MOLECULAR AND CELLULAR BIOLOGY, Apr. 1986, p. 1129–1134
0270-7306/86/041129-06$02.00/0

Vol. 6, No. 4

The *Drosophila melanogaster* Gypsy Transposable Element Encodes Putative Gene Products Homologous to Retroviral Proteins

ROBIN L. MARLOR,† SUSAN M. PARKHURST, AND VICTOR G. CORCES*
Department of Biology, The Johns Hopkins University, Baltimore, Maryland 21218

Received 25 September 1985/Accepted 16 January 1986

We determined the complete nucleotide sequence of the gypsy element present at the forked locus of *Drosophila melanogaster* in the f^1 allele. The gypsy element shares more homology with vertebrate retroviruses than with the copia element of *D. melanogaster* or the Ty element of *Saccharomyces cerevisiae*, both in overall organization and at the DNA sequence level. This transposable element is 7,469 base pairs long and encodes three putative protein products. The long terminal repeats are 482 nucleotides long and contain transcription initiation and termination signals; sequences homologous to the polypurine tract and tRNA primer binding site of retroviruses are located adjacent to the long terminal repeats. The central region of the element contains three different open reading frames. The second one encodes a putative protein which shows extensive amino acid homology to retroviral proteins, including *gag*-specific protease, reverse transcriptase, and DNA endonuclease.

The *Drosophila melanogaster* gypsy transposable element is associated with spontaneous mutations whose phenotype can be reversed by mutations at unlinked suppressor loci (9). This element is transcribed in a temporal specific fashion, giving rise to a major 6.5 kilobase RNA which accumulates at highest levels in 2- to 3-day-old pupae (11). We recently proposed (11, 12) that the mutational activity of the gypsy element on suppressible genes is a direct consequence of the transcriptional properties of this element. The mutagenic effect of the transposable element on these loci is due to transcriptional interference on the genes located nearby.

To understand the molecular basis of this phenomenon, we investigated the DNA structure of the gypsy element. Gypsy is a member of a class of structurally similar transposable elements which contain long terminal repeats (LTRs) (1, 5, 9). Other members of this family are the copia-like elements of *D. melanogaster* (14), the Ty elements of *Saccharomyces cerevisiae* (13), and vertebrate retrovirus proviruses (20). In addition to the conservation in the organization of the different transcription signals between the LTRs of retroviruses and copialike elements, the latter ones also contain nucleotide sequences homologous to the tRNA primer binding site and purine-rich sequences, both necessary for the initiation of DNA synthesis in a retrovirus system (6, 16, 22). The organization of the protein-coding regions of these different elements nevertheless varies. A typical vertebrate retrovirus consists of three genes, termed *gag*, *pol*, and *env*, required for viral infection and replication (see reference 20 for a review). The *gag* region encodes a polyprotein which is cleaved giving rise to several small proteins found in the core of the virus particle (3) and whose exact function is not yet understood. The *pol* gene is expressed as a *gag-pol* polyprotein which is the precursor of the mature form of reverse transcriptase. The N-terminal region of this protein product contains the DNA polymerase and RNase H activities of reverse transcriptase, whereas the DNA endonuclease activity required for provirus integration is located in the carboxy-terminal region (21). In addition, the amino-terminal region of this protein encodes a *gag*-specific protease required for the cleavage of the *gag* polyprotein (3). Finally, the third open reading frame, located at the 3′ end of the retroviral genome, encodes the components of the viral envelope (3, 20).

The copia element of *D. melanogaster* contains only one open reading frame, whereas Ty has two distinct ones, both elements encoding putative protein products homologous to the reverse transcriptase and endonuclease activities of retroviruses (2, 10). Other *Drosophila* copialike elements, such as 17.6, are more similar to vertebrate retroviruses in the organization of their coding potentials, with three different open reading frames, the second one encoding the enzymatic activities (15, 19). Here we present evidence which indicates that the gypsy element contains three different open reading frames and is structurally closer to the *Drosophila* 17.6 element and vertebrate retroviruses than to copia or Ty.

