MOLECULAR AND CELLULAR BIOLOGY, Mar. 1990, p. 863–871
0270-7306/90/030863-09$02.00/0

Vol. 10, No. 3

A Retrotransposable Element from the Mosquito *Anopheles gambiae*

NORA J. BESANSKY†

Department of Biology, Yale University, New Haven, Connecticut 06511, and Malaria Branch, Centers for Disease Control, Atlanta, Georgia 30333

Received 12 September 1989/Accepted 25 November 1989

A family of middle repetitive elements from the African malaria vector *Anopheles gambiae* is described. Approximately 100 copies of the element, designated T1Ag, are dispersed in the genome. Full-length elements are 4.6 kilobase pairs in length, but truncation of the 5′ end is common. Nucleotide sequences of one full-length, two 5′-truncated, and two 5′ ends of T1Ag elements were determined and aligned to define a consensus sequence. Sequence analysis revealed two long, overlapping open reading frames followed by a polyadenylation signal, AATAAA, and a tail consisting of tandem repetitions of the motif TGAAA. No direct or inverted long terminal repeats (LTRs) were detected. The first open reading frame, 442 amino acids in length, includes a domain resembling that of nucleic acid-binding proteins. The second open reading frame, 975 amino acids long, resembles the reverse transcriptases of a category of retrotransposable elements without LTRs, variously termed class II retrotransposons, class III elements, or non-LTR retrotransposons. Similarity at the sequence and structural levels places T1Ag in this category.

A significant fraction of many eucaryotic genomes is mobile, and much of this fraction moves either passively or actively by reverse transcription of RNA to DNA (13, 41, 43). Such sequences, known as retroposons, include processed pseudogenes, retroviruses, and retrotransposable elements. Processed pseudogenes are integrated cDNA copies of processed mRNAs marked by the lack of introns, the presence of poly(A) tails, and variable-length target site duplications. They are accidental by-products of a reverse transcriptase whose source is unknown. Retroviruses encode their own reverse transcriptase under the control of long terminal repeats (LTRs). Poly(A) tails added posttranscriptionally are removed during reverse transcription, and insertion into the chromosome is accompanied by short (4- to 6-base-pair [bp]) target site duplications of fixed length (42). Retrotransposable elements that share these features with retroviruses, such as *copia* of *Drosophila melanogaster* (29) and Ty of *Saccharomyces cerevisiae* (1), are known as viruslike retroposons, or retrotransposons. They differ from retroviruses in the absence of an extracellular stage.

A different class of elements that appear to encode their own reverse transcriptase but lack LTRs has been termed class II retrotransposons (10), class III elements (14), or non-LTR retrotransposons (44). To avoid a terminology muddle, the last, more descriptive designation will be used here. Recognized only recently, these elements are ubiquitous residents of eucaryotic genomes. The best studied is the L1 superfamily of mammals (10, 24), which makes up 8% of genomic sequences. Although not as abundant in other organisms, non-LTR retrotransposon families have been found in fruit flies (8, 9, 11, 33), the silkworm moth (3, 44), trypanosomes (20, 30), frogs (14), and corn (37). The suggestion that these elements transpose through an RNA intermediate is based upon several shared characteristics. These include the dispersion of many copies throughout the genome, long conserved open reading frames (ORFs) with similarity at the amino acid level to reverse transcriptases and nucleic acid-binding domains of retroviruses, frequent 5′ truncation, a polyadenylation signal followed by an oligo(A) or A-rich tail, and target site duplications which, if present, are both long (>7 bp) and variable between element copies. Although many details of the propagation of non-LTR retrotransposons remain to be elucidated, the absence of LTRs, the presence of A-rich tails, and differences in the nature of the target site duplications indicate that different enzymes or mechanisms are involved in reverse transcription and integration from those operating on retrovirus and retrotransposon sequences.

This report describes the identification and characterization of a family of dispersed repeated sequences from the principal vector of malaria in Africa, the mosquito *Anopheles gambiae*. The family was named T1Ag for transposable element family 1 of *A. gambiae* (hereafter referred to as T1). Its features place it in the category of non-LTR retrotransposons.

MATERIALS AND METHODS

Mosquitoes. Field-collected *A. gambiae* were the offspring of adult females collected from villages in Western Kenya as described previously (27). Colony material was obtained from the G3 strain colonized in 1975 from specimens collected in The Gambia and maintained at the Centers for Disease Control.

Preparation of nucleic acids. DNA was prepared from single adult mosquitoes by using a modification of a technique developed for *Drosophila* spp. as described previously (23). Bulk genomic DNA was extracted from adults of the G3 strain as described previously (17). Plasmid DNA was prepared by a standard boiling procedure (19), omitting the phenol-chloroform extractions. Bacteriophage DNA was isolated by layering liquid lysate, from which cellular debris had been removed by centrifugation, onto 40% glycerol in phage dilution buffer and centrifuging it at 48,000 × *g* for 2 h at 4°C. The phage pellet was suspended in 300 μl of buffer (0.3 M NaCl, 0.1 M Tris hydrochloride [pH 8], 1 mM EDTA), phenol-chloroform extracted, and ethanol precipitated.

† Present address: Retrovirus Disease Branch G-19, Centers for Disease Control, Atlanta, GA 30333.

RNA was isolated by a guanidinium thiocyanate method (5) from several grams of adult G3 mosquitoes which had been flash-frozen with liquid nitrogen and stored at −80°C. Poly(A)$^+$ RNA was selected by one pass through an oligo(dT)-cellulose column as described previously (26).

DNA probes. T1-containing phage clone 34 was isolated from an EMBL3 Lambda library (6) of *Sau*3AI partially digested genomic DNA of the G3 strain, with G3 genomic DNA as a probe. Phages 31b, 914a, and 916e were isolated from the G3 library by probing with a 1.5-kilobase-pair (kb) *Sal*I-*Hin*dIII fragment of clone 34 internal to T1 (see Fig. 2 for maps of T1 elements). A 3-kb *Sal*I-*Eco*RI fragment of phage 31b including central and 3′ sequences of T1 as well as some flanking DNA was subcloned into the corresponding sites of Bluescript SK+ (Stratagene Inc.) and designated BSE3.31b. A 1.1-kb *Eco*RI-*Xho*I fragment containing the 5′ end of T1 and some flanking DNA in 916e was subcloned into Bluescript SK+ and designated BEO11.916. Clones 51a and 122b containing the 5′ end of T1 were isolated from the library by screening with the BEO11.916 insert.

