Zam-like element, but, according to Malik *et al.*[6], suggest that a B104-like ancestor has acquired an envelope gene from a baculovirus.

The Env insect retroviral sequences share common features with the Ld130 protein family
The similarities between insect retro viral and Ld130 and Se8 envelope proteins, and evidence indicating that moloney murine leukemia virus pseudotyped with an *env* gene from *DmeGypV* is infectious for Drosophila cells suggest that the Env proteins are fusion proteins. We confirm and extend Rohrmann and Karplus' results [7] that a coiled-coil-like region is present in 9 out of the 13 Env sequences studied, as well as in Ld130 and Se8 using the Leam-coiled-VMF program [21].

Discussion
The IERV monophyletic lineage includes most of the insect endogenous retroviruses
We took advantage of the fact that the sequences of up to 13 insect endogenous retroviruses have been determined to investigate their phylogenetic relationships. The three phylogenetic trees, based on the well conserved Pol domains or on more divergent Gag and Env domains, show the same topology, supported by high bootstrap values. The congruence of these trees strengthens previous conclusions [17] concerning the monophyletic origin of the insect endogenous retroviruses. In all trees, this major clade can be further divided into two main subgroups which also differ by the sequence of their PBS. We propose to use the taxonomic nomenclature of Larsson et al. [22] to name **IERV-K** and **IERV-S** these two major subgroups of **I**nsect **E**ndogenous **R**etro**V**iruses (or **I**nsect **ER**ranti**V**irus, according to the ICTV nomenclature) which respectively use Lys and Ser tRNAs to prime reverse transcription.

In both the Pol and the Gag trees, the *burdock* element is associated with the IERV-K clade. This clustering is strengthened by the fact that it also displays the LTR-PBS overlap typical of IERVs and contains a tRNA^{Ser} PBS. This retrotransposon which, according to the ICTV, is classified as a metavirus may be a former errantivirus which has lost its *env* gene. The fact that *env* is such a modular gene explains why the classification of *Metaviridae* based only on the presence/absence of an *env* gene [2] does not always reflect the actual phylogeny.

Do IERVs Env proteins lack a typical retroviral SU subunit?
Little is known about the structure and function of the IERV Env proteins. The discovery of the relationship be-
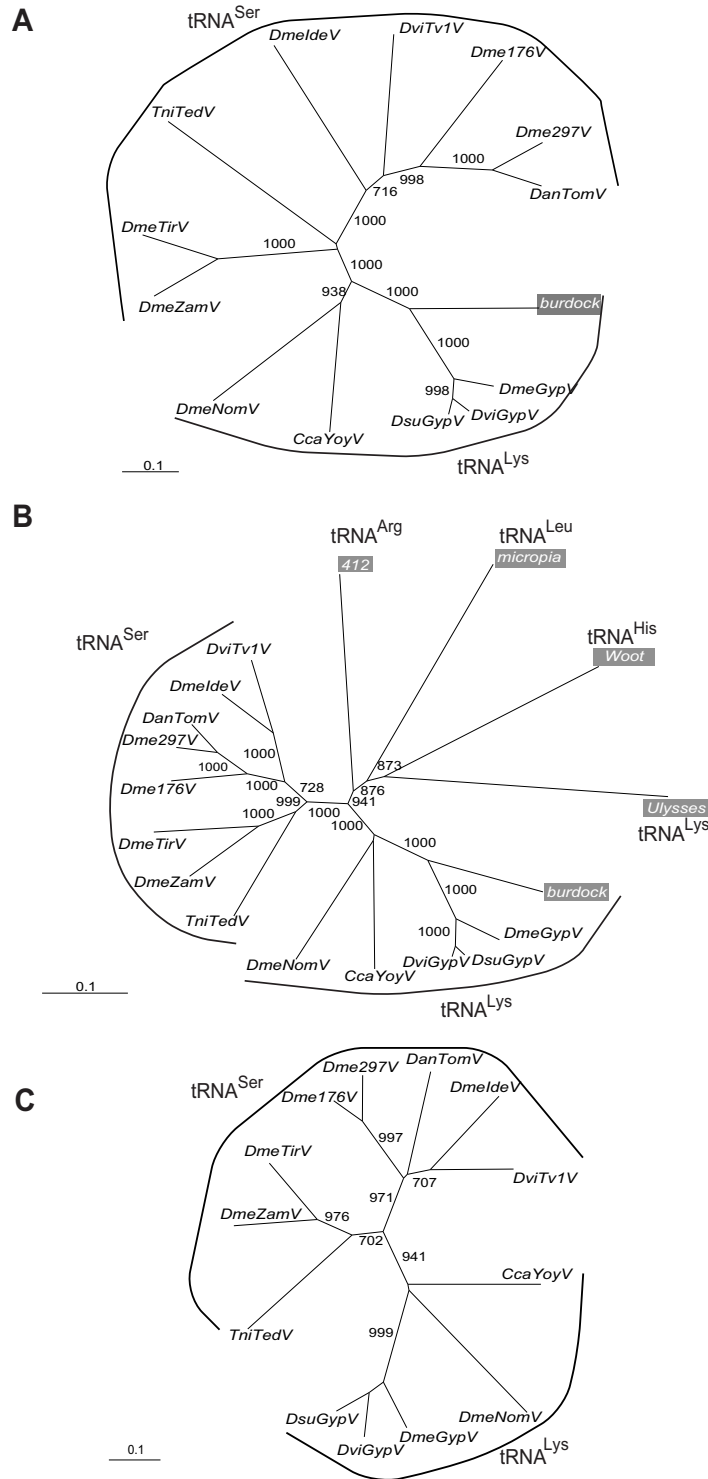
Figure 3
Unrooted phylogenetic trees based on Gag motifs I to III (A), RT and RNaseH (B), and Env motif II (C) domains. Insect *Meta-viridae* devoid of *env*-like genes appear in white lettering on gray background. These sequences (accession numbers) are: from *Drosophila melanogaster*, 412 (X04132) and *burdock* (U89994); *Ulysses* (Z24451) from *D. virilis; micropia* (XI 3304) from *D. hydei; Woot* (U09586) from *Trilobium castaneum.* Bootstrap values (when higher than 70%) and the divergence scales are indicated.

