JOURNAL OF VIROLOGY, June 2003, p. 6153–6166 Vol. 77, No. 11
0022-538X/03/$08.00+0 DOI: 10.1128/JVI.77.11.6153–6166.2003

Boudicca, a Retrovirus-Like Long Terminal Repeat Retrotransposon from the Genome of the Human Blood Fluke *Schistosoma mansoni*

Claudia S. Copeland,[1,2] Paul J. Brindley,[1]* Oliver Heyers,[2] Scott F. Michael,[1] David A. Johnston,[3] David L. Williams,[4] Alasdair C. Ivens,[5] and Bernd H. Kalinna[2]*

Department of Tropical Medicine, School of Public Health and Tropical Medicine, Tulane University Health Sciences Center, New Orleans, Louisiana[1]; Department of Molecular Parasitology, Institute for Biology, Humboldt University, Berlin, Germany[2]; Wolfson Wellcome Biomedical Laboratory, Department of Zoology, The Natural History Museum, London,[3] and Pathogen Sequencing Unit, The Sanger Institute, Hinxton,[5] England; and Department of Biological Sciences, Illinois State University, Normal, Illinois[4]

Received 5 September 2002/Accepted 26 February 2003

The genome of *Schistosoma mansoni* contains a proviral form of a retrovirus-like long terminal repeat (LTR) retrotransposon, designated *Boudicca*. Sequence and structural characterization of the new mobile genetic element, which was found in bacterial artificial chromosomes prepared from *S. mansoni* genomic DNA, revealed the presence of three putative open reading frames (ORFs) bounded by direct LTRs of 328 bp in length. ORF1 encoded a retrovirus-like major homology region and a Cys/His box motif, also present in Gag polyproteins of related retrotransposons and retroviruses. ORF2 encoded enzymatic domains and motifs characteristic of a retrovirus-like polyprotein, including aspartic protease, reverse transcriptase, RNase H, and integrase, in that order, a domain order similar to that of the *gypsy*/*Ty3* retrotransposons. An additional ORF at the 3′ end of the retrotransposon may encode an envelope protein. Phylogenetic comparison based on the reverse transcriptase domain of ORF2 confirmed that *Boudicca* was a *gypsy*-like retrotransposon and showed that it was most closely related to *CsRn1* from the Oriental liver fluke *Clonorchis sinensis* and to *kabuki* from *Bombyx mori*. Bioinformatics approaches together with Southern hybridization analysis of genomic DNA of *S. mansoni* and the screening of a bacterial artificial chromosome library representing ≈8-fold coverage of the *S. mansoni* genome revealed that numerous copies of *Boudicca* were interspersed throughout the schistosome genome. By reverse transcription-PCR, mRNA transcripts were detected in the sporocyst, cercaria, and adult developmental stages of *S. mansoni*, indicating that *Boudicca* is actively transcribed in this trematode.

Schistosomiasis is considered the most important of the human helminthiases in terms of morbidity and mortality (11, 15). *Schistosoma mansoni*, endemic to many areas of Africa and Latin America, is responsible for widespread disease in tropical developing countries in these regions. Symptoms of schistosomiasis include malaise, abdominal pain, diarrhea, fever, and myalgia during the acute phase, followed by a debilitating, sometimes fatal chronic phase characterized by inflammatory and fibrotic responses to schistosome eggs deposited in the liver, spleen, intestinal wall, and other organs (17, 37).

The life cycle of *S. mansoni* involves parasitism of both humans and *Biomphalaria glabrata* snails. Infectious larvae known as cercariae emerge from the snails into a body of water, where they initiate infection by direct penetration of human skin. In the human host, the worms develop into male and female adults which live together within the mesenteric venules of the intestines and release eggs into the bloodstream. To perpetuate the life cycle, the eggs traverse the intestinal wall, facilitated by secreted proteolytic enzymes and their spines, and pass out in the feces to fresh water. Although chemotherapy is available, its effectiveness is limited by continuous reinfection upon subsequent exposure to water containing cercariae. Furthermore, symptoms do not necessarily resolve upon chemotherapeutic cure of the infection, and chronic symptoms of the disease can remain with the patient for life. No vaccine is currently available.

Health education and drug therapy are the cornerstones of the World Health Organization's strategy to combat schistosomiasis. Although the endemic distribution of schistosomiasis has changed in the past 50 years, overall, the estimated number of infected persons and those at risk of infection has not been reduced (6, 11, 69). Moreover, interactions with other infectious diseases can induce increased pathology, as with coinfection with hepatitis C, in which liver damage can be more severe than in patients with either disease alone (26).

Mobile genetic elements appear to be a principal force driving the evolution of eukaryotic genomes (10, 41, 58), and these elements play an important role in the establishment of genome size (51). One of the major categories of mobile genetic elements is the long terminal repeat (LTR) retrotransposable element, i.e., the LTR retrotransposons and the retroviruses (23). These elements are of interest for their potential for horizontal transmission, among other attributes. Among the invertebrate retroviruses, such as *gypsy* (32) and *Tom* (62), acquisition of envelope protein-encoding genes from diverse viruses by unrelated LTR retrotransposons confers the ability to be infectious and thereby facilitates horizontal transmission. Malik et al. (42) theorized that this has occurred independently

* Corresponding author. Mailing address for Paul J. Brindley: Department of Tropical Medicine, SL-17, Tulane University Health Sciences Center, 1430 Tulane Ave., New Orleans, LA 70112-2699. Phone: (504) 988-4645. Fax: (504) 988-6686. E-mail: paul.brindley@tulane.edu. Mailing address for Bernd H. Kalinna: Department of Molecular Parasitology, Institute for Biology, Humboldt University Berlin, Philippstrasse 13, 10115 Berlin, Germany. Phone: 49 30 2093 6055. Fax: 49 30 2093 6051. E-mail: bernd.kalinna@rz.hu-berlin.de.

on several occasions during the evolution of the invertebrate retroviruses.

It is hoped that an enhanced understanding of the schistosome genome can be expected to lead to long-term strategies for the control of schistosomiasis. The genome of schistosomes, blood flukes of the phylum Platyhelminthes, is estimated at ≈270 Mbp per haploid genome (56), arrayed on seven pairs of autosomes and one pair of sex chromosomes (27, 28). Both the evolution and size of this genome may be highly influenced by mobile genetic elements. Indeed, more than half of the schistosome genome appears to be composed of or derived from repetitive sequences, to a large extent from retrotransposable elements (34–36).

Previously characterized schistosome mobile genetic elements include SINE-like retrotransposons (60, 18), LTR retrotransposons (36), and at least two families of non-LTR retrotransposons (35). Although active replication of these elements has not been definitively proven, mRNA transcripts encoding reverse transcriptase and endonuclease have been detected (34, 36), as has reverse transcriptase activity in schistosome extracts (29), suggesting that at least some of these elements are actively mobile within the genome. Indeed, actively replicating mobile genetic elements in other platyhelminths have been described as RNA intermediates (4) and DNA transposons (3, 53). Furthermore, the schistosome mobile genetic elements so far characterized are highly represented within the genome, with copy numbers ranging up to 10,000 per haploid genome (34).

Whereas evidence suggests the presence of a large number of families of retrotransposable elements within the chromosomes of the human blood flukes (20, 25), this is the first description of a full-length LTR-retrotransposon from the genome of the African and neotropical human blood fluke *Schistosoma mansoni*. We have termed this new *S. mansoni* retrotransposon *Boudicca*, after the queen of the Celtic Iceni tribe. In 61 A.D., Boudicca led her Celtic tribesmen in a revolt against the Romans that swept across ancient Britain, culminating in the destruction of Roman London (http://www.athenapub.com/boudicca.htm). The image of Boudicca driving her chariot in battle against the Roman legions, reminiscent of a mobile genetic element moving throughout the genome of the pathogenic *S. mansoni* parasite, inspired the designation of the new retrotransposon.