MATERIALS AND METHODS

DNA isolation and sequencing. Purification of plasmid DNA, digestion with restriction enzymes, and labeling of DNA fragments were done by standard procedures (7). The sequence of the gypsy element was determined on both DNA strands as described by Maxam and Gilbert (8). The analysis of the protein sequence homologies among different retroviral elements was performed with the Bionet computer facility

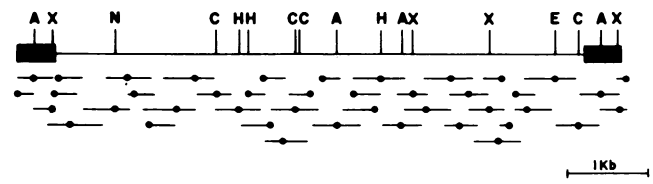

FIG. 1. Restriction map of the gypsy element. The shaded areas represent the LTRs of gypsy, and the central region of the element is indicated by a thin line. The lower part of the figure indicates the strategy used to obtain the DNA sequence of the gypsy element. The symbols used to represent the restriction enzymes are A, *Ava*I; C, *Cla*I; E, *Eco*RI; H, *Hin*dIII; N, *Nco*I; and X, *Xba*I.

* Corresponding author.
† Present address: Department of Biology, Massachusetts Institute of Technology, Cambridge MA 02139.


LTR
a
b
LTR
c

FIG. 2. Nucleotide sequence of gypsy. The gypsy element located at the forked locus in the f^1 allele was cloned as previously reported (11). This complete element was sequenced by the procedure of Maxam and Gilbert (8) and the strategy depicted in Fig. 1. The arrows indicate the starting and ending nucleotides of the LTRs. The different sequences implicated in the transcription and replication of the element are shaded. The CAT box, the Hogness box, the tRNA binding site, the polypurine tract, and the polyadenylation signal have been respectively designated a, b, c, d, and e. Also indicated is the amino acid sequence of the putative gene products encoded by the three different open reading frames located in the central portion of the element.

(IntelliGenetics, Inc., Palo Alto, Calif.) and the SEARCH and ALIGN programs.

RESULTS AND DISCUSSION

General features of the gypsy transposable element. The gypsy element whose sequence is reported in this manuscript was isolated from the forked locus of *D. melanogaster* in the f^1 allele (11). The restriction map of this element and the neighboring sequences of the *f* locus is shown in Fig. 1. The complete sequence of this particular copy of the gypsy element was determined by the sequencing protocol of Maxam and Gilbert (8), following the strategy indicated in the lower part of Fig. 1. This sequence information is shown in Fig. 2. The gypsy element has a total length of 7,469 base pairs (bp) and contains two LTRs 482 bp long. The central part of the element contains three different open reading frames which encode putative proteins whose amino acid sequence is shown in Fig. 2.

Structure of the LTRs. The nucleotide sequence of the LTRs of the gypsy elements responsible for the mutant phenotype in the alleles sc^1, bx^3, and $bx^{34}e$ was determined previously by Freund and Meselson (5). The LTRs of these three different insertions were found to be identical and 482 bp long. In addition, Bayev et al. (1) determined the sequence of the LTRs from two different copies of the gypsy element located at undetermined places in the wild-type Oregon R genome. These LTRs were found to be 479 bp long. In the case of the gypsy element located in the forked locus in f^1, the LTRs are identical and 482 bp long. The sequence we obtained differs from that of Bayev et al. (1) by the addition of three nucleotides at positions 163 (A), 165 (T), and 170 (T), and it differs from that of Freund and Meselson (5) by the deletion of one C at position 67 and the addition of another C at position 102. The observed changes do not affect any of the transcription regulatory sequences of the LTRs. This low divergency rate among different copies of the same transposable element has also been observed in the case of the copia element (10).

The locations on the LTRs of various transcription signals, such as the CAT and TATA boxes and the polyadenylation sequence, are also shown in Fig. 2. Freund and Meselson (5) proposed that the sequence TATATAA, located at the 3′ end of the LTR, could play the role of the Hogness box for initiation of transcription. Because this sequence is located 3′ to the polyadenylation signal, the RNA synthesized would not contain all the information necessary to make a new full copy of the transposable element by reverse transcriptase. In agreement with what is generally the case in retroviruses, in which the TATA box is located 20 to 50 bp upstream of the polyadenylation signal (19), we propose that the sequence AATATT (Fig. 2). serves as the transcription initiation signal for gypsy expression. This sequence is similar to the Hogness box of other retroviruses and retroviruslike elements (20) and is located 30 bp upstream from the polyadenylation signal. In addition, a sequence homologous to the CAT box is present 23 bp usptream of the sequence AATATT in a position similar to that found in other retroviral elements. We do not yet have direct experimental evidence indicating that the sequence AATATT serves as a signal for initiation of transcription, but its location with respect to the different components of the LTR suggests that this might be the case.

