MOLECULAR AND CELLULAR BIOLOGY, July 1985, p. 1630–1638 Vol. 5, No. 7
0270-7306/85/071630-09$02.00/0

Complete Nucleotide Sequence of the *Drosophila* Transposable Element Copia: Homology Between Copia and Retroviral Proteins

STEPHEN M. MOUNT* AND GERALD M. RUBIN

Department of Biochemistry, University of California, Berkeley, California 94720

Received 28 January 1985/Accepted 2 April 1985

We have determined the complete nucleotide sequence of the copia element present at the white-apricot allele of the white locus in *Drosophila melanogaster*. This transposable element is 5,146 nucleotides long and contains a single long open reading frame of 4,227 nucleotides. Analysis of the coding potential of the large open reading frame, which appears to encode a polyprotein, revealed weak homology to a number of retroviral proteins, including a protease, nucleic acid-binding protein, and reverse transcriptase. Better homology existed between another part of the copia open reading frame and a region of the retroviral *pol* gene recently shown to be distinct from reverse transcriptase and required for the integration of circular DNA forms of the retroviral genome to form proviruses. Comparison of the copia sequence with those of the *Saccharomyces cerevisiae* transposable element Ty, several vertebrate retroviruses, and the *D. melanogaster* copia-like element 17.6 showed that Ty was most similar to copia, sharing amino acid sequence homology and organizational features not found in the other genetic elements.

The *Drosophila* transposable element copia is a member of a broad class of structurally homologous genetic elements characterized by the presence of long direct terminal repeats (LTRs). This class of elements includes the transposable Ty elements of *Saccharomyces cerevisiae*, vertebrate retrovirus proviruses, and a number of *Drosophila* elements generally known as copia-like (reviewed in references 56, 57, and 75). Different copia-like elements in *Drosophila*, although structurally homologous, are not closely related at the sequence level and do not cross-hybridize. However, individual copia-like elements are repeated in the genome between 5 and 100 times, and these copies are nearly identical (reviewed in reference 57). At least 11 repetitive elements in the *Drosophila melanogaster* genome have been positively identified as copia-like (28, 38, 42, 57, 70). Sequences of the LTRs of four of these (copia [41], 297 [38], B104 [61], and gypsy [2, 24]) have been published, as has the complete sequence of a 17.6 insertion (59). The transposition of these elements was originally inferred from polymorphism in their number and location between different stocks of *D. melanogaster*, between cell lines and their parent fly stocks, and between *D. melanogaster* and its sibling species *D. simulans* (57). More recently, close association of copia-like elements with recently derived mutations (42) and insertion of copia during laboratory experiments (58) have been observed.

Much work has explored the relationship of LTR-containing transposable elements in both yeasts and fruit flies to retroviral proviruses. All of these elements have short inverted repeats roughly 10 nucleotides long at their termini. Some sequence homology exists between these short repeated sequences; most share the dinucleotide TG at their 5′ termini and CA at their 3′ termini. A short direct repeat, present as only a single copy in the target site before integration, flanks each insertion. The length (3 to 6 nucleotides) but not the sequence of these repeats is characteristic of the particular element. Copia makes 5-base-pair (bp) repeats. Interestingly, the *Drosophila* copia-like elements 297, 17.6, gypsy, and HMS Beagle seem to form a distinct subgroup; they share AGT or AAT at the 5′ end and ACT or ATT at the 3′ end, duplicate 4 bp, and also show insertion site preference for alternating purine-pyrimidine stretches (2, 24, 38, 59, 70).

The similarity between LTR-containing elements extends to their transcription. In all cases examined to date, including copia (6, 16, 21, 79) and Ty (14), these elements give rise to an abundant transcript originating in one LTR and terminating in the other. In retroviruses (75) this RNA serves as the viral genome. After the virus infects a cell, this genomic RNA provides a template for reverse transcription, directing the formation of double-stranded DNA copies of the viral genome with a copy of the LTR at each end. Reverse transcription is primed by a tRNA, the 3′ end of which hybridizes just within the 5′ end of the unique portion of the viral RNA. Once linear DNA is formed, circular DNA containing either one LTR or two tandem LTRs appears. The ability of a 49-bp sequence containing the junction between the two tandem LTRs to direct integration when placed elsewhere in a viral genome (50) has made it clear that circles with two tandem LTRs are precursors to integrated retrovirus DNA.

A number of properties of Ty and *Drosophila* copia-like elements suggest that a scheme like this is followed during their transposition. These transposable elements contain sequences which could serve as tRNA primer-binding sites adjacent to the 5′ LTR and oligo-purine tracts which act in priming second-strand synthesis during retrovirus replication, adjacent to the 3′ LTR (75). Both circular (19, 34, 66; K. G. Mossie and H. E. Varmus, J. Mol. Biol., in press) and linear (18) extrachromosomal forms of copia and other copia-like elements have been observed. Furthermore, sequence analysis of copia circles (which may be abortive products, as none appears to match precisely the tandem LTR structure expected of genuine transposition intermediates) has revealed unusual structures which closely resemble those seen with retroviral circles (20, 69). In addition, abundant virus-like particles containing copia RNA and reverse transcriptase have been observed in the nuclei of *Drosophila* tissue culture cells which have been kept in stationary phase for several days (67). Most recently, Boeke

* Corresponding author.

et al. have shown that Ty transposition in *S. cerevisiae* results in precise removal of an intron, demonstrating the existence of an RNA intermediate (5a).

A retrovirus genome consists of three genes required for a complete cycle of viral infection and replication (75). These are known as *gag*, *pol*, and *env*. *gag*, which is the most 5′, encodes a polyprotein with a molecular weight of 65,000 to 76,000 which is cleaved to form four or five small proteins found in the core of the virus particle. The *pol* gene is expressed as a *gag-pol* polyprotein of 160,000 to 180,000 molecular weight from which a *pol* protein bearing reverse transcriptase and integrase functions is released. The *pol* gene is encoded by a distinct open reading frame (ORF), so that production of the *gag-pol* polyprotein requires a splice, translational frameshift, or nonsense suppression. A similar arrangement of two overlapping ORFs is found in Ty (J. Clare and P. J. Farabaugh, Proc. Natl. Acad Sci. U.S.A., in press; 43b). *env*, the most 3′ retrovirus gene, is translated on membrane-bound ribosomes from a distinct spliced mRNA to form a polyprotein precursor to viral envelope proteins with a molecular weight of 60,000 to 68,000.

The white-apricot mutation, which arose spontaneously in 1923 (43), causes a change in eye color from the red of the wild type to orange-yellow. This allele of white was the first *Drosophila* mutation shown to be associated with the insertion of copia or a related element (4, 5, 25, 27). The mutant phenotype is apparently caused by the presence of copia in a small intron (49, 52), creating a 5-bp duplication of the sequence TAAAG (49). The two most abundant transcripts from the white-apricot allele (other than those which might be derived from sequences entirely within copia) appear to have 5′ termini corresponding to the 5′ end of wild-type white RNA and 3′ termini within one of the other copia LTRs. It is suspected that the residual expression of white is due to rare transcripts which read through copia and give rise to a low level of normally spliced RNA (42, 52). The white-apricot mutation is caused by an insertion within noncoding DNA and can partially revert by homologous recombination between the two LTRs (unpublished data). This is also true of Ty insertions at *HIS4* (56), gypsy insertions within the bithorax complex (3), an ecotropic murine retrovirus insertion at the dilute-coat-color (*d*) locus of the house mouse (8, 33), and a Moloney murine leukemia virus (MoMLV) insertion within a transforming Rous sarcoma virus (RSV) provirus (76). Interestingly, all of these mutations but the last are known to be affected by unlinked recessive suppressor mutations (31, 45, 72, 78).

Here we present the complete sequence of the copia element at the white-apricot allele. It contains one long ORF capable of encoding a polyprotein with several regions of homology to retroviral proteins, including good homology to a region of the *pol* gene recently shown to be involved in the integration of viral DNA to form proviruses. This homology further strengthens the idea that copia transposes by a mechanism like retroviral replication. However, the organization of these coding regions within copia is different from their organization in retroviruses and in the copia-like element 17.6. Rather, copia more closely resembles the Ty transposable elements of *S. cerevisiae*.

MATERIALS AND METHODS

M13 cloning and DNA sequencing. Copia was sequenced by the method of chain termination with dideoxyribonucleotide incorporation as described by Sanger et al. (60). The *Xho-Sal* fragment of lambda w^a5.9 (40) was first subcloned into the *Sal* site of pEMBL9 (10) to generate p3922a15. Digestion of p3922a15 with *Sac*I and *Xba* generated two 3.1-kilobase (kb) fragments, which together contained all of copia and 1.0 kb of white locus DNA. These were gel purified, ligated to form high-molecular-weight DNA in a mixed reaction, and sonicated. The ends of these fragments were repaired with bacteriophage T4 DNA polymerase (PL Biochemicals). Several size-fractionated pools (500 to 1,000 bp) were gel purified and cloned into *Sma*-cut M13mp19 (48). To obtain the sequence across the *Xba* site, separate M13 clones containing the 467-nucleotide *Eco*RI fragment in both orientations were also constructed. One hundred sixty-eight clones were analyzed. The sequence of the entire copia element, except for nucleotides 580 through 680, was determined for both strands. Nucleotides 580 through 680 were unambiguous on the strand sequenced, and the sequence obtained agreed with that previously published (21) for this region. The strains used were TG1 (gift of Toby Gibson, Medical Research Council, Cambridge, England) and DG98 (gift of David Gelfand, patent of Cetus Corp.). Sequencing reactions were performed with [^{35}S]dATP as the radioactive nucleotide and run on gradient gels as described by Bankier and Barrel (1).

