Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 4462–4464, July 1984
Genetics

Long terminal repeat nucleotide sequence and specific insertion of the gypsy transposon

(*Drosophila*/genetic suppression/bithorax/scute)

ROBERT FREUND* AND MATTHEW MESELSON†

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, MA 02138

Contributed by Matthew Meselson, March 16, 1984

ABSTRACT **We have determined the nucleotide sequences of the long terminal repeats of the transposable element gypsy from the cloned mutant alleles sc^1, bx^3, and bx^{34e}. These mutations are suppressible by the suppressor of Hairy-wing, *su(Hw)*. The long terminal repeats are 482 base pairs long and are highly conserved. In each case, gypsy is inserted into the sequence T-A-C-A-T-A and generates a duplication of the sequence T-A-C-A. This was verified by sequencing an empty site in the wild-type *bx* gene. Consideration of the sequence of the long terminal repeats and their surroundings limits the possible explanations for the mechanism of mutation by these gypsy insertions and for their suppression by *su(Hw)*.**

Certain spontaneous mutations of prokaryotes and eukaryotes result from the insertion of transposable genetic elements (for a review, see ref. 1). Some of these insertion mutations are suppressible—that is, expression of the locus at which the transposable element is inserted is partly or fully restored to wild type by mutation of a gene located elsewhere in the genome, termed a suppressor gene. Most of the known suppressible mutations of *Drosophila melanogaster* result from insertions of a 7.3-kilobase transposon with 0.5-kilobase direct terminal repeats called gypsy (2). These mutations are suppressed in flies that lack the function of the gene *su(Hw)*, suppressor of Hairy-wing. It is not known how gypsy insertion disrupts gene function or how the suppressor gene exerts its controlling effect. We have determined the sequence of the long terminal repeats (LTRs) and adjacent DNA of three gypsy insertions, sc^1, bx^3, and bx^{34e}, and also the sequence of wild-type DNA in the region of the bx^{34e} insertion. Gypsy is found to have highly conserved LTRs and to be inserted in each case at the specific sequence T-A-C-A-T-A, generating a duplication of T-A-C-A. The LTR contains multiple nonsense codons in every possible reading frame. Since an LTR remains behind in wild-type reversions of suppressible gypsy mutations (2, 3), we conclude that these gypsy insertions do not lie in coding regions. Also, the LTR does not appear to have functional donor or acceptor splice sites at its ends, suggesting that suppression does not result from splicing gypsy out of transcripts containing it.

RESULTS AND DISCUSSION

Fig. 1 presents the nucleotide sequences of the LTRs of the gypsy elements associated with the mutations sc^1, bx^3, and bx^{34e}. The LTRs are each 482 base pairs and the two LTRs of each insertion are identical. The LTRs of the three gypsys differ at only three positions: 151, 254, and 481. We find no extensive homology to other LTRs, including those of the *Drosophila* transposable elements, 412, 297, B104, MDG1, and MDG3, the yeast transposon TY*1*, or the retrovirus Moloney murine leukemia virus (5–12). Several features distinguish the structure of the gypsy LTR from that of most other *Drosophila* transposons and from the LTRs of retroviruses (13). First, the terminal dinucleotides T-G . . . . C-A found at the ends of most eukaryotic LTRs, including those of *Drosophila* transposons and integrated retroviruses, are not present. Second, the reverse repeats at the ends of each LTR are particularly short. Third, the arrangement of potential promoter and 3′ cleavage sites characteristic of retrovirus LTRs is not present on either strand. In the orientation shown in Fig. 1, T-A-T-A-T-A-A occurs at position 474 and A-A-T-A-A-A lies at position 257. If these sequences function in gypsy as promoter elements and 3′ cleavage sites, respectively, the resulting transcript would lack part of the LTR sequence and therefore could not serve as an RNA intermediate for replication of gypsy by reverse transcriptase.

Fig. 2 shows sequences flanking the LTRs of the three gypsy insertions. The first 20 nucleotides at each end of gypsy adjacent to the LTRs are fully conserved in all three cases and are ≈ 50% G-C. The host DNA (lowercase letters) flanking each gypsy insertion is rich in A-T. In each case, the gypsy element is inserted into host DNA, in one orientation or the other, at the sequence *t-a-c-a-t-a*. Insertion appears to cause a repeat of the tetranucleotide *t-a-c-a*. This is in accord with the sequence of the empty site of the bx^{34e} insertion, shown in Fig. 2. In contrast to this high specificity of gypsy insertion, other transposons and retroviruses have not been found to integrate at precisely specific sequences, with the possible exception of the *Drosophila* element 297. This element is reported to insert preferentially at the sequence A-T-A-T, or possibly T-A-T-A-T-A. Element 297 also lacks the terminal dinucleotides T-G . . . . . C-A (7, 14). Some transposable elements, including the bacterial transposon Tn*10* (15) and the *Drosophila* P-element (16), insert at sites bearing enough homology to define an approximate consensus sequence. Others, such as *copia* (17) and retroviruses (18), integrate with no apparent specificity.

