The Retrotransposon Osvaldo from Drosophila buzzatii Displays All Structural Features of a Functional Retrovirus

Alexandros Pantazidis,1 Mariano Labrador,2 and Antonio Fontdevila
Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Spain

The Osvaldo retrotransposon has shown a high transposition rate in some strains of Drosophila buzzatii and in hybrids between D. buzzatii and its sibling D. koepferae. In order to understand the molecular basis of this phenomenon, we developed a procedure to clone a recently transposed copy with the aim of characterizing an active, full-length Osvaldo element. The complete nucleotide sequence of Osvaldo, obtained from a recent insertion site, was determined. Osvaldo is 9,045 bp long and is composed of a central coding region flanked by identical long terminal repeats (LTRs) of 1,196 bp each. Sequences homologous to the polypurine tract and tRNA-primer-binding site of retroviruses are located adjacent to the 3′ and 5′ LTRs, respectively. The internal region of Osvaldo contains three long open reading frames (ORFs 1, 2, and 3), comparable in size and location to gag, pol, and env retroviral genes. The conceptual translation of Osvaldo ORF1 exhibits sequence homology to HIV1 and SIV capsid (p24) and nucleocapsid (p7) mature proteins. ORF2 encodes the putative protease (PR), reverse transcriptase/ribonuclease H (RT/RH), integrase (IN), and a significant portion of the surface envelope (ENV) protein that is interrupted by a putative intron. A third ORF encodes the remaining part of the ENV protein. The predicted 62-kDa ENV protein shares several general features with membrane glycoproteins, including a potential signal peptide, a transmembrane domain near the C-terminus that could function as a membrane anchor, four consensus N-linked glycosylation motifs, and, finally, a potential protease cleavage site. The phylogenetic relationships of Osvaldo are explored, and they suggest that Osvaldo may constitute a new family of retroviruses in insects, distantly related to the previously described group of gypsy retroviruses.

Introduction

Long terminal repeat (LTR) retrotransposons have been identified in a wide variety of organisms, including fungi (Boeke and Sandmeyer 1991), plants (Flavell et al. 1992; Kumar 1996), invertebrates (Springer, Davidson, and Britten 1991; Britten 1995), and vertebrates (Flavell and Smith 1992). Many of these LTR retrotransposons can be sorted into two superfamilies named after the copia and gypsy elements from Drosophila.

Copia-like and gypsy-like elements are structurally similar to retroviruses, and it is widely considered that they (especially the gypsy group) are close relatives of the vertebrate retroviruses (Eickbush 1994; Bucheton 1995). For that reason, the abbreviations used for proteins or other structural features in retroviruses have been adopted (Leis et al. 1988). The phylogenetic relationships of LTR retrotransposons have been described elsewhere (Xiong and Eickbush 1990; Eickbush 1994), particularly the relationships of elements within the copia-like (Flavell et al. 1992) and gypsy-like families (Springer and Britten 1993).

The major difference between most retrotransposons and retroviruses is that the retrotransposons lack a gene encoding the envelope protein (ENV). This protein is required by retroviruses for cell-to-cell infectious processes. For that reason, retrotransposons have been distinguished from retroviruses by the apparently exclusive intracellular cycles of replication and integration. However, this distinction has been eroded in recent years by two discoveries: that retroviruses can undergo limited cycles of intracellular replication and integration (Tchennio and Heidmann 1991) and that gypsy encodes envelope proteins and is transmitted extracellularly (Kim et al. 1994; Pelisson et al. 1994; Song et al. 1994; 1997).

The gypsy-like elements for which the full sequence is known include 176 (Saigo et al. 1984), 297 (Inouye, Yukie, and Saigo 1986), Tom (Tanda, Mullor, and Corces 1994), 412 (Yuki et al. 1986), gypsy (Marlor, Parkhurst, and Corces 1986), and Zam (Leblanc et al. 1997) from Drosophila and TED from the moth Trichoplusia ni (Friessen and Nissen 1990). As with RT, the protein sequence predicted from the first open reading frames (ORFs) of 176, 297, and Tom show that these three elements form a gypsy subfamily (Inouye, Yuki, and Saigo 1986; Tanda et al. 1988; Springer and Britten 1993). Although the biochemistry of ORF2 proteins is mostly unknown, reverse transcriptase (RT) activity has been associated with gypsy particles (Song et al. 1994). The predicted proteins encoded in the third ORF contain hydrophobic, potentially membrane-spanning regions and appear to constitute envelope proteins (Kim et al. 1994; Song et al. 1994, 1997; Pelisson et al. 1994; Tanda, Mullor, and Corces 1994).