Sequence assembly and analysis. The copia sequence was assembled from the sequence of random clones and analyzed on a Bion workstation (Intelligenetics). Parameters for searches with the SEARCH program were generally minmatch = 5, percentmatch = 20, aftermismatch = 1, and loopout = 0. This combination of search parameters generated an excess of output, which was then analyzed as described below. Parameters for Needleman-Wunch homology searches (47) with the ALIGN program were chosen to suit weakly homologous sequences: mismatch penalty = 0, gappenalty = 2, gapsizepenalty = 0.1, overlap = 50%, segmentsize = 40, and minalign = 1. This program was used primarily to generate the alignments shown in Fig. 3 and 4. The sequence of the putative translation product of the large ORF was submitted to the National Biomedical Research Foundation Protein Identification Resource for comparison with their protein sequence database with the program FASTP (43a).

RESULTS AND DISCUSSION

General features. The nucleotide sequence of the copia element at the white-apricot allele is shown in Fig. 1. This element was 5,146 nucleotides long and rich in A and T, containing 36.4% A, 30.6% T, 18.9% G, and 14.1% C residues. Major structural features included the 276-nucleotide LTRs and an internal 108-nucleotide tandem repeat. The two LTRs were identical to each other and to the LTRs on a previously partially sequenced copia element (that in cDm 2056 [41]). In fact, there were only two differences between these two copias in the 1,130-nucleotide stretch reported by Flavell et al. (21) (Fig. 1). One of these was a single-base substitution 16 nucleotides downstream from the 5′ LTR, in a region which would be the site of binding of a tRNA primer if copia indeed transposes by a mechanism similar to the mechanism of retroviral replication. The other difference between these two copias was an insertion or deletion at nucleotide 575, which created a translational stop in the cDm 2056 copia that was not present in the white-apricot copia. A slightly greater divergence (but still only about 1%) was observed between the copia at white-apricot and that in clone DM311 in the 633-nucleotide region sequenced by Fouts et al. (23), in which there were eight changes (Fig. 1, nucleotides 2302 through 2934). Both of these comparisons are consistent with the observed

```
                                                                          70
TGTTGGAATA TACTATTCAA CCTACAAAAA TAACGTTAAA CAACACTACT TTATATTTGA TATGAATGGC

CACACCTTTT ATGCCATAAA ACATATTGTA AGAGAATACC ACTCTTTTTA TTCCTTCTTT CCTTCTTGTA
                                                                         210
CGTTTTTTGC TGTGAGTAGG TCGTGGTGCT GGTGTTGCAG TTGAAATAAC TTAAAATATA AATCATAAAA
                                                                         280
CTCAAACATA AACTTGACTA TTTATTTATT TATTAAGAAA GGAAATATAA ATTATAAATT ACAACAGGTT
                                                                  ←5' LTR
            G                                                            350
ATGGGCCCAG TCCATGCCTA ATAAACAATT AAATTGTGAA TTAAAGATTG TGAAAATAAA TTGTGAAATA
                                                                         420
GCATTTTTTC ACATTCTTGT GAAATAGCTT TTTTTTTCAC ATTCTTGTGA AATTATTTCC TTCTCAGAAT

TTGAGTGAAA A ATG GAC AAG GCT AAA CGT AAT ATT AAG CCG TTT GAT GGC GAG AAG 476
             MET Asp Lys Ala Lys Arg Asn Ile Lys Pro Phe Asp Gly Glu Lys  15

TAC GCG ATT TGG AAA TTT AGA ATT AGG GCT CTT TTA GCC GAG CAA GAT GTG CTT  530
Tyr Ala Ile Trp Lys Phe Arg Ile Arg Ala Leu Leu Ala Glu Gln Asp Val Leu   33
                                                     -
AAA GTA GTT GAT GGT TTA ATG CCT AAC GAG GTA GAT GAC TCC TGG AAA AAG GCA  584
Lys Val Val Asp Gly Leu MET Pro Asn Glu Val Asp Asp Ser Trp Lys Lys Ala   51

GAG CGT TGT GCA AAA AGT ACA ATA ATA GAG TAC CTA AGC GAC TCG TTT TTA AAT  638
Glu Arg Cys Ala Lys Ser Thr Ile Ile Glu Tyr Leu Ser Asp Ser Phe Leu Asn   69

TTC GCA ACA AGC GAC ATT ACG GCG CGT CAG ATT CTT GAG AAT TTG GAC GCC GTT  692
Phe Ala Thr Ser Asp Ile Thr Ala Arg Gln Ile Leu Glu Asn Leu Asp Ala Val   87

TAT GAA CGA AAA AGT TTG GCG TCG CAA CTG GCG CTG CGA AAA CGT TTG CTT TCT  746
Tyr Glu Arg Lys Ser Leu Ala Ser Gln Leu Ala Leu Arg Lys Arg Leu Leu Ser  105

CTG AAG CTA TCG AGT GAG ATG TCA CTA TTA AGC CAT TTT CAT ATT TTT GAC GAA  800
Leu Lys Leu Ser Ser Glu MET Ser Leu Leu Ser His Phe His Ile Phe Asp Glu  123

CTT ATA AGT GAA TTG TTG GCA GCT GGT GCA AAA ATA GAA GAG ATG GAT AAA ATT  854
Leu Ile Ser Glu Leu Leu Ala Ala Gly Ala Lys Ile Glu Glu MET Asp Lys Ile  141

TCT CAT CTA CTG ATC ACA TTG CCT TCG TGT TAC GAT GGA ATT ATT ACA GCG ATA  908
Ser His Leu Leu Ile Thr Leu Pro Ser Cys Tyr Asp Gly Ile Ile Thr Ala Ile  159

GAG ACA TTA TCT GAA GAA AAT TTG ACA TTG GCG TTT GTG AAA AAT AGA TTG CTG  962
Glu Thr Leu Ser Glu Glu Asn Leu Thr Leu Ala Phe Val Lys Asn Arg Leu Leu  177

GAT CAA GAA ATT AAA ATT AAA AAT GAC CAC AAC GAT ACA AGC AAG AAA GTT ATG 1016
Asp Gln Glu Ile Lys Ile Lys Asn Asp His Asn Asp Thr Ser Lys Lys Val MET  195

AAC GCG ATC GTG CAC AAC AAT AAT AAC ACT TAT AAA AAT AAT TTG TTT AAA AAT 1070
Asn Ala Ile Val His Asn Asn Asn Asn Thr Tyr Lys Asn Asn Leu Phe Lys Asn  213

CGG GTA ACT AAA CCA AAG AAA ATA TTC AAG GGA AAT TCA AAG TAT AAA GTC AAG 1124
Arg Val Thr Lys Pro Lys Lys Ile Phe Lys Gly Asn Ser Lys Tyr Lys Val Lys  231

TGT CAC CAC TGT GGC AGA GAA GGC CAC ATT AAA AAA GAT TGT TTC CAT TAT AAA 1178
Cys His His Cys Gly Arg Glu Gly His Ile Lys Lys Asp Cys Phe His Tyr Lys  249

AGA ATA TTA AAT AAT AAA AAT AAA GAA AAT GAA AAA CAA GTT CAA ACT GCA ACA 1232
Arg Ile Leu Asn Asn Lys Asn Lys Glu Asn Glu Lys Gln Val Gln Thr Ala Thr  267

TCA CAC GGC ATT GCG TTT ATG GTA AAA GAA GTG AAT AAT ACT TCA GTG ATG GAC 1286
Ser His Gly Ile Ala Phe MET Val Lys Glu Val Asn Asn Thr Ser Val MET Asp  285

AAC TGC GGG TTT GTC CTT GAT TCT GGT GCT AGT GAC CAT CTT ATA AAT GAT GAG 1340
Asn Cys Gly Phe Val Leu Asp Ser Gly Ala Ser Asp His Leu Ile Asn Asp Glu  303

TCG CTG TAT ACC GAC AGT GTG GAG GTT GTG CCT CCA CTT AAG ATT GCA GTG GCC 1394
Ser Leu Tyr Thr Asp Ser Val Glu Val Val Pro Pro Leu Lys Ile Ala Val Ala  321

AAG CAA GGC GAA TTT ATT TAT GCC ACT AAG CGT GGT ATT GTC CGA CTA CGG AAT 1448
Lys Gln Gly Glu Phe Ile Tyr Ala Thr Lys Arg Gly Ile Val Arg Leu Arg Asn  339

GAC CAT GAG ATT ACA CTG GAG GAT GTA CTC TTT TGT AAG GAA GCT GCT GGT AAT 1502
Asp His Glu Ile Thr Leu Glu Asp Val Leu Phe Cys Lys Glu Ala Ala Gly Asn  357

TTG ATG TCC GTA AAG CGT CTC CAA GAG GCA GGA ATG TCG ATC GAA TTT GAC AAA 1556
Leu MET Ser Val Lys Arg Leu Gln Glu Ala Gly MET Ser Ile Glu Phe Asp Lys  375
                                                                 *
AGC GGT GTA ACC ATT TCG AAA AAT GGG TTA ATG GTT GTC AAA AAT TCA GGT ATG 1610
Ser Gly Val Thr Ile Ser Lys Asn Gly Leu MET Val Val Lys Asn Ser Gly MET  393

TTA AAC AAT GTA CCT GTG ATC AAT TTT CAA GCA TAT TCT ATA AAT GCT AAG CAT 1664
Leu Asn Asn Val Pro Val Ile Asn Phe Gln Ala Tyr Ser Ile Asn Ala Lys His  411

AAA AAT AAT TTT CGT TTA TGG CAT GAG AGG TTT GGC CAT ATA AGC GAT GGC AAA 1718
Lys Asn Asn Phe Arg Leu Trp His Glu Arg Phe Gly His Ile Ser Asp Gly Lys  429

TTA TTA GAA ATA AAA CGA AAG AAT ATG TTT AGT GAT CAA AGT CTT CTA AAC AAC 1772
Leu Leu Glu Ile Lys Arg Lys Asn MET Phe Ser Asp Gln Ser Leu Leu Asn Asn  447

TTA GAG TTA TCA TGT GAA ATT TGT GAA CCC TGT TTA AAT GGT AAA CAG GCA AGA 1826
Leu Glu Leu Ser Cys Glu Ile Cys Glu Pro Cys Leu Asn Gly Lys Gln Ala Arg  465

CTT CCT TTT AAA CAA TTG AAA GAT AAG ACC CAT ATT AAA AGA CCA CTT TTT GTA 1880
Leu Pro Phe Lys Gln Leu Lys Asp Lys Thr His Ile Lys Arg Pro Leu Phe Val  483

GTA CAC TCA GAT GTC TGT GGG CCT ATT ACT CCA GTT ACT TTA GAT GAT AAA AAT 1934
Val His Ser Asp Val Cys Gly Pro Ile Thr Pro Val Thr Leu Asp Asp Lys Asn  501

TAT TTT GTG ATC TTT GTT GAT CAG TTT ACA CAT TAT TGT GTA ACT TAT TTA ATT 1988
Tyr Phe Val Ile Phe Val Asp Gln Phe Thr His Tyr Cys Val Thr Tyr Leu Ile  519

AAA TAT AAA TCT GAT GTG TTT AGC ATG TTT CAA GAT TTT GTA GCC AAG AGT GAA 2042
Lys Tyr Lys Ser Asp Val Phe Ser MET Phe Gln Asp Phe Val Ala Lys Ser Glu  537

GCT CAT TTT AAT TTA AAG GTT GTG TAC TTA TAC ATT GAC AAT GGT AGA GAA TAC 2096
Ala His Phe Asn Leu Lys Val Val Tyr Leu Tyr Ile Asp Asn Gly Arg Glu Tyr  555

TTG TCA AAT GAG ATG AGA CAA TTT TGT GTT AAG AAA GGA ATT TCT TAT CAC TTA 2150
Leu Ser Asn Glu MET Arg Gln Phe Cys Val Lys Lys Gly Ile Ser Tyr His Leu  573

ACA GTG CCA CAT ACA CCT CAG TTA AAT GGT GTT TCT GAG AGA ATG ATA AGA ACC 2204
Thr Val Pro His Thr Pro Gln Leu Asn Gly Val Ser Glu Arg MET Ile Arg Thr  591