tween IERV Env and baculoviral envelope fusion proteins described previously [6,7] leads us to reconsider previous analyses based on a possible analogy with the glycoproteins of vertebrate retroviruses. The latter are known to be cleaved in two subunits by a furin-like cellular endopeptidase. The fusion peptide is located in the extracellular terminus of the transmembrane subunit (TM). The surface subunit (SU) is linked to TM by disulfide bonds and is responsible for the attachment of the virus to the cellular receptor [23]. Only the envelope proteins of *DmeGypV, DanTomV* and *TniTedV* have been studied by Western blotting [3,15,24]. Bands tentatively corresponding to the processing of *DmeGypV* and *DanTomV* envelopes at a more or less degenerate cleavage site could be visualized but in lower amounts than major bands of higher molecular weights. In the three cases, the latter were inferred to correspond to the unprocessed precursor. However, since the putative cleavage site of the motif II is located very close to the NH2 terminus of all three proteins, the sizes of the major bands are also compatible with a preferential processing of the precursor at this site. Moreover, the exact size is also difficult to predict because the Env proteins are likely glycosylated as has been shown for *DmeGypV*[3] and *TniTedV*[24]. One can also note that the only furin-like cleavage sites that can be predicted from the sequences of the other IERV envelopes are those located in the NH2 terminus of the motif II. Whether they are processed or not, these species of high molecular weight would look like large TM-like fusion proteins. A general feature of vertebrate retroviral Env proteins is that SU is always larger than TM. Since the SU-like peptide upstream of the fusion peptide of IERV Env is much smaller than TM, we wonder whether the attachment function is encoded by another (cellular?) protein or is definitely dispensable. Further studies of the replication of *DmeGypV,* the IERV paradigm, will hopefully provide deeper insights into these questions.

Conclusions

Previous studies have shown a relationship between envelope proteins of some errantiviruses and baculovirus envelope fusion proteins [6,7]. We extended these results to a larger sample of errantiviruses, and we showed that these elements form a monophyletic group. Our analysis provide a basis for the understanding of the evolutionary origin of IERV in insect genomes.