Recently, Labrador and Fontdevila (1994) investigated the transposition of Osvaldo, a new LTR retrotransposon from Drosophila buzzatii. A D. buzzatii inbred line in which all insertion sites were known was used to measure Osvaldo transposition rates in hybrids between this line and its sibling Drosophila koepferae from the repleta group (Ruiz and Wasserman 1993). Considering new insertions only in euchromatic posi-
tions, Labrador and Fontdevila (1994) presented evidence that Osvaldo transposition occurs at high rates, suggesting that transposition can be enhanced by interspecific hybridization. These results have been reproduced in a recent work showing that the Osvaldo transposition rate is significantly higher in hybrids, but transposition rate is also high in strains recently extracted from natural populations (Labrador et al. 1999). The characterization of the cDb314 clone, which contains the sequence used as a probe for in situ hybridization, showed that Osvaldo belongs to the gypsy group. However, the sequence analysis of cDb314 revealed that this clone contained a defective heterochromatic copy of the element, and no further conclusions such as the presence or absence of a third ORF of Osvaldo could be obtained (Labrador and Fontdevila 1994).

The high transposition activity associated with Osvaldo and its relation to the gypsy family suggest that this element could be an excellent system with which to study the evolution and the molecular biology of these retroviruses, first considered to be retrotransposons. As an essential step toward the understanding of Osvaldo transposition mechanisms and the investigation of a potential third ORF; we isolated and sequenced a new, apparently intact, copy of the element. This copy has been cloned from the euchromatic region 2F4f, which has already been described as the site of a recent insertion event in the second chromosome of D. buzzatii (see Labrador and Fontdevila 1994). According to our data, Osvaldo consists of 9,045 bp, and its LTRs measure 1,196 bp. The remaining central part has three ORFs that correspond to the gag, pol, and env retroviral genes, respectively. Here, we present the whole nucleotide sequence of this intact element and compare its structural features and its putative translation products with their homologs from other gypsy-related Drosophila retrotransposons and vertebrate retroviruses. We suggest that Osvaldo could constitute a new family of insect retroviruses that is not closely related to gypsy.

Materials and Methods
Genomic Library

Total genomic DNA was extracted according to Piniol et al. (1988) from the D. buzzatii inbred line BU-30/4 (Labrador and Fontdevila 1994). The extracted DNA was partially digested with MboI and treated with RNase and alkaline phosphatase according to Maniatis, Fritsch, and Sambrook (1982). Then, 1.6 μg of genomic DNA was ligated with 4 μg of lambda DASH II arms, digested with BamHI. Recombinant DNA was packaged using the Gigapack II XL-4 extracts from Stratagene.

DNA Hybridization

Forty thousand plaque-forming units (pfu’s) of the genomic library were screened for Osvaldo sequences by hybridization with a 2.1-kb KpnI fragment from cDb314 containing only Osvaldo-coding regions (Labrador and Fontdevila 1994). Probes were obtained by random-primed labeling of 300 ng DNA with digoxigenine (DIG)-11-dUTP (Boehringer Mannheim). Hybridization and posthybridization procedures are detailed in the protocol provided by the DIG supplier. Hybridization was performed overnight at 42°C with 50% formamide, and the maximum-stringency posthybridization washes were carried out at 50°C with 0.1 SSC.

Lambda phage lysates were prepared as described in Patterson and Dean (1987). Recombinant phage DNA was extracted by using the Promega Wizard Lambda Preps purification system. Southern procedures were performed as described in Maniatis, Fritsch, and Sambrook (1982). For reverse genomic hybridization (Macrin, Labrador, and Fontdevila 1992), approximately 1 μg of total genomic DNA from D. buzzatii BU-30/4 female adults was labeled with (DIG)-11-dUTP according to the random priming procedure and used as a probe. Hybridization conditions, posthybridization washes, and development procedures were as described above. In situ hybridizations were carried out as in Labrador and Fontdevila (1994). We used the polytene chromosomes from the strain BU-42 (63/7), which lacks euchromatic positions of Osvaldo, in order to map by in situ hybridization the cytological position of the euchromatic genomic flanking regions from the selected phages. BU-30/4 was used to confirm hybridization of the new Osvaldo sequences to euchromatic copies of Osvaldo (Labrador and Fontdevila 1994).

Sequencing of Osvaldo

Nuclease Bal 31 was used to obtain overlapping series of DNA fragments and subcloned in vector pTZ18U for double-stranded DNA sequencing reactions. High-quality double-stranded DNA was obtained using QIAGEN Midi preps. Sequencing was performed completely in both strands by the dideoxy-chain method (Sanger, Nicklen, and Coulson 1977), using both Pharmacia ALF and Applied Biosystems 373A automatic sequencing systems. DNA and protein sequences were analyzed using the PCGENE package of programs and compared with sequences in the databases accessible through the National Center for Biotechnology Information at the National Institute of Health. The BLAST program was used to establish DNA and protein simple alignments between different sequences, and the CLUSTAL W (Higgins, Thompson, and Gibson 1996) program was used to establish the protein multiple alignments.