ATT ACG GAA AAA GCT CGA ACC ATG GTT AGT GGT GCA AAG CTA GAT AAA AGC TTT 2258
Ile Thr Glu Lys Ala Arg Thr MET Val Ser Gly Ala Lys Leu Asp Lys Ser Phe  609
                                                                  T
TGG GGC GAA GCA GTA TTA ACT GCT ACT TAT TTA ATC AAC AGA ATT CCT AGC AGA 2312
Trp Gly Glu Ala Val Leu Thr Ala Thr Tyr Leu Ile Asn Arg Ile Pro Ser Arg  627

GCA CTT GTT GAT AGT TCA AAG ACC CCA TAT GAG ATG TGG CAC AAT AAG AAG CCA 2366
Ala Leu Val Asp Ser Ser Lys Thr Pro Tyr Glu MET Trp His Asn Lys Lys Pro  645

TAC TTA AAA CAT TTG AGA GTG TTT GGT GCA ACT GTT TAT GTG CAT ATT AAA AAC 2420
Tyr Leu Lys His Leu Arg Val Phe Gly Ala Thr Val Tyr Val His Ile Lys Asn  663

AAA CAA GGA AAG TTT GAT GAT AAA TCA TTT AAA AGT ATT TTT GTG GGC TAT GAA 2474
Lys Gln Gly Lys Phe Asp Asp Lys Ser Phe Lys Ser Ile Phe Val Gly Tyr Glu  681

CCC AAT GGT TTT AAG TTG TGG GAT GCT GTA AAT GAA AAA TTT ATT GTC GCA AGA 2528
Pro Asn Gly Phe Lys Leu Trp Asp Ala Val Asn Glu Lys Phe Ile Val Ala Arg  699

GAT GTT GTT GTC GAT GAA ACC AAT ATG GTT AAT TCT AGA GCT GTT AAA TTT GAA 2582
Asp Val Val Val Asp Glu Thr Asn MET Val Asn Ser Arg Ala Val Lys Phe Glu  717
                 C
ACA GTG TTC CTG AAA GAT AGT AAG GAA AGT GAA AAT AAA AAT TTT CCG AAT GAC 2636
Thr Val Phe Leu Lys Asp Ser Lys Glu Ser Glu Asn Lys Asn Phe Pro Asn Asp  735

AGT AGG AAA ATA ATA CAA ACA GAG TTC CCG AAT GAG AGT AAG GAA TGC GAC AAC 2690
Ser Arg Lys Ile Ile Gln Thr Glu Phe Pro Asn Glu Ser Lys Glu Cys Asp Asn  753

ATA CAA TTC CTG AAA GAT AGT AAG GAA AGT GAA AAT AAA AAT TTT CCG AAT GAC 2744
Ile Gln Phe Leu Lys Asp Ser Lys Glu Ser Glu Asn Lys Asn Phe Pro Asn Asp  771

AGT AGG AAA ATA ATA CAA ACA GAA TTC CCG AAT GAG AGT AAG GAA TGC GAC AAC 2798
Ser Arg Lys Ile Ile Gln Thr Glu Phe Pro Asn Glu Ser Lys Glu Cys Asp Asn  789
                                                    A
ATA CAA TTC CTG AAA GAT AGT AAG GAA AGT AAT AAA TAT TTT CTG AAT GAG AGT 2852
Ile Gln Phe Leu Lys Asp Ser Lys Glu Ser Asn Lys Tyr Phe Leu Asn Glu Ser  807
                            A                                   GG
AAG AAA AGA AAG CGA GAT GAT CAC CTG AAT GAA AGT AAG GGA TCA GGC AAC CCG 2906
Lys Lys Arg Lys Arg Asp Asp His Leu Asn Glu Ser Lys Gly Ser Gly Asn Pro  825
                          G G
AAT GAG AGT AGG GAA AGT GAA ACA GCA GAG CAC TTA AAA GAA ATT GGA ATT GAT 2960
Asn Glu Ser Arg Glu Ser Glu Thr Ala Glu His Leu Lys Glu Ile Gly Ile Asp  843

AAT CCA ACT AAA AAT GAT GGC ATA GAA ATT ATT AAT AGA AGA AGT GAG AGA TTA 3014
Asn Pro Thr Lys Asn Asp Gly Ile Glu Ile Ile Asn Arg Arg Ser Glu Arg Leu  861

AAG ACT AAG CCT CAG ATA TCC TAT AAT GAA GAG GAT AAC AGT CTA AAT AAA GTT 3068
Lys Thr Lys Pro Gln Ile Ser Tyr Asn Glu Glu Asp Asn Ser Leu Asn Lys Val  879

GTT CTA AAT GCT CAC ACT ATA TTT AAC GAT GTC CCA AAT TCA TTT GAT GAA ATT 3122
Val Leu Asn Ala His Thr Ile Phe Asn Asp Val Pro Asn Ser Phe Asp Glu Ile  897

CAA TAT AGG GAT GAT AAA TCT TCT TGG GAA GAA GCC ATC AAT ACA GAG TTA AAT 3176
Gln Tyr Arg Asp Asp Lys Ser Ser Trp Glu Glu Ala Ile Asn Thr Glu Leu Asn  915

GCT CAT AAA ATT AAT AAT ACT TGG ACA ATT ACA AAA AGG CCT GAA AAC AAA AAT 3230
Ala His Lys Ile Asn Asn Thr Trp Thr Ile Thr Lys Arg Pro Glu Asn Lys Asn  933

ATT GTA GAT AGC AGA TGG GTA TTT TCT GTT AAA TAT AAT GAA CTT GGA AAT CCA 3284
Ile Val Asp Ser Arg Trp Val Phe Ser Val Lys Tyr Asn Glu Leu Gly Asn Pro  951

ATT AGA TAC AAA GCT AGA TTG GTT GCA CGA GGA TTC ACT CAA AAA TAC CAA ATA 3338
Ile Arg Tyr Lys Ala Arg Leu Val Ala Arg Gly Phe Thr Gln Lys Tyr Gln Ile  969

GAC TAT GAA GAG ACA TTT GCT CCT GTA GCT AGA ATT TCA AGT TTC CGA TTT ATA 3392
Asp Tyr Glu Glu Thr Phe Ala Pro Val Ala Arg Ile Ser Ser Phe Arg Phe Ile  987

TTG TCA TTA GTA ATA CAG TAT AAC TTG AAA GTC CAT CAA ATG GAT GTA AAA ACA 3446
Leu Ser Leu Val Ile Gln Tyr Asn Leu Lys Val His Gln MET Asp Val Lys Thr 1005

GCT TTC TTA AAT GGC ACG TTA AAA GAG GAA ATT TAT ATG AGA CTT CCT CAA GGT 3500
Ala Phe Leu Asn Gly Thr Leu Lys Glu Glu Ile Tyr MET Arg Leu Pro Gln Gly 1023

ATA TCG TGT AAT AGT GAC AAT GTG TGT AAA TTG AAT AAG GCA ATT TAC GGA CTC 3554
Ile Ser Cys Asn Ser Asp Asn Val Cys Lys Leu Asn Lys Ala Ile Tyr Gly Leu 1041

AAG CAA GCG GCT AGA TGC TGG TTT GAA GTA TTT GAG CAA GCA TTG AAA GAG TGT 3608
Lys Gln Ala Ala Arg Cys Trp Phe Glu Val Phe Glu Gln Ala Leu Lys Glu Cys 1059

GAG TTT GTA AAC TCT TCA GTT GAT CGC TGT ATA TAT ATT TTA GAC AAA GGT AAC 3666
Glu Phe Val Asn Ser Ser Val Asp Arg Cys Ile Tyr Ile Leu Asp Lys Gly Asn 1077

ATC AAT GAA AAC ATA TAT GTA TTA TTA TAT GTA GAT GAT GTG GTT ATA GCT ACA 3716
Ile Asn Glu Asn Ile Tyr Val Leu Leu Tyr Val Asp Asp Val Val Ile Ala Thr 1095

GGA GAT ATG ACA AGA ATG AAT AAC TTC AAA AGG TAT TTA ATG GAA AAG TTT AGG 3770
Gly Asp MET Thr Arg MET Asn Asn Phe Lys Arg Tyr Leu MET Glu Lys Phe Arg 1113

ATG ACT GAC CTA AAT GAA ATA AAA CAT TTT ATT GGA ATT AGG ATA GAG ATG CAG 3824
MET Thr Asp Leu Asn Glu Ile Lys His Phe Ile Gly Ile Arg Ile Glu MET Gln 1131

GAA GAT AAA ATC TAT TTA AGC CAA TCT GCA TAT GTT AAA AAA ATT TTA AGT AAA 3878
Glu Asp Lys Ile Tyr Leu Ser Gln Ser Ala Tyr Val Lys Lys Ile Leu Ser Lys 1149

TTT AAC ATG GAA AAT TGT AAT GCA GTT AGT ACT CCT TTA CCT AGT AAA ATA AAT 3932
Phe Asn MET Glu Asn Cys Asn Ala Val Ser Thr Pro Leu Pro Ser Lys Ile Asn 1167

TAT GAA TTA CTT AAT TCA GAT GAA GAC TGC AAT ACC CCA TGC CGT AGC CTC ATA 3986
Tyr Glu Leu Leu Asn Ser Asp Glu Asp Cys Asn Thr Pro Cys Arg Ser Leu Ile 1185

GGA TGT TTA ATG TAC ATA ATG CTT TGT ACA CGC CCA GAT TTA ACT ACT GCA GTA 4040
Gly Cys Leu MET Tyr Ile MET Leu Cys Thr Arg Pro Asp Leu Thr Thr Ala Val 1203

AAT ATC TTG AGC AGA TAT AGT AGC AAA AAT AAC TCC GAA TTA TGG CAG AAC TTA 4094
Asn Ile Leu Ser Arg Tyr Ser Ser Lys Asn Asn Ser Glu Leu Trp Gln Asn Leu 1221

AAA AGA GTT CTT AGA TAT TTG AAG GGC ACT ATC GAT ATG AAA TTG ATT TTT AAA 4148
Lys Arg Val Leu Arg Tyr Leu Lys Gly Thr Ile Asp MET Lys Leu Ile Phe Lys 1239

AAG AAC TTG GCA TTT GAA AAT AAA ATT ATT GGT TAT GTG GAT TCT GAT TGG GCT 4202
Lys Asn Leu Ala Phe Glu Asn Lys Ile Ile Gly Tyr Val Asp Ser Asp Trp Ala 1257

GGT AGT GAA ATT GAT AGA AAA AGT ACA ACA GGG TAT TTA TTC AAA ATG TTT GAT 4256
Gly Ser Glu Ile Asp Arg Lys Ser Thr Thr Gly Tyr Leu Phe Lys MET Phe Asp 1275

TTT AAT CTC ATT TGT TGG AAT ACA AAG AGA CAG AAC TCA GTA GCA GCC TCA TCA 4310
Phe Asn Leu Ile Cys Trp Asn Thr Lys Arg Gln Asn Ser Val Ala Ala Ser Ser 1293

ACT GAA GCT GAG TAT ATG GCC CTA TTT GAA GCC GTG AGA GAA GCT CTA TGG CTT 4364
Thr Glu Ala Glu Tyr MET Ala Leu Phe Glu Ala Val Arg Glu Ala Leu Trp Leu 1311

AAA TTT TTA TTA ACT AGT ATT AAC ATT AAA CTA GAA AAC CCC ATT AAA ATT TAC 4418
Lys Phe Leu Leu Thr Ser Ile Asn Ile Lys Leu Glu Asn Pro Ile Lys Ile Tyr 1329

GAA GAC AAT CAA GGC TGT ATT AGC ATA GCA AAC AAT CCC TCA TGT CAT AAA CGA 4472
Glu Asp Asn Gln Gly Cys Ile Ser Ile Ala Asn Asn Pro Ser Cys His Lys Arg 1347

GCT AAA CAT ATT GAT ATT AAA TAT CAT TTT GCC AGA GAG CAA GTT CAG AAT AAT 4526
Ala Lys His Ile Asp Ile Lys Tyr His Phe Ala Arg Glu Gln Val Gln Asn Asn 1365

GTG ATT TGT CTT GAG TAT ATT CCT ACA GAG AAT CAA CTG GCT GAC ATA TTT ACA 4580
Val Ile Cys Leu Glu Tyr Ile Pro Thr Glu Asn Gln Leu Ala Asp Ile Phe Thr 1383

AAA CCG TTG CCT GCT GCG AGA TTT GTG GAG TTA CGA GAC AAA TTG GGT TTG CTG 4634
Lys Pro Leu Pro Ala Ala Arg Phe Val Glu Leu Arg Asp Lys Leu Gly Leu Leu 1401
                                                                            4701
CAA GAC GAC CAA TCG AAT GCT GAA TGA AATTTTTTAT ATATATTTTT CAAATTTAAA TTCCTGTAAA
Gln Asp Asp Gln Ser Asn Ala Glu     1409

CATATTTTGT TACAATGATC TGATCGGGTT TTTCTGGGTT TTCCCCGTAT CCTCGCA GCA AAT  4764
                                                               Ala Asn   393

GCT GGA TCA GTT AAC ACT TCC CAG AAT GCA CAC CAC CCA CAT TTG ATA GTT ACT 4818
Ala Gly Ser Val Asn Thr Ser Gln Asn Ala His His Pro His Leu Ile Val Thr  411
                                                                  3' LTR→
AAT GAA TAT TAT TGT TAT GTT TTT AAT TAT AGA CGT TAT TTT TGA GGGGGCGTGT  4873
Asn Glu Tyr Tyr Cys Tyr Val Phe Asn Tyr Arg Arg Tyr Phe      425
                                                                         4943
TGGAATATAC TATTCAACCT ACAAAAATAA CGTTAAACAA CACTACTTTA TATTTGATAT GAATGGCCAC
                                                                         5013
ACCTTTTATG CCATAAAACA TATTGTAAGA GAATACCACT CTTTTTATTC CTTCTTTCCT TCTTGTACGT
                                                                         5083
TTTTTGCTGT GAGTAGGTCG TGGTGCTGGT GTTGCAGTTG AAATAACTTA AAATATAAAT CATAAAACTC
                                                                   5146
AAACATAAAC TTGACTATTT ATTTATTTAT TAAGAAAGGA AATATAAATT ATAAATTACA ACA
```

