MOLECULAR AND CELLULAR BIOLOGY, Jan. 1988, p. 114–123 Vol. 8, No. 1
0270-7306/88/010114-10$02.00/0

The Site-Specific Ribosomal DNA Insertion Element R1Bm Belongs to a Class of Non-Long-Terminal-Repeat Retrotransposons

YUE XIONG AND THOMAS H. EICKBUSH*

Department of Biology, River Campus, University of Rochester, Rochester, New York 14627

Received 9 June 1987/Accepted 5 October 1987

Two types of insertion elements, R1 and R2 (previously called type I and type II), are known to interrupt the 28S ribosomal genes of several insect species. In the silkmoth, *Bombyx mori*, each element occupies approximately 10% of the estimated 240 ribosomal DNA units, while at most only a few copies are located outside the ribosomal DNA units. We present here the complete nucleotide sequence of an R1 insertion from *B. mori* (R1Bm). This 5.1-kilobase element contains two overlapping open reading frames (ORFs) which together occupy 88% of its length. ORF1 is 461 amino acids in length and exhibits characteristics of retroviral *gag* genes. ORF2 is 1,051 amino acids in length and contains homology to reverse transcriptase-like enzymes. The analysis of 3′ and 5′ ends of independent isolates from the ribosomal locus supports the suggestion that R1 is still functioning as a transposable element. The precise location of the element within the genome implies that its transposition must occur with remarkable insertion sequence specificity. Comparison of the deduced amino acid sequences from six retrotransposons, R1 and R2 of *B. mori*, I factor and F element of *Drosophila melanogaster*, L1 of *Mus domesticus*, and Ingi of *Trypanosoma brucei*, reveals a relatively high level of sequence homology in the reverse transcriptase region. Like R1, these elements lack long terminal repeats. We have therefore named this class of related elements the non-long-terminal-repeat (non-LTR) retrotransposons.

Transposable elements are discrete DNA segments, found in both procaryotes and eucaryotes, which are capable of movement to many locations throughout the genome (53). In eucaryotes, most of these elements have been described as retrotransposable elements or retrotransposons because, like retroviruses, they appear to propagate via an RNA intermediate (18). This classification was originally based on the amino acid homology of their deduced open reading frames (ORFs) to the reverse transcriptases of retroviruses (8, 39, 48). Subsequently, in the *Saccharomyces cerevisiae* retrotransposon Ty, mobility via an RNA intermediate was directly established (6). A second feature of retroviruses that is shared by a majority of retrotransposons is the presence of long terminal repeats (LTRs). In retroviruses, these LTRs have been shown to function in the initiation of transcription, polyadenylation, the initiation of DNA replication, regeneration of a full-length virus, and integration of the virus into the host genome (reviewed in references 40, 58, and 61).

A fraction of the 28S ribosomal genes in several insect species are interrupted by specific non-ribosomal DNA (non-rDNA) insertions (12, 16, 21, 43, 45, 57). Two types of these elements have been identified by their insertion site in the 28S gene (Fig. 1A). Type I elements interrupt the 28S genes at a location approximately two-thirds of the distance from the 5′ end of the gene and are flanked by a 14-base-pair (bp) duplication of rDNA sequences. Type II elements interrupt the 28S gene 75 bp upstream of the type I element and are not flanked by duplications of rDNA sequences. In *Drosophila melanogaster*, interrupted rDNA units are transcribed at a significantly lower level than noninterrupted rDNA units (27, 28, 33). To be consistent with the single-letter nomenclature currently used to describe many repetitive and transposable elements, we refer to these elements as the R1 (type I) and R2 (type II) elements (R referring to the rDNA units) (7).

Recently we determined the complete nucleotide sequence of the 4.2-kilobase (kb) R2 element of *Bombyx mori* (R2Bm). This analysis revealed the presence of a 1,151-amino-acid ORF. Because the central region of this ORF contains homology to reverse transcriptase-like enzymes, we suggested that the origin of R2 was that of a retrotransposon (7). We further suggested that if this element retains its ability to transpose, this could explain why insects are unable to rid their rDNA loci of these elements despite the ability of rDNA units to undergo sequence turnover by unequal crossovers.

In this paper we report the nucleotide sequence of a complete 5.1-kb R1 insertion element from *B. mori* (R1Bm) which suggests that the origin of this element was also that of a retrovirus or retrotransposon. Junction sequences from independently isolated R1 copies within the ribosomal locus provide evidence that R1 remains a retrotransposable element. Comparison of reverse transcriptase regions reveals that R1 and R2 belong to a distinct group of retrotransposons which lack LTRs.

MATERIALS AND METHODS

Genomic library screening and DNA hybridization. Clone B78 was isolated from a Charon 4 lambda library (15) and has been previously described (16). Other genomic lambda clones described in this paper were isolated from a Charon 35 library (32) made from partially *Sau*3AI-digested *B. mori* genomic DNA (a gift of B. Hibner). Clones containing R1 elements were isolated from this library by hybridization to a 2.5-kb *Kpn*I fragment isolated from the R1 element of clone B78. DNA hybridizations were carried out by the method of Eickbush and Kafatos (15) at 68°C in 0.6 M saline (0.6 M NaCl, 0.12 M Tris hydrochloride [pH 8], 4 mM EDTA).

Subcloning and sequencing. The region of clone B78 extending from the *Bam*HI restriction site 380 bp 5′ of the R1 insertion site to the *Xba*I site approximately 1.6 kb 3′ of the R1 insertion site was subcloned into pUC18 on four overlapping segments (Fig. 1B). From a detailed restriction map

* Corresponding author.