Probes were made radioactive by release of the insert or fragments of the insert with the appropriate enzymes, separation from the vector by electrophoresis through a 1.2% low-melting-point agarose gel (SeaPlaque; FMC Corp.), excision of the band of interest, and incorporation of [^{32}P]dCTP by random primer extension (12). Filter lifts were performed on nitrocellulose disks (Millipore Corp.).

Southern hybridization. Single-mosquito DNA was digested with 10 U of restriction enzyme, electrophoresed on a 0.7% agarose gel (Bethesda Research Laboratories, Inc.), and transferred to GeneScreen Plus membrane (Du Pont, NEN Research Products). Hybridization was performed overnight in 4× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–5× Denhardt solution–1% sodium dodecyl sulfate at 65°C. The membrane was washed in prewarmed 0.1× SSC–0.1% sodium dodecyl sulfate at 65°C and autoradiographed.

Copy number determination. G3 DNA was diluted to 10 μg/ml in TE (10 mM Tris chloride [pH 7.6] plus 1 mM EDTA [pH 8.0]), denatured with an equal volume of 0.5 N NaOH at 55°C for 10 min, and chilled on ice. Five samples of G3 DNA were removed, with the amount successively doubled from 6.25 to 100 ng. These were adjusted to a total DNA concentration of 100 ng with denatured salmon sperm DNA and a final volume of 100 μl with 0.125 N NaOH–0.125× SSC. Three replicate spots for each dilution were loaded onto GeneScreen Plus membrane in a dot blot apparatus (Bio-Rad Laboratories). To provide a standard scale for comparison, samples of denatured BSE3.31b were applied to the membrane in triplicate spots, with eight samples doubled successively from 0.1 to 12.8 ng, adjusted to 100 ng with denatured salmon sperm DNA. The membrane was hybridized to a 1.7-kb *Sal*I-*Hin*dIII fragment of T1 from BSE3.31b. The extent of hybridization was measured by liquid scintillation counting (Optifluor; Packard Instrument Co., Inc.) of spots cut out of the membrane.

DNA sequencing. Restriction fragments of phages 31b and 34 were subcloned into M13mp19; those of 914a, 916e, 51a, and 122b were subcloned into Bluescript SK+. Nested deletions were generated from M13 clones (Cyclone I Biosystem; IBI, Inc.) and Bluescript clones (ExoIII/Mung Bean Nuclease Deletion Kit; Stratagene). Sequencing was performed by using the dideoxy-chain termination method (36) by extending from the universal and reverse primers (U.S. Biochemical Corp.) as well as synthetic 18-mers with modified T7 DNA polymerase (Sequenase; U.S. Biochemical Corp.). Sequence data were analyzed by using GENEPRO 4.2 (Hoefer Scientific Instruments) and the Genetics Computer Group Sequence Analysis Package (Version 6.0) of the University of Wisconsin (J. Devereux, P. Haeberli, and O. Smithies, Nucleic Acids Res. **12**:387, 1984) on the VAX network of the Division of Host Factors, Centers for Disease Control.

RESULTS

Isolation of a dispersed repetitive sequence. Phages containing middle repetitive sequences from the *A. gambiae* genome were identified by screening the G3 library with nick-translated G3 genomic DNA and picking moderately to weakly hybridizing plaques. Back hybridizations were performed to Southern blots of DNA from *A. gambiae* and four of its sibling species digested with various restriction enzymes. One phage, phage 34, contained a repeated element later denoted T1 and was selected for further analysis because it produced an array of bands whose pattern differed among strains of a given species as well as among the sibling species (results not shown). Cross-hybridizing sequences were not detected from the congener *A. stephensi*.

The copy number of T1 in the genome of *A. gambiae* was estimated by a dot blot experiment in which dots containing defined amounts of G3 DNA and the T1-containing plasmid BSE3.31b DNA were hybridized with an internal segment of T1 (the 1.7-kb *Sal*I-*Hin*dIII fragment from BSE3.31b). Radioactive counts hybridized per dot of BSE3.31b were plotted against the amount of plasmid DNA per dot (data not shown). Assuming a genome size of 0.25 pg (N. J. Besansky, Ph.D thesis, Yale University, New Haven, Conn., 1990), the number of counts hybridized per 100-ng dot of G3 DNA indicated that T1 is present in about 100 copies per haploid genome.

Hybridization of a tandemly repeated sequence to genomic Southern blots typically produces a ladder composed of multimers of a basic repeating unit, if that unit is sensitive to the restriction enzyme used to digest the genomic DNA. The absence of a ladder and the polymorphic pattern of bands observed among species and strains on Southern blots by using several restriction enzymes (results not shown) suggested that multiple copies of T1 are dispersed in the genome and implied that their location is polymorphic between strains. An additional Southern hybridization experiment strengthened this conclusion (Fig. 1a). DNA from different isofemale families of field-collected *A. gambiae* (see Materials and Methods) was digested with restriction enzymes that do not cleave the 5′ end of T1. A probe was prepared from the 5′ end of T1 (Fig. 1b). It detected bands of various sizes within a lane, reflecting various distances from different T1 elements to the nearest restriction site in flanking DNA sequences in an individual. If T1 were arranged identically in different individuals, the hybridization patterns among lanes would be the same (ignoring restriction site polymorphism). The heterogeneous pattern of bands observed among lanes is interpreted as polymorphism in the genomic location of T1 among individuals. Although restriction site polymorphism would be expected to contribute to these differences, its role must be minor, since heterogeneous patterns were observed for all four enzymes.

In situ hybridization of T1 to ovarian polytene chromosomes of the G3 strain revealed elements on all six chromosome arms in addition to the chromocenter. About 50 euchromatic sites per complement were detected (Besansky, Ph.D. thesis).