Materials and Methods
Sequence identification and retrieval
Insect retroviral sequences previously identified [2,25] were obtained from the EMBL/Genbank database using their accession number (Table 1). We observed that *DviTv1V, DviGypV, DsuGypV* and *DmeNomV* Env sequences display premature stop codons due to a couple

of indels. Hence we introduced 2 or 3 frameshifts to restore full-length Env coding capacity. The full-length "restored" sequences were then used in our analyses.

Alignment and phylogenetic analyses
The MEME program [http://meme.sdsc.edu/meme/website/] was used to identify the most conserved amino acid motifs of insect *Metaviridae* Gag and Env proteins. MEME is a motif-based method which detects conserved patterns or motifs among a set of sequences [26]. This method provides a robust way to construct multiple alignments based on segment-to-segment (or local) comparisons rather than the residue-to-residue (or global) comparisons. This approach is specially adequate when one considers sequences with low levels of global similarities [27], which is the case for Gag and Env proteins [28,29]. The selected domains were then manually aligned. Phylogenetic analyses were performed on the multiple sequence alignments using Neighbor-Joining distance methods used by CLUSTAL X [30]. Bootstrap N-J commands (N = 1000) of CLUSTAL X were used to generate bootstrapped trees. PHI-BLAST (Pattern-Hit Initiated BLAST; [http://www.ncbi.nlm.nih.gov/blast/]) is a search program that combines matching of regular expressions with local alignments surrounding the match [31]. All trees generated were visualised with Tree View 1.5.2 [32] [http://taxonomy.zoology.gla.ac.uk/rod/treeview.html] . The putative coiled-coil-like regions were detected using the LearnCoil-VMF program [http://web.wi.mit.edu/kim] , which is designed for viral membrane-fusion proteins [21].

References
1. Xiong Y, Eickbush TH: **Origin and evolution of retroelements based upon their reverse transcriptase sequences.** *Embo J* 1990, **9**:3353-62
2. Boeke JD, Eickbush TH, Sandmeyer SB, Voytas DF: **Metaviridae.** *In: in Virus Taxonomy: ICTV VIIth Report Edited by Murphy FA. New York: Springer-Verlag* 2000
3. Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG: **An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus.** *Genes Dev* 1994, **8**:2046-57
4. Kim A, Terzian C, Santamaria P, Pelisson A, Purd'homme N, Bucheton A: **Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster.** *Proc Natl Acad Sci USA* 1994, **91**:1285-9
5. Teysset L, Burns JC, Shike H, Sullivan BL, Bucheton A, Terzian C: **A Moloney murine leukemia virus-based retroviral vector pseudotyped by the insect retroviral gypsy envelope can infect Drosophila cells.** *J Virol* 1998, **72**:853-6
6. Malik HS, Henikoff S, Eickbush TH: **Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses.** *Genome Res.* 2000, **10**::1307-1318
7. Rohrmann GF, Karplus PA: **Relatedness of baculovirus and gypsy retrotransposon envelope proteins.** *BMC Evol Biol* 2001, **1**:1
8. Pearson MN, Groten C, Rohrmann GF: **Identification of the lymantria dispar nucleopolyhedro virus envelope fusion pro-**

tein provides evidence for a phylogenetic division of the bac-
uloviridae. *J Virol* 2000, **74**:6126-31

9. IJkel WF, Westenberg M, Goldbach RW, Blissard GW, Vlak JM, Zui-
dema D: **A novel baculovirus envelope fusion protein with a
proprotein convertase cleavage site.** *Virology* 2000, **275**:30-41

10. Vogt VM: **Retroviral virions and genomes.** *In: Retroviruses Edited
by Coffin JM, Hughes SH, Varmus HE. NY: Cold Spring Harbor Laboratory
Press* 199727-69

11. Alberola TM, de Frutos R: **Molecular structure of a gypsy ele-
ment of Drosophila subobscura (gypsyDs) constituting a de-
generate form of insect retroviruses.** *Nucleic Acids Res* 1996,
24:914-23

12. Robbins J, Dilworth SM, Laskey RA, Dingwall C: **Two interdepend-
ent basic domains in nucleoplasmin nuclear targeting se-
quence: identification of a class of bipartite nuclear targeting
sequence.** *Cell* 1991, **64**:615-23