FIG. 1. Sequence of the copia insertion at the white-apricot allele. The strand having the same polarity as the copia RNAs (63) is shown. The translation of the long ORF from the methionine codon at nucleotide 432 to the termination codon at nucleotide 4659 and a shorter ORF from a potential 3′ splice site (*) at 4760 to the termination codon at 4861 is shown on the bottom line. Nucleotides and amino acids are numbered separately, and the amino acids in the second ORF are numbered based on the conjectured splice between nucleotides 1605 and 4760 (*). Nucleotides between positions 1 and 1130 which differed between this copia and that in cDm 2056, partially sequenced by Flavell et al. (21), are shown above the line, as are nucleotides between 2302 and 2934 which differed from those in a copia partially sequenced by Fouts et al. (23). The site of copia RNA 5′ termini is indicated by a bold overline (21). The seven occurrences of TGTGAA or its complement, TTCACA, between positions 310 and 410 are underlined.

homogeneity of copia elements at the level of conservation of restriction sites (53). The possibility that mutations may have arisen during the process of transposition or after means that any particular copia element, including the one whose sequence is discussed in this paper, may not be capable of transposition. This fact makes the differences between these copias difficult to interpret.

The most abundant RNAs encoded by copia are a 5-kb RNA, which runs from the 5′ LTR to the 3′ LTR, and a 2-kb RNA (6, 16, 22, 63). No information exists about the 3′ ends of either RNA, except that the 5-kb RNA is known to extend into the 3′ LTR (63). The heterogeneous 5′ ends of these two RNAs are shared and map in the region between nucleotides 127 and 147 (21). Neither of these RNAs is spliced between the 5′ end and the *Pvu*II site at nucleotide 820 (21). The first potential initiation codon in both of these RNAs was the ATG at nucleotide 294. The sequence flanking this ATG did not conform to the consensus sequence for efficiently used initiation codons (36), and initiation here would lead to termination after the translation of only 17 amino acids. The second ATG (nucleotide 432), which did conform to the consensus for efficiently used initiation codons, was the second codon of an ORF extending from nucleotide 429 to 4658. Translation of this ORF would give rise to a 1,409-amino-acid translation product of 163,000 daltons. We consider it likely that this large ORF is indeed translated.

Possible regulatory site. The sequence between the 5′ LTR and the beginning of the large ORF had some interesting features. In particular, the 22-nucleotide sequence TTTTTTCACATTCTTGTGAAAT occurred twice, at nucleotide positions 354 to 375 and 382 to 403. This sequence displayed dyad symmetry (underlined), a property frequently observed in sites recognized by DNA-binding proteins. Furthermore, sequences identical to one half of this symmetrical stretch, TGTGAAA, occurred at positions 329 to 335 and 342 to 348. A similar sequence, TGTGGAAA, occurs in the simian virus 40 enhancer, and mutations in these nucleotides of the simian virus 40 enhancer have been shown to interfere with enhancer function (77). The presence of an enhancer-like activity at a similar position within Ty was suggested by the observation that a single-base change in this region within a Ty inserted upstream of *HIS4* significantly affects the influence of Ty on *HIS4* (R. Pearlman, A. Rose, and S. Roeder, personal communication). Interestingly, the region of Ty involved is similar to the same region of the simian virus 40 enhancer. Whether this 22-nucleotide repeated sequence within copia is indeed an enhancer is a question for further study. It seems likely that such a sequence plays some role in the expression of copia.