Northern (RNA) blotting experiments failed to detect the presence of T1 homologous sequences in G3 adult total or poly(A)$^+$ RNA by using M13 probes containing the plus and

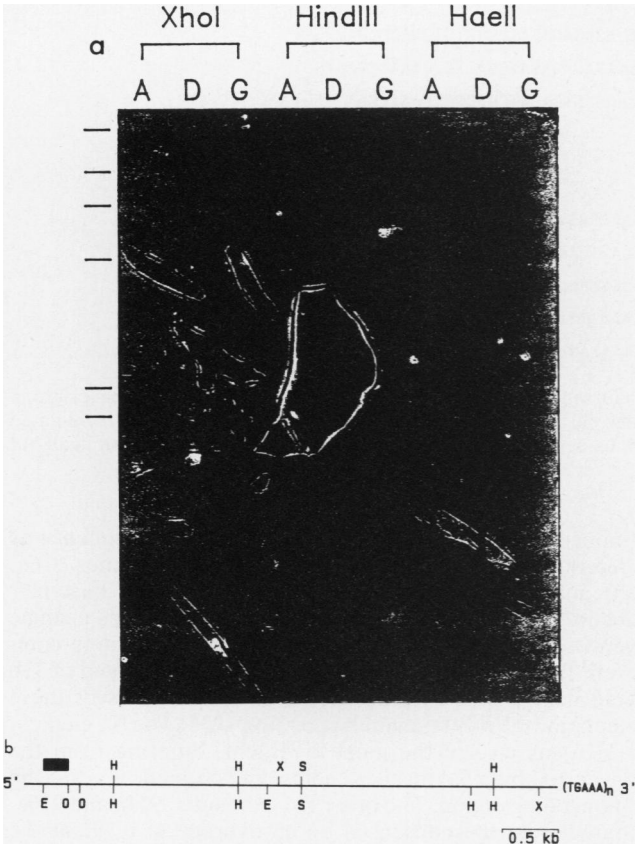


FIG. 1. (a) Heterogeneity between isofemale families of field-collected *A. gambiae*. The DNA was digested with *Xho*I, *Hin*dIII, or *Hae*II and probed with a 5′ segment of T1. Lanes A, D, and G contain the DNA of one daughter from the female collected at field sites of the same designation (27). The bars to the left represent the positions of *Hin*dIII fragments of lambda DNA of 23.6, 9.6, 6.7, 4.3, 2.2 and 1.9 kb. (b) T1 consensus maps. The horizontal line represents T1. Restriction sites: E, *Eco*RI; H, *Hin*dIII; O, *Xho*I; S, *Sal*I; X, *Xba*I. Sites typical of one subfamily are placed above, and sites typical of the other subfamily are placed below the horizontal line; some sites are shared. The black bar indicates the source of the probe used to identify T1 in panel a.

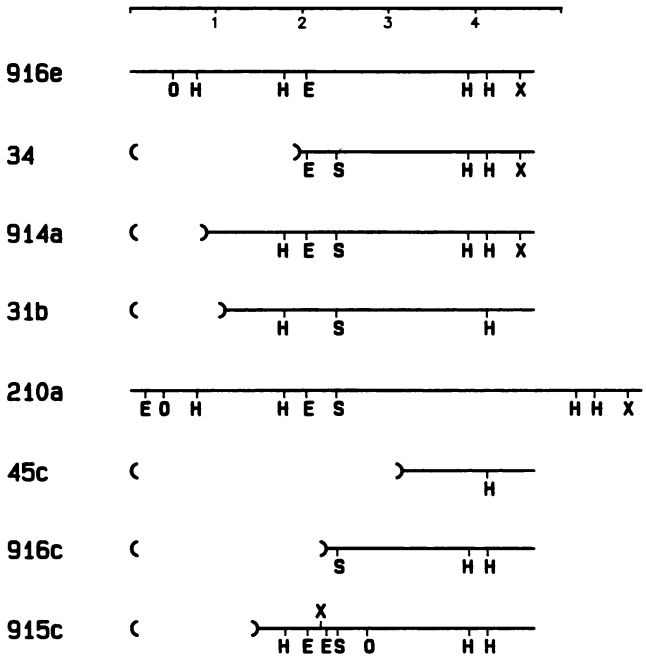


FIG. 2. Restriction maps of cloned T1s. The horizontal lines represent different T1s from the phage clones indicated to the left. The parentheses represent deletions with respect to full-length T1 sequences. Restriction sites are the same as those in Fig. 1.

minus strands of the 2.2-kb *Hin*dIII segment of T1 from phage 31b.

Structure of genomic and cloned T1 elements. Restriction mapping by Southern analysis of genomic copies showed the complete T1 element to be at least 4.4 kb in length. A T1 consensus restriction map (Fig. 1b) was established by single and double digestion of G3 genomic DNA with a variety of enzymes followed by sequential probing with internal segments of a cloned T1 element (data not shown). Interestingly, this analysis identified two consensus maps which share some but not all enzyme sites, suggesting the presence of subfamilies (N. J. Besansky, submitted for publication). Deviation from the consensus was revealed by multiple faint bands of hybridization above and below the predominant ones.

Both nucleotide sequence and length heterogeneities were found in comparisons of eight cloned T1 elements randomly selected from the G3 library (Fig. 2). The elements were restriction mapped and roughly delimited by single and double digestions with five restriction endonucleases, followed by agarose gel electrophoresis, blotting to nylon, and cross-hybridization with various labeled T1 segments. In addition, the elements in 916e, 914a, and 31b were completely sequenced along with portions of 34, including the 3′ end (see below). The endpoints of the other elements must be regarded as approximations to the nearest restriction site. Evaluation of the clones with respect to the consensus restriction maps (Fig. 1b) revealed restriction site polymorphisms in some of the elements, for example, the lack of a 5′ *Eco*RI site and a *Sal*I site in the T1 copy from 916e and the additional restriction sites present in 915c. Sequence analysis showed the missing sites in 916e to result from single-base-pair changes. The extra *Xba*I, *Eco*RI, and *Xho*I sites in 915c could have been created by single-base-pair substitutions, as judged from the T1 consensus nucleotide sequence (see below). The T1 element in 210a has an undefined insertion of approximately 1.2 kb between the *Sal*I and *Hin*dIII sites.

Another source of variation is the tendency for T1 members to be truncated at the 5′ end. This was observed to various extents in six of eight clones analyzed. The generality of this phenomenon is indicated by the fraction of plaques from the G3 library which hybridize to different T1 probes. Only 0.4% of plaques from the library hybridize with a probe containing the 5′ end (BEO11.916), in contrast with 1.6% of plaques which hybridize with a probe containing central to 3′ sequences, the 1.5-kb *Sal*I-*Hin*dIII fragment of phage 34. Thus, sequences from the 5′ end are four times less abundant in the genome than sequences from the 3′ end.