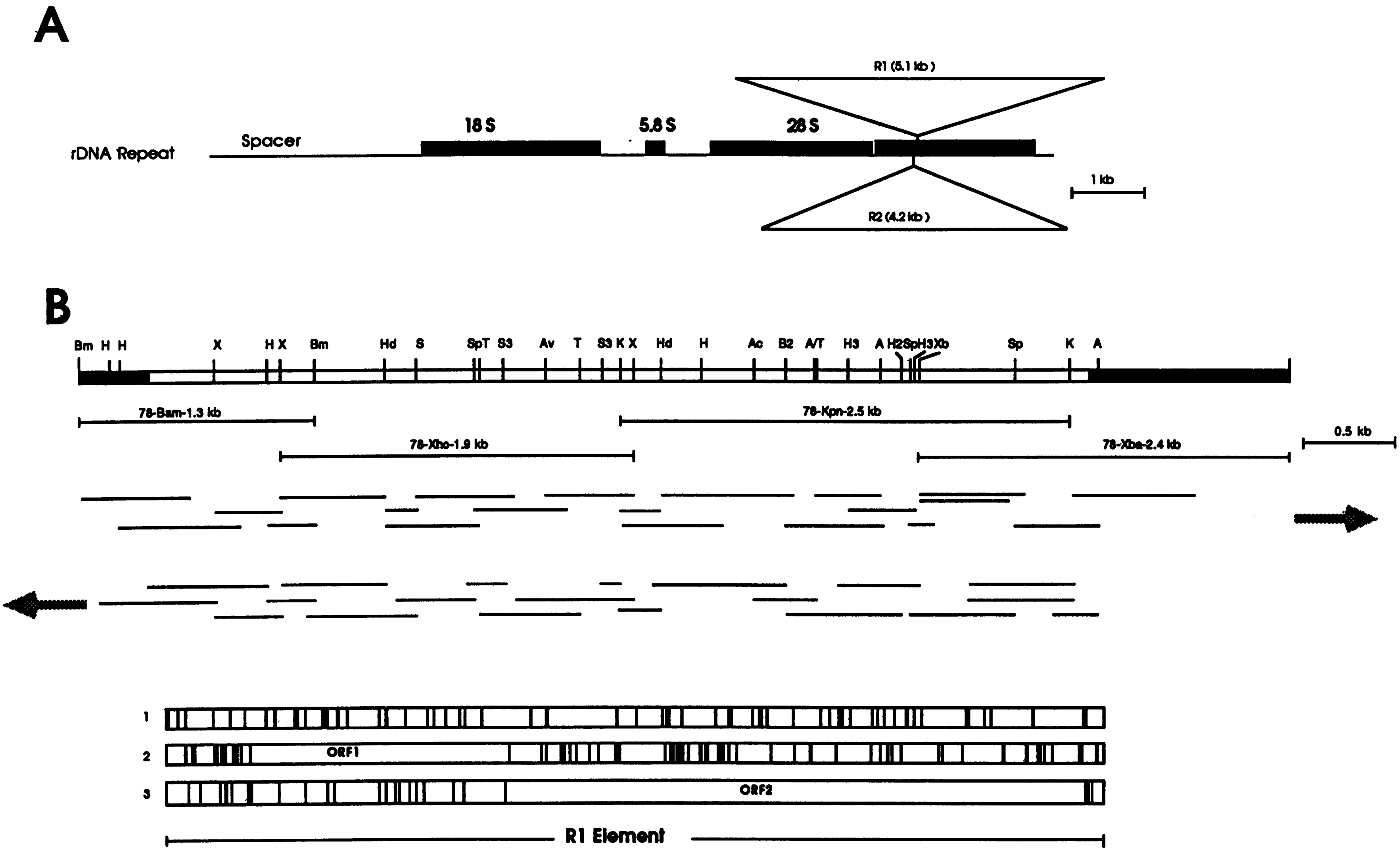

FIG. 1. Location and restriction map of the R1 element within the rDNA units of *B. mori*. (A) Diagram of an rDNA unit showing the location of the R1 and R2 insertions. Solid bars, rRNA genes; thin lines, transcribed and nontranscribed spacer regions. The R1 and R2 insertion sites within the 28S gene are 74 bp apart. A 30-bp hidden break is found near the center of the 28S gene (20). (B) Restriction map, subclones, sequencing strategy, and termination codon map of R1 element. Solid bar, 28S coding region; open bar, R1 insertion. Abbreviations for restriction sites used in the nucleotide sequence determination: Bm, *Bam*HI; H, *Hin*cII; X, *Xho*I; Hd, *Hin*dIII; S, *Sst*I; Sp, *Sph*I; Av, *Ava*I; K, *Kpn*I; B2, *Bgl*II; S3, *Sau*3AI; A, *Alu*I; Xb, *Xba*I; T, *Taq*I; H2, *Hpa*II; H3, *Hae*III. Shown below the restriction map are the fragments subcloned into pUC plasmids and the direction and extent of the sequence determination. The bottom of the figure indicates the location of termination codons in the three reading frames of R1. The large ORFs in frames 2 and 3 overlap by 19 nucleotides.

```
                                              1
rDNA  GCGCATGAATGGATTAACGAGATTCCCACTGTCCCTATCTACT
B78   GCGCATGAATGGATTAACGAGATTCCCACTGTCCCTATCTACTTGACTTCGCCGTCGGCCTTGGTCGAGGACAGACGTGCGTTCCGTTATTTCTTTATTTTCCGTCATTTAAGTGTATTGTGTTTCTATTGGTGTATCGGACC  100

CTCTCGTTTCGGCTTGAGGTTTAAGTCATAAGACGCCGCGGCCATCTTGCTGTGTGAGCGGTGTGACGAGTGCGAAGGCGGAGTTTAGCTCGACGTGGAGTCGGCCCCTCTCGCTTCCTCTTGGGTGCCGGTCCATATAGGTCGGTGTCC  250

ATATTGGATTGCGTGTGAGACGGCCGATTTGCGTGAGGGCGGACCCGCCATTTAGGTCTGTGACAGTGACACTAGTGTGCGATAGTGACGTTTTATAATTTGCTGTGGGCGGTAGCCGCCATTTTGTGATAGTGACACTCGAGGATGCGA  400

CAGTGACGTTGGTTTGTGTTTGTGTTCGTGTGCGTGTGTGAATATCTTTGCGTGATAGTAATATAATTTGGAACAATGTCGGAGGAGGAGAGGGAGCTATTTTCCCCTCGGGTCTCTTTGGCGCGCTCGCCGCCTAGGGGACCGACTGCA  550
                                                          ORF1  F  G  T  M  S  E  E  E  R  E  L  F  S  P  R  V  S  L  A  R  S  P  P  R  G  P  T  A   28

CCTCTGCCGGCGCCTATGCCGCCGCCGGCCCCCCGGGGGGTTAGGGCGGGAGCGGCTATGAGTGCCAAAGGCCGGAGGACCGGCCTGCATATCCCGGTTGCTTCTGGACTGTCGTCCTCGGCGCCTGTCACCCCGGTTGACCCAGTGGCT  700
 P  L  P  A  P  M  P  P  P  A  P  R  G  V  R  A  G  A  A  M  S  A  K  G  R  R  T  G  L  H  I  P  V  A  S  G  L  S  S  S  A  P  V  T  P  V  D  P  V  A   78

GGTGTCCCTAGCTTCCCCATTCCTGTTGCGTCGGGGGCTCCGACTGGACCCCTTGATGTAGCAGCGCAGGGGAGGCTGGAGCTGCTCGAGCGGGCAAACCGAGCGGTGCGCGGGATCATGTCGGTGGCGACGGCGGCATCTAAGCTCAAT  850
 G  V  P  S  F  P  I  P  V  A  S  G  A  P  T  G  P  L  D  V  A  A  Q  G  R  L  E  L  L  E  R  A  N  R  A  V  R  G  I  M  S  V  A  T  A  A  S  K  L  N  128

AAGTCGGAGGTCAATCTAATCTCTGAACTCGGTCGCGATATCCTGGCCGTAGTTGGGGCCCTCGGGATCCAGCTCTCTGATAAGGAGCTGGTGGTCGAAAGGCTCCGCTCGGCAGAGGCCTCTGCACGGCGTGACTCCGTGGCGCCTTCG 1000
 K  S  E  V  N  L  I  S  E  L  G  R  D  I  L  A  V  V  G  A  L  G  I  Q  L  S  D  K  E  L  V  V  E  R  L  R  S  A  E  A  S  A  R  R  D  S  V  A  P  S  178

ATGGCTGCTGGTGCGGCGGCCGGGTCTGGGACTGGGCCTGCGACCTTCGCCACTGTCCTCAGGACGGGCCCTGGTGGGGTCCCGAGGTCCATTGGGGCCTCTCAGGGGCCTTCCTTGGCATTCTACCCGTCCGAGGGCAATGCGGAACTG 1150
 M  A  A  G  A  A  A  G  S  G  T  G  P  A  T  F  A  T  V  L  R  T  G  P  G  G  V  P  R  S  I  G  A  S  Q  G  P  S  L  A  F  Y  P  S  E  G  N  A  E  L  228

AAGACAGCTGAGGACACAAAGAAAGAGGTAAAAAAGGCCATTGACCCTAAGTCAATGGCTATTGGAATACAGAGCGTAAGGAAGGTTGGGAATGCGGGGGTAGTGGTGCAGACCACGTCCCCCGGGGCCGCAGTCAAGCTTAGAAATGCT 1300
 K  T  A  E  D  T  K  K  E  V  K  K  A  I  D  P  K  S  M  A  I  G  I  Q  S  V  R  K  V  G  N  A  G  V  V  V  Q  T  T  S  P  G  A  A  V  K  L  R  N  A  278

GCCCCGCCATCACTGAGAGTCACCGAACCCAGGCGCCGACAGCCCTTAGTAGCTGTCAATGGCGTGGAGGGCGACCCCTCCTTCGAGGAGGTAATCGAGTGCCTGGCTAGCCAGAACCTCGACCCGGAGGAGTGGCCCCTCACTAGGGTA 1450
 A  P  P  S  L  R  V  T  E  P  R  R  R  Q  P  L  V  A  V  N  G  V  E  G  D  P  S  F  E  E  V  I  E  C  L  A  S  Q  N  L  D  P  E  E  W  P  L  T  R  V  328

CGAGCGGAGCTCACGGGGGCGTTCAAAAAGGGGAGGCGACAATCCAACAACACTACTGTGGTGTTTAACGCCTCCCCTCGCATCAGGGACGCCCTCGTGAAGATTGGCAGAGTGTATGTGGGGTGGGTCGCCTGTGAGGTCACGGACTTT 1600
 R  A  E  L  T  G  A  F  K  K  G  R  R  Q  S  N  N  T  T  V  V  F  N  A  S  P  R  I  R  D  A  L  V  K  I  G  R  V  Y  V  G  W  V  A  C  E  V  T  D  F  378

GTCCGGGTGACCTGCTGCAATAAGTGTCAGCAATATGGTCACCCGGAGAAATTTTGCCGGGCCAAGGAGGCCACCTGTGGCCGATGTGGAGAGGATGGCCACCGAATGGAGGCCTGTAAGGCAGCCTCCGCGTGCTGTGCGACCTGCCGG 1750
 V  R  V  T  C  C  N  K  C  Q  Q  Y  G  H  P  E  K  F  C  R  A  K  E  A  T  C  G  R  C  G  E  D  G  H  R  M  E  A  C  K  A  A  S  A  C  C  A  T  C  R  428

CGATTCCGTCGCGAGGCTATGCACCCGACGGCCTCGCGCGACTGTCCGGCGCGCCGGCATGCGGAGGAGCGCTTCCTAAATCAGGTCGAGTATGGATATTAGGCCCCGACTTCGTATTGGCCAAATCAATCTGGGTGGTGCAGAGGATGC 1900
 R  F  R  R  E  A  M, H  P  T  A  S  R  D  C  P  A  R  R  H  A  E  E  R  F  L  N  Q  V  E  Y  G  Y---  461
                                                                                ORF2  I  R  S  S  M  D  I  R  P  R  L  R  I  G  Q  I  N  L  G  G  A  E  D  A   24

GACGAGGGAGCTACCCTCCATTGCACGGGATCTCGGCCTGGATATTGTTCTTGTACAGGAACAATATTCCATGGTCGGGTTCCTAGCCCAATGTGGAGCACACCCCAAGGCGGGTGTGTATATCCGCAATAGGGTGCTCCCCTGCGCGGT 2050
 T  R  E  L  P  S  I  A  R  D  L  G  L  D  I  V  L  V  Q  E  Q  Y  S  M  V  G  F  L  A  Q  C  G  A  H  P  K  A  G  V  Y  I  R  N  R  V  L  P  C  A  V   74

TCTGCACCACCTTAGCAGCACACATATAACGGTAGTGCACATTGGGGGGTGGGACTTATATATGGTGTCTGCGTACTTCCAGTATAGTGACCCTATTGACCCATACCTGCACCGGCTCGGGAATATTCTTGACCGGCTGCGGGGGGCTCG 2200
 L  H  H  L  S  S  T  H  I  T  V  V  H  I  G  G  W  D  L  Y  M  V  S  A  Y  F  Q  Y  S  D  P  I  D  P  Y  L  H  R  L  G  N  I  L  D  R  L  R  G  A  R  124

GGTCGTTATCTGCGCAGACACTAATGCCCACTCGCCATTGTGGCACTCGCTGCCCAGGCACTACGTCGGTCGGGGTCAGGAAGTGGCTGACCGCCGCGCCAAGATGGAGGATTTCATTGGGGCGAGGCGGTTGGTCGTCCATAACGCGGA 2350
 V  V  I  C  A  D  T  N  A  H  S  P  L  W  H  S  L  P  R  H  Y  V  G  R  G  Q  E  V  A  D  R  R  A  K  M  E  D  F  I  G  A  R  R  L  V  V  H  N  A  D  174