***gag* and reverse transcriptase homologies.** The putative protein encoded by the large ORF in copia was compared with retrovirus gene products (62, 65, 68) and putative proteins encoded by ORFs present within Ty912 (Clare and Farabaugh, in press) and 17.6 (59). The first set of comparisons was done with a computer program (SEARCH) capable of finding short stretches of strong homology, but incapable of looking past mismatched regions of more than a few amino acids to find the most meaningful alignments. Homologies are judged to be meaningful when a sequence element is shared by three or more of the proteins involved. One example (Fig. 2) was the homology between amino acids 232 through 245 of the copia ORF and a 14-amino-acid region conserved among retroviral nucleic acid-binding proteins (9). These small proteins are products of the *gag* polyprotein and are found bound to RNA within the virion. An absolutely conserved spacing of three cysteine residues observed in all of these proteins was associated with a clustering of other well-conserved amino acids, several of which were seen in the copia sequence. This sequence cannot be found in Ty or 17.6.

The sequence Leu-Asp-Ser-Gly-Ala occurs in RSV, copia, and Ty (Fig. 2). The sequence occurs at a comparable location in the three genomes, and the relative spacing between this homology and the nucleic acid-binding homology is similar in RSV and copia. In RSV this block lies within p15, a protease with a molecular weight of 15,000 that is involved in cleaving viral polyproteins. It is significant that this pentapeptide is centered around the tripeptide Asp-Ser-Gly, which is present at the active site of enzymes in the trypsin-protease family (35). This homology supports the idea that the translation product of the 1,409-amino-acid ORF is a polyprotein by identifying a potential copia-encoded function capable of processing the polyprotein into smaller functional proteins.

```
NUCLEIC ACID BINDING

MoMLV (504)      CAYCKEKGHWAKDC
HTLV  (357)      CFACGKAGHWSADC...
      (380)      CPLCQDPTHWKADC
RSV   (509)      CYTCGSPGHYQAQC...
      (535)      CQLCNGMGHNAKQC
                 *  **  **  ***
Copia (232)      CHHCGREGHIKKDC

PROTEASE

RSV     (612)    LLDSGADITII
                 ******   **
Copia   (290)    VLDSGASDHLI
                 *******  **
Ty-ORF2 (32)     LLDSGASRTLI

POLYMERASE

HTLV  (152)  VLPQGF.25.TILQYVDDILLASP
RSV   (145)  VLPQMT.24.CMLHYMDDLLLAAS
MoMLV (307)  ALPQGF.25.ILLQYVDDLLLAAT
             *****      ** ********
Copia (1019) ALPQGI.58.YVLLYVDDVVIATG
```

FIG. 2. Homologies to short stretches of conserved amino acids found in retroviruses. In each case, agreement between the putative copia protein and the retrovirus proteins is indicated by an asterisk. Numbers in parentheses indicate the amino acid number in the appropriate protein or ORF of MoMLV (68), human T-cell leukemia virus (HTLV) (65), RSV (62), Ty (Clare and Farabaugh, in press), or copia. For this figure only, agreement means either identity or interconversion between the long-chain hydrophobic residues leucine, isoleucine, and valine. Amino acids: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; and Y, tyrosine.

The most basic requirement of an RNA-based transposition mechanism for copia is that copia encode, or at least make use of, an RNA-templated DNA polymerase (reverse transcriptase). A number of sequence features conserved among a wide variety of reverse transcriptases have recently been identified (74; R. Patarca and W. A. Haseltine, Letter, Nature [London] **309**:288, 1984), and one of the most highly conserved was found at amino acids 1019 through 1024 of the copia ORF (Fig. 2). A second conserved block in this region (Tyr-Met or Val-Asp-Asp followed by three strictly hydrophobic long-chain amino acids and an alanine) was found at amino acids 1087 through 1094 of the copia ORF. Sequences related to this homology core are found not only in reverse transcriptases, but also in RNA-templated virus and phage polymerases encoded by viruses as diverse as influenza virus, tobacco mosaic virus, and bacteriophage MS2 (37).

Integrase homology. A comparison of the putative copia polyprotein sequence with the National Protein Information Resource protein sequence database revealed matches between amino acids 547 through 594 of the copia ORF and sequences in the carboxy-terminal portions of MoMLV (67), AKV murine leukemia virus (32), squirrel monkey retrovirus (7), and simian sarcoma virus (11) as the four highest-scoring homologies. These amino acids were part of a larger 180-amino-acid domain recognized by Chiu et al. (7) as the portion of the *pol* gene that is best conserved among a variety of vertebrate retroviruses. Figure 3 presents a comparison of the MoMLV *pol* gene, the copia ORF, and the second Ty ORF in this region. Copia shared 27 amino acids with MoMLV and 39 amino acids with Ty in the 150-amino-acid stretch shown. Although this was only 18 and 26% agreement, respectively, the agreement between copia and MoMLV was 8.9 standard deviations above the mean score for agreement between randomized sequences of the same composition.

This well-conserved domain is present in a distinct 32,000-dalton protein (p32) produced by cleavage of one subunit of the avian retrovirus *pol* dimer (26, 55). Analysis of the activities of p32 have shown that it has endonuclease activity (29); it also binds the LTRs (44) and can introduce single-strand breaks near the boundaries of the LTRs (13, 29). These properties are those expected of an integrase, the putative enzyme responsible for directing the joining of circular retrovirus DNA to chromosomal DNA to form a provirus. The idea that this domain acts as an integrase has been studied by using site-directed mutagenesis in three laboratories. Transfecting cells with DNA containing a mutation of Arg-905 in the MoMLV *pol* gene (corresponding to Arg-479 in the copia ORF) led to the release of virus particles capable of producing an infection in which all DNA forms arose normally, but no integrated proviruses were found (12). Similar results have been reported for a deletion of 91 bp which would be expected to remove all amino acids 3′ of Ala-889 from the MoMLV *pol* protein (64) and for a mutant carrying a mutation in this region of the spleen necrosis virus *pol* gene (51). These are precisely the results expected for mutations in an integrase function. They at least establish that this portion of the *pol* gene plays an essential role in retrovirus infection which is independent of reverse transcription. We therefore conclude that the strongest homology between copia and retroviruses is to an integrase.

Genome organization. Two domains of roughly 25% amino acid identity between copia and Ty proteins lie at amino acids 429 to 667 and 903 to 1397 in the copia ORF, and we argue that these correspond to integrase (Fig. 3) and reverse transcriptase (Fig. 2 and 4) functions, respectively. The assumption of a rough conservation, of both domain sizes and the location of conserved features within each domain, generates a remarkably good correspondence between the two domains of conservation between copia and Ty and the two peptides generated by cleavage of the avian retrovirus *pol* protein. The order of these two domains in retrovirus *pol* genes is 5′-reverse transcriptase-integrase-3′. However, there was a different order in copia and Ty: 5′-integrase-?-reverse transcriptase-3′ (Fig. 5).

```
RVcon  (50)     *    **   *   *  *  *              *  **  *
MoMLV  (898) VKQGTAV-AG HAPGTHWEID FTEIKPGLYG YKYLLVFIDT FSGWIEAF-P
              *  *   *                    *     *   * *  *
Copia  (471) LKDKTHIKAP LFVVHSDVCG PI-TPVTLDD KNYFVIFVDQ FTHYCVTY-L
             **       *     * *  * *            **   * *   *     * *
TyORF  (229) LKYQNSY-EP FQYLHTDIFG PV-HNLPNSA PSYFISFTDE TTKFAWVYPL


RVcon  (97)                            * *     ***
MoMLV  (946) TKKETA---K VVTKKLLEEI FPAFGMPQVL GTDNGPAFVS KVSQTVADLL
              *                              *   ***    *
Copia  (519) IKYKSD-VFS MFQDFVAKSE AHFNLKVVYL YIDNGAEYLS NEMAQFCVKK
                  *      *    *      *   *      * * **        *  *
TyORF  (277) HDAAEDSILD VFTTILAFIK NQFQASVLVI QMDAGSEYTN ATLHKFLEKN


RVcon  (143)  *                ***      *
MoMLV  (993) GIDWKLHCAY APQSSGQVER MNATIKETLT KLTLATGSAD WVLLLPLALY
             **   *      **  *  ** * *** *
Copia  (568) GISYHLTVPH TPQLNGVSEA MIATITEKAA TMVSGAKLDK SFWGEAVLTA
             **    *         ** **   **     * *      *     *  *
TyORF  (327) GITPCYTTTA DSAAHGVAEA LNATLLDDCA TQLQCSGLPN YLWFSAIEFS
```

FIG. 3. Homology to the integrase region of the *pol* gene. Top line (labeled RVcon) indicates (*) amino acid positions which are invariant between homologous regions of the *pol* genes of MoMLV (68) (second line), human T-cell leukemia virus (65), squirrel monkey retrovirus (7), mouse mammary tumor virus (7, 54), and RSV (62). Identity between amino acids in the copia and MoMLV sequences and the copia and Ty (Clare and Farabaugh, in press) sequences are also indicated. Numbers in parentheses indicate the amino acid number in Fig. 5 of reference 7, the *pol* gene of MoMLV, the large copia ORF, or the second ORF in Ty912. For amino acid abbreviations, see Fig. 2 legend.

It is interesting to compare copia and 17.6, two copia-like elements in *Drosophila*, with respect to gene order and sequence conservation (Fig. 5). 17.6 bore the least resemblance to copia of any of the LTR-containing elements considered in this paper. The best homology between 17.6 and retroviruses was in the reverse transcriptase domain (59), whereas the best homology between copia and retroviruses was in the integrase domain; the order of the two domains was retrovirus-like in 17.6 and Ty-like in copia. Also, 17.6 may encode a protein homologous to retrovirus *env* proteins (59), whereas copia did not (see below). Clearly, residence within the same organism does not imply that these two retrovirus-like genetic elements are more similar to each other than either is to elements in other, widely divergent, species.