Nucleotide sequence of T1 elements. The 5′ and 3′ ends of T1 family members were defined by the end of nucleotide sequence similarity between them. Both ends of one complete element (916e) and two 5′ and four 3′ ends of other elements were sequenced and are aligned in Fig. 3. At the 5′ end, sequence similarity among the elements begins at the same nucleotide. The 3′ end is characterized by a polyadenylation signal followed immediately by a variable-length tail consisting of the sequence $(TGAAA)_n$. Sequence similarity degenerates after the tail. The ends lack direct or inverted

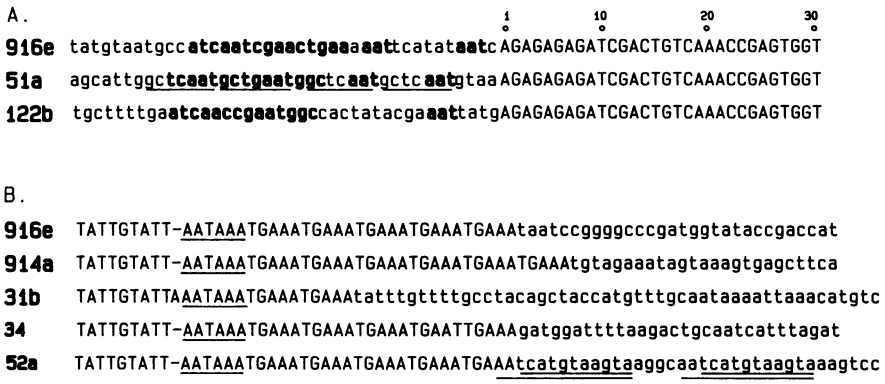


FIG. 3. Alignment of sequences at the 5′ and 3′ termini of T1 elements. T1 sequences are in capital letters, and flanking sequences are in lowercase letters. (A) 5′ end. Each T1 copy begins at position 1 with the same base, A. Similarities among flanking sequences are highlighted by bold type. The 5′ direct repeats of 51a are underlined. (B) 3′ end. The polyadenylation signal is underlined. The preceding dash in all but 31b indicates a 1-bp gap introduced to optimize the alignment. The 3′ direct repeats of 52a are underlined.

LTRs. The regions surrounding different elements share no obvious features, suggesting that insertion does not depend on a specific target sequence. However, a 15-bp region near the left end of 51a is identical to 916e at 8 positions and to 122b at 11 positions (Fig. 3).

The full-length element in 916e is flanked by the sequence TAATC, but two truncated elements in 31b and 914a lack flanking target site duplications. It is interesting, however, that a T1 member in 52a contains a 12-bp duplication separated by 5 bp that abuts the 3′ end of the element (Fig. 3). The duplication could be extended to 14 bp separated by 3 bp if the last two bases of the element were included. A T1 element from clone 51a has four tandem duplications of a 7-bp sequence occurring 4 bp from its 5′ end. The 51a 5′ and 52a 3′ duplications have no apparent sequence similarity. Although rare among transposable elements, duplications on the 5′ side of a fraction of R1Bm and R2Bm ribosomal insertion elements have been reported (3, 44). In R2Bm, tandem duplications at the 5′ end always involve the same 24 bp of 28S rDNA.

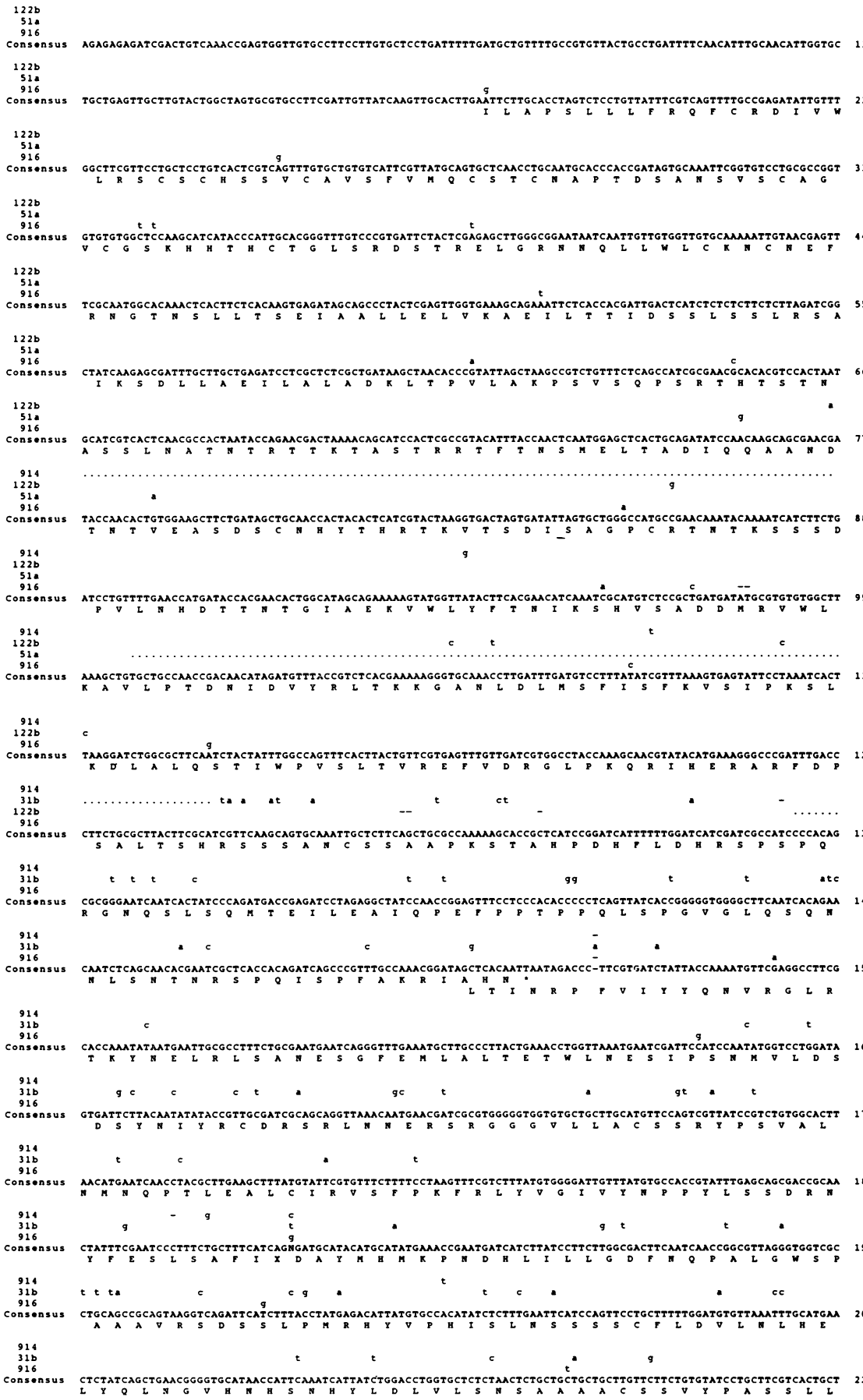
No two T1 elements that have been examined are identical. Sequence polymorphisms noted during the sequencing of different T1s prevented the unambiguous identification of characteristic T1 structure. Therefore, a consensus nucleotide sequence of T1 was determined from an alignment of the complete sequences of one full-length (916e) and two truncated (31b, 914a) elements as well as the left ends of elements in 51a and 122b. Thus, the consensus is derived from at least three T1 copies throughout (Fig. 4). At every position in the alignment except one, two or more nucleotides are identical. Position 1900 of the 4,634-bp consensus is ambiguous because it is occupied by a different nucleotide in all three elements that were compared. It is the third position of the codon AGN. Where N is C or T, as in 914a and 31b, respectively, the codon specifies the same amino acid, serine. In 916e, where N is G, the codon specifies arginine.