Other conserved sequences present in retroviral elements and necessary for the completion of their life cycle are also found in the gypsy element. Immediately 3′ of the left LTR there is a sequence similar to the tRNA binding site of other retroviruses (Fig. 2) which is involved in the synthesis of the first strand of viral DNA (20). This sequence has a strong homology with *Drosophila* tRNAs ($tRNA^{Lys}$). In addition, the region located immediately 5′ to the second LTR contains a 10-bp purine-rich tract similar to the sequence highly conserved among retroviruses which appears to serve as primer for the synthesis of plus-strand DNA (20).

Insertion site of the gypsy element. Freund and Meselson reported previously that the gypsy element shows a sequence specificity at the insertion point in the *Drosophila* genome. In all three mutations sc^1, bx^3, and $bx^{34}e$, the gypsy element inserts at the sequence TACA*TA, in which * denotes the insertion site (5). To determine the recognition sequence for gypsy insertion in the f^1 mutation, we sequenced both the surrounding region of the forked locus, located outside the gypsy element in f^1, and the corresponding wild-type sequences from a Canton S strain (11). From the comparison of the mutant and wild-type sequences (data not shown), we determined that the recognition sequence for gypsy insertion in this particular case is TATA*TA. This is also the insertion site for one of the gypsy elements characterized by Bayev et al. (1). Because the sequence TACATA is present in the forked locus a few nucleotides upstream of the insertion site of the gypsy element (data not shown), it seems that this particular sequence might not be required for the insertion of the transposable element and that the consensus sequence for gypsy insertion is TA(C/T)A*TA. This sequence is similar to that reported by Ikenaga and Saigo for the insertion of the 17.6 element (6). As a consequence of the insertion of the gypsy element, there is a 4-bp duplication at the insertion site.

Gypsy encodes putative gene products homologous to retroviral proteins. The central region of the gypsy element contains three different open reading frames (Fig. 2) organized in a fashion similar to those of the *Drosophila* 17.6 transposable element (19) and vertebrate retroviruses (20). Using the ALIGN program of the Bionet computer facility, we did not detect any significant homology between putative proteins encoded by the first and third open reading frames (designated ORF1 and ORF3, respectively) and gene products encoded by retroviruses or other transposable elements. These two open reading frames presumably encode gene products respectively similar to the *gag* and *env* proteins which are expected to be different among various retroviruses. The second open reading frame, however, encodes a putative protein (ORF2) highly homologous to proteins typically encoded by retroviruses. It is interesting to note that the putative gene products ORF2 and ORF3 are encoded in the same reading frame and separated by only one translation termination codon (Fig. 2). Thus, a 1-bp change in this codon will yield a gypsy element in which the ORF2 and ORF3 gene products are translated as a single polypeptide. To check whether the existence of this termination codon is a general structural characteristic of the gypsy element or a particular property of the gypsy located in the forked locus in the f^1 allele, we have also sequenced the corresponding DNA region of a different gypsy element responsible for the mutant phenotype of y^2 (12). This gypsy element is organized in the same fashion as that in the f^1 mutant, suggesting that gypsy encodes three different putative gene products, the second and third ones being separated only by a protein termination codon.

The amino acid sequence in three different regions of the putative gypsy ORF2-encoded protein is shown in Fig. 3. These regions are respectively homologous to the retroviral *gag*-specific protease (Fig. 3A), the reverse transcriptase