MATERIALS AND METHODS

Analysis of bacterial artificial chromosomes. Le Paslier et al. (39) described the construction and characterization of a bacterial artificial chromosome (BAC) library of the *Schistosoma mansoni* genome. The library, constructed in the BAC plasmid vector pBeloBac11 with genomic DNA from cercariae of a Puerto Rican strain of *S. mansoni* partially digested with *Hin*dIII, consists of ≈21,000 clones, with an average insert size ranging from 120 to 170 kb, providing ≈8-fold coverage of the schistosome genome. Numerous BAC end sequences determined from randomly selected clones from this library are now in the public databases. The sequence of the insert of one of the BAC clones of Le Paslier et al. (39), clone number 53-J-5, was determined recently in its entirety (133.5 kb) at the Sanger Institute, and this sequence is available from the Sanger Institute *Schistosoma* Genome Project site (ftp://ftp.sanger.ac.uk/pub/databases/Trematode/S.mansoni/BACs/53J5/).

Plasmid DNA from BAC clone 53-J-5 was prepared from liquid cultures of *Escherichia coli* with the PhasePrep BAC DNA kit (Sigma). Examination of the nucleotide sequence of BAC 53-J-5 by BlastX searches suggested that it included a retrotransposable element bearing a reverse transcriptase-encoding domain (not shown). A sequence of 5,858 nucleotides encoding a degenerate copy of the *Boudicca* element was located between residues 109662 and 115518 in the reverse orientation of clone 53-J-5, as detailed below. Subsequently, the sequences of the LTRs and of predicted open reading frame (ORFs) of the novel 53-J-5-associated retrotransposon were employed as the query in Blast searches to interrogate the TIGR database of *S. mansoni* genomic DNA BAC end sequences at http://tigrblast.tigr.org/euk-blast/index.cgi?project=sma1.

Sequences identified with strong matches were obtained subsequently from GenBank, as follows: LTR-specific BACs BAC 39-I-11 (GenBank accession no. BH201890), BAC 41-N-21 (BH203925), BAC 42-I-15 (BH200403), BAC 47-B-16 (BH206071), BAC 50-J-17 (BH210181), BAC 55-G-14 (BH204242), BAC 57-M-13 (BH210591), BAC 58-C-4 (BH202081), BAC 60-C-2 (BH203189), BAC 60-J-19 (BH203616), and BAC 62-M-3 (BH205111); ORF1-specific BAC 45-H-5 (BH206669), BAC 46-G-15 (BH209250), BAC 62-G-23 (BH211091), and BAC 62-N-16 (BH204125); ORF2 protease domain-specific BAC 43-P-17 (BH20912); ORF2 integrase, zinc finger, and DDE domain-specific BAC 53-C-10 (BH 200708), BAC 49-G-18 (BH 202479), BAC 45-E-19 (BH 211202), BAC 53-L-22 (BH 201551), and BAC 61-H-12 (BH 210420); and ORF2 integrase COOH-terminal domain BAC 45-L-19 (BH 199683), BAC 47-D-15 (BH 210140), BAC 48-A-14 (BH 208816), and BAC 51-L-17 (BH 207502), BAC56-N-2 (BH 200029). (We did not target BAC clones with high identity to the reverse transcriptase domain of ORF2 because the strong conservation of reverse transcriptase in evolutionary terms [68, 41, 45] makes it problematic to ensure that the Blast-identified BAC clones represented *Boudicca* rather than some other unidentified LTR retrotransposon from the genome of *S. mansoni*.)

Both the contiguous 53-J-5 copy of *Boudicca* and a composite of genomic sequence fragments assembled from these BAC ends were used to generate the consensus sequences. Alignments were established with ClustalW and MacVector software. Analysis of potential ORFs was accomplished with MacVector software, with parameters set to use stop codons as ends, an ATG codon as the beginning of ORF1, and codons after stops as the beginnings of subsequent ORFs. Since ORFs downstream of ORF1 in retrotransposons generally begin as a ribosome slip event rather than the beginning of a new mRNA transcript, downstream ORFs do not necessarily begin with an ATG (66). Structural analysis of predicted polypeptides for the presence of signal peptides and transmembrane domains was carried out with the on-line tools at http://www.cbs.dtu.dk/services/SignalP/(46) and http://www.cbs.dtu.dk/services/TMHMM-2.0/(33), respectively.

Phylogenetic analysis. In order to characterize the phylogenetic relationship of *Boudicca* to other mobile genetic elements, phylograms based on the reverse transcriptase domain of retroviral Pol were generated and rooted with reverse transcriptase from members of the *copia* family as the outgroup (61). Alignments of the reverse transcriptase sequences and generation of the bootstrapped phylogenetic tree (54, 47) were accomplished with ClustalX (63) and Njplot (50) software, with manual adjustment of gap size in *Ty1*/*copia* and *gypsy* so that the YVDD active sites aligned. Branch length ratios were preserved upon transfer into CorelDraw diagrams for display.

Sequences of reverse transcriptases used in the phylogenetic analysis were *nomad* (AF039416), *gypsy* (GNFFG1), *yoyo* (U60529), *Tom* (CAA80824), *Ted* (M32662), *ZAM* (CAA04050), *kabuki* (BAA92689), *CsRn1* (AAK07486), *Ty3* (S53577), *grasshopper* (M77661), *Maggy* (L35053), *Hsr1* (X92487), *sushi* (AAC33526), *Dea1* (T07863), *Osvaldo*, (CAB39733), *Ulysses* (X56645), *Woot* (U09586), *micropia* (X14037), *Blastopia* (Z27119), *Cyclops* (AJ000640), *Gulliver* (F243513), *Mag* (S08405), *Cer1* (U15406), feline leukemia virus (NP047255), human immunodeficiency virus type 1 (PO4585), human immunodeficiency virus type 2 (J04542), simian immunodeficiency virus (AAA47606), mouse mammary tumor virus (GNMVMM), *Ty1* (P47100), and *copia* (OFFFCP) and were obtained from the GenBank, EMBL, and PIR databases. Where possible, protein sequences were used directly from the database; otherwise, reverse transcriptase sequences were predicted by translation of ORF2, followed by removal of protease, RNase H, and integrase to leave an amino acid sequence for reverse transcriptase.

Developmental stages of schistosomes; isolation of schistosome nucleic acids. *Schistosoma mansoni* (Puerto Rican NMRI strain) was propagated by infecting BALB/c mice by intradermal injection of 70 cercariae collected from *Biomphalaria glabrata* snails that were maintained in the laboratory. Adult worms were recovered from infected mice by portal perfusion at 6 to 7 weeks after infection. Genomic DNA was extracted from adult worms and from cercariae. About 30 mixed-sex, adult *S. mansoni* worms were lysed in 0.1% sodium dodecyl sulfate–100 mM NaCl–50 mM Tris–20 mM EDTA (pH 8)–proteinase K at 500 μg/ml. Genomic DNA was extracted from the lysate by sequential partition against phenol-chloroform and chloroform-isoamyl alcohol, digestion with RNase A, a second partition against phenol-chloroform, and precipitation in ethanol in the

presence of sodium acetate. The *S. mansoni* genomic DNA was dissolved in 10 mM Tris–1 mM EDTA, pH 8.0.

For Southern blot analysis, genomic DNA was isolated from *S. mansoni* cercariae (NMRI strain), with a wet pellet of cercariae of ≈1 ml in volume, with the AquaPure genomic DNA kit from Bio-Rad. Schistosome eggs were isolated from the livers of infected mice at 8 weeks postinfection. For the initiation of in vitro cultures, miracidia were transformed into primary (mother) sporocysts by overnight culture in MEMSE-J with 10% fetal calf serum (31) or medium F with 10% bovine serum albumin (29) at 26°C and 5% CO_2. Shed ciliated plates were washed away.