Most of the variation results from single-base-pair substitutions, which appear to be evenly distributed. The majority of insertions and deletions that occur involve single nucleotides. A notable exception is the 26-bp insertion in the 3′-noncoding region of 31b. With the consensus sequence as a reference, the percent divergence of T1 in clones 916e, 31b, and 914a was calculated by summing all base-pair substitutions, counting each insertion or deletion as a single event, and dividing by the total nucleotide positions compared. It is important to note that the truncated 5′ end of 31b could not be defined precisely. For the purposes of these calculations it was taken to be position 1235. The T1 element in 31b was clearly the most divergent, differing from the consensus by 5% in the coding region and 9% in the 3′-noncoding region. The other two elements, 914a and 916e, differed from the consensus by an average of 0.7% in the coding and 0.9% in the 3′-noncoding region. By contrast, intraspecific divergence values of L1 clones derived from rats, mice, and humans range from an average of 6% in coding regions to 12% in 3′-noncoding regions (40).

Coding regions of T1. The T1 consensus contains two long ORFs, together occupying 92% of its length. These are indicated in Fig. 4. All three of the completely sequenced T1s appear to be pseudogene copies of this potentially functional element. The 5′ ORF is truncated in 914a (through position 879), and the 3′ ORF is shifted to a different frame relative to the consensus by a 1-bp deletion at position 1884. In 31b, ORF1 is truncated (through position 1229) and ORF2 shifts frames owing to a 1-bp insertion at position 1505. A stop codon interrupts ORF1 at position 955 in 916e, and ORF2 is shifted to a different frame by a 4-bp deletion beginning at position 3334.

In the T1 consensus, ORF1 begins 169 bp from the 5′ terminus. It comprises 1,325 bp and is followed by two consecutive stop codons. Translation from the first ATG codon, 34 codons from the beginning of the ORF, to the first stop codon would produce a 408-amino-acid protein. Nucleic acid-binding domains in the form of cysteine-rich motifs ($CX_2CX_4HX_4C$) are characteristic of retroviral *gag* genes and have been identified in one or both ORFs of most non-LTR retrotransposons described to date (9, 11, 14, 33, 37, 44), with the exception of L1 elements. In a typical "Cys" motif, the residue preceding histidine is glycine and the spacing between cysteine residues is well conserved (7).

FIG. 4. Nucleotide sequence alignment used to derive the T1 consensus and the deduced amino acid sequence. In the alignment, only differences from the consensus are indicated, in lowercase letters. To highlight the beginning of 5′-truncated 914a and 31b, dots are used before the first base. Dots are used after the last sequenced base of 51a and 122b and after the last base of 31b and 916e. Dashes indicate gaps introduced to improve the alignment. N represents a no-consensus base; X represents an arginine or a serine.