The sequence *a-t-g-t-a*-A(A/G)T, resembling the consensus donor sequence A-G|G-T(A/G)A-G-T for RNA splicing (19), overlaps the host DNA–LTR junction (lower strand, right-hand junction in Fig. 2). No sequence resembling the consensus acceptor site $(T/C)_nN(C/T)A\text{-}G|G$ is seen, however, anywhere near the other end of the gypsy element either in the LTR or in the adjacent host DNA. Moreover, the gypsy elements of bx^3 and bx^{34e} are oriented in parallel (2, 20) but are inserted in *t-a-c-a-t-a* in opposite orientations so that the potential donor site we find, being on opposite strands in these two insertions, could not be functional in both. It is therefore unlikely that suppression occurs by splicing out gypsy sequences from longer transcripts.

The largest open reading frames surrounding the bx^3 and bx^{34e} insertions are only 90 and 144 nucleotides long, respectively, or 69 and 93, considering the orientation reported for

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: LTR, long terminal repeat.

*Present address: Department of Pathology, Harvard Medical School, 25 Shattuck Street, Boston, MA 02115.
†To whom reprint requests should be addressed.

```
          10         20         30         40         50         60
          *          *          *          *          *          *
AGTTAACAACTAACAATGTATTGCTTCGTAGCAACTAAGTAGCTTTGTATGAACAATGCT

          70         80         90        100        110        120
          *          *          *          *          *          *
GACGCGCCCAGAATTGGGTTCAACGCTCCACGCGAAGAATGCTGGCAGCGGAAAGCTGAC

         130        140        150        160        170        180
          *          *          *          *          *          *
ACTTCCTACCGGGAGTGTTGCTTCACGCTGTAAGAAATGCTGAGTCGGCTTGCCGACTTG
                                 C
                                 C
         190        200        210        220        230        240
          *          *          *          *          *          *
TGGCGGCGCGATGCATTGCTCGAGGGTAAACTTAGTTTTCAATATTGTCTTCTACTCAGT

         250        260        270        280        290        300
          *          *          *          *          *          *
TCAAATCTTGTGTTGAAATAAACCACAGCTTGCTCCGGCTCATTGCCGTTAAACATCATT
               C
         310        320        330        340        350        360
          *          *          *          *          *          *
GTTCTTATTTACAATCAAATCGCTATCGCCACAAGGCTAGTGATAATAACTAAGGGGGCG

         370        380        390        400        410        420
          *          *          *          *          *          *
AAGTCAAGCCCTCCAACCTAATCTCCATAAACAGTGTCTAAGACGAACCTCAGCGAAAGA

         430        440        450        460        470        480
          *          *          *          *          *          *
AGGAAGATCTCTAGACCTACTGGAAATAACATAACTCTGGACCTATTGGAACTTATATAATT
                                                                 C
```

FIG. 1. Nucleotide sequences of the gypsy LTRs of bx^{34e}, bx^3, and sc^1. The complete sequence of the bx^{34e} LTR is shown, with differences in the bx^3 and sc^1 LTRs, printed one or two lines below, respectively. Sequencing was done by the method of Maxam and Gilbert (4).