Results and Discussion
Cloning Osvaldo from a Recent Insertion Site

Twenty genomic clones with strong hybridization signals were selected after the screening of the gene library. These clones contained Osvaldo sequences from both euchromatic and heterochromatic positions present in the BU-30/4 line. We expected that some of these sequences would correspond to the recently described transposition in this line (Labrador and Fontdevila 1994). Each clone was individually hybridized in situ to polytene chromosomes of the D. buzzatii BU-42 (63/7) stock, in which no euchromatic copies of Osvaldo have ever been detected (Labrador and Fontdevila 1994; un-
published data). For that reason, any single hybridization signal in euchromatin was attributed to nonrepetitive DNA sequences flanking the TE, and consequently the corresponding clones were qualified as euchromatic. By this procedure, 7 out of the 20 isolated clones were characterized as euchromatic Osvaldo clones and became the subjects of further molecular analysis. Four out of the seven euchromatic clones \( (A_{15}, B_4, B_{10}, \text{ and } B_{16}) \) were cloned from the 2F4f band of chromosome 2, while the rest \( (A_{17}, B_8, \text{ and } B_{20}) \) were cloned from the bands F1f (chromosome 5), A4a (chromosome 2), and A4a (chromosome 3), respectively.

In the clone \( A_{15} \), a large part (~6 kb) of the TE was cloned just next to the right arm of our cloning vector. Consequently, the isolation of the entire TE was carried out by means of standard chromosomal walking procedure. We used the reverse genomic hybridization method (Pirrotta 1986) to determine the repetitive DNA sequences within \( A_{15} \). This method uses total genomic DNA as a probe against the DNA of recombinant clones. Under appropriate conditions, only those fragments that carry repetitive DNA show hybridization signal (see, e.g., Pirrotta, Hadfield, and Pretorius 1983; Baumann et al. 1987; Healy, Russell, and Miklos 1988; Marin, Labrador, and Fontdevila 1992). Using as probes the unique DNA fragments adjacent to the repetitive sequences of the \( A_{15} \) clone, we rescanned the original \textit{D. buzzatii} BU-30/4 genomic library and isolated a new clone, named \( A_{15(10)} \), containing a copy of the TE flanked on both sides by genomic DNA. Most interestingly, this copy has been cloned from the euchromatic region F4f of the second \textit{D. buzzatii} chromosome, which has already been described as a site for a recent insertional event (Labrador and Fontdevila 1994).

General Features of Osvaldo

\textit{Osvaldo} is an LTR retrotransposon of \textit{D. buzzatii}, whose insertion creates a 4-bp target site duplication of the host sequence. The restriction map, a schematic representation of the ORFs and the location of the genes encoding the putative protein products are shown in figure 1. The \textit{Osvaldo} element has a total length of 9,045 bp and contains two LTRs of 1,196 bp. The complete nucleotide sequence of \textit{Osvaldo}, together with the duplicated 4-bp target site, is presented in figure 2. The central part of the element contains three different ORFs that encode putative proteins, whose amino acid sequence is also shown in figure 2. Some of the \textit{Osvaldo} general characteristics are presented in table 1 and compared with their analogs from members of the \textit{Drosophila} gypsy family.

Structure of the LTRs and the Adjacent Sequences

LTRs in retroviruses and retrotransposons are in direct orientation at each end, and within these are U3, R, and U5 regions, which contain signals for initiation and termination of transcription. Transcription begins at the U3-R junction in the 5' LTR and proceeds through the U5 portion of the 5' LTR, the large central coding portion of the transposon, and the U3 portion of the 3' LTR and finally ends at the R-U5 junction in the 3' LTR, where it is terminated (polyadenylated). The locations on the \textit{Osvaldo} LTRs of various transcription signals are also shown in figure 2. We do not yet have experimental evidence indicating that the sequence TTCAGAC, located at the 5' LTR nucleotide positions 667–673, is the Inr element of Osvaldo, but its homology to the experimentally determined initiator elements of other LTR retrotransposons (see table 1 in Arkhipova and Ilyin 1992) with respect to the presence of a putative A\textsubscript{ACAC} +30 element at nucleotide positions 692–695 suggests that this might be the case.

Another common feature of the LTRs in retrotransposons is the presence of the AATAAA sequence that corresponds to the AAUAAA polyadenylation signal of the RNA transcript, which in \textit{Osvaldo} is located at nucleotide positions 8,967–8,972 of the 3' LTR. The positions of the RNA start site and the polyadenylation site define the length of the U3, R, and U5 domains in each LTR. The total length of each \textit{Osvaldo} LTR is 1,196 bp, and, according to our predictions, the putative RNA start site is situated about 450 nt upstream of the putative polyadenylation signal. Therefore, \textit{Osvaldo} transcripts must have a terminal redundancy (R) domain much longer than the U5 domain, which probably does not exceed 80 nt. If our prediction is correct, the putative U3 and the R+U5 domains of each \textit{Osvaldo} LTR must not exceed 666 and 530 nt, respectively, and the total premature \textit{Osvaldo} RNA transcript would measure ~8.3 kb.