Southern blots and BAC library screening. *S. mansoni* genomic DNA (≈33 μg per lane) and BAC clone 53-J-5 (insert size, 133.5 kb) were digested with restriction enzymes, separated through a 0.8% agarose gel by electrophoresis, transferred to nylon (Zeta-Probe GT; Bio-Rad) by capillary action (59), and cross-linked to the nylon by UV light. A *Boudicca*-specific probe was produced by PCR amplification with BAC 53-J-5 as the template and 5′-AACTGCAGA TGCACGGAATCACGGACT (forward) and 5′-GCTCTAGACTAAGATTCA GTCGGCAGATGC (reverse) primers, with restriction sites for *Pst*I and *Xba*I added to facilitate cloning into plasmid vectors. The probe, targeting part of the *gag* gene of *Boudicca*, was 385 bp in length, spanning residues 622 to 1006 of the 5,858 nucleotides of the 53-J-5 copy of *Boudicca.*

For Southern hybridization, the North2South Direct horseradish peroxidase labeling and detection kit (Pierce, Rockford, Ill.) was employed, with the washing and stringency conditions recommended by the manufacturer. This system uses direct labeling of the probe with horseradish peroxidase and a chemiluminescent signal detected with X-ray film (Fuji). To screen the high-density nylon filters representing the *S. mansoni* genomic DNA BAC library (39), the 385-bp probe (above) was labeled with digoxigenin with the digoxigenin labeling system from Roche (Indianapolis, Ind.). Hybridizations of the BAC high-density filters were carried out at 42°C overnight in the hybridization solution from Roche's digoxigenin labeling system, after which the nylon membranes were washed at 68°C for 30 min in 0.5× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–0.1% sodium dodecyl sulfate. Development of the digoxigenin-labeled signal was accomplished by immunoblotting with antidigoxigenin-alkaline phosphatase-conjugated immunoglobulins (Roche), after which CSPD [disodium 3-(4-methoxyspiro(1,2-dioxetane-3,2′ (5′-chloro)tricyclo[3,3.1.1$^{3.7}$]decan)-4-yl) phenylphosphate; (Roche] was used as the substrate for development of the digoxigenin chemiluminescence, which in turn was detected on X-ray film.

Retrotransposon gene copy number analysis. Comparative estimates of the copy number of *Boudicca* were obtained by a bioinformatics approach, wherein Blast analysis of the BAC end database of *S. mansoni* genomic sequences targeted better-characterized retrotransposable elements from *S. mansoni* for which copy numbers have been reported, including the non-LTR retrotransposons *SR1* and *SR2* (19, 20) and the SINE-like element Smα (60). *S. manosni* BAC end sequences from the Institute for Genomic Research and the Centre National de Sequençage were obtained from the TIGR ftp site (ftp.tigr.org/) and from Raymond Pierce (Institut Pasteur, Lille, France), respectively. Standalone Blast queries of the known repeat sequences against the BAC end sequences were performed.

In addition, the copy number estimate obtained from this bioinformatics approach was supported by Southern hybridization analysis of restriction enzyme-digested genomic DNAs alongside titrations of increasing quantities of *Hin*dIII-digested BAC 53-J-5. Densitometric analysis of Southern hybridization signals was accomplished with the Versa-Doc gel documentation system (Bio-Rad) and accompanying Quantity-One software (Bio-Rad). Densitometric data for the genomic DNA- and BAC-containing lanes in the Southern hybridization were used to estimate the copy number for *Boudicca* according to the formula $[(A/B) \times C]/E = F$. This formula was derived from two equations: $(A/B) \times C = D$ and $D/E = F$, where A is the number of copies of *Boudicca* in the BAC 53-J-5 lane (Fig. 8, lane 14), B is the density volume of the BAC 53-J-5 lane in units of optical density per mm^2, C is the density volume of the *S. mansoni* genomic DNA lanes in units of optical density per mm^2, D is the total number of copies of *Boudicca* per lane, E is the number of haploid genomes in each genomic DNA lane, and F is the total number of copies of *Boudicca* per haploid genome.

Finally, our copy number estimates were further supported by hybridization analysis of high-density filters of the BAC library, which represents a ≈8-fold coverage of the of *S. mansoni* genome (39).

RT-PCR. Twenty adult worms, 1,000 sporocysts, or 2,000 cercariae of *S. mansoni* were homogenized by mechanical disruption in lysis buffer (RNeasy RNA extraction kit; Qiagen). RNA was extracted according to the manufacturer's instructions under RNase-free conditions. After the RNA isolation, any residual DNA contamination was removed by digestion with RNase-free DNase (Promega). The RNA was precipitated to remove the DNase, dissolved in RNase-free water, and used as a template for oligo(dT)-primed reverse transcriptions (RT) with Moloney murine leukemia virus H$^-$ point mutant reverse transcriptase (Promega). The resulting cDNA was used as the template in PCR experiments with primers specific for the reverse transcriptase of *Boudicca* (forward, 5′-CCCTAAAAAGGACAGCAACGATTG; reverse, 5′-GGTTCCGATTTGG CATTTCTG, 447-bp product) and with primers forward (5′ ATGCACGGAA TCACGGAC) and reverse (5′-GAGTGATGATGGCGGTTTTAGG) spanning the region from *gag* to the reverse transcriptase (1,571-bp product overlapping ORF1 and ORF2 of *Boudicca*) (see Fig. 8A below).

PCR amplification was performed under the following cycling conditions: 94°C for 5 min; 35 cycles of 94°C for 1 min, annealing for 1 min at primer-dependent temperature (see below), and extension at 72C for 1 min; and a final extension at 72°C for 10 min. A nested PCR was subsequently performed on the 447-bp reverse transcriptase product (primers: forward, 5′-CCAACTGATGTTTATCG GCGTC; reverse, 5′-GAGTGATGATGGCGGTTTTAGG, 183-bp product) to validate the specificity of the first round of PCR. Annealing temperatures were 62°C for the 447- and 183-bp reverse transcriptase fragments and 58°C for the 1,571-bp fragment spanning both ORF1 and ORF2. The transcription reaction was validated with primers forward (5′-GATTTGCGCTATGGCTTC) and reverse (5′-GGCCATCACCATACTAGC, 342-bp product) specific for the *S. mansoni* cytochrome *c* oxidase subunit 1 gene (GenBank accession no. AF101196) (49). Negative control reactions were carried out with DNase-treated RNA as the template. Genomic DNA from adult *S. mansoni* was used as the positive control (64).

Nucleotide sequence accession numbers. Nucleotide sequences reported here have been assigned GenBank accession numbers as follows: consensus sequences of the *Boudicca* LTR, accession no. BK000439; *gag* region, BK000440; integrase-zinc finger-DDE regions, BK000441; and integrase carboxy terminus, BK000444.

RESULTS

***Boudicca*, an LTR retrotransposon from the genome of *S. mansoni*.** *Boudicca* appears to be the first full-length LTR retrotransposon from *Schistosoma mansoni* to be described. Figure 1A is a schematic diagram outlining the structural features of this new retrotransposon, which is ≈5.8 kb in length. Whereas a full, contiguous copy of this element is contained within the BAC clone 53-J-5, this copy appears to have been degraded by a number of mutations (Fig. 1B and 2). In order to reconstruct the genomic structure of a putatively active *Boudicca* element, we examined sequences of BAC ends from *S. mansoni* genomic DNA (39) that are available at the TIGR website (above). Fragments from a number of discrete copies of *Boudicca* were aligned to assemble regions of consensus that resolved mutations evident in the 53-J-5 copy of the element (see Fig. 3).