TGGCCACCTGCCGACCTTCAGTACGGCGAACGGAGAATCTTATGTCGATGTCACGCTGTCTACGCGGGGAGTACGCGTGTCTGAATGGCGTGTAACTAATGAATCATCGAGCGATCACCGGCTCATTGTGTTTGGGGTGGGGGGCGGTAC 2500
 G  H  L  P  T  F  S  T  A  N  G  E  S  Y  V  D  V  T  L  S  T  R  G  V  R  V  S  E  W  R  V  T  N  E  S  S  S  D  H  R  L  I  V  F  G  V  G  G  G  T  224

AACAGGGGAGCGGGACGAGGACGAGGAGGCGCGGAGCGATTTGAGGCCGGGCGAGCCGTGCGCACTGCGTCGGTACCGGGACCGTGGGGTGGATTGGGACCTCTTCAGATCGCGTATCCACGAGCGAATGGGGAGTCTGGACCTCGAGGA 2650
 T  G  E  R  D  E  D  E  E  A  R  S  D  L  R  P  G  E  P  C  A  L  R  R  Y  R  D  R  G  V  D  W  D  L  F  R  S  R  I  H  E  R  M  G  S  L  D  L  E  E  274

ACCTGTGGCTGCCCTTTGCGAAAAATTTACCGGGGTTATAACTCGCACAGCTGAAGAATGCTTAGGATCACTGAAAGCAGATAGAACTGACAGGGGTTATGAGTGGTGGACCCCAGTACTCGATAAGCTTAGGGTAGCCCAGGGTAGGGC 2800
 P  V  A  A  L  C  E  K  F  T  G  V  I  T  R  T  A  E  E  C  L  G  S  L  K  A  D  R  T  D  R  G  Y  E  W  W  T  P  V  L  D  K  L  R  V  A  Q  G  R  A  324

CAGGCGTCGATGGCAAAAGGCCCGCCGAACGGGGGGTGAGGAAGAGGAGCAGTCTGGGAGAGTCTTCCGCGACCGCAGGCGCGAGTATCGAAGAGCGATGCATGACGCTGAGACCGCCTTTTACCGGGAGATCGCTGAAGGGGGAAACCG 2950
 R  R  R  W  Q  K  A  R  R  T  G  G  E  E  E  E  Q  S  G  R  V  F  R  D  R  R  R  E  Y  R  R  A  M  H  D  A  E  T  A  F  Y  R  E  I  A  E  G  G  N  R  374

TGACCCGTGGGGACTAGCGTATCGGACGGCGAGTGGTAGGCGACGCGCACCAACTAATGTGGTTAACGGGGTGGAGTATGCGGGGCGGTGCTCGGATGATGTGAGTGGGGCCATGCGCACCTTGATGTGGGCGCTGTGTCCGGATGACTA 3100
 D  P  W  G  L  A  Y  R  T  A  S  G  R  R  R  A  P  T  N  V  V  N  G  V  E  Y  A  G  R  C  S  D  D  V  S  G  A  M  R  T  L  M  W  A  L  C  P  D  D  Y  424

TATGTCTCGGGACACTCCGTACCATGCGCGGGTGCGTATCATGGCCGCGCTCCCCCCATCCGGGCGGGACGCAGACCCGCTGAGCAAAGACTCCCTTCGTGCCATAATTGGCTCACTGAAGAATACCGCACCGGGCATCGACGGTTTAAC 3250
 M  S  R  D  T  P  Y  H  A  R  V  R  I  M  A  A  L  P  P  S  G  R  D  A  D  P  L  S  K  D  S  L  R  A  I  I  G  S  L  K  N  T  A  P  G  I  D  G  L  T  474

GGCGCGCATTATCAAAAAGGCACTTCCGGCTGCTGAGGCCGAGTTCGTGGCCGTATACGCACGGTGCGTTGTGGAGGGGACCTTCCCGCCGGTGTGGAAGGATGGCCGCCTACTTGTTCTGCCAAAGGGGAATGGCAGGCCCTTAACGGA 3400
 A  R  I  I  K  K  A  L  P  A  A  E  A  E  F  V  A  V  Y  A  R  C  V  V  E  G  T  F  P  P  V  W  K  D  G  R  L  L  V  L  P  K  G  N  G  R  P  L  T  D  524

CCCTAAGGCGTATCGCCCGGTCACCTTGCTGCCGGTCCTGGGAAAGATCTTGGAGAAGGTATTATTGCAGTGTGCTCCTGGCCTCACCCATAGTATTAGTCCGCGCCAGCACGGGTTCTCTCCTGGACGCTCAACGGTGACGGCGCTGCG 3550
 P  K  A  Y  R  P  V  T  L  L  P  V  L  G  K  I  L  E  K  V  L  L  Q  C  A  P  G  L  T  H  S  I  S  P  R  Q  H  G  F  S  P  G  R  S  T  V  T  A  L  R  574

AACTCTGCTGGACGTGTCGCGCGCCTCGGAGCAGAGGTACGTAATGGCCATATTCTTGGACATCAGTGGAGCTTTCGATAACGCGTGGTGGCCCATGATAATGGTGAAGGCCAAGCGGAACTGTCCGCCCAACATCTATCGGATGCTGAC 3700
 T  L  L  D  V  S  R  A  S  E  Q  R  Y  V  M  A  I  F  L  D  I  S  G  A  F  D  N  A  W  W  P  M  I  M  V  K  A  K  R  N  C  P  P  N  I  Y  R  M  L  T  624

GGACTATTTCCGCGGACGCCGTATTGCCGTTGTCGCGGGGGAATGTGCGGAATGGAAGGTGTCCACGATGGGCTGTCCGCAGGGCTCAGTGCTCGGGCCGACGCTCTGGAACGTTCTGATGGATGACCTGCTCGCCTTGCCGCAGGGGAT 3850
 D  Y  F  R  G  R  R  I  A  V  V  A  G  E  C  A  E  W  K  V  S  T  M  G  C  P  Q  G  S  V  L  G  P  T  L  W  N  V  L  M  D  D  L  L  A  L  P  Q  G  I  674

AGAGGGAACAGAGATGGTCGCCTATGCCGATGACGTGACGGTACTGGTTAGGGGTGACTCACGGGCGCAGCTTGAGAGGAGAGCGCACGCCGTGCTAGGACTCGCAGAGGGGTGGGCGAGCAGGAATAAGCTCGATTTTGCCCCGGCGAA 4000
 E  G  T  E  M  V  A  Y  A  D  D  V  T  V  L  V  R  G  D  S  R  A  Q  L  E  R  R  A  H  A  V  L  G  L  A  E  G  W  A  S  R  N  K  L  D  F  A  P  A  K  724

GTCCCGATGCATAATGCTGAGGGGAAAGTTTCAGCGTCCCCCTATAGTCCGGTACGGCAGTCATGTCATTCGGTTCGAGAACCAGGTGACGGTGTTGGGCGTCTCTTCGACGATTGCCTCTCTTTCGCGGCATGCGGCGGCCATTGGCGA 4150
 S  R  C  I  M  L  R  G  K  F  Q  R  P  P  I  V  R  Y  G  S  H  V  I  R  F  E  N  Q  V  T  V  L  G  V  S  S  T  I  A  S  L  S  R  H  A  A  A  I  G  E  774

GAGGGCGAGCAGGTGCTTCGGCAAGATGTCTAGAGTTTCGGCTTCGGCTTGGGGGCTGCGATATAGGGCTTTGCGTGTCTTGTACATGGGCACTTATGTTACAACCCTTACCTATGCGGCGGCCGTATGGTATTTGCGGGCTGCTGTGCA 4300
 R  A  S  R  C  F  G  K  M  S  R  V  S  A  S  A  W  G  L  R  Y  R  A  L  R  V  L  Y  M  G  T  Y  V  T  T  L  T  Y  A  A  A  V  W  Y  L  R  A  A  V  H  824

CGTCGTGCGCAGCGTGCTGCTTAGGACGCAGCGCCCGTCGTTGACGCTGCTAACGAAGGCCTACCGTTCGTGCAGCACGGCTGCTTTGCCGGTGTTGGCGGGCGTCCTGCCGGCGGACCTGGAGGTGACTCGTGCTGGACGGAGGATGCG 4450
 V  V  R  S  V  L  L  R  T  Q  R  P  S  L  T  L  L  T  K  A  Y  R  S  C  S  T  A  A  L  P  V  L  A  G  V  L  P  A  D  L  E  V  T  R  A  G  R  R  M  R  874

GGAGTGCGAAGGATTGGCGCGGGAGTTGGCGGCGGAGAGACGACGACGGATCGACGGCGATGTCTTGGTAGTTTGGCAGAACAGGTGGGTGTCTGAGGGTAAGGGGAGGGAACTGTACAAGTTCTTTCCCGATGTTGCGGACAGGAAGAA 4600
 E  C  E  G  L  A  R  E  L  A  A  E  R  R  R  R  I  D  G  D  V  L  V  V  W  Q  N  R  W  V  S  E  G  K  G  R  E  L  Y  K  F  F  P  D  V  A  D  R  K  K  924

GGCAACGTGGATGGAGCCGGACTATCAGACCTCGCAGATCCTCACGGGTCATGGGATCTTTAATAAGCGGTTGGCGGATATGCGACTGAGGGAGGGGCATGCTTGCGACTGCGGGGCGGTTGAGGAGGATAGGGACCATGTCCTGTGGGA 4750
 A  T  W  M  E  P  D  Y  Q  T  S  Q  I  L  T  G  H  G  I  F  N  K  R  L  A  D  M  R  L  R  E  G  H  A  C  D  C  G  A  V  E  E  D  R  D  H  V  L  W  E  974

GTGTCCTCTCTATGACGAAATCCGGGGCAGGATGCTCGATGGAATCTCGCGGTCTGAGGTGGGCCCAGTTTACCACGCGGACCTGGTCAGGGACGAGAAAAATTTTCGGCTCTTGCGCGAGTTCGCGATACATGGCACACGGCGCGCACT 4900
 C  P  L  Y  D  E  I  R  G  R  M  L  D  G  I  S  R  S  E  V  G  P  V  Y  H  A  D  L  V  R  D  E  K  N  F  R  L  L  R  E  F  A  I  H  G  T  R  R  A  L 1024

GCGCGCGAATCACGAGGACTGCCTGGCGACGAAGAATGTGGGTGATGTTAGCAGCCCCAGAACGGGGAGAATGCTAGTGGAGTGAAGGGGACGGTGTTAGTTTGTAGAACCGATCGGTACCTTGGTGCCGTGAAGTTCATGCTTCGGTCC 5050
 R  A  N  H  E  D  C  L  A  T  K  N  V  G  D  V  S  S  P  R  T  G  R  M  L  V  E---  1051

                                        5092
                                                       ATCTAGCGAAACCACAGCCAAGGGAACGGGCTTGGGAGAATCAGCGGGGAAAGAAGACCCTGTTGAGCTTGACTCTAGTCTGGCATTGTAAGGA rDNA
TAATAACCGCAAGGTTGGTGGACCATGGGAGGTGGTGGGAATGTCCCTATCTACTATCTAGCGAAACCACAGCCAAGGGAACGGGCTTGGGAGAATCAGCGGGGAAAGAAGACCCTGTTGAGCTTGACTCTAGTCTGGCATTGTAAGGA B78
```