Each LTR of the \textit{Osvaldo} element is internally bordered by the short inverted repeat TGT\textsubscript{A}G\textsubscript{G}]-[CgTt-
Fig. 2.—*Osvaldo* complete nucleotide and deduced protein sequences (DDBJ/EMBL/GenBank accession number: AJ133521). The ranges of the LTRs are shown by L-shaped arrowheads labeled “LTR.” The open boxes externally adjacent to the LTRs indicate the duplicated 4-bp target insertion site of the host DNA. Underlined nucleotides at the limits of each LTR show the inverted repeats that belong to *Osvaldo* sequence. L-shaped arrowheads labeled “ORF” show the ranges of the three ORFs. An asterisk designates the termination codon of each ORF.

Putative motifs implicated in several important molecular processes are shaded and underlined: 

(a) sequence almost identical to that experimentally determined in the case of the 17.6 element around the 5′ RNA start site and (b) its 3′ promoter essential element downstream of the putative RNA start site (see also table 2); (c) polyadenylation signal; (d) minus DNA strand primer-binding site (PBS); (e) plus DNA strand primer-binding site (PBS); (f) RNA start site of the 17.6 element; (g) internal T-rich motif (TIM); (h) internal C-rich motif (ICM).
FIG. 2 (Continued) primer-binding site (polypurine track, PPT); (f) heptanucleotide sequence of the putative ribosomal frameshift site identical to that determined for the feline immunodeficiency virus; (g) splicing donor sequence near the 3′ end of the ORF2; (h) splicing acceptor sequence near the 5′ end of the ORF3. The underlining of the first codon permits the easy localization of the gag (MA, CA, and NC) and pol (PR, RT, RH, and IN) Osvaldo polyproteins. SP and shadowed amino acids denote the signal peptide from the predicted amino acid sequence of the ORF2–ORF3 Osvaldo env products (see also fig. 6).

ACA, and externally adjacent to the LTRs are genomic short direct repeats (fig. 2). Two base pairs after the 3′ end of the left LTR, there is a 18-bp sequence similar to the tRNA binding sites of other retroviruses which is involved in the synthesis of the first (minus) strand of viral DNA. This sequence is homologous to the Droso-

The Coding Region of Osvaldo

The central part of Osvaldo consists of 6,653 bp and contains three different ORFs (ORFs 1, 2, and 3) comparable in size and location with those observed in other Drosophila retrotransposons or vertebrate retrovirus (fig. 1B). Translation of these ORFs results in protein sequences of 413, 1,147, and 440 amino acids, respectively (fig. 2) with high degrees of similarity to gag (fig. 3) and RT of pol (fig. 4) retroid gene products. The first ORF starts from nucleotide position 1,248 (ACA), one codon upstream of the first methionine residue that is probably used as a translation initiation site to make the putative Osvaldo ORF1–ORF2 fusion protein (fig. 2). The putative translational product of ORF1 exhibits sequence homology to p17 (matrix), p24 (capsid), and p7 (nucleocapsid) gag polyprotein domains of human immunodeficiency virus type 1 (HIV-1) and the simian immunodeficiency virus (SIV) (fig. 3). Although the homology corresponding to the nucleocapsid and partially to the capsid is reliable and was determined after a search of the data banks using BLASTP, the remaining GAG homologies are the result of a posterior alignment using CLUSTAL W1.6, and their significance...
is more difficult to establish. The identity that we show here is surprisingly high considering the low conservation usually found between GAG proteins, which are very distinct even between close retroviruses or retrotransposons. The significance of this homology is not clear, particularly considering that HIV is a human retrovirus. It is even more surprising that no other putative insect retrovirus described so far has shown such a perfectly defined zinc-finger domain. The existence of these perfect retroviral GAG proteins argues in favor of our claim that Osvaldo is a true retrovirus.

As in the cases of 17.6, 297, 412, and gypsy, ORFs 1 and 2 in Osvaldo overlap with each other. Although the lengths of these overlapping regions vary from one element to another (table 1), in all cases ORFs 1 and 2 were found to be out of phase by −1. Thus, the putative Osvaldo ORF1-ORF2 fusion protein is likely to be produced by a specific ribosomal frameshifting event. A potential signal (GGGAAAC) for such an event was found at nucleotide positions 2420–2426 near the 3′ end of the gag-pol 292-bp overlapping region. Similar sequences have already been identified to play such a role in gag/pol or gag/pro (for retroviral gene arrangement, see Varmus 1988) overlapping regions of several retroviruses (table 2). Interestingly, the heptanucleotide GGGAAAC is located very close to the 3′ end of the gag-pol and 292-bp overlapping region. Several retroviruses have already been identified to play such a role in gag/pol or gag/pro (for retroviral gene arrangement, see Varmus 1988) overlapping regions of several retroviruses (table 2). Interestingly, the heptanucleotide GGGAAAC is located very close to the 3′ end of the gag-pol and 292-bp overlapping region. Similar sequences have already been identified to play such a role in gag/pol or gag/pro (for retroviral gene arrangement, see Varmus 1988).