Internal repeat. It seems likely that amino acids 700 to 900 of the copia ORF encode a distinct, unsuspected function. This region was a break in the amino acid homology between copia and Ty and was larger by ca. 200 amino acids in Ty. However, the potential translation products of this region of both copia and Ty were rich in polar amino acids. In copia, it included a tandem repeat of 36 amino acids, or 108 nucleotides. The 108-nucleotide repeat, which lay in a short region of copia previously sequenced by Fouts et al. (23),

```
Copia  (901) DDKSSWEEAI NTELNAHKIN NTWTITKAPE NKNIVDSAWV FSVKYNELGN
               *    **    * *       **         * *   *    *
TyORF  (818) KEKEKYIEAY HKEVNQLLKN KTWDTDEYYD AKEIDPKAVI NSNFIFNKKA

Copia  (951) PIAYKAALVA AGFTQKYQID YEETFAPVAA ISSFAFILSL VIQYNLKVHQ
                 * * ** **  *                        ***     *    *
TyORF  (868) DGTHKAAFVA AGDIQHPDTY DSGNQSNTVH HYALNTSLSL ALDNNYYITQ

Copia (1001) NDVKTAFLNG TLKEEIYNAL PQGISCNSDN VCKLNKAIYG LKQAAACWFE
              *   * *     *** * *  *     * *     * *  *  ***    * *
TyORF  (918) LDISSAYLYA DIKEELYIAP PPHLGNN-DK LIALKKSLYE LKQSGANWYE

Copia (1051) VFEQALKECE FVNSSVDACI YILDKGNINE NIYVLLYVDD VVIATGDNTA
                  *                        *       * ***  *
TyORF  (967) TIKSYLIQQC GNE-----EV AGWSCVFKNS QVTICLFVDD NVLFSKN---

Copia (1101) NNNFKAYLNE KFANTDLNEI KHFIGIAIEN QEDKIYLSQS AYVKKILSKF
              *  **   * *  *     *         *  * * *       *      *
TyORF (1009) LNSNKA-IIE KLKNQYDTKI INLGESDEEI QYD-ILGLEI KYQAGKYNKL

Copia (1151) NNEN-----C NAVSTPLPSK ---------- ---INYELLN SDEDCNT---
              ***            **  *               *    *   ***
TyORF (1057) GNENSLTEKI PKLNVPLNPK GAKLSAPGQP GLYIDQDELE IDEDEYKEKV

Copia (1180) -PCASLIGCL NYINLCTAPD LTTAVNILSA YSSKNNSELW QNLKAVLAYL
                  ***    *     * * *    * *
TyORF (1107) HENQKLIGLA SYVGYKFAFD LLYYINTLAQ HILFPSAQVL DNTYELIQFA

Copia (1229) KGTIDNKLIF KKNLAFENKI IGYVDSDW-A GSEIDAKSTT GYLFKNFDFN
               * *  **   **             **    *     **   *  *
TyORF (1157) WDTADKQLIW HKNKPTKPDN KLVAISDASY GNQPYYKSQI GNIFLLNGKV

Copia (1278) LICWNTKAQN SVAASSTEAE YNALFEAVAE ALWLKFLLTS INIKLENPIK
                  **        * ****   *  **       *  *     * *
TyORF (1207) IGGKSTKASL TC-TSTTEAE IHAVSEAIPL LNNLSHLVQE LNKK-PIIKG

Copia (1328) IYEDNQGCIS IANNPSCHK- AAKHIDIKYH FAAEQVQNNV ICLEYIPTEN
                *    ** *       *  *      *     *  *  *      ** *
TyORF (1255) LLTDSASTIS IIKSTNEEKF ANAFFGTKAN ALADEVSGNN LYVYYIETKK

Copia (1377) QLADIFTKPL PAAAFVELAD KLGLLQDDQS NAE.
               **  **** *  **  *   *
TyORF (1305) NIADVNTKPL PIKAFKLLTN KWIH.
```

FIG. 4. Homology between copia and Ty912 ORF in the putative reverse transcriptase domain. Asterisks indicate homology. Numbers in parentheses indicate amino acid number in the large ORF of copia or the second ORF of Ty912 (Clare and Farabaugh, in press). For amino acid abbreviations, see Fig. 2 legend.

was present twice between nucleotides 2589 and 2804. The only difference between the two copies of the repeat was between nucleotides 2660 (which was a G) and 2768 (an A). The repeat could also be considered to continue beyond nucleotide 2804 in that the 58 nucleotides 2805 through 2862 differed from nucleotides 2697 through 2754 (and nucleotides 2589 through 2646) in only five positions, all of which would cause amino acid substitutions.

Mode of expression of copia proteins. Translation of copia RNA in vitro in a rabbit reticulocyte lysate leads to the synthesis of proteins ranging in molecular weight from 18,000 to 51,000 (17, 22). All of these products can be made from RNA in the 2-kb size range, which appears to be translated more efficiently in vitro than the 5-kb RNA (22). The discrepancy between these small proteins and the presence of a single large ORF would be explained if it were supposed that copia, like retroviruses, produces a *gag*-like polyprotein in molar excess relative to a larger polyprotein and that some cleavage of the polyprotein could occur in the reticulocyte lysate. In this case the abundant 51,000-

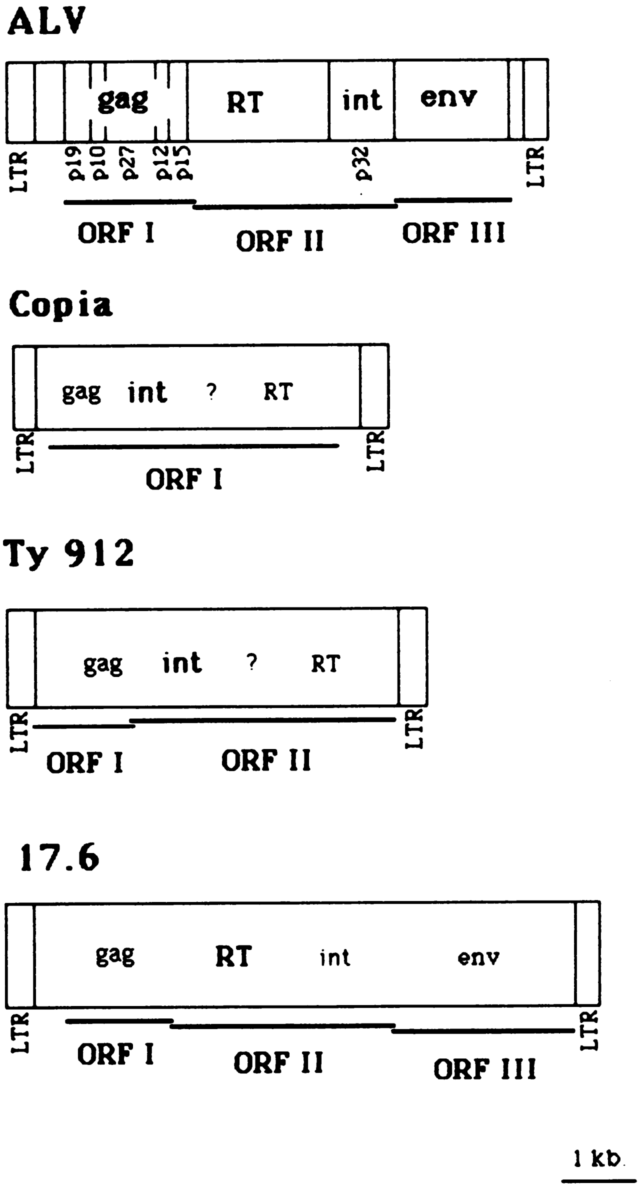

FIG. 5. Overall organization of avian leukosis virus (ALV), copia, Ty912, and 17.6. The scale is shown at bottom right. The organization and sequence features of an avian leukosis virus were inferred from the sequence of RSV described previously (62). p19, p10, etc., represent protein products cleaved from the *gag* polyprotein precursor. p32 represents the 32,000-dalton protein corresponding to the integrase domain (see text). RT, reverse transcriptase; int, integrase. The organization and sequence features of Ty912 and 17.6 are taken from Clare and Farabaugh (in press) and reference 59, respectively. Dark lines under each drawing represent ORFs. Bold type is used to indicate the most reliable homologies between the retrovirus protein and potential transposable-element gene products (i.e., integrase for copia and Ty and reverse transcriptase for 17.6).

molecular-weight protein would correspond to the *gag* polyprotein. Production of a 2-kb RNA which encoded only the 5′ portion of the large ORF would be the mechanism for producing *gag* products in excess over *pol* products, whereas translation of the 1,409-amino-acid ORF on the 5-kb RNA would produce a large protein corresponding to the *gag-pol* fusion protein.

The precise structure of the 2-kb RNA is not known (21, 22, 63, 79), but this RNA does not hybridize with restriction fragments from the region between the *Eco*RI site at position 2300 and the *Hin*f site at position 4556 (63). If 2-kb RNA were produced simply by polyadenylation at a site roughly 2 kb into the element there would be no stop codon; in this case the ORF would be translated into the poly(A) tail, yielding polylysine (which we consider unlikely). A likely alternative is that the 2-kb RNA is spliced to shift the frame of translation, resulting in termination of translation before the poly(A) tail is reached. This could involve either a small intron or a splice to a site near the 3′ LTR, so that the 2- and 5-kb RNAs would share their 3′ termini. Interestingly, good matches to 5′ and 3′ splice site consensus sequences (46) occurred at nucleotides 1605 and 4760, respectively (Fig. 1). Splicing the 5-kb RNA between these two sites would result in an RNA roughly 2 kb in size, which could encode a 425-amino-acid protein of 48,000 daltons, the last 34 amino acids of which would be encoded by a distinct ORF lying between nucleotides 4760 and 4860.