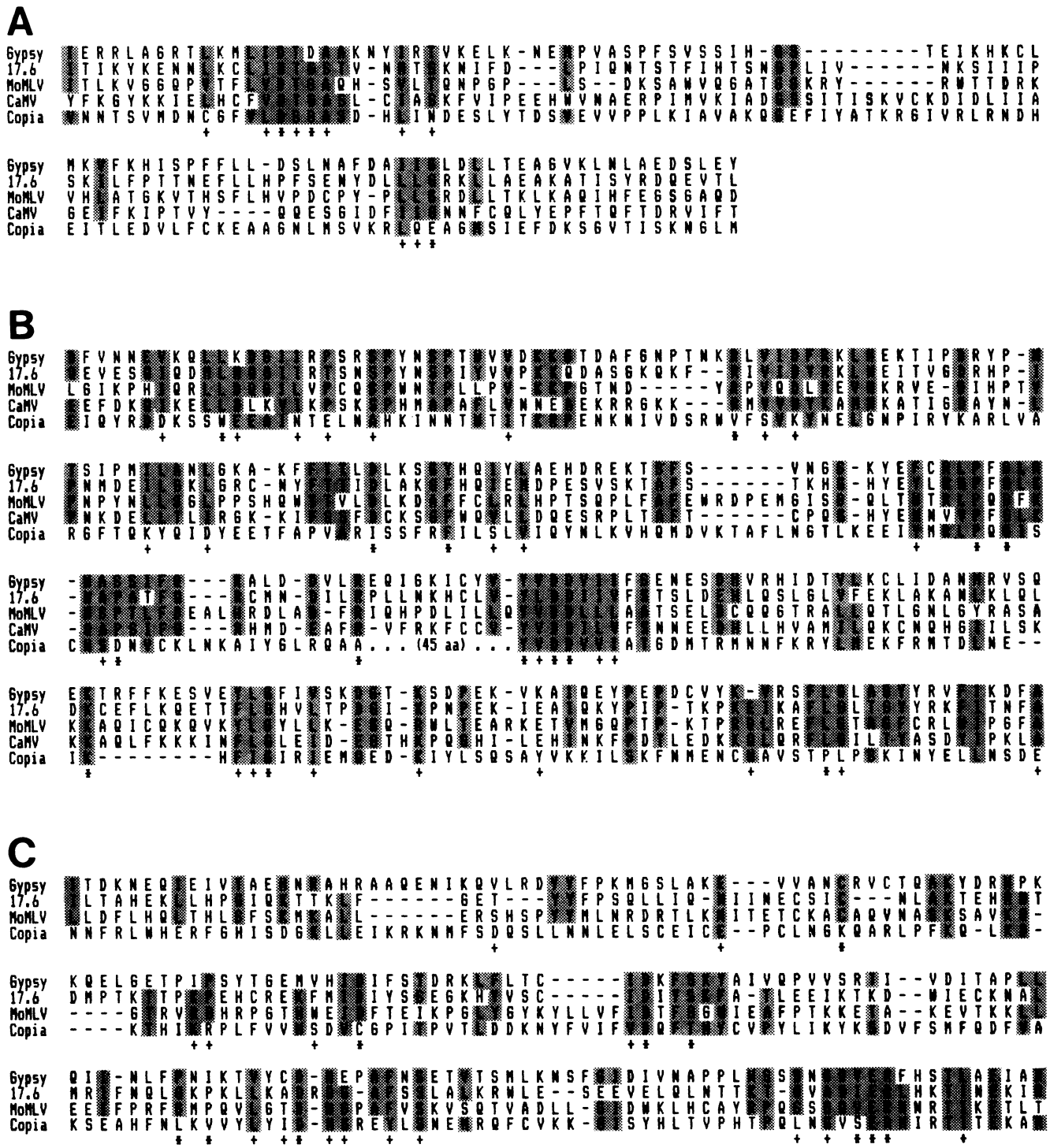

FIG. 3. Alignment of the amino acid sequences of the protease (A), reverse transcriptase (B), and endonuclease (C) activities encoded by the gypsy and other retroviral elements. The positions at which four or more of the sequences share identical or chemically similar amino acids are shaded. The symbol * indicates that the position is occupied by identical residues, and + indicates that similar amino acids are present at that position. Chemically similar amino acids are grouped as follows (17): A, S, T, P, and G; N, D, E, and Q; H, R, and K; M, L, I, and V; F, Y, and W.

(Fig. 3B), and the endonuclease activity required for provirus integration (Fig. 3C). Also shown in Fig. 3 is a comparison of the amino acid sequence of the ORF2 gypsy gene product with homologous proteins from cauliflower mosaic virus (4), Moloney murine leukemia virus (18), and the *Drosophila* 17.6 (15) and copia (10) elements. The homology among these different gene products is concentrated in specific regions of the protein, and it is apparent

that the gypsy-encoded protein contains the same or a conserved amino acid at those locations in the protein where the sequence has been shown to be invariant among many retroviruses and retroviral elements (19). These locations are indicated in the figure by the symbol * if the different proteins contain the same amino acid at that position or + if there is a conservative change. By comparing the various sequences represented in Fig. 3 at those specific locations, we concluded that the gypsy element is evolutionarily closer to the 17.6 element and vertebrate retroviruses than to the *Drosophila* copia element.

ACKNOWLEDGMENTS

This work was supported by Public Health Service award GM32036 from the National Institutes of Health and by Biomedical Research support grant S07 RR07041. S.M.P was funded by National Institutes of Health training grant GM07231.