**Table 1**

| Structural Properties of the Retrotransposon Osvaldo Compared with other Drosophila Long Terminal Repeat (LTR) Retrotransposons Related to the gypsy Retrovirus |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| **Osvaldo**     | **gypsy**       | **17.6**        | **297**         | **412**         | **17.6**        | **297**         |
| **Element/LTR** | **Element/LTR** | **Element/LTR** | **Element/LTR** | **Element/LTR** | **Element/LTR** | **Element/LTR** |
| 9,045/1,196     | 7,469/482       | 7,439/512       | 6,995/414       | 7,060/474       | 7,340/514       | 7,440/481       |
| **IR**          | **IR**          | **IR**          | **IR**          | **IR**          | **IR**          | **IR**          |
| 4               | 4               | 4               | 4               | 4               | 4               | 4               |
| **Zinc Fingers**| **Zinc Fingers**| **Zinc Fingers**| **Zinc Fingers**| **Zinc Fingers**| **Zinc Fingers**| **Zinc Fingers**|
| Yes             | Yes             | Yes             | Yes             | Yes             | Yes             | Yes             |
| **Duplication** | **Duplication** | **Duplication** | **Duplication** | **Duplication** | **Duplication** | **Duplication** |
| 88              | 78              | 58              | 46              | 40              | 88              | 88              |
| **Spleen**      | **Spleen**      | **Spleen**      | **Spleen**      | **Spleen**      | **Spleen**      | **Spleen**      |
| 101             | 101             | 80              | 60              | 60              | 101             | 101             |
| **Zinc Dimer**  | **Zinc Dimer**  | **Zinc Dimer**  | **Zinc Dimer**  | **Zinc Dimer**  | **Zinc Dimer**  | **Zinc Dimer**  |
| 69              | 69              | 78              | 58              | 58              | 69              | 69              |
| **PR/RT**       | **PR/RT**       | **PR/RT**       | **PR/RT**       | **PR/RT**       | **PR/RT**       | **PR/RT**       |
| 92              | 88              | 58              | 46              | 40              | 88              | 88              |
| **RH/IN**       | **RH/IN**       | **RH/IN**       | **RH/IN**       | **RH/IN**       | **RH/IN**       | **RH/IN**       |
| 123             | 122             | 123             | 123             | 123             | 123             | 123             |
| **GAG**         | **GAG**         | **GAG**         | **GAG**         | **GAG**         | **GAG**         | **GAG**         |
| Yes             | Yes             | Yes             | Yes             | Yes             | Yes             | Yes             |
| **pol**         | **pol**         | **pol**         | **pol**         | **pol**         | **pol**         | **pol**         |
| No              | No              | No              | No              | No              | No              | No              |

**References:** This works; gypsy (Marie, Parkhurst, and Cores 1980); 17.6 (Saigo et al. 1984); 297 (Inouye, Yuki, and Saigo 1986); Tom (Tamura et al. 1988; Tamura, Muldo, and Cores 1974–1984; Yuki et al. 1985).

**Notes:**

- **Sequences at both ends of LTRs.** Uppercase denotes part of the repeat structure; lower case denotes not part of the repeat structure.
- **Presumed plus strand primer-binding site (polypurine tract, PPT) adjacent to 3′ LTR sequence.**
- **GAG metal-finger motif is a Cys/His sequence of the type C(2X)C(4X)H(4X)C found in the retroviral nucleocapsid proteins.**
- **Number of amino acid residues between the protease (PR) and the reverse transcriptase (RT) products, which could not be aligned.**
- **Numbers of amino acid residues between the ribonuclease H (RH) and the integrase (IN) pol products, which could not be aligned.**

**Downloaded from https://academic.oup.com/mbe/article-abstract/16/7/909/2925465 by guest on 03 March 2019**
Fig. 3.—Multiple-sequence alignment of the putative gag products matrix (MA), capsid (CA), and nucleocapsid (NC) encoded by Osvaldo (this work), HIV-1, MN isolate (Gurgo et al. 1988), and SIVsmE543 (Hirsch et al. 1997). The positions at which Osvaldo shares identical or chemically similar amino acids with one or more sequences are shaded. The open boxes of the nucleocapsid shows the characteristic CXXCXXXXHXXXXC metal finger motif. Numbers below the sequences refer to biochemically similar amino acids (Blosum62): (1) D, N; (2) E, Q; (3) S, T; (4) K, R; (5) F, Y, W; (6) L, I, V, M.

A comparison of this sequence of Osvaldo and the cDb314 sequence described in Labrador and Fontdevila (1994) reveals a high degree of identity—close to 99% along the conserved GAG and POL proteins (data not shown). However, the former sequence corresponds to a defective copy of Osvaldo in which the protease was included in the ORF1, a stop codon was found in the middle of the RT, and no LTRs or envelope-coding sequences were identified. Additional data suggesting that the described intron is a structural part of the element comes from the finding that the intervening sequence is also present in the previous sequence of Osvaldo (fig. 6). Although the sequence cDb314 corresponds to a heterochromatic copy of Osvaldo and therefore is structurally degenerated, our alignment clearly shows stretches of the intron conserved in both copies of the element. This conservation can only be explained if the intron is a structural part of the element. Otherwise, it would not be present in two copies as highly divergent as cDb314 and the Osvaldo sequence described here. Experimental evidence of the presence of this intron and of the resulting putative fusion proteins or of a second intron suggested by our analysis will be necessary, however, to confirm our predictions.