of these subclones, specific restriction fragments were subcloned into M13mp18 and M13mp19 vectors (62) for sequence determination (49). The restriction sites used in this procedure and the direction and extent of sequence determination from these sites are shown in Fig. 1B. Nucleotide sequences of the 5′ and 3′ regions of R1 elements isolated from the Charon 35 genomic library were determined by directly subcloning 0.8-kb *Hin*cII and 2.5-kb *Kpn*I fragments, respectively, from the lambda clones into the M13mp18 vector. DNA sequences and deduced protein sequences were analyzed with the help of computer programs developed by Pustell and Kafatos (41, 42).

RESULTS

Nucleotide sequence of an R1 element. The complete nucleotide sequence of the R1 element of B78 was determined as shown in Fig. 1B and is presented in Fig. 2. The total length of this element was 5,092 bp. Neither short nor long direct or inverted repeats were present at the extreme ends of the element. The most striking feature apparent in the nucleotide sequence was the presence of two ORFs which together occupied 88% of its length (Fig. 1B, bottom). Both ORFs were in the same orientation as the direction of transcription of the ribosomal units. ORF1 was 461 amino acids long, starting from bp 467 and ending at position 1849. The relative molecular mass of a polypeptide initiated from the first ATG codon (position 4) was 48 kilodaltons.

ORF2 was 1,051 amino acids long, starting at bp 1830 and ending at position 4982. ORF2 overlapped ORF1 by 19 nucleotides in the +1 reading frame. ORFs of these sizes and degree of overlap have been found in a number of retroviruses and retrotransposable elements (8, 25, 35, 36, 61). In certain of these elements, the proteins encoded by both reading frames are produced from the same mRNA by readthrough frameshifting (8, 25, 36). While a similar mechanism may exist for R1, it should be noted that an ATG codon was present at position 5 of ORF2. This codon was flanked by an A at position −3 and a G at position +4 and was therefore a good match to Kozak's optimum initiation sequence for eucaryotic protein translation (30). This suggests that protein translation might initiate at this ATG codon of ORF2 instead of readthrough frameshifting from ORF1. The predicted relative molecular mass of a polypeptide initiated from this ATG was 117 kilodaltons.

ORF1 of R1 is similar to the *gag* genes of retroviruses. Most retroviruses have a common structure in which three essential genes, *gag*, *pol*, and *env*, are flanked by LTRs (61). Nearly all retroviral *gag* genes contain a conserved series of amino acids called Cys motifs, located near their carboxyl-terminal ends (Fig. 3A and B). The conserved amino acid positions within these motifs are three cysteine residues invariably located at positions 1, 4, and 14 and a histidine residue located at position 9 (11). Many retroviral *gag* genes contain two Cys motifs separated by a short segment of from 5 to 21 amino acids, while others contain only one motif (Fig. 3A). These Cys motifs are part of the protein p12, which is cleaved from the entire polypeptide encoded by the *gag* gene. p12 has been shown to have nucleic acid-binding ability and is required in the assembly of retroviral core particles (37, 51, 52). Cys motifs (also called metal-binding

A

R1Bm	CNKCQQYGHPEKFC	6 aa	CGRCGEDGHRMEAC		
RSV	CYTCGSPGHYQAQC	12 aa	CQLCNGMGHNAKQC		
HIV	CFNCGKEGHIARNC	7 aa	CWKCGKEGHQMKNC		
BLV	CYRCLKEGHWARDC	20 aa	CPICKDPSHWKRDC		
Mo-MuLV	CAYCKEKGHWAKDC				
I Factor	CKKCLRFGHPTPIC	5 aa	CINCSETKHTNDGE		
F Element	CTNCQEYGHTRSYC	5 aa	CVVCGDLHDSKQC	11 aa	CNNCGGNHTANYRGC
Copia	CHHCGREGHIKKDC				

B

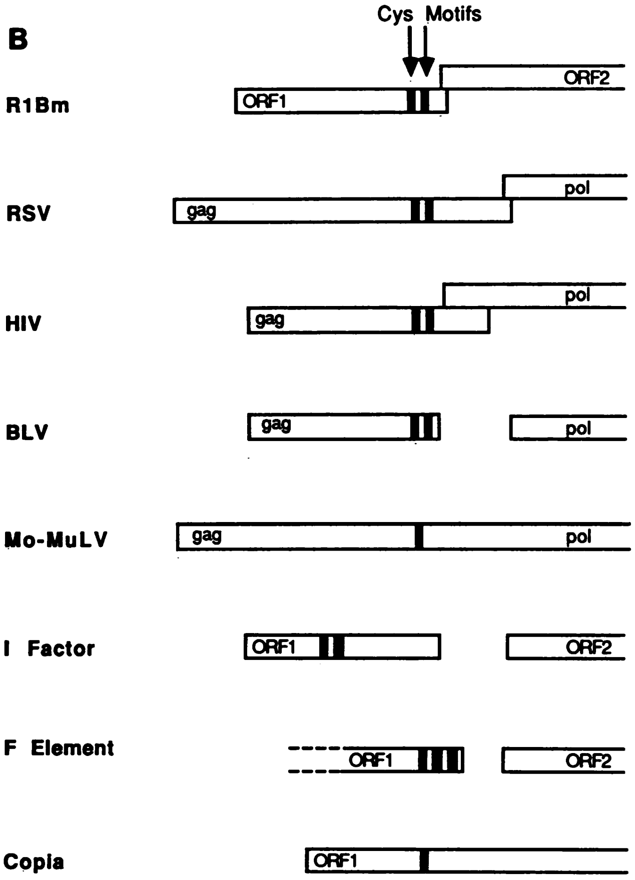

FIG. 3. Comparison of the Cys motifs found in R1 with those of retroviruses and retrotransposons. Sequences are taken from Rous sarcoma virus (RSV) (50), Mo-MuLV (55), BLV (47), HIV (44), copia (39), I factor (17), and F element (13). (A) Comparison of the Cys motif sequences. The number of amino acids (aa) separating multiple Cys motifs within the same element is indicated. Conserved C and H residues are shown in boldface. (B) Comparison of the Cys motif location in *gag* genes and ORFs. Horizontal open bars correspond to ORFs or *gag* and *pol* genes of the elements indicated at left (amino terminus at left). The 5′ terminus of the F element is truncated in the sequenced copy (13). In certain of these elements the reading frames overlap, while in others they are separated by a distance. Each vertical solid bar corresponds to one Cys motif. The stippled bars in ORF1 of I and F indicate imperfect Cys motifs.

FIG. 2. Nucleotide sequence of R1. The complete nucleotide sequence of the R1 element from the lambda clone B78 is shown, together with the deduced amino acid sequence of the two overlapping ORFs. The flanking 28S gene sequence from the uninserted rDNA unit of clone B108 (16) is shown above the sequence determined from B78. Boxed regions are the 14-bp duplication sequence of the 28S gene.