The predicted structure of *Boudicca*, based on the consensus sequence, is shown in Fig. 1A. Both the 5′ and 3′ LTRs of *Boudicca* begin with TGT and end with TCA, forms of the generalized LTR start and end motifs, TGN and NCA, respectively, of LTR retrotransposons, known as the direct inverted repeats (7, 16) (Fig. 1). Two other hallmarks of LTRs, motifs for promoter initiation, are apparent in the *Boudicca* LTRs; a CAAT box at positions 27 to 30 and two TATA boxes at positions 188 to 193 and 208 to 213 (Fig. 2). Based on examination of the LTRs in seven discrete BAC clones, the 5′ and 3′ LTRs of *Boudicca* are 328 bp in length and appear to be direct repeats. The consensus sequence of the LTR of *Boudicca* has been assigned GenBank accession no. BK000439. The 5′ and 3′ LTRs of the *Boudicca* element in BAC clone 53-J-5 are 88.4% identical to one another, indicating that the copy in BAC clone 53-J-5 is partially degraded. The difference in size between the two (the 3′ LTR is 301 residues long and the 5′ LTR is 327 residues long, 1 bp shorter than the consensus sequence length of 328) appears to be due to the deletion

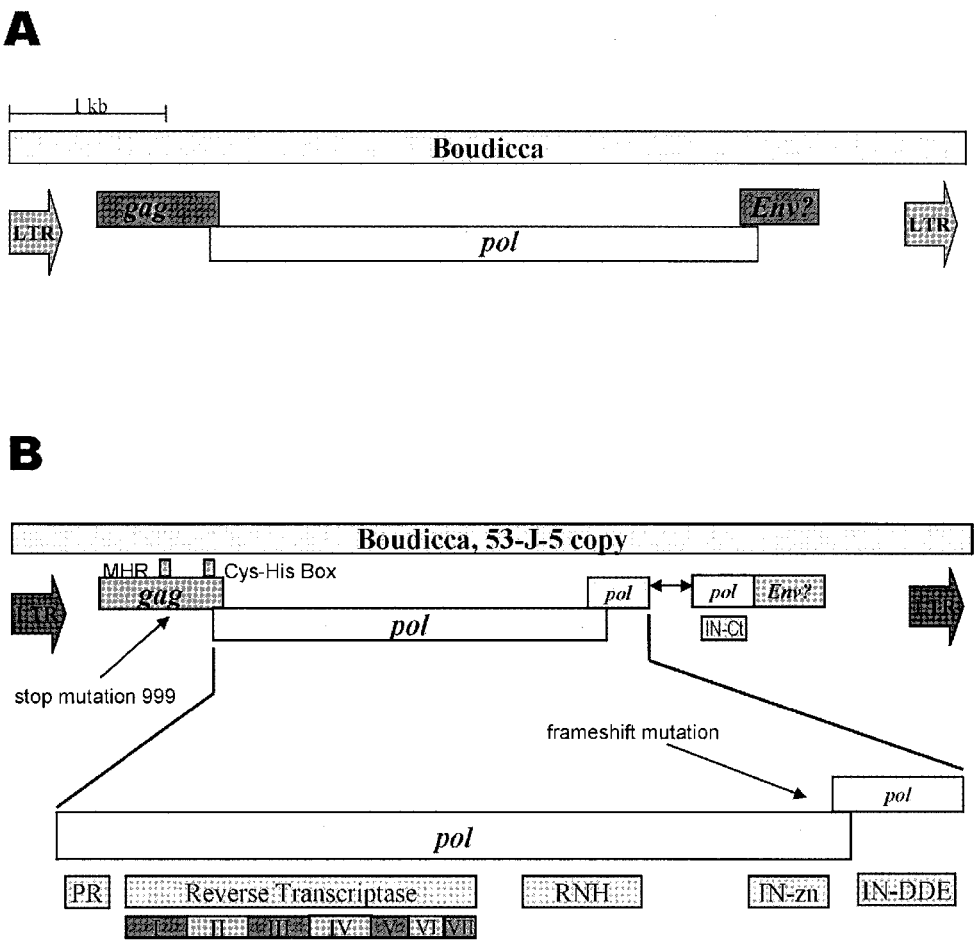


FIG. 1. (A) Schematic representation of the predicted structure of the *Schistosoma mansoni* retrotransposon *Boudicca.* (B) Schematic representation of a contiguous but degraded copy of *Boudicca*, found in the insert of BAC 53-J-5 of the *S. mansoni* BAC library of Le Paslier et al. (39). MHR, major homology region; PR, protease, RNH, RNase H; IN-zn, integrase zinc finger domain; IN-DDE, integrase DDE domain; IN-Ct, integrase C-terminal domain, env?, possible envelope protein.

of CTTTCCTACCTCTTCTCGTCTGACTTCTGATTTC after residue 214 of the 3′ LTR and the insertion or substitution of ATATTATA at the same location (after residue 214 of the 3′ LTR). This results in a net loss of 26 bp from the 3′ LTR relative to the 5′ LTR (Fig. 2 and Fig. 3A).

A search of the nonredundant database at GenBank with the consensus sequence of the *Boudicca* LTR did not reveal significant matches to LTRs of other retrotransposons or indeed any significant matches at all. However, BlastN searches of the Genome Survey Sequences (GSS) database at GenBank returned 185 matches with significant scores of 170 and greater, over 150 bp and longer, all of which were from *S. mansoni* (mostly to BAC end sequences) (not shown). Furthermore, a search of the nonhuman, nonmouse expressed sequence tag database revealed four significant matches to *S. mansoni* expressed sequence tags and no other matches (not shown). The length of the LTRs of most retrotransposons ranges in size from about 200 to about 600 bp (45): for comparison of sizes, the LTRs of related retrotransposons are *Gulliver*, 259 bp; *kabuki*, 272 bp; *gypsy*, 526 bp; *ZAM*, 594 bp; *Ted*, 273 bp; and *Tom*, 474 bp.

***gag* gene of *Boudicca* encodes a distinctive Cys-His box motif, CHCC.** Downstream of its 5′ LTR, *Boudicca* exhibits two, and possibly three, protein-encoding reading frames, ORF1, ORF2, and ORF3, and terminates in the 3′ LTR (Fig. 1 and 2). ORF1 of the BAC 53-J-5 copy included a stop codon (TGA) at position 999 that would result in a truncated Gag precursor. However, the consensus sequence from four other BACs spanning that particular region revealed that the putatively active *Boudicca* element contained an intact ORF1 encoding Gag (Fig. 3B). This *gag* gene, encoding 276 amino acids, starts with an ATG (methionine) codon at nucleotide positions 505 to 507 of *Boudicca*, ends at nucleotide position 1332, and encodes at least two conserved domains of the retroviral structural proteins encoded by *gag* (matrix, capsid, and nucleocapsid) (Fig. 1 and 2).

Near its NH_2 terminus, the Gag protein included a major homology region, orthologous in sequence to the major homology region from numerous retroviruses (Fig. 4A). The major homology region is a region within the capsid protein that is highly conserved among several retroviruses as well the yeast retrotransposon *Ty3* (13, 14). The major homology region is required for infectivity and is involved in capsid assembly (8, 52). Carboxy terminal to the major homology region, the Gag