The Envelope Protein of Osvaldo

No appreciable homology was found between the third ORF of Osvaldo and any retroviral env gene. However, the predicted Osvaldo ENV protein encoded by the spliced RNA between ORF2 and ORF3 shows several structural characteristics of retroviral envelope proteins (fig. 7). The putative Osvaldo env gene product contains a putative signal peptide required for targeting for the endoplasmic reticulum and secretion for the membrane. Two potential cleavage sites that conform well to the (−3, −1) rule (von Heijne 1986) follow the signal peptide and are indicated by arrows in figure 7. The predicted processed protein would be of 566 residues and would have a molecular weight of 62,137. In a process that is required for infectivity, retroviral ENV proteins are cleaved at dibasic motifs (Arg/Lys-Xaa-Lys-Arg) by a host endopeptidase to generate surface and transmembrane peptides (for a review, see Swanstrom and Wills 1997). Cleavage of Osvaldo ENV proteins at a putative protease site (Lys-Thr-Lys-Arg) near residue 404 would give rise to two polypeptides with molecular weights of 44,614 and 17,541 that could correspond to the surface and transmembrane proteins of retroviruses, respectively. The putative transmembrane protein contains a 22-residue hydrophobic transmembrane domain and hydrophobic amino acids in the amino-terminal end charac-

---

The entire text is a detailed analysis of the genetic and structural characteristics of the retrovirus Osvaldo, focusing on its gag, capsid, and nucleocapsid proteins, as well as its envelope protein, and comparing it with other retroviral sequences.
FIG. 4.—Multiple-sequence alignment of reverse transcriptases. Empty gaps are between the conserved domains (boxes) of reverse transcriptases established by Xiong and Eickbush (1990). Numbers indicate the amino acids not used in the alignment. Chemically similar amino acids are as in figure 3. Elements are:

- Woot, Tribolium castaneum (Beeman et al. 1996);
- Jockey, Drosophila melanogaster (Priimaegi, Mizrokhi, and Ilyin 1988);
- Ty1, D. melanogaster (Boeke et al. 1988);
- Copia, D. melanogaster (Mount and Rubin 1985);
- Mlv, D. melanogaster (Shinnick, Lerner, and Sutcliffe 1981);
- HTLV1, human (Seiki et al. 1983);
- RSV, human (Broome and Gilbert 1985);
- HIV1, human (accession number 1123014PID);
- SIV, human (Tsujimoto et al. 1989);
- Mag from Caenorhabditis elegans (accession number 726387);
- Mag from Bombyx mori (Michaille et al. 1990);
- Ulysses, Trichoplusia ni (Friesen and Nissen 1990);
- Cer-1, C. elegans (Britten 1995);
- CaMV, Cauliflower mosaic virus (Chenault and Melcher 1993);
- CoYMV, Tomato yellow fleck virus (Medberry, Lockhart, and Olszewski 1990);
- D. melanogaster (Yuki et al. 1986);
- gypsy, D. melanogaster (Marlor, Parkhust, and Corces 1986);
- gypsy, Ceratitis capitata (accession number 1402848);
- 17.6 D. melanogaster (Saigo et al. 1984);
- Tom D. ananassae (accession number 422418);
- 297 D. melanogaster (Inouye, Yuki, and Saigo 1986);
- Ted Tricoplusia ni (Friesen and Nissen 1990);
- Zam D. melanogaster (Leblanc et al. 1997);
- Micropia D. melanogaster (Lankenau et al. 1988);
- Mdg3 D. melanogaster (Avedisov and Il'in 1995).

Ty3, yeast S. cerevisiae (Hansen, Chalker, and Sandmeyer 1988); TF-1, Schizosaccharomyces pombe (Levin, Weaver, and Boeke 1990); Del-1, retrotransposon (Smyth et al. 1989).

Downloaded from https://academic.oup.com/mbe/article-abstract/16/7/909/2925465 by guest on 03 March 2019
teristic of the fusion peptide which could mediate penetration of the host cell membrane after receptor binding (fig. 7). In addition, both proteins contain several putative glycosylation sites for addition of N-linked carbohydrate side chains and Cys residues that could mediate attachment between the surface and transmembrane proteins via disulfide bonds, as in the case of vertebrate retroviruses (Varmus and Brown 1989).