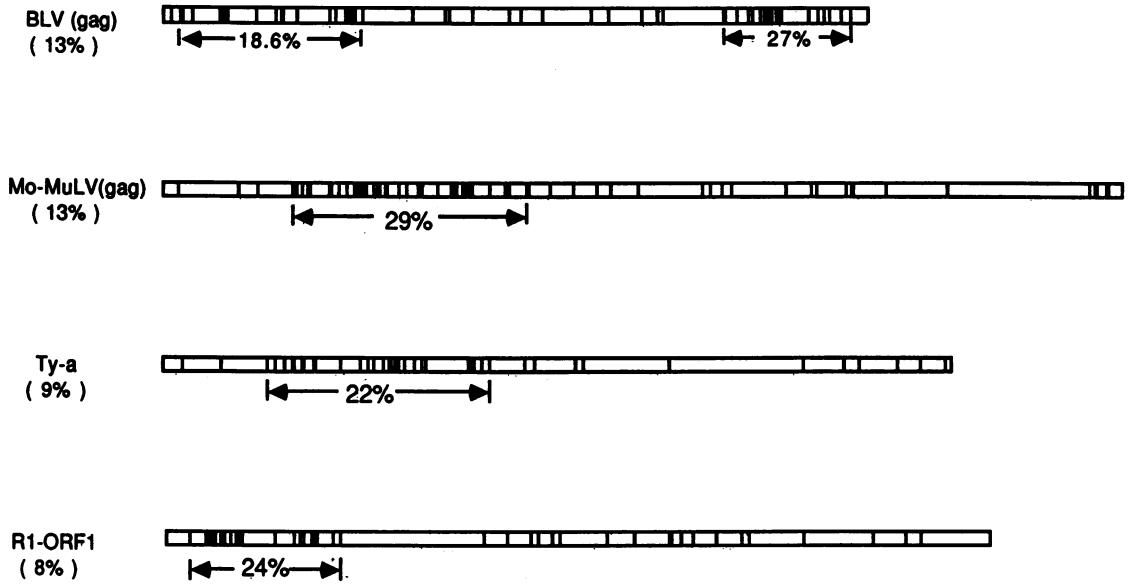

FIG. 4. Distribution of proline residues in ORF1 of R1 compared with that in other elements. Horizontal bars correspond to ORFs (amino terminus at left). The vertical lines represent the positions of proline residues. The segments with a high concentration of proline residues are marked with arrows. The moles percent proline residues are indicated for the entire *gag* gene or ORF (at left) and for each region of high proline concentration. Ty-a, *S. cerevisiae* transposable element Ty, first ORF (8); BLV (47); Mo-MuLV (55).

domains) with a somewhat different spacing of Cys and His residues have also been found in a number of eucaryotic DNA-binding proteins that are regulatory in nature (5).

Among transposable elements, Cys motifs have been reported in copia, F element, and I factor (13, 17, 39). Copia contains one motif near the amino terminus of its single large ORF. The I factor and F element contain two and three motifs in their first ORF, respectively (Fig. 3A and B). The second Cys motif of I is missing the Cys at position 14, while both the second and third Cys motifs of F contain an unusual location of the His and final Cys residues. As shown in Fig. 3, ORF1 or R1 was more retroviruslike than any other known transposable element. It contained two perfect Cys motifs separated by 6 amino acids. Both the sequence and the location of the two Cys motifs were similar to those of viral *gag* genes.

A second feature of the R1 ORF1 which was similar to retroviral *gag* proteins was its high content (8%) and nonuniform distribution of proline residues. Within an 82-amino-acid region near the amino terminus of ORF1 were located 20 proline residues (24% proline) (Fig. 4). A similar distribution of proline residues was found in certain retroviruses and retrotransposons. For example, the proline content is 29% in a 135-amino-acid region of the Moloney murine leukemia virus (Mo-MuLV) *gag* protein (55), 19% in a 101-amino-acid region of the bovine leukemia virus (BLV) *gag* protein (47), and 22% in a 124-amino-acid region of ORFa of the retrotransposon Ty (8). The *gag* protein of BLV contains a second proline-rich region near its carboxyl terminus, the same region containing the Cys motifs. While the function of these proline-rich domains is not known, it has recently been suggested that, at least in the Ty element, this domain is involved in the formation of viruslike particles (1).

These similarities between retroviral *gag* genes and R1 ORF1 strongly suggest that the R1 element has a common origin with retroviruses and retrotransposable elements. They also indicate that either R1 forms viruslike particles or there are unknown additional functions for the *gag* gene in the life cycle of these mobile elements.

ORF2 of R1 contains homology to reverse transcriptase-like enzymes. A common property of the genomes of all retroviruses and retrotransposons is the existence of a reverse transcriptase-like coding region. Based on a comparison of several retroviral *pol* genes and the large ORF of retrotransposon 17.6, Toh et al. (59, 60) identified a series of amino acid segments common to all elements, in which 33 positions are either invariant or contain chemically similar amino acids. These conserved amino acids have become a standard means of identifying homology to reverse transcriptase in ORFs from a variety of sources. In Fig. 5 the central region of the R1 ORF2 is compared with the *pol* gene of three retroviruses (Mo-MuLV, Rous sarcoma virus, and human immunodeficiency virus [HIV]), three copialike retrotransposons found in *D. melanogaster* (17.6, gypsy, and copia), and five recently discovered elements which appear to be retrotransposons but lack LTRs (R2 of *B. mori*, I factor and F element of *D. melanogaster*, L1 of *M. domesticus*, and Ingi of *T. brucei*).

These amino acid sequence comparisons are presented in Fig. 5 as eight segments which correspond to the regions of highest homology between the elements. These eight regions include most (27 of 33) of the conserved residues identified by Toh et al., including 12 of the 13 positions that were originally scored as invariant (59, 60). R1 contained all but one of these invariant residues and thus had homology to reverse transcriptase. Indeed, the residue not found in R1 (a Leu at the beginning of region 2) was also not found in a number of more recently sequenced elements, including copia and HIV, and thus should no longer be regarded as invariant.

A second conclusion that can be derived from the sequence comparisons is that the six retrotransposons grouped at the top show greater similarity to each other than they do to the retroviruses or copialike retrotransposable elements. This greater homology can be seen (i) in the similarity in

```
                          1                                                 2
                             L     PG D           F      G  P            K           YRP  L     K        R
R1Bm   ..(449)..PLSKDSLRAIIGSLK-NTAPGIDGL..(18)..VYARCVVEGTFPPVWKDGRLLVLPKGNGRPLTDPKAYRPVTLLPVLGKILEKVLLQCA..(10)..
R2Bm   ..(426)..PISVEEIKASRFDWR--TSPGPDGI..(15)..MFNAWMARGEIPEILRQCRTVFVPKVERP--GGPGEYRPISIASIPLRHFHSILARRL..( 9)..
L1Md   ..(477)..PISPKEIEAVINSLPTKKSPGPDGF..(18)..LFHKIEVEGTLPNSFYEATITLIPKPQKD-PTKIENFRPISLMNIDAKILNKILANRI..(11)..
I      ..(298)..NITYLELSSALQTLK-GCAPGLNRI..(18)..LFNE-IFNSHIPQAYKTSLIIPILKPNTD-KTKTSSYRPISLNCCIAKILDKIIAKRL..(13)..
Ingi   ..(133)..PITMAELRRSIKLLPSGSAAGPDCL..(18)..LFNESLRTGVVPPAWKTGVIIPILKAGKK-AEDLDSYRPVTLTSCLCKVMERIIAARP..(11)..
F      ..(435)..FRPKEITKIIKDNLSPKKSPGYDLI..(18)..LFNAITKLGYFPQRWKMMKIIMIPKPGKN-HTVASSYRPISLLSCISKLFEKCLLIRL..(13)..

17.6   ..................(227)...................QDMLNQGIIRTSNSPYNSPIWVVPKKQD--ASGKQKFRIVIDYRKLNEITVGDRHPIP........
Gypsy  ..................(201)...................KQLLKDGIIRPSRSPYNSPTWVVDKKGTD-ASGNPNKRLVIDFRKLNEKTIPDRYPMP........
Copia  ..................(920)...................NTWTITKRPENKNIVDSRRWVFSVKYNELGNPIRYKARLVARGFTQKYQIDYEETFAP........

MuLV   ..................(198)...................QRLLDQGILVPCQSPWNTPLLPVKKPGTN------DYRPVQDLREVNKRVEDIHPTVP........
RSV    ..................( 37)...................EKELQLGHIEPSLSCWNTPVFVIRKASGS-------YRLLHDLRAVNAKLVPFGAVQQ........
HIV    ..................(196)...................EKEGKISKIGPEN-PYNTPVFAIKKKDST------KWRKLVDFRELNKRTQDFWEVQL........
                                                    L                                 R


           3                  4                         5                      6                 7               8
       Q GF     T        LD   AFD              G  QG      L        L           ADD                    K          LGV
R1Bm   QKGFSPGRST..(23)..LDISGAFDNAWW..(40)..TMGCPQGSVLGPTLWNVLMDDLLA..(11)..AYADDVTVLV..(20)..GWA-SRNK..(36)..TVLGVSS
R2Bm   QRGFICADGT..(25)..LDFAKAFDTVSH..(43)..GRGVRQGDPLSPILFNVVMDLILA..(18)..AYADDLVLLA..(20)..GLRLNCRK..(39)..RYLGVDF
L1Md   QVGFIPGMQG..(26)..LDAEKAFDKIQH..(43)..KSGTRQGCPLSPYLFNIVLE-VLA..(22)..LFADDMIVYI..(21)..GYKINSNK..(29)..KYLGVTL
I      QFGFKKGKST..(22)..LDFSRAFDRVGV..(43)..FNGIPQGSPISVILFFIAFN-KLS..(12)..AYADDFFLII..(26)..GASLSLSK..(29)..KILGRTL
Ingi   QSGFRPGCST..(23)..VDYEKAFDTVDH..(43)..ERGVPQGTVPGSIMFIIVMN-SLS..(12)..FFADDLTLLA..(26)..SVNVAKTK..(27)..KLLGVTF
F      QFGFRESHGT..(25)..LDVSQAFDKVWL..(42)..EAGVPQGSVLGPTLYLIYTADIPT..( 6)..TFADDTAILS..(26)..RIKVNEQK..(30)..TYLGVHL

17.6   ............(24)..IDLAKGFHQIEV..(18)..YLRMPFGLKNAPATFQRCMNDILR..( 8)..VYLDDIIVFS..(20)..NLKLQLDK..( 8)..TFLGHVL
Gypsy  ............(24)..LDLKSGYHQIYL..(18)..FCRLPFGLRNASSIFQRALDDVLR..( 8)..VYVDDVIIFS..(20)..NMRVSQEK..( 8)..EYLGFIV
Copia  ............(21)..MDVKTAFLNGTL..( 4)..YMRLPQGISCNSDNVCKLNKAIYG..(45)..LYYDDVVIAT..( 5)..FKRYLMEK..( 9)..KFIGIRI

MuLV   ............(25)..LDLKDAFFCLRL..(24)..WTRLPQGFKNSPTLFDEALHRDLA..(12)..QYVDDLLLAA..(20)..GYRASAKK..( 8)..KYLGYLL
RSV    ............(23)..LDLKDCFFSIPL..(25)..WKVLPQGMTCSPTICQLVVGQVLE..(12)..HYMDDLLLAA..(20)..GFTISPDK..( 7)..QYLGYKL
HIV    ............(23)..LDVGDAYFSVPL..(25)..YNVLPQGWKGSPAIFQSSMTKILE..(12)..QYMDDLYVGS..(21)..GLT-TPDK..( 8)..LWMGYEL
                          D    F                 P G    P                     Y DD             G      K           G
```