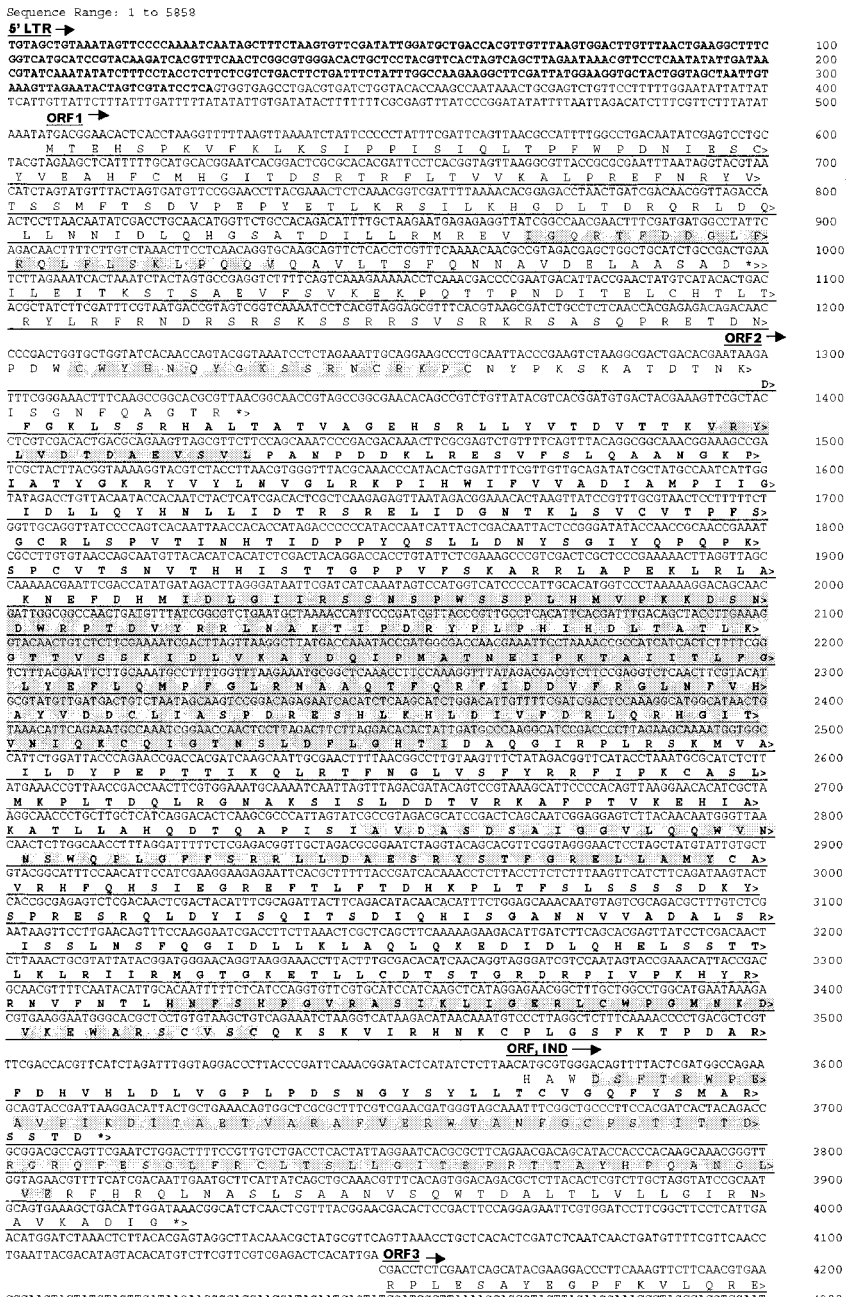
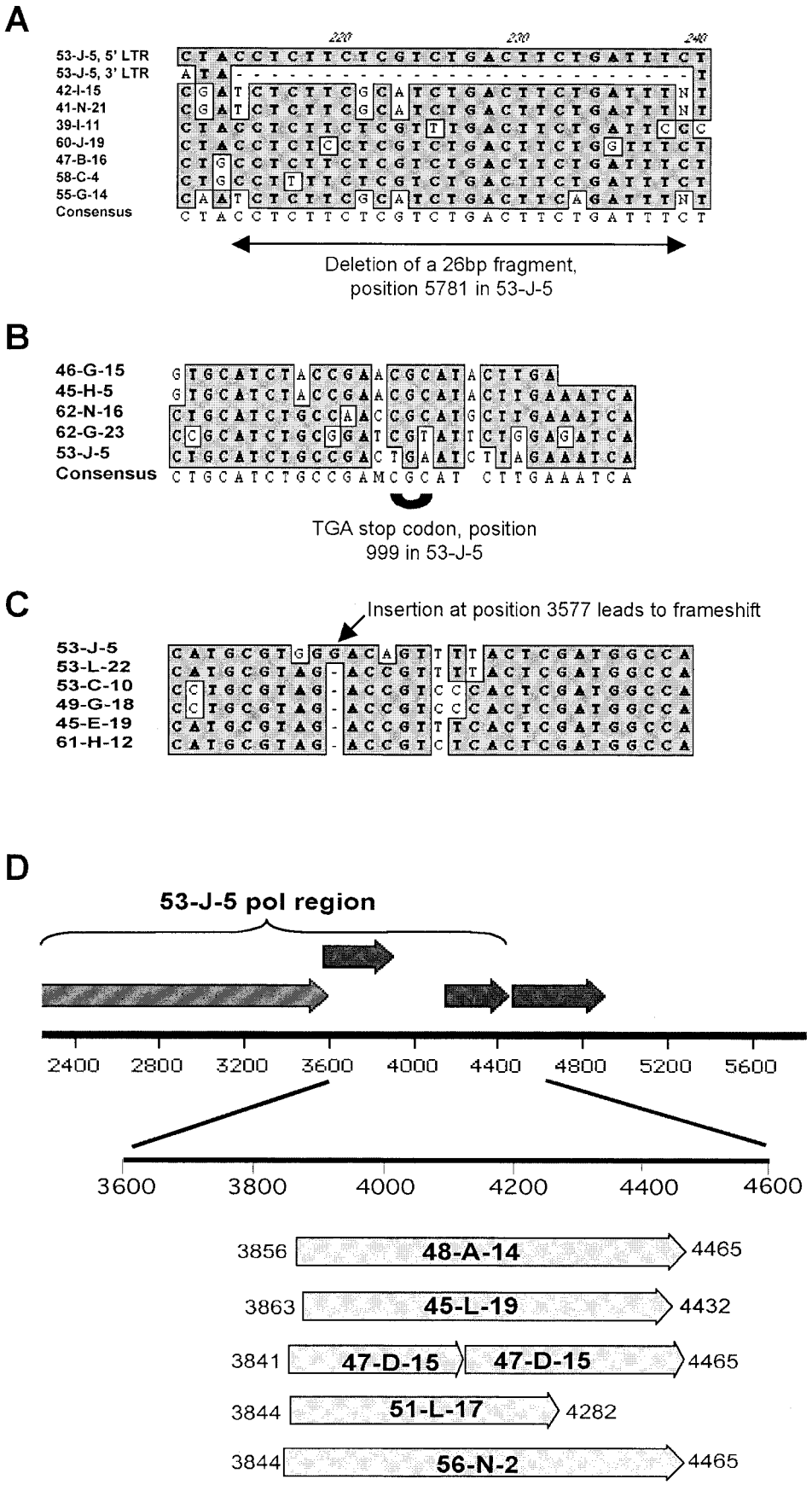


FIG. 2. Annotated sequence of *Boudicca* from BAC clone 53-J-5. Functional domains and conserved regions of protein domains are shaded and labeled to the left of the sequence. Amino acid sequences are shown in frame below their corresponding nucleic acid sequence. The amino acid sequence of the +3 reading frame (ORF2, *pol*) is shown in bold; amino acids in the +1 reading frame (all other proteins) are shown in lightface. LTRs are also shown in bold. MHR, major homology region; CHB, Cys-His box; PR, protease; RT, reverse transcriptase; RNH, RNase H; INZ, integrase zinc finger domain; IND, integrase DDE domain.


A
220
230
240
53-J-5, 5' LTR
53-J-5, 3' LTR
42-I-15
41-N-21
39-I-11
60-J-19
47-B-16
58-C-4
55-G-14
Consensus
Deletion of a 26bp fragment,
position 5781 in 53-J-5
B
46-G-15
45-H-5
62-N-16
62-G-23
53-J-5
Consensus
TGA stop codon, position
999 in 53-J-5
C
Insertion at position 3577 leads to frameshift
53-J-5
53-L-22
53-C-10
49-G-18
45-E-19
61-H-12
D
53-J-5 pol region
2400
2800
3200
3600
4000
4400
4800
5200
5600
3600
3800
4000
4200
4400
4600
3856
48-A-14
4465
3863
45-L-19
4432
3841
47-D-15
47-D-15
4465
3844
51-L-17
4282
3844
56-N-2
4465

protein included a conserved domain of the nucleocapsid protein, involved in RNA binding, known as a Cys-His box (21, 44). As shown in the alignment in Fig. 4B, orthologous Cys-His domains are present in the *CsRn1* and *Kabuki* retrotransposons. This *Boudicca*/*CsRn1*/*Kabuki* CHCC motif, specifically $\mathbf{CX_2HX_9CX_3C}$, is dissimilar to the more usual $\mathbf{CX_2CX_4HX_4C}$ (i.e., CCHC) zinc finger motif seen with Cys-His nucleic acid binding motifs of nucleocapsid proteins of most other retrotransposons and retroviruses (4, 21, 44).