**Osvaldo Is Not Closely Related to gypsy Retrovirus**

Gypsy is the only retrotransposon so far known to be able to produce infectious virus particles (Song et al. 1994, 1997). Other retrotransposons related to gypsy also display the structural requirements for a retrovirus, such as the capability to encode an envelope protein. These retrotransposons are Ted, 17.6, 297, Tom, and Zam. Some of them actually produce a transcript that corresponds to an ORF3 that encodes for envelope proteins (Tanda, Mullor, and Corces 1994; Leblanc et al. 1997). All these elements are closely related and clustered in a single group inside the more diverse gypsy/ty3 class of retrotransposons. In addition to the envelope, Osvaldo also displays a very well conserved capsid and nucleocapsid, closely related to retrovirus GAG proteins, together with an unusual intron interrupting the env gene. In view of these observations, and for a further understanding of the relationship of Osvaldo with other retrotransposons and retroviruses, we performed multiple alignments using the RT, a protein that previously showed seven motifs highly conserved between all retroelements (Xiong and Eickbush 1990; Eickbush et al. 1997). Other retrotransposons related to Osvaldo, together with an unusual intron interrupting the env gene. In view of these observations, and for a further understanding of the relationship of Osvaldo with other retrotransposons and retroviruses, we performed multiple alignments using the RT, a protein that previously showed seven motifs highly conserved between all retroelements (Xiong and Eickbush 1990; Eickbush et al. 1997). Other retrotransposons related to Osvaldo, such as the capability to encode an envelope protein and the structural requirements for a retrovirus, such as the 5' nucleotide of the heptamer sequence and the 3' end of its overlap (as delineated by the first nucleotide of the codon). Sequences are grouped according to their final three nucleotides; these constitute a codon in the upstream (e.g., gag) gene. References: (1) Jacks et al. (1988) and references therein; (2) Inouye, Yuki, and Saigo (1986); (3) Tanda, Mullor, and Corces (1994); (4) Morikawa and Bishop (1992).

**Table 2**

<table>
<thead>
<tr>
<th>Retrovirus or Retrotransposon</th>
<th>Overlap</th>
<th>Sequence</th>
<th>Distance from 3' End of Overlap (bp)</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIV ................................</td>
<td>gag/pol</td>
<td>GGU UUU UUA GGC</td>
<td>213</td>
<td>1</td>
</tr>
<tr>
<td>HIV-1 ............................</td>
<td>gag/pol</td>
<td>AUU UUU UGG GGG</td>
<td>198</td>
<td>1</td>
</tr>
<tr>
<td>gypsy ............................</td>
<td>gag/pol</td>
<td>AUA UUU UGG GGG</td>
<td>51</td>
<td>1</td>
</tr>
<tr>
<td>17.6 ................................</td>
<td>gag/pol</td>
<td>GAA AUA UUU CAG</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>297 ................................</td>
<td>gag/pol</td>
<td>GAA AUA UUU CAG</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Tom ................................</td>
<td>gag/pol</td>
<td>GAA AUA UUU CAG</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>Osvaldo .........................</td>
<td>gag/pol</td>
<td>UCG GGA AAC GAC</td>
<td>60</td>
<td>This work</td>
</tr>
<tr>
<td>FIV ................................</td>
<td>gag/pol</td>
<td>UCG GGA AAC UGG</td>
<td>93</td>
<td>1</td>
</tr>
<tr>
<td>SRV-1 ............................</td>
<td>gag/pro</td>
<td>CAG GGA AAC GAC</td>
<td>147</td>
<td>1</td>
</tr>
<tr>
<td>MPMV .............................</td>
<td>gag/pro</td>
<td>CAG GGA AAC GGG</td>
<td>147</td>
<td>1</td>
</tr>
<tr>
<td>Visma .............................</td>
<td>gag/pol</td>
<td>CAG GGA AAC AAC</td>
<td>45</td>
<td>1</td>
</tr>
</tbody>
</table>

Note.—Common heptanucleotide sequence motifs are present in all retroviral gag/pol or gag/pro overlapping regions known or presumed to contain frameshifting sites. The heptanucleotides are shown in boldface type along with their neighboring sequences and the distance (in nucleotides) between the 5' nucleotide of the heptameric sequence and the 3' end of its overlap (as delineated by the first nucleotide of the codon). Sequences are grouped according to their final three nucleotides; these constitute a codon in the upstream (e.g., gag) gene. References: (1) Jacks et al. (1988) and references therein; (2) Inouye, Yuki, and Saigo (1986); (3) Tanda, Mullor, and Corces (1994); (4) Morikawa and Bishop (1992).