LITERATURE CITED

1. **Bayev, A. A., N. V. Lyubomirskaya, E. B. Dzhumagaliev, E. V. Ananiev, I. G. Amiantova, and Y. V. Ilyin.** 1984. Structural organization of transposable element mdg4 from *Drosophila melanogaster* and a nucleotide sequence of its long terminal repeats. Nucleic Acids Res. **12**:3707–3723.
2. **Clare, J., and P. Farabaugh.** 1985. Nucleotide sequence of a yeast Ty element: evidence for an unusual mechanism of gene expression. Proc. Natl. Acad. Sci. USA **82**:2829–2833.
3. **Dickson, C., R. Eisenman, H. Fan, E. Hunter, and N. Teich.** 1982. Protein biosynthesis and assembly. Cold Spring Harbor Monogr. Ser. **10C**:513–648.
4. **Franck, A., H. Guilley, G. Jonard, K. Richards, and L. Hirth.** 1980. Nucleotide sequence of cauliflower mosaic virus DNA. Cell **21**:285–294.
5. **Freund, R., and M. Meselson.** 1984. Long terminal repeat nucleotide sequence and specific insertion of the gypsy transposon. Proc. Natl. Acad. Sci. USA **81**:4462–4464.
6. **Ikenaga, H., and K. Saigo.** 1982. Insertion of a mobile genetic element, *297*, into the TATA box from the H3 histone gene in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **79**: 4143–4147.
7. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
8. **Maxam, A. M., and W. Gilbert.** 1980. Sequencing end-labeled DNA with base-specific chemical cleavages. Methods Enzymol. **65**:499–560.
9. **Modolell, J., W. Bender, and M. Meselson.** 1983. *Drosophila melanogaster* mutations suppressible by the *suppressor of Hairy-wing* are insertions of 7.3 kilobase mobile element. Proc. Natl. Acad. Sci. USA **80**:1678–1682.
10. **Mount, S. M., and G. M. Rubin.** 1985. Complete nucleotide sequence of the *Drosophila* transposable element copia: homology between copia and retroviral proteins. Mol. Cell. Biol. **5**:1630–1638.
11. **Parkhurst, S. M., and V. G. Corces.** 1985. *forked*, gypsys and suppressors in *Drosophila*. Cell **41**:429–437.
12. **Parkhurst, S. M., and V. G. Corces.** 1986. Interactions among the gypsy transposable element and the yellow and the suppressor of Hairy-wing loci in *Drosophila melanogaster*. Mol. Cell. Biol. **6**:47–53.
13. **Roeder, G. S., and G. R. Fink.** 1983. Transposable elements in yeast, p. 299–328. *In* J. A. Shapiro (ed.), Mobile genetic elements. Academic Press, Inc., Orlando, Fla.
14. **Rubin, G. M.** 1983. Dispersed repetitive DNAs in *Drosophila*, p. 329–361. *In* J. A. Shapiro (ed.), Mobile genetic elements. Academic Press, Inc., Orlando, Fla.
15. **Saigo, K., W. Kugimiya, Y. Matsuo, S. Inouye, K. Yoshioka, and S. Yuki.** 1984. Identification of the coding sequence for a reverse transcriptase-like enzyme in a transposable genetic element in *Drosophila melanogaster*. Nature (London) **312**:659–661.
16. **Scherer, G., C. Tschudi, J. Perera, H. Delius, and V. Pirrota.** 1982. *B104*, a new dispersed repeated gene family in *Drosophila melanogaster* and its analogies with retroviruses. J. Mol. Biol. **157**:435–451.
17. **Schwartz, R. M., and M. O. Dayhoff.** 1978. Matrices for detecting distant relationships, p. 353–358. *In* M. O. Dayhoff (ed.), Atlas of protein sequence and structure, vol. 5. National Biomedical Research Foundation, Washington, D.C.
18. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukemia virus. Nature (London) **293**:543–548.
19. **Toh, H., R. Hikuno, H. Hayashida, T. Miyata, W. Kugimiya, S. Inouye, S. Yuki, and K. Saigo.** 1985. Close structural resemblance between putative polymerase of a *Drosophila* transposable genetic element 17.6 and *pol* gene product of Moloney murine leukemia virus. EMBO J. **4**:1267–1272.
20. **Varmus, H. E.** 1983. Retroviruses, p. 411–503. *In* J. A. Shapiro (ed.), Mobile genetic elements. Academic Press, Inc., Orlando, Fla.
21. **Varmus, H. E., and R. Swanstrom.** 1982. Replication of retroviruses. Cold Spring Harbor Monogr. Ser. **10C**:369–512.
22. **Will, B. M., A. A. Bayev, and D. J. Finnegan.** 1981. Nucleotide sequence of terminal repeats of *412* transposable element of *Drosophila melanogaster*. A similarity to proviral long terminal repeats and its implication for the mechanism of transposition. J. Mol. Biol. **153**:897–915.