***gypsy*-like *pol* ORF2.** The second open reading frame, ORF2, started at about residue 1299 in the +3 frame and encoded a retrovirus-like polyprotein, polymerase, of ≈1,060 amino acid residues (Fig. 1, 2, and 5). The *Boudicca* polymerase included four enzymatic domains, protease, reverse transcriptase, RNAse H, and integrase, in that order (Fig. 1, 2, and 5), characteristic of the domain order and structure of the polyprotein of retroviruses and *Ty3*/*gypsy* LTR retrotransposons. Part of the sequence of the protease domain of *Boudicca*, in particular the region of the active-site triad of residues including the catalytic Asp residue, is shown as a ClustalW-generated alignment in Fig. 5. Strong conservation with the active-site regions of proteases from other *gypsy*-like retrotransposons was apparent, although the more usual DT/SG motif was mutated to DTD in the 53-J-5 copy of *Boudicca*. However, the protease of active *Boudicca* elements may have DTG, since the *Boudicca* protease in BAC end BAC 43-P-17 (BH20912) had DTG (not shown). The processing sites for protease have yet to be determined, and hence we do not yet know the lengths of the mature forms of each of the protease, reverse transcriptase, RNase H, and integrase domains of the *Boudicca* polyprotein. Protease enzymes from retroviruses are ≈100 amino acids in length (30), and it can be anticipated based on identity that the protease of *Boudicca* will be of similar dimensions.

The reverse transcriptase of LTR retrotransposons has the key role of reverse transcribing the RNA genome into the proviral DNA form of the element within the cytoplasm of the host cell for subsequent integration into a chromosome of the host genome. The seven conserved domains of reverse transcriptase, as described by Eickbush and colleagues (43, 68), were present in *Boudicca* and included ≈180 amino acid residues (Fig. 1, 2, and 5). These conserved domains are presented in amino acid alignment with the orthologous reverse transcriptase regions of *CsRn1*, *kabuki*, and related LTR retrotransposons, where strong identity with these other reverse transcriptases was apparent (Fig. 5). Furthermore, we focused on the sequence of the reverse transcriptase residues of *Boudicca* to examine its phylogenetic relationships to other retrotransposons (below).

In conjunction with reverse transcriptase, RNase H is involved in transcription of the RNA genome to the proviral DNA genome; RNase H cleaves the original RNA template of the RNA-DNA hybrid. Also, it generates a polypurine tract, the primer for plus-strand DNA synthesis, and it removes the RNA primers from newly synthesized minus and plus strands of the proviral DNA (41). RNase H enzymes exhibit a characteristic active site that includes four conserved carboxylate residues, three Asps and a Glu; all four appear to be conserved in the *Boudicca* RNase H (Fig. 5 and not shown) (41). The locations of these active sites indicated that the RNase H enzyme

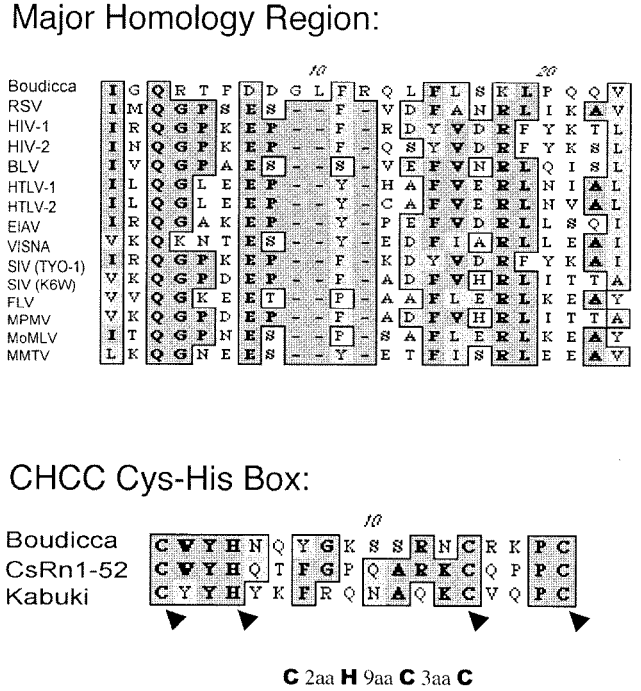


FIG. 4. Alignments of *Boudicca* Gag functional domains with those of LTR-type retrotransposons and retroviruses. Two conserved domains of the Gag protein, the Cys-His box and the major homology region, were located within the first open reading frame of *Boudicca*. Alignments with related elements (Cys-His box) and retroviruses (major homology domain) are shown. For the Cys-His box comparisons, translations of DNA sequences rather than direct protein sequences were used. Accession numbers, followed by location within the nucleotide sequence in parentheses, are *CsRn1*, AY013561 (1261 to 1329), and *kabuki*, AB032718 (6047 to 6100). Retroviral major homology regions were obtained from reference 12. aa, amino acids; RSV, Rous sarcoma virus; HIV, human immunodeficiency virus; BLV, bovine leukemia virus; HTLV, human t-cell lymphotropic virus; EIAV, equine infectious anemia virus; VISNA, visna virus; SIV, simian immunodeficiency virus; FLV, feline leukemia virus; MPMV, Mason-Pfizer monkey virus; MoMLV, Moloney murine leukemia virus; MMTV, mouse mammary tumor virus.

FIG. 3. Resolution of specific mutated regions of the 53-J-5 copy of *Boudicca*. (A) Alignment of seven BAC ends containing the *Boudicca* LTR sequence indicate that the short 3′ LTR in 53-J-5 is due to a deletion specific to that LTR. (B) Alignment of four BAC ends containing the middle portion of the *Boudicca gag* gene indicated that the stop codon at position 999 of 53-J-5 was a mutation unique to that copy. (C) Alignment of five BAC ends containing the region at the end of the long, +3 *pol* ORF indicate that the frameshift at the end of this ORF in 53-J-5 is due to an insertion mutation at position 3577 in the 53-J-5 copy. (D) ORF analysis of five BAC ends spanning the region between the integrase DDE domain and the integrase C-terminal domain indicated that a large gap of noncoding sequence in the 53-J-5 copy (positions 3925 to 4149) corresponds to an intact *pol* open reading frame in the putatively active copy of *Boudicca*. Nomenclature for the BAC ends is that used in the TIGR database to classify the *Schistosoma mansoni* BAC library clones (http://tigrblast.tigr.org/euk-blast/index.cgi?project=sma1).