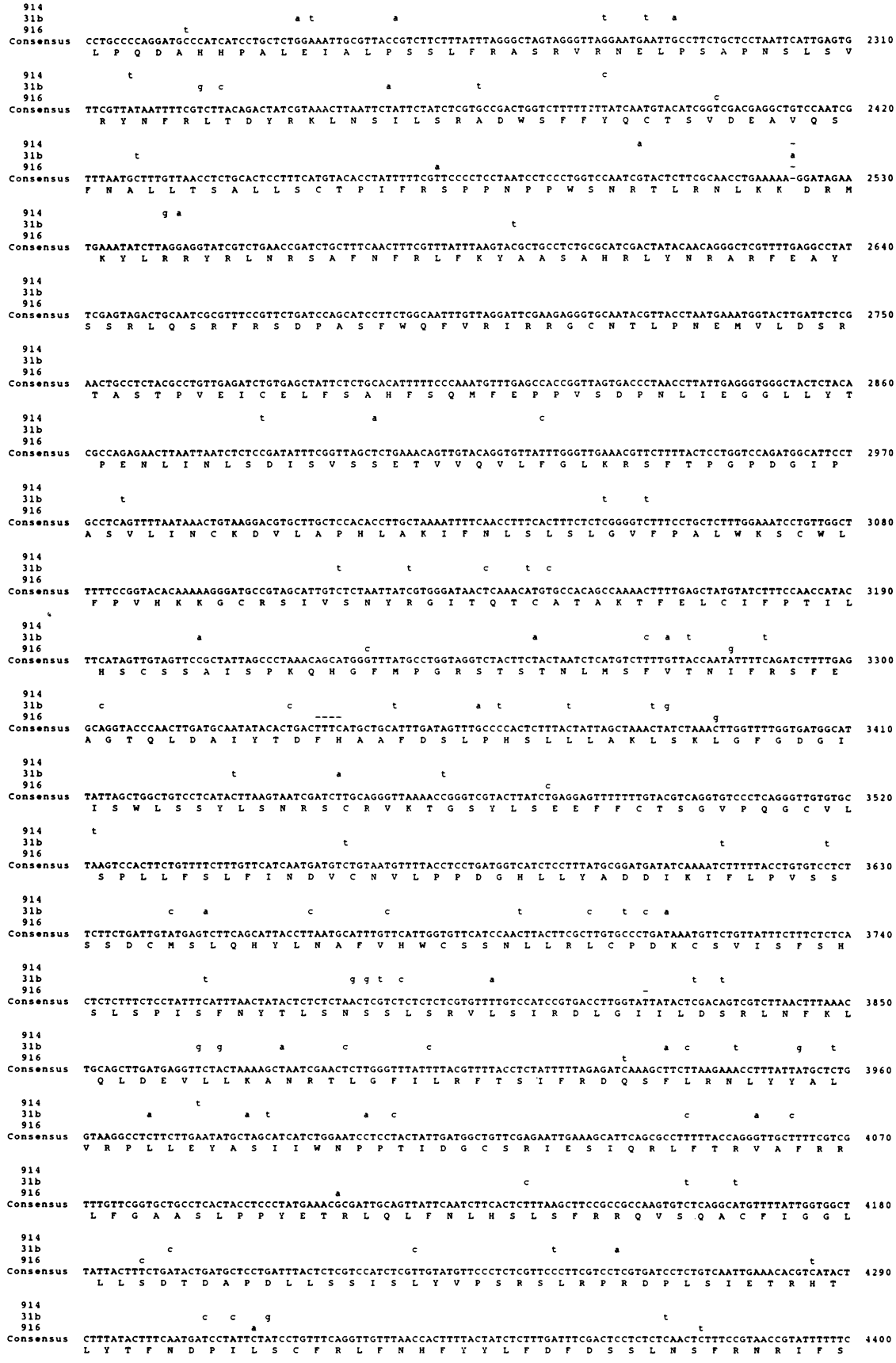


FIG. 4—*Continued.*

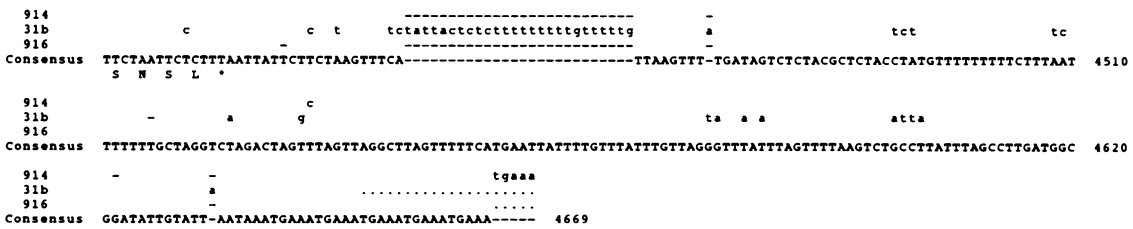


FIG. 4—*Continued.*

Furthermore, the motifs are generally located near the carboxyl terminus of ORF1. There is a sequence in the protein predicted by ORF1 of the T1 element (beginning at residue 52) which resembles a Cys motif but is unusual in its placement at the amino terminus, imperfect in the spacing of residues, and missing the glycine residue. No Cys motifs were observed in any other potential reading frames. The amino terminus of the ORF1 protein is itself unusual in that 7 of the first 54 residues are cysteine.

ORF2 (2,925 bp) overlaps ORF1 by 8 bp in the +1 reading frame and ends 223 bp before the polyadenylation signal. It potentially encodes a protein of 941 amino acids from the first ATG codon, 35 codons from the beginning of the ORF. Part of the expected protein is similar to the reverse transcriptase-like region of non-LTR retrotransposons. In Fig. 5, most of this region of T1 is aligned with those of nine non-LTR retrotransposons from a diverse group of organisms based on previously identified conserved domains (14, 44) that are separated by highly divergent regions. The last domain of the reverse transcriptase region, present in the T1 consensus, is not shown. Among the most highly conserved residues are those which make up the "YXDD box," typical of all reverse transcriptases and necessary for their activity (22). The alanine at position X instead of a bulky hydrophobic residue, as in viral reverse transcriptases, characterizes all non-LTR retrotransposons, including T1.

DISCUSSION

The notion that T1 is or once was mobile in the *A. gambiae* genome is supported by several observations. First, there are approximately 100 copies of T1 scattered throughout the genome in different locations both within and among sibling species. Variation in genomic location among field populations of *A. gambiae* further implies that migration has been recent and is ongoing. Second, a copy of T1 has been found inserted near the 3′ end of the 28S coding region of rDNA. Third, structural and sequence similarities place it in a class of known mobile elements, including the I factor involved in a system of hybrid dysgenesis in *D. melanogaster* (2). The presence or absence of non-LTR retrotransposons causes allelic variation at several genetic loci in their respective hosts (2, 21). Although functional reverse transcriptase has not been isolated, nor has it been directly demonstrated that non-LTR retrotransposons transpose through an RNA intermediate, the presence of extensive and conserved ORFs with similarity to retroviral reverse transcriptases suggests that this is the case. In addition,

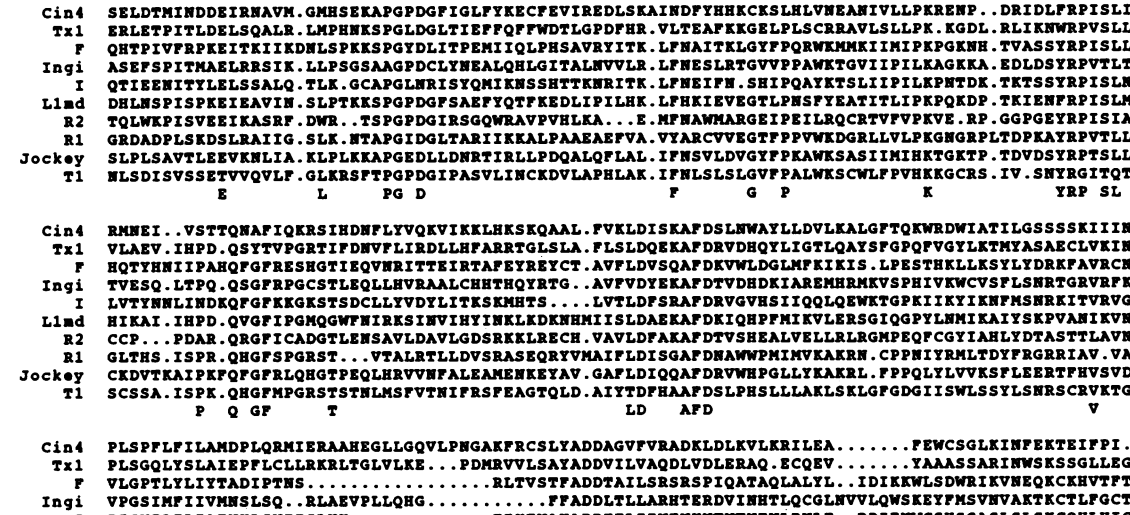


FIG. 5. Amino acid alignment including reverse transcriptase-like sequences from various non-LTR retrotransposons. The alignment is based upon previously identified regions of conservation (14, 44). Not shown is the last conserved domain. Amino acid sequences were extracted from GenBank (version 6.0) entries except for jockey (33), R1Bm (44), and Tx1 (14) which were taken from the literature. Dots represent gaps introduced to improve the alignment. Capital letters below the alignment indicate positions that are identical in at least 7 of the 10 elements.

functional expression of the endonuclease encoded by R2Bm which cleaves the target site has been demonstrated (45).