13. Pelisson A, Song SU, Prud'homme N, Smith PA, Bucheton A, Corces
VG: **Gypsy transposition correlates with the production of a
retroviral envelope-like protein under the tissue-specific
control of the Drosophila flamenco gene.** *EMBO J* 1994,
13:4401-11

14. Leblanc P, Desset S, Dastugue B, Vaury C: **Invertebrate retrovi-
ruses: ZAM a new candidate in D.melanogaster.** *EMBO J* 1997,
16:7521-31

15. Tanda S, Mullor JL, Corces VG: **The Drosophila torn retrotrans-
poson encodes an envelope protein.** *Mol Cell Biol* 1994, **14**:5392-
401

16. Whalen JH, Grigliatti TA: **Molecular characterization of a ret-
rotransposon in Drosophila melanogaster, nomad, and its
relationship to other retrovirus-like mobile elements.** *Mol
Gen Genet* 1998, **260**:401-9

17. Lerat E, Capy P: **Retrotransposons and retroviruses: analysis
of the envelope gene.** *Mol Biol Evol* 1999, **16**:1198-207

18. Steiner DF: **The proprotein convertases.** *Curr Opin Chem Biol*
1998, **2**:31-9

19. Malik HS, Eickbush TH: **Modular evolution of the integrase do-
main in the Ty3/Gypsy class of LTR retrotransposons.** *J Virol*
1999, **73**:5186-90

20. Frame IG, Cutfield JF, Poulter RT: **BEL-like LTR-retrotrans-
posons in Fugu rubripes, Caenorhabditis elegans, and Dro-
sophila melanogaster.** *Gene* 2001, **263**:219-30

21. Singh M, Berger B, Kim PS: **LearnCoil-VMF: computational evi-
dence for coiled-coil-like motifs in many viral membrane-fu-
sion proteins.** *J Mol Biol* 1999, **290**:1031-41

22. Larsson E, Kato N, Cohen M: **Human endogenous proviruses.**
Curr Top Microbiol Immunol 1989, **148**:115-32

23. Swanstrom VM, Wills JM: **Synthesis, assembly and processing of
viral proteins.** *In: Retroviruses Edited by Coffin JM, Hughes SH, Varmus
HE. pp. 263–334. NY: Cold Spring Harbor Laboratory Press;* 1997263-334

24. Ozers MS, Friesen PD: **The Env-like open reading frame of the
baculovirus-integrated retrotransposon TED encodes a ret-
rovirus-like envelope protein.** *Virology* 1996, **226**:252-9

25. Bucheton A, Terzian C, Pelisson A: **Retroviruses of Drosophila:
the gypsy paradigm.** *In: Encyclopedia of Virology. Edited by Press A.
London* 19991526-1530

26. Bailey TL, Elkan C: **Fitting a mixture model by expectation
maximization to discover motifs in biopolymers.** *In: Proceed-
ings of the Second International Conference on Intelligent Systems for Mo-
lecular Biology; Menlo Park, California.* 199428-36

27. Thompson JD, Plewniak F, Poch O: **A comprehensive compari-
son of multiple sequence alignment programs.** *Nucleic Acids Res*
1999, **27**:2682-90

28. McClure MA, Johnson MS, Feng DF, Doolittle RF: **Sequence com-
parisons of retroviral proteins: relative rates of change and
general phylogeny.** *Proc Natl Acad Sci US A* 1988, **85**:2469-73

29. Lewe G, Flugel RM: **Comparative analysis of the retroviral pol
and env protein sequences reveal different evolutionary
trees.** *Virus Genes* 1990, **3**:195-204

30. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The
CLUSTAL_X windows interface: flexible strategies for mul-
tiple sequence alignment aided by quality analysis tools.** *Nu-
cleic Acids Res* 1997, **25**:4876-82

31. Zhang Z, Schaffer AA, Miller W, Madden TL, Lipman DJ, Koonin EV,
Altschul SF: **Protein sequence similarity searches using pat-
terns as seeds.** *Nucleic Acids Res* 1998, **26**:3986-90

32. Page RD: **Tree View: an application to display phylogenetic
trees on personal computers.** *Comput Appi Biosci* 1996, **12**:357-8