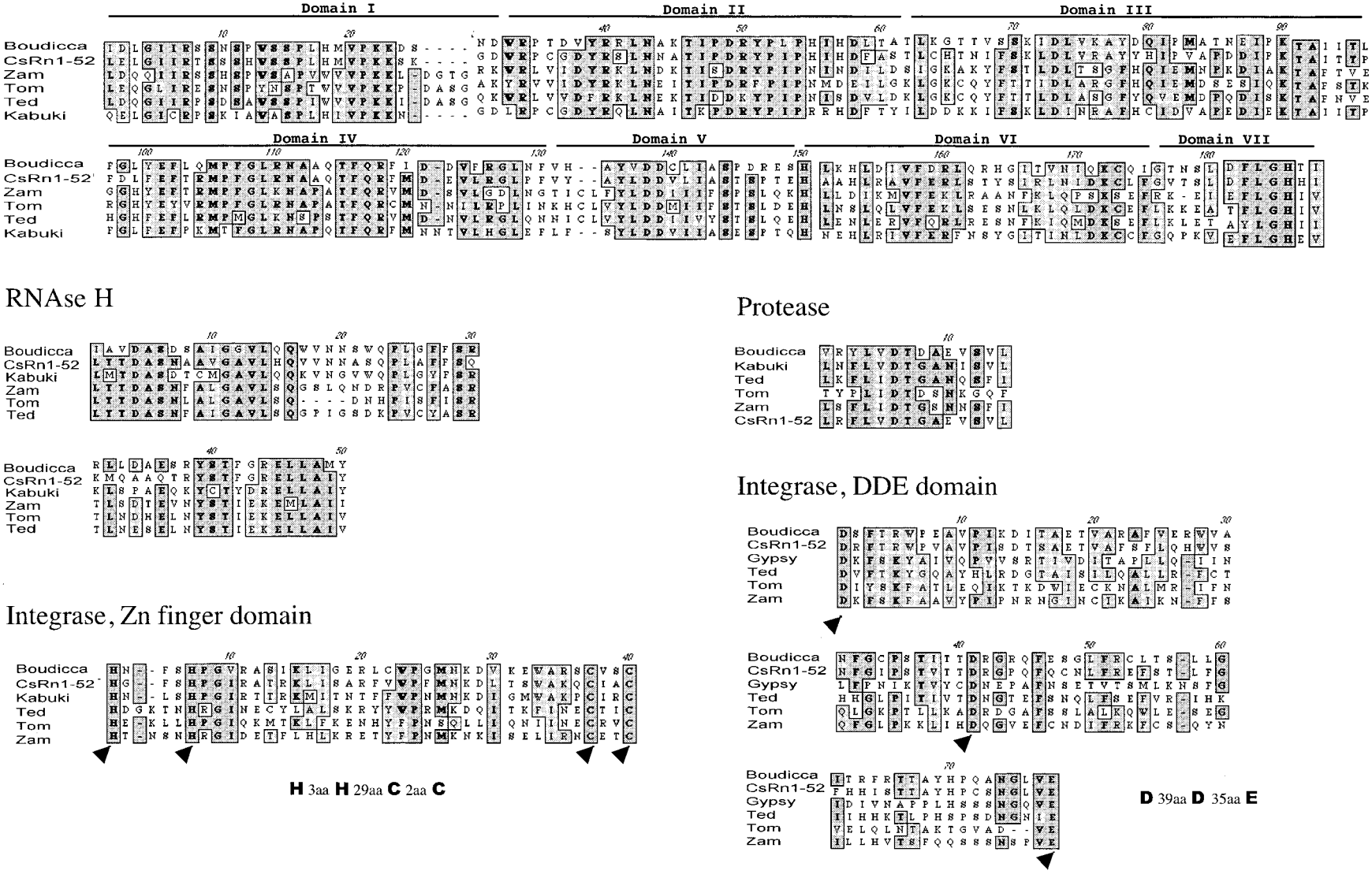


FIG. 5. Alignments of *Boudicca* polymerase functional domains with those of five other LTR-type retrotransposons. Five of six *gypsy*-like LTR retrotransposons (*CsRn1*, *kabuki*, *ZAM*, *Ted*, *Tom*, and *gypsy*) were used to find conserved regions of *pol* in *Boudicca*. ClustalW alignments are presented for conserved regions of reverse transcriptase, RNase H, aspartic protease, integrase, zinc finger domain, and integrase DDE (catalytic) domain. The conserved integrase motifs HHCC (for the zinc finger domain) and DDE (for the central catalytic domain) are marked with arrowheads. Identical amino acids appear in boldface and are shaded dark gray; similar amino acids are in lightface and are shaded light gray. Accession numbers are as follows: *CsRn1*, AAK07486; *kabuki*, BAA92690; *ZAM*, CAA04050; *Ted*, B36329; *Tom*, CAA80824; and *gypsy*, GNFFG1. aa, amino acids.

domain of *Boudicca* extended for ≈140 residues COOH-terminal to reverse transcriptase (not shown).

The integrase of *Boudicca* resided at the COOH terminus of the polymerase precursor, like other *Ty3/gypsy* like retrotransposons. Integrase mediates integration of the DNA provirus into the host chromosome. Integrase is composed of three domains: the NH_2-terminal zinc-binding domain, the central catalytic domain (DDE domain), and a COOH-terminal, nonspecific DNA binding domain. Alignments of the *Boudicca* integrase zinc-binding domain, with its characteristic $\mathbf{HX_3HX_{29}CX_2C}$ motif, and of the catalytic DDE domain with its D-39aa-D-35aa-E spacing, where Xaa represents the indicated number of amino acids, are presented in Fig. 5. It was clear that the integrase domain of *Boudicca* was orthologous to that of *kabuki*, *CsRn1*, *Ted*, and other LTR retrotransposons. Within the 53-J-5 copy of *Boudicca*, the insertion of a G nucleotide at position 3577 led to a frameshift mutation that disrupted ORF2 in the region encoding the integrase (Fig. 1, 2, and 3). However, this frameshift is apparently absent from active copies of *Boudicca* because it was not seen in five other copies that we examined (Fig. 3, panel C).

Furthermore, within the 53-J-5 copy of *Boudicca*, an additional open reading frame starts at position 4150 and extends to 4923, with the stop codon TAG at 4462 to 4464 (Fig. 1B and 2). The 5′ half of this reading frame shows identity to the integrase of *CsRn1* (AAK07485, AAK07486, and AAK07487) and several other elements (e.g., BAA92695.1, NP_178653, and BAA92696) encoding the carboxy-terminal portion of integrase. This leaves 225 bp of noncoding sequence interrupting the coding region of the integrase (Fig. 1B). To determine whether this mutation was present in other copies of *Boudicca*, BAC ends homologous to this region were examined for open reading frames (Fig. 3D). In contrast to the 53-J-5 copy, four of five other *Boudicca* sequences spanning this site contained uninterrupted coding sequences, indicating that the *pol* gene of the putatively active *Boudicca* is composed of a single ORF spanning from 1299 to 4464 (Fig. 1A, 1B, 2, and 3D). Furthermore, the reading frame of the fifth *Boudicca* positive BAC end, 47-D-15, was interrupted by only a single stop codon.

Envelope glycoprotein encoded by ORF3 of *Boudicca*? The region from positions 4465 to 4923 (Fig. 2) forms another, discrete open reading frame (Fig. 1, 2, and 3). Although in the same reading frame (+1) as the carboxy terminus of the integrase, this coding region appears to encode a polypeptide that, so far, does not show identity to other known proteins. By comparison with the genome organization of related retroelements, including the errantiviruses (41, 42), this 3′-situated ORF may encode a retrovirus-like envelope protein (Fig. 1, 2, and 3D). Whereas BlastX searches with this sequence did not reveal significant matches, the envelope proteins encoded by other LTR retrotransposons and retroviruses exhibit low or little identity, making identification by sequence alignment difficult (38). However, many envelope functions appear to have similar structural characteristics, including signal peptides, transmembrane domains, glycosylation, and Cys bridges (40). Indeed, the deduced polypeptide of *Boudicca* was predicted to include a signal peptide of ≈50 amino acids, with cleavage at TLG*CS (Fig. 2), according to the algorithm of Nielsen et al. (46). Also, it has a predicted transmembrane domain in the vicinity of amino acid residues 78 and 79, according to the algorithm of Krogh et al. (33), and two potential disulfide bridges. These structural features usually occur in viral envelopes (48).

Interestingly, all three BAC end sequences which span the region 3′ of nucleotide 4465 on the 53-J-5 copy exhibit a stop codon at positions 4465 to 4467 (Fig. 3). In addition, clone 58-D-2 also includes positions 4465 to 4467 and also has this stop codon. Therefore, unlike the premature stop mutation at position 999 in the *gag* domain of the 53-J-5 copy (Fig. 1, 2, and 3), which is clearly a mutation in a nonfunctional copy of *Boudicca*, the stop codon downstream of nucleotide 4465 appears to be a characteristic of active *Boudicca* elements. This also suggests that the region after this stop codon encodes a third polypeptide, discrete from Gag and Pol. Based on the structures of related retrotransposons, including the errantiviruses, the logical candidate for this region would be an envelope protein.