Codon usage. The choice among degenerate codons (codon usage) is generally not random, but highly skewed in favor of some codons and against others. Codon usage varies among taxonomic groups (30) and can be used as a good indicator of whether a pontential ORF is likely to be translated (71). However, it has long been recognized that codon usage differs between viruses and their hosts (30). Codon usage in copia (data not shown) was very different from codon usage in a group of 11 *Drosophila* genes containing 2,758 codons, compiled by Ken Burtis (personal communication). For example, TTA was used in only 2 of 204 leucine codons in that sample, but in 45 of 123 leucine codons in the copia ORF. Conversely, GCC was used in 128 of 219 alanine codons in the *Drosophila* gene sample, but in only 10 of 72 alanine codons in copia. In marked contrast with the situation in most genes, the frequency of particular trinucleotides in copia was roughly equal in the ORF and the two noncoding frames. Termination codons were an obvious exception to this; another was the occurrence of more glutamic and aspartic acid codons in the ORF than the overall trinucleotide frequency would predict. Both of these exceptions are easily seen as responses to selection for functional protein products. Although unusual, the even use of codons may not be surprising if it is supposed that copia transposition involves replication by an error-prone polymerase (such as a reverse transcriptase). In this case selection against the use of undesirable codons is likely to be insufficient to counterbalance the error rate of reverse transcription. The similarity of copia to viruses in this additional respect is worthy of note.

Absence of an *env* homolog. There are a number of reasons for believing that no homolog to the retrovirus *env* protein is encoded by copia. Because the second largest ORF on the transcribed strand of copia was less than 80 amino acids long and *env* proteins are generally much larger, it seems likely that if there were an *env*-like protein it would be encoded by a portion of the large ORF. *env* proteins are translated on membrane-bound ribosomes and are inserted in the plasma membrane, which buds off to form the lipid coat of mature virus particles. Because the large ORF also appears to encode a number of cytoplasmic proteins, this same ORF, or parts of it, would have to be translated on both free and membrane-bound ribosomes. A search for homology to sequences conserved between the *env* proteins of several retroviruses (R. Patarca and W. A. Haseltine, Letter, Nature [London] **312**:496, 1984) revealed no homologous regions within the large ORF. Also, because *env* proteins are transmembrane proteins present within the lipid coat of retroviral virions, they should have hydrophobic transmembrane domains. Such domains usually consist of at least 19 generally hydrophobic residues (15). A hydropathicity plot (39) of the large ORF revealed no regions with transmembrane potential. Finally, the total coding capacity of copia (and Ty) was less than that of a retrovirus by roughly the amount needed to encode the *env* proteins. Interestingly, other copia-like elements are larger (57), and 17.6 is reported to have *env* homology (59).

Retroviruses gain entry to cells by a specific interaction between their envelope glycoproteins and cellular receptors; accordingly, these *env* proteins are the primary determinants of viral host range. The absence from copia of any *env*-like protein therefore strongly supports the view that copia is not a virus, but rather a transposable element whose mechanism of transposition is very similar to the mechanism of replication of a retrovirus. Copia and similar transposable elements are likely to share a distant common ancestor with retroviruses. Whether that ancestor was a virus or a transposable element remains a matter for speculation (73, 75). It seems to us that the acquisition or loss of an *env* gene together with the associated capacity of horizontal transfer which distinguishes viruses from transposable elements could have occurred repeatedly during the evolution of this broad class of genetic elements. In our view the distinction between retrovirus-like transposable elements and those which transpose through a DNA rather than an RNA intermediate is more basic than the distinction between viruses and transposable elements.

ACKNOWLEDGMENTS

We are extremely grateful to Philip Farabaugh for communicating results before publication. We thank Cetus Corp. for allowing us to use the patented bacterial strain DG98. We thank Harold Varmus and Jeremy Thorner for discussions. We also thank Daniel Kalderon, Mary Rykowski, Harold Varmus, Danielle Thierry-Mieg, and Robert Levis for their careful readings of the manuscript.

This work was supported by a grant from the American Cancer Society to G.M.R. S.M.M. was a National Institutes of Health postdoctoral fellow.

ADDENDUM IN PROOF

Emori et al. (Y. Emori, T. Shiba, S. Kanaya, S. Inouye, S. Yuki, and K. Saigo, Nature, in press) have also sequenced a copia element. Their element differs from ours by four silent single bp substitutions within the ORF and three single bp deletions 3′ of the ORF.

LITERATURE CITED

1. **Bankier, A. T., and B. G. Barrel.** 1983. Shotgun DNA sequencing, p. 1–33. *In* R. A. Flavell (ed.), Techniques in nucleic acid biochemistry, vol. B5. Elsevier Scientific Publishing, New York.
2. **Bayev, A. A., N. V. Lyvbomirskaya, E. B. Dzhumagdiev, E. V. Ananiev, I. G. Amiant, 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.

3. **Bender, W., M. Akam, F. Karch, P. A. Beachy, M. Peifer, P. Spierer, E. B. Lewis, and D. S. Hogness.** 1983. Molecular genetics of the bithorax complex in *Drosophila melanogaster*. Science **221**:23–29.
4. **Bingham, P. M., and B. H. Judd.** 1981. A copy of the copia transposable element is very tightly linked to the w^a allele at the white locus of *D. melanogaster*. Cell **25**:705–711.
5. **Bingham, P. M., R. Levis, and G. M. Rubin.** 1981. Cloning of DNA sequences from the white locus of *D. melanogaster* by a novel and general method. Cell **25**:693–704.

5a. **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.

6. **Carlson, M., and D. Brutlag.** 1978. One of the copia genes is adjacent to satellite DNA in *Drosophila melanogaster*. Cell **15**:733–742.
7. **Chiu, I.-M., R. Callahan, S. R. Tronick, J. Schlom, and S. A. Aaronson.** 1984. Major *pol* gene progenitors in the evolution of oncoviruses. Science **223**:364–370.
8. **Copeland, N. G., K. W. Hutchinson, and N. A. Jenkins.** 1983. Excision of the DBA ecotropic provirus in dilute-coat-color revertants of mice occurs by homologous recombination involving the viral LTRs. Cell **33**:379–387.
9. **Copeland, T. D., S. Oroszlan, V. S. Kalyanaraman, M. G. Sarngadharan, and R. C. Gallo.** 1983. Complete amino acid sequence of human T-cell leukemia virus structural protein p15. FEBS Lett. **162**:390–395.
10. **Dente, L. G., G. Cesareni, and R. Cortese.** 1983. pEMBL: a new family of single-stranded plasmids. Nucleic Acids Res. **11**:1645–1655.
11. **Devare, S. G., E. P. Reddy, J. D. Law, K. C. Robbins, and S. A. Aaronson.** 1983. Nucleotide sequence of the simian sarcoma virus genome: demonstration that its acquired cellular sequences encode the transforming gene product $p28^{sis}$. Proc. Natl. Acad. Sci. U.S.A. **80**:731–735.
12. **Donehower, L. A., and H. E. Varmus.** 1984. A mutant murine leukemia virus with a single missense codon in *pol* is defective in a function affecting integration. Proc. Natl. Acad. Sci. U.S.A. **81**:6461–6465.
13. **Duyk, G., J. Leis, M. Longiaru, and A. M. Skalka.** 1983. Selective cleavage in the avian retroviral long terminal repeat sequence by the endonuclease associated with the alphabeta form of avian reverse transcriptase. Proc. Natl. Acad. Sci. U.S.A. **80**:6745–6749.
14. **Elder, R. T., E. Y. Loh, and R. W. Davis.** 1983. RNA from the yeast transposable element Tyl has both ends in the direct repeats, a structure similar to retrovirus RNA. Proc. Natl. Acad. Sci. U.S.A. **80**:2431–2436.
15. **Engleman, D. M., A. Goldman, and T. A. Steitz.** 1982. The identification of helical segments in the polypeptide chain of bacteriorhodopsin. Methods Enzymol. **88**:81–89.
16. **Falkenthal, S., M. L. Grahm, E. L. Korn, and J. A. Lengyel.** 1982. Transcription, processing and turnover of RNA from the *Drosophila* mobile genetic element copia. Dev. Biol. **92**:294–305.
17. **Falkenthal, S., and J. A. Lengyel.** 1980. Structure, translation and metabolism of the cytoplasmic copia ribonucleic acid of *Drosophila melanogaster*. Biochemistry **19**:5842–5850.
18. **Flavell, A. J.** 1984. Role of reverse transcription in the generation of extrachromosomal copia mobile genetic elements. Nature (London) **310**:514–515.
19. **Flavell, A. J., and D. Ish-Horowicz.** 1981. Extrachromosomal circular copies of the eukaryotic transposable element copia in cultured *Drosophila* cells. Nature (London) **292**:591–595.
20. **Flavell, A. J., and D. Ish-Horowicz.** 1983. The origin of extrachromosomal circular copia elements. Cell **34**:415–419.
21. **Flavell, A. J., R. Levis, M. A. Simon, and G. M. Rubin.** 1981. The 5′ termini of RNAs encoded by the transposable element copia. Nucleic Acids Res. **9**:6279–6291.
22. **Flavell, A. J., S. W. Ruby, J. T. Toole, B. E. Roberts, and G. M. Rubin.** 1980. Translation and developmental regulation of RNA encoded by the eukaryotic transposable element copia. Proc. Natl. Acad. Sci. U.S.A. **77**:7107–7111.
23. **Fouts, D. L., J. E. Manning, G. M. Fox, and C. W. Schmid.** 1981. A complex repeated DNA sequence within the *Drosophila* transposable element copia. Nucleic Acids Res. **9**:7053–7064.
24. **Freund, R., and M. Meselson.** 1984. Long terminal repeat nucleotide sequence and specific insertion of the gypsy transposon. Proc. Natl. Acad. Sci. U.S.A. **81**:4462–4464.
25. **Gehring, W. J., and R. Paro.** 1980. Isolation of a hybrid plasmid with homologous sequences to a transposing element of *Drosophila*. Cell **19**:897–904.
26. **Gibson, W., and I. M. Verma.** 1974. Studies on the reverse transcriptase of RNA tumor viruses. Structural relatedness of two subunits of avian RNA tumor viruses. Proc. Natl. Acad. Sci. U.S.A. **71**:4991–4994.
27. **Goldberg, M. L., R. Paro, and W. J. Gehring.** 1982. Molecular cloning of the white locus region of *Drosophila melanogaster* using a large transposable element. EMBO J. **1**:93–98.
28. **Goldberg, M. L., J.-Y. Sheen, W. J. Gehring, and M. M. Green.** 1983. Unequal crossing over associated with asymmetrical synapsis between nomadic elements in the *Drosophila melanogaster* genome. Proc. Natl. Acad. Sci. U.S.A. **80**:5017–5021.
29. **Grandgenett, D. P., A. C. Vora, and R. D. Schiff.** 1978. A 32,000 dalton nucleic acid binding protein from avian retrovirus cores possesses DNA endonuclease activity. Virology **89**:119–132.
30. **Grantham, R., C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier.** 1981. Codon catalogue usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9**:r43–r74.
31. **Green, M. M.** 1959. Spatial and functional properties of pseudoalleles at the white locus in *Drosophila melanogaster*. Heredity **13**:303–315.
32. **Herr, W., V. Corbin, and W. Gilbert.** 1982. Nucleotide sequence of the 3′ half of AKV. Nucleic Acids Res. **10**:6931–6944.
33. **Hutchinson, K. W., N. G. Copeland, and N. A. Jenkins.** 1984. Dilute-coat-color locus of mice: nucleotide sequence analysis of the d^{+2J} and d^{+Ha} revertant alleles. Mol. Cell. Biol. **4**:2899–2904.
34. **Ilyin, Y. V., N. G. Schuppe, N. V. Lyvbomirskaya, I. V. Gorelova, and J. R. Arkhipova.** 1984. Circular copies of mobile dispersed genetic elements in cultured *Drosophila melanogaster* cells. Nucleic Acids Res. **12**:7517–7531.
35. **James, M. N. G.** 1980. An X-ray crystallographic approach to enzyme structure and function. Can. J. Biochem. **58**:251–271.
36. **Kozak, M.** 1982. How do ribosomes recognize the unique AUG initiator codon in messenger RNA? Biochem. Soc. Symp. **47**:113–128.
37. **Kramer, G., and P. Argos.** 1984. Primary structural determinants of RNA-dependent polymerases from plant, animal and bacterial viruses. Nucleic Acids Res. **12**:7269–7282.
38. **Kugimiya, W., H. Ikenaga, and K. Saigo.** 1983. Close relationship between the long terminal repeats of avian leukosis-sarcoma virus and copia-like movable genetic elements of *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. **80**:3193–3197.
39. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.
40. **Levis, R., P. M. Bingham, and G. M. Rubin.** 1982. Physical map of the white locus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. U.S.A. **79**:564–568.
41. **Levis, R., P. Dunsmuir, and G. M. Rubin.** 1980. Terminal repeats of the *Drosophila* transposable element copia: nucleotide sequence and genomic organization. Cell **21**:581–588.
42. **Levis, R., K. O'Hare, and G. M. Rubin.** 1984. Effects of transposable element insertions on RNA encoded by the white gene of *Drosophila*. Cell **38**:471–481.
43. **Lindsley, D. L., and E. H. Grell.** 1968. Genetic variations of *Drosophila melanogaster*. Publication no. 627. Carnegie Institution, Washington, D.C.