One of the hallmarks of mobile DNA is the flanking target site duplication produced upon insertion. However, its absence from T1 does not preclude a mobile origin. Neither of the retrotransposons described from slime molds, the DIRS1 elements of *Dictyostelium discoideum* and *Hpa*II repeats of *Physarum polycephalum* (4, 32), is flanked by target site duplications, nor are the non-LTR type II ribosomal insertion elements of *Bombyx mori* (R2) and *D. melanogaster* (3, 35, 44). L1 elements are also not always surrounded by direct repeats (10). Flanking repeats are generated by insertion of a sequence into staggered chromosomal breaks at the host target site which are filled in. If integration were to occur at a blunt cut or involve deletions or target sequence, no duplications would be created.

An integration mechanism has been proposed that accounts for the 5′ tandem duplications associated with some *B. mori* R2 elements based on the involvement of a specific 28S rDNA target sequence in the transposition intermediate (45). The tandem duplication of this sequence depends upon how the transposition intermediate is resolved. Although 5′ direct repeats are sometimes found with R1 elements as well, the integration model proposed for R2 is unlikely to apply to R1, in which the direct repeats are not precisely in tandem, are variable in length, and have little or no similarity to the target site. For similar reasons, the insertion mechanism that potentially gave rise to the 5′ duplications of 51a and 3′ duplications of 52a probably differs from that proposed for R2; it is currently unknown.

Non-LTR retrotransposons have in common the absence of LTRs that carry the transcriptional control signals in retroviruses and retrotransposons and maintain their integrity through cycles of replication. However, the control of transcription and maintenance of non-LTR retrotransposons almost certainly differ among families of elements. It is presumed that R1 and R2 of *B. mori* lack promoters and are processed from RNA polymerase I transcripts of rDNA (44). Rat L1 elements have functional promoters located within a CpG-rich region of tandem repeats at the 5′ end (31). Human L1 cDNAs have a clustering of CpGs at the 5′ end, but lack repeats or any currently identifiable promoter (38). An internal RNA polymerase II promoter has been predicted for the I factor and demonstrated for jockey (11, 28). No promoterlike regions have been identified either internal or external to T1 elements. If T1 elements are transcribed, they may be processed from readthrough RNA polymerase II transcripts. Alternatively, the three T1 clones that begin at the 5′ end with the same base (916e, 51a, and 122b) may lack promoter regions because they are integrated cDNA copies of true T1 genes.

There is no direct evidence that T1 elements are transcribed, since no transcripts were found from *A. gambiae* adult RNA. It is possible that transcripts were present at such low levels that they escaped detection. Another possibility is that T1 may be expressed exclusively at another developmental stage. Complete and specific human L1 transcripts have been isolated only from a cell line exhibiting an embryonal carcinoma morphology (39). Embryonic, larval, and pupal RNA from *A. gambiae* has yet to be tested for the presence of T1 transcripts.

One unusual feature of T1 elements is the tandem repetition of the sequence TGAAA which follows a polyadenylation signal and defines the 3′ terminus. Most other non-LTR retrotransposons terminate with oligo(A) tracts of variable lengths, which presumably represent the poly(A) of the processed mRNAs. Exceptions to this are the I factors which feature TAA repeats at the 3′ end and the strict $(A)_4$ with which R2Bm elements terminate. The latter case may reflect the specific cleavage of a longer poly(A) tail (3). The origin of these repeating tail motifs, which also occur in other retroposons (34), is not certain. However, the similarity of some tails to telomeric repeats has been noted (34; A. M. Weiner, personal communication). This similarity raises the interesting possibility that synthesis of the tail depends upon a telomere terminal transferase (telomerase), the enzyme responsible for telomere maintenance. It has been shown for *Tetrahymena thermophila* that telomerase is capable of adding telomeric repeats one nucleotide at a time to the 3′ end of single-stranded sequences (15). This enzyme has an essential RNA component that carries its own template which is complementary to the telomeric motifs (16). If the T1 tail is like *Anopheles* telomeres, T1 elements may co-opt the telomerase for their own ends. The *Anopheles* telomere sequences have yet to be determined, so a direct comparison cannot be made. However, it should be possible to test the model by in situ hybridization to polytene chromosomes by using as a probe oligonucleotides containing tail repeats. An appealing aspect of this model is that it can explain how cDNA synthesis is primed without another DNA fragment or the eventual loss of unique 3′ sequences. Telomeric sequences are able to form hairpin structures through non-Watson-Crick base pairing (18, 25). Thus, reverse transcription might be self-primed by nonstandard base pairing between tail sequences.

Despite structural and sequence similarities, non-LTR retrotransposon families differ in several respects. From what is known so far, these include the mechanism, level, and stage of transcription as well as the target site specificity, the presence of a polyadenylation signal, the 3′ tail structure, and the presence of flanking repeats. Two families from *Xenopus laevis*, Tx1 and Tx2, are most unusual in that they occur as composites with smaller elements that are found independently (14). Whether similarities among non-LTR retrotransposons are due to conservation of a common progenitor gene or to horizontal transfer, the differences among them indicate that they should be viewed as an opportunistic group which has been able to exploit various niches in the genomes of a broad array of organisms.

ACKNOWLEDGMENTS

I am grateful to Frank Collins for providing the mosquitoes, for advice and encouragement in this work, and for comments on the manuscript; Brian Holloway and the staff of the CID Biotechnology Core Facility for synthesizing the sequencing primers; Diane Hamm and Ira Goldman for their help; and Sarah McKneally of the CID Division of Host Factors, Centers for Disease Control, for assistance with the VAX.

This work was supported by Public Health Service training grant GM07499-11 from the National Institutes of Health and by Public Health Service grant GM36648 to J. R. Powell, Department of Biology, Yale University, from the National Institutes of Health.