FIG. 5. Comparison of amino acid sequences of known and putative reverse transcriptase regions from various origins. For each element the first number in parentheses indicates the distance (in amino acids) from the beginning of the *pol* genes or the ORFs. Other numbers in parentheses indicate the number of amino acids omitted from the sequence shown. The letters on the bottom indicate the invariant residues identified by Toh et al. (59, 60). The letters on the top indicate conserved residues shared by at least five of the six non-LTR retrotransposons. The sequences were taken from the literature as follows: R2 (7), I (17), L1 (31, 54), Ingi (29), F (13), 17.6 (48), gypsy (35), copia (39), Mo-MuLV (55), Rous sarcoma virus (50), and HIV (44).

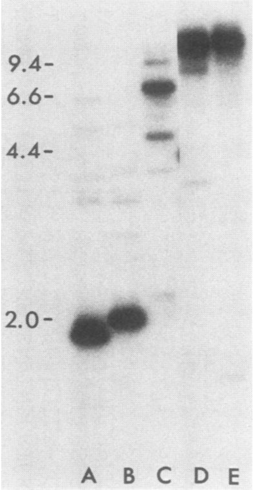

FIG. 6. Genomic blot hybridization of *B. mori* DNA probed with R1 sequences. For each lane, 3 μg of genomic DNA was digested with restriction enzymes, fractionated on a 1% agarose gel, and transferred to nitrocellulose paper. The blot was hybridized with a nick-translated 1.1-kb *Sst*I-*Kpn*I fragment isolated from plasmid 78-Xho-1.9 kb (see Fig. 1B). The restriction endonucleases used for each genomic DNA digestion were *Xho*I (lane A), *Sst*I and *Bgl*II (lane B), *Sst*I (lane C), *Eco*RI (lane D), and *Pst*I (lane E). Numbers at left indicate length (in kilobases) of DNA standards.

spacing between the regions of conserved sequence, (ii) in the presence of two segments (1 and 3) which were highly conserved in these retrotransposons but for which corresponding regions could not be found in retroviruses or copialike elements, and (iii) in the numerous amino acid positions which were shared by at least five of the six elements in this group but not found in retroviruses or copialike elements. Certain of these conserved residues appeared to be highly diagnostic. For example, in region 6 (Fig. 5), which contained the highly conserved YXDD box found in all reverse transcriptase-like enzymes, X was a hydrophobic residue in all the retroviruses and copialike retrotransposons but was an invariant residue, A, in this group of retrotransposons. The greater conservation of sequence for these six retrotransposable elements suggests that they form a distinct group with a common origin. Since they also share the property of not containing LTRs (7, 13, 17, 29, 31, 54), we suggest that these mobile elements be referred to as the non-LTR class of retrotransposable elements. In fact, as we originally noted for R2Bm (7), the reverse transcriptase regions of all six non-LTR retrotransposons show somewhat higher homology to the ORFs found in certain class II introns of fungi (38) than to those of the LTR-containing retrotransposons and retroviruses (data not shown). A comprehensive sequence comparison of all elements with reverse transcriptase homology and the generation of a phylogenetic tree are in progress (Xiong and Eickbush, in preparation).

Amino acid sequence homology has previously been reported between I and L1 (17), Ingi and L1 (29), R2 and L1 (7), and F, I, and L1 (13). Certain of these reports have indicated that homology is found beyond the region we have defined as segment 1 in Fig. 5 (7, 13, 29) and in a region midway between segment 8 and the 3′ end of the ORF (7). However, when all six non-LTR sequences were compared, these regions of homology were clearly not as conserved as the reverse transcriptase regions shown in Fig. 5. More importantly, we have not been able to detect homology between these regions and the protease or integrase regions of retroviruses and other retrotransposons (35, 39, 60).

Number and distribution of R1 element within the genome. To determine the number and location of R1 elements within the *B. mori* genome, a series of genomic blotting experiments were conducted with internal fragments of the R1 element from B78 used as the probe. An example of these experiments with the 1.1-kb *Sst*I-*Kpn*I fragment purified from subclone 78-Xho-1.9 as probe (Fig. 1B) is shown in Fig. 6. The genomic DNA for this experiment was digested with *Xho*I (lane A) or *Sst*I and *Bgl*II (lane B). Most of the DNA within the *B. mori* genome that hybridized with the R1 probe was located on the 1.9-kb and 2.0-kb fragments predicted from the restriction map of B78. The intensity of the hybridization to these bands was approximately equal to 25 copies per genome. In addition to this intense band, a series of much fainter hybridizing bands were also seen in both lanes. These bands, which hybridized at a level roughly equivalent to 1 copy per genome, were higher in molecular weight than the main band and thus represented R1 elements which had lost one or both of the restriction sites used in genomic DNA digestion.

To determine what fraction of the R1 elements present within the *B. mori* genome were located within rDNA units, genomic DNA was digested with *Sst*I alone (Fig. 6, lane C). Most of the genomic DNA that hybridized to the R1 probe was located on 7.5-kb fragments, the size predicted of R1 elements located within rDNA units (the distance from the *Sst*I site within the R1 element to the *Sst*I site within the 18S rRNA gene of the adjacent rDNA unit). Hybridizing fragments of lengths other than 7.5 kb could correspond to the absence of the *Sst*I site within the R1 element, the presence of an *Sst*I site within the spacer region of certain rDNA units, or the location of R1 elements outside normal rDNA units. To more accurately estimate the number of R1 elements located outside the rDNA locus, genomic DNA was digested with *Eco*RI or *Pst*I (Fig. 6, lanes D and E, respectively). These enzymes do not cleave within R1 elements and only once within the rDNA unit. The size of the major hybridizing band in both lanes of the genomic blot was approximately 15 kb, corresponding to the length of an rDNA unit (10.6 kb) with an R1 element (5.1 kb). Only a few, faint, lower-molecular-weight bands are seen in lanes D and E. Since R1 elements located outside the rDNA units should on average be located on restriction fragments less than 15 kb long, these lower-molecular-weight bands represent our best estimate of the number of R1 elements not located within rDNA units.

Based on this and other series of genomic blots (not shown), we conclude that, as in the case of R2 elements (7), R1 sequences are present in approximately 10% of the estimated 240 rDNA units of *B. mori* (22) and only a few copies of R1 exist outside the rDNA repeats.

Sequence variation at the R1-rDNA boundaries. To determine whether R1 elements located within the ribosomal locus exhibited sequence variations at their borders with the 28S gene, 10 lambda clones containing R1 elements located within the ribosomal locus were isolated from the Charon 35 library. The sequence of five 3′ junctions and all 10 5′ junctions revealed that all R1 elements were flanked by an identical 14-bp duplication of 28S gene sequences. As in the case of R2 elements (7), no variation was detected at the 3′ junctions of the R1 elements with the 28S gene. On the other hand, as shown in Fig. 7, a total of five different junctions were observed in the immediate vicinity of the 5′ border of

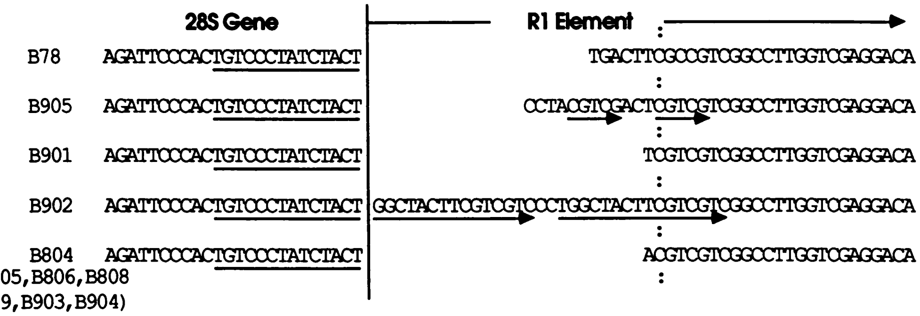
28S Gene
R1 Element
B78 AGATTCCCACTGTCCCTATCTACT TGACTTCGCCGTCGGCCTTGGTCGAGGACA
B905 AGATTCCCACTGTCCCTATCTACT CCTACGTCGACTCGTCGTCGGCCTTGGTCGAGGACA
B901 AGATTCCCACTGTCCCTATCTACT TCGTCGTCGGCCTTGGTCGAGGACA
B902 AGATTCCCACTGTCCCTATCTACT GGCTACTTCGTCGTCCCTGGCTACTTCGTCGTCGGCCTTGGTCGAGGACA
B804 AGATTCCCACTGTCCCTATCTACT ACGTCGTCGGCCTTGGTCGAGGACA
(B805,B806,B808
B809,B903,B904)