43a. **Lipman, D. J., and W. R. Pearson.** 1985. Rapid and sensitive protein similarity searches. Science **227**:1435–1441.

43b. **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 Tyl. Nature (London) **313**:243–246.

44. **Misra, T. K., D. P. Grandgenett, and J. T. Parsons.** 1982. Avian retrovirus pp32 DNA-binding protein. I. Recognition of specific sequences on retrovirus DNA terminal repeats. J. Virol. **44:**330–343.

45. **Modolell, J., W. Bender, and M. Meselson.** 1983. *Drosophila melanogaster* mutations suppressible by the suppressor of hairy-wing are insertions of a 7.3 kilobase mobile element. Proc. Natl. Acad. Sci. U.S.A. **80:**1678–1682.

46. **Mount, S. M.** 1982. A catalogue of splice junction sequences. Nucleic Acids Res. **10:**459–472.

47. **Needleman, S. B., and C. D. Wunch.** 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. **48:**443–453.

48. **Norrander, J., T. Kempe, and J. Messing.** 1983. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. Gene **26:**101–106.

49. **O'Hare, K., C. Murphy, R. Levis, and G. M. Rubin.** 1984. DNA sequence of the white locus of *Drosophila melanogaster*. J. Mol. Biol. **180:**437–455.

50. **Panganiban, A. T., and H. M. Temin.** 1984. Circles with two tandem LTRs are precursors to integrated retroviral DNA. Cell **36:**673–679.

51. **Panganiban, A. T., and H. M. Temin.** 1984. The retrovirus *pol* gene encodes a product required for DNA integration: identification of a retrovirous *int* locus. Proc. Natl. Acad. Sci. U.S.A. **81:**7885–7889.

52. **Pirrotta, V., and C. Brockl.** 1984. Transcription of the *Drosophila* white locus and some of its mutants. EMBO J. **3:**563–568.

53. **Potter, S. S., W. J. Brorein, Jr., P. Dunsmuir, and G. M. Rubin.** 1979. Transposition of elements of the 412, copia and 297 dispersed repeated gene families in *Drosophila*. Cell **17:**415–427.

54. **Redman, S. M. S., and C. Dickson.** 1983. Sequence and expression of the mouse mammary tumor virus *env* gene. EMBO J. **2:**125–131.

55. **Rho, H. M., D. P. Grandgenett, and M. Green.** 1979. Sequence relatedness between the subunits of avian myeloblastosis virus reverse transcriptase. J. Biol. Chem. **250:**5278–5280.

56. **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.

57. **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.

58. **Rubin, G. M., M. G. Kidwell, and P. M. Bingham.** 1982. The molecular basis of hybrid dysgenesis: the nature of induced mutations. Cell **29:**987–994.

59. **Saigo, K., W. Kugimiya, Y. Matsuo, S. Inouye, K. Yoshioka, and S. Yuki.** 1984. Identification of the coding sequence for reverse transcriptase-like enzyme in a transposable genetic element, 17.6, in *Drosophila melanogaster*. Nature (London) **312:**659–661.

60. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. **74:**5463–5467.

61. **Scherer, G., C. Tschudi, J. Perera, H. Delius, and V. Pirrotta.** 1982. B104, a new dispersed repeated gene family in *Drosophila melanogaster* and its analogies with retroviruses. J. Mol. Biol. **157:**435–451.

62. **Schwartz, D. E., R. Tizard, and W. Gilbert.** 1983. Nucleotide sequence of Rous sarcoma virus. Cell **32:**853–869.

63. **Schwartz, H. E., T. J. Lockett, and M. W. Young.** 1982. Analysis of transcripts from two families of nomadic DNA. J. Mol. Biol. **157:**49–68.

64. **Schwartzberg, P., J. Colicelli, and S. P. Goff.** 1984. Construction and analysis of deletion mutants in the *pol* gene of Moloney murine leukemia virus: a new viral function required for productive infection. Cell **37:**1043–1052.

65. **Seiki, M., H. Seisuke, Y. Hirayama, and M. Yoshida.** 1983. Human adult T-cell leukemia virus: complete nucleotide sequence of the provirus genome integrated in leukemia cell DNA. Proc. Natl. Acad. Sci. U.S.A. **80:**3618–3622.

66. **Shepherd, B. M., and D. J. Finnegan.** 1984. Structure of circular copies of the 412 transposable element present in *Drosophila melanogaster* cells, and isolation of a free 412 long terminal repeat. J. Mol. Biol. **180:**21–40.

67. **Shiba, T., and K. Saigo.** 1983. Retrovirus-like particles containing RNA homologous to the transposable element copia in *Drosophila melanogaster*. Nature (London) **302:**119–124.

68. **Shinnick, T. M., R. A. Lerner, and J. G. Sutcliffe.** 1981. Nucleotide sequence of Moloney murine leukemia virus. Nature (London) **293:**543–548.

69. **Shoemaker, C., J. Hoffman, S. P. Goff, and D. Baltimore.** 1981. Intramolecular integration within Moloney murine leukemia virus DNA. J. Virol. **40:**164–172.

70. **Snyder, M. P., D. Kimbrell, M. Hunkapiller, R. Hill, J. Fristrom, and N. Davidson.** 1982. A transposable element that splits the promoter region inactivates a *Drosophila* cuticle protein gene. Proc. Natl. Acad. Sci. U.S.A. **79:**7430–7434.

71. **Staden, R., and A. D. McLauchlin.** 1982. Codon preference and its use in identifying protein regions in long DNA sequences. Nucleic Acids Res. **10:**141–156.

72. **Sweet, H. O.** 1984. Dilute suppressor, a new suppressor gene in the house mouse. J. Hered. **74:**305.

73. **Temin, H. M.** 1980. The origin of retroviruses from cellular moveable genetic elements. Cell **21:**599–600.

74. **Toh, H., H. Hayashida, and T. Miyata.** 1983. Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. Nature (London) **305:**827–829.

75. **Varmus, H. E.** 1983. Retroviruses, p. 411–503. *In* J. A. Shapiro (ed.), Mobile genetic elements. Academic Press, Inc., Orlando, Fla.

76. **Varmus, H. E., N. Quintrell, and S. Ortiz.** 1981. Retroviruses as mutagens: insertion and excision of a nontransforming provirus alter expression of a resident transforming provirus. Cell **25:**23–36.

77. **Weiher, H., M. Konig, and P. Gruss.** 1983. Multiple point mutations affecting the simian virus 40 enhancer. Science **219:**626–631.

78. **Winston, F., D. T. Chaleff, B. Valert, and G. R. Fink.** 1984. Mutations affecting Ty-mediated expression of the *HIS4* gene of *Saccharomyces cerevisiae*. Genetics **107:**179–197.

79. **Young, M. W., and H. E. Schwartz.** 1981. Nomadic gene families in *Drosophila*. Cold Spring Harbor Symp. Quant. Biol. **45:**629–640.