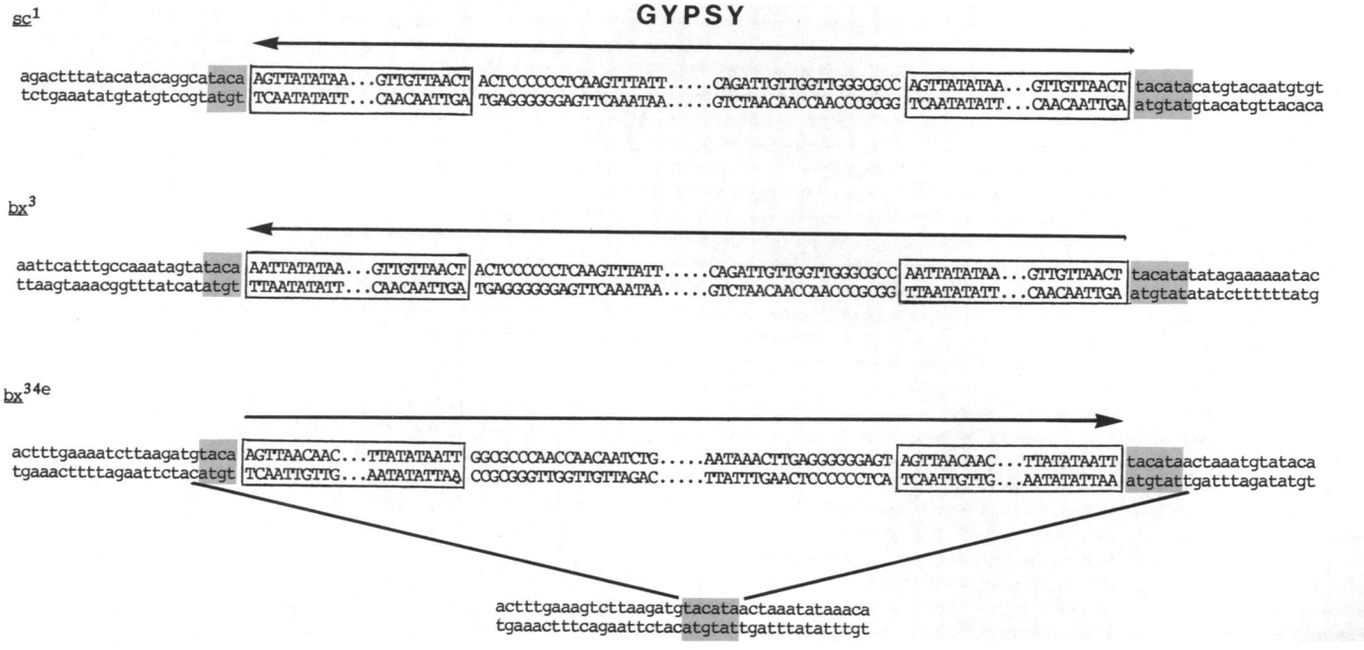

FIG. 2. Nucleotide sequences flanking the LTRs of gypsy elements at sc^1, bx^3, and bx^{34e} and the empty site of the bx^{34e} insertion. LTR sequences are enclosed in boxes. Host DNA is shown in lowercase letters. The recognition site T-A-C-A-T-A and the duplication of T-A-C-A created by gypsy insertion are shaded and are presented in the same orientation in each case. Arrows delimit the gypsy elements and show their orientation relative to the LTR sequence of Fig. 1. The sequence of the empty bx^{34e} insertion site was determined from a clone of the Canton S strain.

the *bx* transcript (20). Although a gypsy LTR remains in wild-type revertants (2, 3), we find that the LTR contains multiple stop codons in all six possible reading frames. We conclude that such gypsy insertions do not lie in exons and that their mutagenic effect does not result from the interruption of sequences coding for protein.

This, however, leaves open the question of how gypsy insertions disrupt gene function. One possibility is that such disruption results from some local effect, such as the provision of sequences that, when the $su^+(Hw)$ gene product is present, terminate or otherwise interfere with transcription or affect RNA processing. Another possibility is that the interaction of the $su^+(Hw)$ product with the gypsy element disrupts gene expression by altering chromatin structure over a considerable distance, for example by altering DNA superhelicity throughout a chromosomal domain.

We thank W. Bender, M. Peifer, and J. Modolell for some of the clones used in this study and J. Pustell for providing computer programing for DNA sequence analysis. We are grateful to the National Institutes of Health for support.

1. Shapiro, J., ed. (1983) *Mobile Genetic Elements* (Academic, New York).
2. Modolell, J., Bender, W. & Meselson, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 1678–1682.
3. Bender, W., Akam, M., Karch, F., Beachy, P. A., Peifer, M., Spierer, P., Lewis, E. B. & Hogness, D. S. (1983) *Science* **221,** 23–29.
4. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
5. Levis, R., Dunsmuir, P. & Rubin, G. M. (1980) *Cell* **21,** 581–588.
6. Will, B. M., Bayev, A. A. & Finnegan, D. J. (1981) *J. Mol. Biol.* **153,** 897–915.
7. Ikenaga, H. & Saigo, K. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4143–4147.
8. Scherer, G., Tschudi, C., Perera, J., Delius, H. & Pirrotta, V. (1982) *J. Mol. Biol.* **157,** 435–451.
9. Kulguskin, V. V., Ilyin, Y. V. & Georgiev, G. P. (1981) *Nucleic Acids Res.* **9,** 3451–3464.
10. Bayev, A. A., Jr., Krayev, A. S., Lyubomirskaya, N. V., Ilyin, Y. V., Skryabin, K. G. & Georgiev, G. P. (1980) *Nucleic Acids Res.* **8,** 3263–3273.
11. Roeder, G. S., Farabaugh, P. J., Chaleff, D. T. & Fink, G. R. (1980) *Science* **209,** 1375–1380.
12. Van Beveren, C., Goddard, J. G., Berns, A. & Verma, I. M. (1980) *Proc. Natl. Acad. Sci. USA* **77,** 3307–3311.
13. Temin, H. M. (1981) *Cell* **27,** 1–3.
14. Rubin, G. (1983) in *Mobile Genetic Elements*, ed. Shapiro, J. A. (Academic, New York), pp. 329–361.
15. Halling, S. M. & Kleckner, N. (1982) *Cell* **28,** 155–163.
16. O'Hare, K. & Rubin, G. M. (1983) *Cell* **34,** 25–35.
17. Dunsmuir, P., Brorein, W. J., Jr., Simon, M. A. & Rubin, G. M. (1980) *Cell* **21,** 575–579.
18. Varmus, H. E. (1982) *Science* **216,** 812–820.
19. Mount, S. M. (1982) *Nucleic Acids Res.* **10,** 459–472.
20. Akam, M. (1983) *Trends Biochem. Sci.* **89,** 173–177.