FIG. 7. Sequence variation at the 5′ border of R1 with the 28S gene. The 5′ junction sequence of 11 independently isolated R1 copies was determined. The 14-bp sequence of the 28S gene which is duplicated at the 3′ end of the elements is underlined. Direct repeats of sequence within the R1 element are indicated by arrows beneath the sequence. The dotted vertical line defines the extent of the 5′ variation.

the R1 element with the 28S gene. Identity of all clones began with the sequence CGTCGT (B78 has a C instead of a T at the third position of this sequence). The five different variants contained from 1 to 26 bp between the region of sequence identity and the 28S gene. Although there was some similarity between these 5′ sequence variants, they did not appear to have been generated by any simple series of recombination events. The major variant was a single A 5′ of the region of sequence identity (7 of 11 independent isolates). Of the other four variants, two involved duplications of a small segment of DNA. Clone B905 contained a 5-bp duplication separated by 3 bp. Clone B902 contained a 5-bp duplication separated by 3 bp. The B902 duplication could be extended to 17 bp separated by 1 bp if the last 2 bases of the 28S gene (CT) were included. While duplications of DNA sequences on the same side of a retrovirus or transposable element are not common, such a duplication was found in a fraction of the R2 elements we have analyzed from *B. mori* (7). Duplication of DNA sequences on the 5′ side of the insertion may reflect an unusual integration mechanism for these sequence-specific elements or even of the integration of non-LTR elements in general.

No ribosomal sequence deletion was found associated with the insertion of R1 elements in *B. mori*. In *D. melanogaster*, 23 bp of rDNA flanking the 5′ end of R1 elements are absent in the major 5.0-kb R1 variant (43).

DISCUSSION

The presence of the R1 elements in approximately 10% of the rDNA units of *B. mori* could be the result of either repeated transposition of the element, its expansion by recombinational means such as unequal crossovers, or a combination of these processes (7, 9, 14, 56). Several considerations suggest that the retention of R1 within the 28S genes is at least partly due to its ability to actively transpose. First, R1 insertions have been found in both lepidopteran and dipterian species (3, 4, 16, 21, 46). This would suggest that R1 is very ancient or that it has independently been introduced multiple times in insects. In each of these insect species, R1 elements occupy only a fraction of the rDNA units. Unequal crossover events acting alone would eventually lead to the fixation or elimination of such variants. Second, the very fact that two large ORFs exist within the R1 element of *B. mori* suggests that this element is still a functional mobile element. These ORFs would not be maintained in the course of evolution unless they are under selective constraint as protein-coding sequences. Finally, analysis of 10 independently isolated R1 elements revealed sequence variation at their 5′ junction with the 28S gene. Unequal crossover or gene conversion mechanisms cannot explain this variation or the fact that it is limited to the 5′ end.

Retrotransposable elements and retroviruses are able to insert into many sites of the host chromosome. While it does not limit their insertion to a particular region of the genome, some of these mobile elements have been shown to have a degree of insertion sequence specificity or preference (e.g., ATAT for 17.6 [24], TATATA for 297 [23], and TA[C/A]TA for gypsy [19, 35]). If R1 is a transposable element, then its insertion into the genome must be highly site specific. All or nearly all copies of R1 are located at the identical position within the 28S gene. This degree of site specificity in the insertion of R1, and as previously described in R2 (7), is similar to that of the mobile rl intron present in the large rRNA gene of *S. cerevisiae* mitochondria (26, 34). The specificity of this intron's transposition has been shown to be a result of an endonuclease encoded by its 240-amino-acid ORF (10). It is interesting that a 4.5-kb element was found to interrupt a fraction of the 26S ribosomal genes of *Ascaris lumbricoides* (2). These elements are located 34 bp upstream of the insect R1 elements and are flanked by 13-bp duplications of target sequence. It is possible that this element corresponds to the R1 elements of insects which has shifted its insertion sequence specificity within the large rRNA gene. Experiments to determine whether the ORFs of R1 and R2 encode sequence-specific endonucleases (integrases) are currently in progress.

Sequence comparison of the reverse transcriptase region indicates that R1 is most closely related to the R2 of *B. mori*, I factor and F element of *D. melanogaster*, L1 of *M. domesticus*, and Ingi of *T. brucei*. Like R1, all of these elements lack LTRs. The sequence comparison of all six elements presented in this paper strongly suggests that these elements represent a distinct related group rather than a variety of separate elements that have independently lost their LTRs. We refer to this group as the non-LTR retrotransposons. These elements all have a large ORF (range, 859 to 1,259 amino acids) which contains reverse transcriptase homology. Except for R2, they also contain a shorter ORF, similar in size and location to the retroviral *gag* gene. The absence of this ORF in R2 suggests that this element may utilize the *gag*-related ORF of R1. This possibility is supported by the fact that these two elements are specifically located in the 28S genes of the same organism and, as discussed below, probably have similar methods of transcription. Surprisingly, if R1 and R2 do have a common *gag* protein, they exhibit no other features suggesting that they are evolving in concert. Protein and DNA homology between these two elements was only detected in the reverse transcriptase region, where it was similar to that of the other non-LTR elements.

The essential role of LTRs in the life cycle of retroviruses and most retrotransposable elements suggests that substantially different mechanisms must have evolved for the propagation of the non-LTR retrotransposons. Central to this issue is the manner in which these elements are transcribed and how these transcripts are able to give rise to complete reverse transcripts. In the case of the I factor, Fawcett et al. have suggested the possible existence of an internal promoter (17), while Loeb et al. have suggested that the promoter for L1 is located within a series of tandem repeats at its 5′ end (31, 54). We suggest that R1 and R2 might be able to use the heterologous polymerase I promoter of the rDNA unit for their transcription. There are two findings to support this suggestion. First, the coding strand of all R1 and R2 elements located within the ribosomal locus is oriented in the same direction as ribosomal gene transcription. Second, the R1 and R2 transcripts detected at a low level in *D. melanogaster* cells are linked to the rRNA sequences (27, 28, 33). Since R1 and R2 transcripts are not abundant in either *D. melanogaster* (27, 28, 33) or *B. mori* (Xiong and Eickbush, unpublished observations), transcription through the R1 and R2 elements may be very inefficient or may be limited to particular tissues and developmental periods (e.g., germ cells). Alternatives to certain of the other major functions of LTRs, such as providing a primer-binding site for reverse transcription and aiding in the production of unit-length progeny, must also have evolved in R1 and the other non-LTR retrotransposons. Understanding these processes, which may even have evolved prior to LTR function, is one of the major challenges posed by the non-LTR elements.

ACKNOWLEDGMENTS

We thank Bill Burke for many stimulating discussions and for critically reading the manuscript.

The investigation was supported in part by a Public Health Service grant from the National Institutes of Health.