LITERATURE CITED

1. **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.
2. **Bucheton, A., R. Paro, H. M. Sang, A. Pelisson, and D. J. Finnegan.** 1984. The molecular basis of I-R hybrid dysgenesis in *Drosophila melanogaster*: identification, cloning, and properties of the I factor. Cell **38**:153–163.
3. **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.
4. **Cappello, M., S. M. Cohen, and H. F. Lodish.** 1984. *Dictyostelium* transposable element DIRS-1 preferentially inserts into DIRS-1 sequences. Mol. Cell. Biol. **4**:2207–2213.
5. **Chomczynski, P., and N. Sacchi.** 1987. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal. Biochem. **162**:156–159.
6. **Collins, F. H., M. A. Mendez, M. O. Rasmussen, P. C. Mehaffey, N. J. Besansky, and V. Finnerty.** 1987. A ribosomal RNA gene probe differentiates member species of the *Anopheles gambiae* complex. Am. J. Trop. Med. Hyg. **37**:37–41.
7. **Covey, S. N.** 1986. Amino acid homology in *gag* region of reverse transcribing elements and the coat protein gene of cauliflower mosaic virus. Nucleic Acids Res. **14**:623–633.
8. **Di Nocera, P. P.** 1988. Close relationship between non-viral retrotransposons in *Drosophila melanogaster*. Nucleic Acids Res. **16**:4041–4052.
9. **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.
10. **Fanning, T. G., and M. F. Singer.** 1987. LINE-1: a mammalian transposable element. Biochim. Biophys. Acta **910**:203–212.
11. **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.
12. **Feinberg, A. P., and B. Vogelstein.** 1984. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **137**:266–267.
13. **Finnegan, D. J.** 1985. Transposable elements in eukaryotes. Int. Rev. Cytol. **93**:281–326.
14. **Garrett, J. E., D. S. Knutzon, and D. Carroll.** 1989. Composite transposable elements in the *Xenopus laevis* genome. Mol. Cell. Biol. **9**:3018–3027.
15. **Greider, C. W., and E. H. Blackburn.** 1985. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. Cell **43**:405–413.
16. **Greider, C. W., and E. H. Blackburn.** 1989. A telomeric sequence in the RNA of Tetrahymena telomerase required for telomere repeat synthesis. Nature (London) **337**:331–337.
17. **Grimwade, B. G., M. A. T. Muskavitch, W. J. Welshons, B. Yedvobnick, and S. Artavanis-Tsakonas.** 1985. The molecular genetics of the *Notch* locus in *Drosophila melanogaster*. Dev. Biol. **107**:503–519.
18. **Henderson, E., C. Hardin, S. Wolk, I. Tinoco, and E. H. Blackburn.** 1987. Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine-guanine base pairs. Cell **51**:899–908.
19. **Holmes, D. S., and M. Quigley.** 1981. A rapid boiling method for the preparation of bacterial plasmids. Anal. Biochem. **114**: 193–197.
20. **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.
21. **Lakshmikumaran, M. S., E. D'Ambrosio, L. A. Laimins, D. T. Lin, and A. V. Furano.** 1985. Long interspersed repeated DNA (LINE) causes polymorphism at the insulin 1 locus. Mol. Cell. Biol. **5**:2197–2203.
22. **Larder, B. A., D. J. M. Purifoy, K. L. Powell, and G. Darby.** 1987. Site-specific mutagenesis AIDS virus reverse transcriptase. Nature (London) **327**:716–717.
23. **Livak, K.** 1984. Organization and mapping of a sequence on the *Drosophila melanogaster* X and Y chromosomes that is transcribed during spermatogenesis. Genetics **107**:611–634.
24. **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.
25. **Lyamichev, V. I., S. M. Mirkin, O. N. Danilevskaya, O. N. Voloshin, S. V. Balatskaya, V. N. Dobrynin, S. A. Filippov, and M. D. Frank-Kamenetskii.** 1989. An unusual DNA structure detected in a telomeric sequence under superhelical stress and at low pH. Nature (London) **339**:634–637.
26. **Maniatis, T., E. F. Fritsch, and J. Sambrook.** 1982. Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
27. **McLain, D. K., F. H. Collins, A. D. Brandling-Bennett, and J. B. O. Were.** 1989. Microgeographic variation in rDNA intergenic spacers of *Anopheles gambiae* in western Kenya. Heredity **62**:257–264.
28. **Mizrokhi, L. J., S. G. Georgieva, and Y. V. Ilyin.** 1988. Jockey, a mobile *Drosophila* element similar to mammalian LINEs, is transcribed from the internal promoter by RNA polymerase II. Cell **54**:685–691.
29. **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.
30. **Murphy, N. B., A. Pays, P. Tebabi, H. Coquelet, M. Guyaux, M. Steinert, and E. Pays.** 1987. *Trypanosoma brucei* repeated element with unusual structural and transcriptional properties. J. Mol. Biol. **195**:855–871.
31. **Nur, I., E. Pascale, and A. V. Furano.** 1988. The left end of rat L1 (L1Rn, long interspersed repeated) DNA which is a CpG island can function as a promoter. Nucleic Acids Res. **16**: 9233–9251.
32. **Pearston, D. H., M. Gordon, and N. Hardman.** 1985. Transposon-like properties of the major, long repetitive sequence family in the genome of *Physarum polycephalum*. EMBO J. **4**:3557–3562.
33. **Priimagi, A. F., L. J. Mizrokhi, and Y. V. Ilyin.** 1988. The *Drosophila* mobile element jockey belongs to LINEs and contains coding sequences homologous to some retroviral proteins. Gene **70**:253–262.
34. **Rogers, J.** 1985. The origin and evolution of retroposons. Int. Rev. Cytol. **93**:187–279.
35. **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.
36. **Sanger, F., S. Nicklen, and A. R. Coulson.** 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74**:5463–5467.
37. **Schwarz-Sommer, Z., L. Leclercq, E. Gobel, and H. Saedler.** 1987. Cin4, an insert altering the structure of the A1 gene in *Zea mays*, exhibits properties of nonviral retrotransposons. EMBO J. **6**:3873–3880.
38. **Skowronski, J., T. G. Fanning, and M. F. Singer.** 1988. Unit-length Line-1 transcripts in human teratocarcinoma cells. Mol. Cell. Biol. **8**:1385–1397.
39. **Skowronski, J., and M. F. Singer.** 1986. The abundant LINE-1 family of repeated DNA sequences in mammals: genes and pseudogenes. Cold Spring Harbor Symp. Quant. Biol. **5**:457–464.
40. **Soares, M. B., E. Schon, and A. Efstratiadis.** 1985. Rat LINE1: the origin and evolution of a family of long interspersed middle repetitive DNA elements. J. Mol. Evol. **22**:117–133.
41. **Temin, H. M.** 1985. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Mol. Biol. Evol. **2**:455–468.
42. **Varmus, H. E.** 1983. Retroviruses, p. 411–503. *In* J. A. Shapiro (ed.), Mobile genetic elements. Academic Press, Inc., Orlando, Fla.
43. **Weiner, A. M., P. L. Deininger, and A. Efstratiadis.** 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. **55**:631–661.
44. **Xiong, Y., and T. H. Eickbush.** 1988. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. Mol. Cell. Biol. **8**:114–123.
45. **Xiong, Y., and T. H. Eickbush.** 1988. Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. Cell **55**:235–246.