![Fig. 5.—A. Potential hairpin loop of Osvaldo RNA downstream of the putative frameshift site (underlined). A partial amino acid sequence of the ORF1-ORF2 overlap region is also presented. The L-shaped arrow between the Asn and Arg residues shows the presumable ORF1-ORF2 fusion site. B. Putative donor and acceptor sequences found in the boundary regions between ORFs 2 and 3. Vertical arrows in the genomic sequence indicate the proposed splicing sites. Consensus sequences (Mount 1982) of exon-intron borders of eukaryotic genes are also shown above the putative splice sites of the Osvaldo element. Terminal nucleotide sequence from each exon and the resulting amino acid residues are shown in bold letters. The numerals in both A and B show the nucleotide positions.](https://academic.oup.com/mbe/article-abstract/16/7/909/2925465/103)
**FIG. 6.** Alignments of two independent sequences of a 501-bp fragment from the putative intron found in *Osvaldo*. The cDb314 sequence was obtained from an *Osvaldo* defective copy cloned from a heterochromatic position (Labrador and Fontdevila 1994). The lA10(15) sequence corresponds to the intron of the recently transposed *Osvaldo* copy described in this work. Asterisks denote identical nucleotides. Mismatches, insertions, and deletions reflect the divergence between both copies.

**FIG. 8.** Phylogenetic relationship of *Osvaldo*. The retrotransposons and retroviruses represented in the tree are the same as the ones shown in the alignments in figure 4. Numbers in the branching points indicate the number of times out of 1,000 that each branch was supported by the bootstrap. The presence of a third ORF encoding ENV is indicated.

The tree in figure 8 is rooted using *jockey* as out-group. This is in agreement with all strategies used to root retroelement trees, which locate non-LTR retrotransposons as outgroup in relation to the LTR retrotransposon and retrovirus branches (Eickbush 1994). This phylogenetic tree indicates that for most gypsy/Ty3 elements, bootstrapping poorly supports the branching. This observation, already noted by Eickbush (1994), makes it difficult to establish either a reliable relationship between different elements or the exact position of *Osvaldo*. However, our phylogenetic tree is strongly supported at the branching point that separates retroviruses from the gypsy/Ty3 group. The tree basically coincides with the one presented by Xiong and Eickbush (1990), with three main branches well supported (*P* > 0.95): one represented by the Ty1/Copia family, the next represented by retroviruses, and, finally, a monophyletic group composed of all gypsy/Ty3 retrotransposons. This group is clearly divided into two different branches (*P* > 0.95): Mag retrotransposons and a group of elements that contains gypsy and gypsy-like retroviruses, Ty3-like, caulimoviruses, and *Osvaldo*. The overall localization of *Osvaldo* and the main groups described here is stable regardless of the method used (neighbor joining or maximum parsimony) (not shown). We also used the RNAseH and integrase protein sequences with both methods, and the results were, in general, consistent...
with the tree shown here, although the supporting values at the branching points are also very low (data not shown). Although Osvaldo is always more closely related to woot, and sometimes to Ulysses (using different proteins and methods), than to any other element in the tree, its location is not very well supported by the bootstrapping procedure. Osvaldo is, however, unequivocally not closely related to the gypsy group of retroviruses represented by all described retrotransposons containing a third ORF encoding for an envelope protein. None of the multiple trees constructed includes Osvaldo among these elements, otherwise always very well supported in a monophyletic group. Attempts to find any homology between the Osvaldo envelope and the envelope from gypsy and gypsy-related retrotransposons were unsuccessful. Interestingly, some homologies can be found between the envelope proteins from the gypsy family of elements (Leblanc et al. 1997). For that reason, and in view of the particular features described here, it is possible that Osvaldo belongs to a new family of insect retroviruses. However, we cannot discard the possibility that Osvaldo is a distantly related member of the gypsy family of retroviruses.

If we assume that Osvaldo belongs to a new family of retroviruses in insects, independent of gypsy, the origin of retroviruses can be explained considering that the envelope is an ancestral character for both retroviruses and gypsy/Ty3 elements and that it has been lost in many elements in the course of evolution. Those elements would be most of the gypsy/Ty3 retrotransposons and the endogenous retroviruses. This would be in agreement with the broad distribution of these elements in the phylogenetic scale. The alternative explanation would be that retroviruses were originated in vertebrates and subsequently spread by horizontal transmission to all the other phyla, where they frequently lost the envelope gene (see Eickbush 1994). Before the characterization of retroviruses in insects, the first hypothesis was difficult to sustain due to ambiguities in the phylogenetic trees. However, the presence of true retroviruses like gypsy and other related elements in insects, together with the existence of retroviruses like Osvaldo, could indicate that the envelope appeared before the divergence of retroviruses and gypsy/Ty3 retrotransposons. The recent characterization of a putative envelope in plant LTR retrotransposons also distantly related to gypsy (Wright and Voytas 1998) strongly suggests that this could be the case.

Acknowledgments

We thank Anna Pons and Antonio Casamayor for their contribution with the automatic sequencing systems at the Consejo Superior de Investigaciones Cientificas and the UAB Universitat Autonoma de Barcelona, respectively. This work has been supported by Spanish Direcccion General de Investigacion Cientifica y Tecnica (DGICYT) grants PB90/0711, PB93/0843, and PB96-1136 to A.P. and by personal fellowships from the DGICYT to A.P. (Ref: SB93-AOI064823).

LITERATURE CITED


PIERRE CAPY, reviewing editor

Accepted March 22, 1999