LITERATURE CITED

1. **Adams, S. E., J. Mellor, K. Gull, R. B. Sim, M. F. Tuite, S. M. Kingsman, and A. J. Kingsman.** 1987. The functions and relationships of Ty-VLP proteins in yeast reflect those of mammalian retroviral proteins. Cell **49**:111–119.
2. **Back, E., E. van Meir, F. Muller, D. Schaller, H. Nehaus, P. Aeby, and H. Tobler.** 1984. Intervening sequences in the ribosomal RNA genes of *Ascaris lumbricoides*: DNA sequences at the junctions and genomic organization. EMBO J. **3**:2523–2529.
3. **Barnett, T., and P. M. M. Rae.** 1979. A 9.6 kb intervening sequence in *D. virilis* rRNA, and sequence homology in rDNA interruptions of diverse species of *Drosophila* and other diptera. Cell **16**:763–775.
4. **Beckingham, K., and R. White.** 1980. The ribosomal DNA of *Callipora erythrocephala*: an analysis of hybrid plasmids containing ribosomal DNA. J. Mol. Biol. **137**:349–373.
5. **Berg, J. M.** 1986. Potential metal-binding domains in nucleic acid binding proteins. Science **232**:485–487.
6. **Boeke, J. D., D. J. Garfinkel, C. A. Styles, and G. R. Fink.** 1985. Ty elements transpose through an RNA intermediate. Cell **40**: 491–500.
7. **Burke, W. D., C. C. Calalang, and T. H. Eickbush.** 1987. The site-specific ribosomal insertion element type II of *Bombyx mori* (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. Mol. Cell. Biol. **7**:2221–2230.
8. **Clare, J., and P. Farabaugh.** 1985. Nucleotide sequence of a yeast Ty element: evidence for a novel mechanism of gene expression. Proc. Natl. Acad. Sci. USA **82**:2829–2833.
9. **Coen, E. S., T. Strachan, and G. Dover.** 1982. Dynamics of concerted evolution of ribosomal DNA and histone gene families in the *melanogaster* species subgroup of *Drosophila*. J. Mol. Biol. **158**:17–35.
10. **Colleaux, L., L. d'Auriol, M. Betermier, G. Cottarel, A. Jacquier, F. Galibert, and B. Dujon.** 1986. Universal code equivalent of a yeast mitochondrial intron reading frame is expressed in *E. coli* as a specific double strand endonuclease. Cell **44**: 521–533.
11. **Covey, S. N.** 1986. Amino acid sequence homology in *gag* region of reverse transcriptase elements and the coat protein gene of cauliflower mosaic virus. Nucleic Acids Res. **14**:623–633.
12. **Dawid, I. B., and M. L. Rebbert.** 1981. Nucleotide sequence at the boundaries between gene and insertion regions in the rDNA of *Drosophila melanogaster*. Nucleic Acids Res. **9**:5011–5020.
13. **Di Nocera, P. P., and G. Casari.** 1987. Related polypeptides are encoded by *Drosophila* F elements, I factors, and mammalian L1 sequences. Proc. Natl. Acad. Sci. USA **84**:5843–5847.
14. **Dover, G., and E. Coen.** 1981. Springcleaning ribosomal DNA: a model for multigene evolution. Nature (London) **290**:731–732.
15. **Eickbush, T. H., and F. C. Kafatos.** 1982. A walk in the chorion locus of *Bombyx mori*. Cell **29**:633–643.
16. **Eickbush, T. H., and B. Robins.** 1985. *Bombyx mori* 28S genes contain insertion elements similar to the type I and II elements of *Drosophila melanogaster*. EMBO J. **4**:2281–2285.
17. **Fawcett, D. H., C. K. Lister, E. Kellett, and D. J. Finnegan.** 1986. Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINES. Cell **47**: 1007–1015.
18. **Finnegan, D. J.** 1985. Transposable elements in eukaryotes. Int. Rev. Cytol. **93**:281–326.
19. **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.
20. **Fujiwara, H., and H. Ishikawa.** 1986. Molecular mechanism of introduction of the hidden break into the 28S rRNA of insects: implication based on structural studies. Nucleic Acids Res. **14**:6393–6401.
21. **Fujiwara, H., T. Ogura, N. Takada, N. Miyajima, H. Ishikawa, and H. Maekawa.** 1984. Introns and their flanking sequences of *Bombyx mori* rDNA. Nucleic Acids Res. **12**:6861–6869.
22. **Gage, L. P.** 1974. Polyploidization of the silkgland of *Bombyx mori*. J. Mol. Biol. **86**:97–108.
23. **Ikenaga, H., and K. Saigo.** 1982. Insertion of a movable genetic element, 297, into the T-A-T-A box for the H3 histone gene in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **79**:4143–4147.
24. **Inouye, S., S. Yuki, and K. Saigo.** 1984. Sequence-specific insertion of the *Drosophila* transposable genetic element 17.6. Nature (London) **310**:332–333.
25. **Jacks, T., and H. E. Varmus.** 1985. Expression of the Rous sarcoma virus *pol* gene by ribosomal frameshifting. Science **230**:1237–1242.
26. **Jacquier, A., and D. Dujon.** 1985. An intron-encoded protein is active in a gene conversion process that spreads an intron into a mitochondrial gene. Cell **41**:383–394.
27. **Jamrich, M., and O. L. Miller.** 1984. The rare transcripts of interrupted rRNA genes in *Drosophila melanogaster* are processed or degraded during synthesis. EMBO J. **3**:1541–1545.
28. **Kidd, S. J., and D. M. Glover.** 1981. *Drosophila melanogaster* ribosomal DNA containing type II insertions is varibly transcribed in different strains and tissues. J. Mol. Biol. **151**: 645–662.
29. **Kimmel, B. E., O. K. ole-Moiyoi, and J. R. Young.** 1987. Ingi, a 5.2-kilobase dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Mol. Cell. Biol. **7**:1465–1475.
30. **Kozak, M.** 1986. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell **44**:283–292.

31. **Loeb, D. D., R. W. Padgett, S. C. Hardies, W. R. Shehee, M. B. Comer, M. H. Edgell, and C. A. Hutchison III.** 1986. The sequence of a large L1Md element reveals a tandemly repeated 5′ end and several features found in retrotransposons. Mol. Cell. Biol. **6**:168–182.
32. **Loenen, W. A. M., and F. R. Blattner.** 1983. Lambda Charon vectors (Ch32, 33, 34 and 35) adapted for DNA cloning in recombination-deficient hosts. Gene **26**:171–179.
33. **Long, E. O., and I. B. Dawid.** 1979. Expression of ribosomal DNA insertions in *Drosophila melanogaster*. Cell **18**:1185–1196.
34. **Macreadie, I. G., R. M. Scott, A. R. Zinn, and R. A. Butow.** 1985. Transposition of an intron in yeast mitochondria requires a protein encoded by that intron. Cell **41**:395–402.
35. **Marlor, R. L., S. M. Parkhurst, and V. G. Corces.** 1986. The *Drosophila melanogaster* gypsy transposable element encodes putative gene product homologous to retroviral proteins. Mol. Cell. Biol. **6**:1129–1134.
36. **Mellor, J., S. M. Fulton, M. J. Dobson, W. Wilson, S. M. Kingsman, and A. J. Kingsman.** 1985. A retrovirus-like strategy for expression of a fusion protein encoded by yeast transposon Ty1. Nature (London) **313**:243–246.
37. **Meric, C., and P. F. Spahr.** 1986. Rous sarcoma virus nucleic acid-binding protein p12 is necessary for viral 70S RNA dimer formation and packing. J. Virol. **60**:450–459.
38. **Michel, F., and B. F. Lang.** 1985. Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. Nature (London) **316**:641–643.
39. **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.
40. **Panganiban, A. T.** 1985. Retroviral DNA integration. Cell **42**: 5–6.
41. **Pustell, J., and F. C. Kafatos.** 1982. A convenient and adaptable package of DNA sequence analysis programs for microcomputers. Nucleic Acids Res. **10**:51–60.
42. **Pustell, J., and F. C. Kafatos.** 1984. A convenient and adaptable package of computer programs for DNA and protein sequence management, analysis and homology determination. Nucleic Acids Res. **12**:643–655.
43. **Rae, P. M. M., B. D. Kohorn, and R. P. Wade.** 1980. The 10 kb *Drosophila virilis* 28S rDNA intervening sequence is flanked by a direct repeat of 14 base pairs of coding sequence. Nucleic Acids Res. **8**:3491–3505.
44. **Ratner, L., W. Haseltine, R. Patarca, K. J. Livak, B. Starcich, S. F. Josephs, E. R. Doran, J. A. Rafalski, E. A. Whitehorn, K. Baumeister, L. Ivanoff, S. R. Petteway, Jr., M. L. Pearson, J. A. Lautenberger, T. S. Papis, J. Ghrayeb, N. T. Chang, R. C. Gallo, and F. Wong-Staal.** 1985. Complete nucleotide sequence of the AIDS virus, HTLV-III. Nature (London) **313**:277–284.
45. **Roiha, H., J. R. Miller, L. C. Woods, and D. M. Glover.** 1981. Arrangements and rearrangements of sequences flanking the two types of rDNA insertion in *D. melanogaster*. Nature (London) **290**:749–753.
46. **Roiha, H., C. A. Read, M. J. Browne, and D. M. Glover.** 1983. Widely differing degrees of sequence conservation of two types of rDNA insertion within the *melanogaster* species sub-group of *Drosophila*. EMBO J. **2**:721–726.
47. **Sagata, N., T. Yasunaga, J. Tsuzuku-Kawamura, K. Ohishi, Y. Ogawa, and Y. Ikawa.** 1985. Complete nucleotide sequence of the genome of bovine leukemia virus: its evolutionary relationship to other retroviruses. Proc. Natl. Acad. Sci. USA **82**:677–681.
48. **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.
49. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
50. **Schwartz, D. E., R. Tizard, and W. Gilbert.** 1983. Nucleotide sequence of Rous sarcoma virus. Cell **32**:853–869.
51. **Sen, A., C. J. Sherr, and G. J. Todaro.** 1977. Phosphorylation of murine type C viral p12 protein regulates their extent of binding to the homologous viral RNA. Cell **10**:489–496.
52. **Sen, A., and G. J. Todaro.** 1977. The genome-associated, specific RNA binding proteins of avian and murine type C viruses. Cell **10**:91–99.
53. **Shapiro, J. A. (ed.).** 1983. Mobile genetic elements. Academic Press, Inc., Orlando, Fla.
54. **Shehee, W. R., S.-F. Chao, D. D. Loeb, M. B. Comer, C. A. Hutchison III, and M. H. Edgell.** 1987. Determination of a functional ancestral sequence and definition of the 5′ end of A-type mouse L1 elements. J. Mol. Biol. **196**:757–767.
55. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukaemia virus. Nature (London) **293**:543–548.
56. **Smith, G. P.** 1976. Evolution of repeated DNA sequences by unequal crossovers. Science **191**:528–535.
57. **Smith, V. L., and K. Beckingham.** 1984. The intron boundaries and flanking rRNA coding sequences of *Calliphora erythrocephala* rDNA. Nucleic Acids Res. **12**:1707–1724.
58. **Temin, H. M.** 1982. Function of the retrovirus long terminal repeat. Cell **28**:3–5.
59. **Toh, H., H. Hayashida, and T. Miyata.** 1983. Sequence homology between retroviral transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature (London) **305**:827–829.
60. **Toh, H., R. Kikuno, T. Hayashida, T. Miyata, W. Kugimiya, S. Inouye, S. Yuki, and K. Saigo.** 1985. Close structure 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.
61. **Varmus, H. E.** 1983. Retroviruses, p. 411–503. *In* J. A. Shapiro (ed.), Mobile genetic elements. Academic Press, Inc., Orlando, Fla.
62. **Yanisch-Perron, C., J. Vieira, and J. Messing.** 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. Gene **33**:103–119.