***Boudicca* is related to *kabuki* and to *gypsy*.** The predicted reverse transcriptase domain of *Boudicca* was aligned with the orthologous domains of numerous other LTR retrotransposons from the *Ty3/gypsy* and *Ty1/copia* families and the vertebrate retroviruses. Phylogenetic comparison of the reverse transcriptase domains of these diverse elements revealed that *Boudicca*'s closest relatives were *kabuki* from *Bombyx mori* (1) and *CsRn1* from the related trematode parasite *Clonorchis sinensis* (4) (Fig. 6), placing *Boudicca* in the newly characterized *Kabuki/CsRn1* clade of *gypsy*-like retrotransposons (4). Although not as closely related as to *kabuki* and *CsRn1*, *Boudicca* was also closely related to *gypsy* and related errantiviruses from insects, including *nomad*, *ZAM*, and *Tom* from species of *Drosophila*, *Ted* from *Trichoplusia ni*, and *yoyo* from *Ceratitis capitata* (41). These closely related retroelements, along with the vertebrate retroviruses and several other *gypsy*-like LTR retrotransposons, form a group of LTR elements distinct (bootstrap = 99.5%) from the *Ty1/copia* assemblage.

The identity, number, and order of domains encoded by the ORFs of *Boudicca* are, in general, similar to those of members of the *Ty3/gypsy* family and dissimilar to those of members of the *Ty1/copia* family, where, among other differences, integrase precedes reverse transcriptase in *pol* (41). As described above, like the errantiviruses such as *gypsy* (32) and *ZAM* (38), *Boudicca* may also have a third ORF encoding an envelope and, if so, may likewise be capable of horizontal transmission as an extracellular infectious particle. *Boudicca* was clearly distinct from *Gulliver*, an LTR retrotransposon reported previously from the Asian schistosome *Schistosoma japonicum* (36) (Fig. 6).

Numerous copies of *Boudicca* in the schistosome genome. A Southern hybridization analysis of *S. mansoni* genomic DNA digested with *Bam*HI, *Hin*dIII, *Pst*I, and *Kpn*I probed with a *Boudicca* ORF1-specific probe was carried out. *Bam*HI cuts once, *Hin*dIII cuts twice, and neither *Kpn*I nor *Pst*I cuts within the 5,858 bp of the *Boudicca* copy in BAC 53-J-5. The probe does not contain restriction sites for these enzymes. In adjacent lanes, *Hin*dIII-digested BAC 53-J-5, of which the insert is 133.5 kb and contains a single copy of *Boudicca*, was serially diluted to contain 3.8, 38, 380, 3,800,3.8×10^4, 3.8×10^5, 3.8×10^6, 3.8×10^7, 3.8×10^8, and 3.8×10^9 copies of *Boudicca*, respectively. Densitometric analysis of the resulting Southern hybridization signals provided copy number estimates for the *Boudicca* retrotransposon of 2,630, 2,582, 2,099, and 2,026 per haploid genome for the

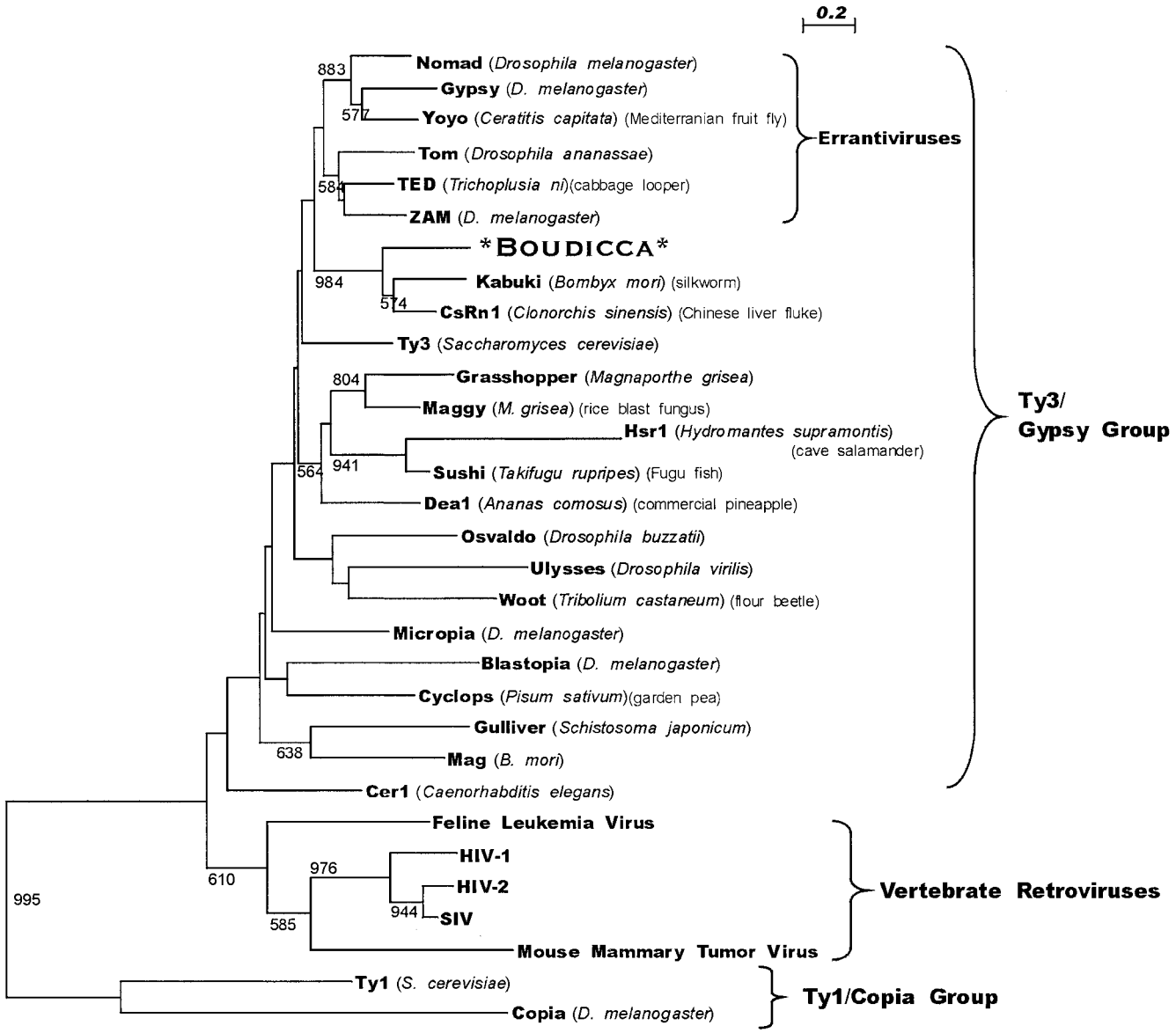


FIG. 6. Phylogeny of *Boudicca* and other retrotransposons and retroviruses based on their reverse transcriptase domains (68). Bootstrapped tree rooted with the phylogenetically distinct LTR retrotransposons *Ty1* and *copia*. The tree was constructed with ClustalW and Nj Plot and subsequently imported into CorelDraw for annotation, where branch lengths were preserved. Bootstrap values for 1,000 replicates are shown, where values of greater than 500 were obtained.

*Bam*HI-, *Hin*dIII-, *Pst*I-, and *Kpn*I-digested preparations, respectively (Fig. 7). Together, these Southern hybridization signals indicated that there were between 2,000 and 3,000 copies of *Boudicca* in the schistosome genome.

Next, BlastN searches (version 2.2.3) of the *S. mansoni* BAC end sequences in the TIGR and CNS databases, i.e., 26,284 GSS (as of 12 July 2002) with the 5,858 bp of the *Boudicca* copy in BAC clone 53-J-5 as the query and 1,000 descriptions, returned more than 743 hits with a score of >100. In comparison, queries with the non-LTR retrotransposons *SR1* (partial sequence, 2,337-bp consensus, U66331) and *SR2* (AF025673, 3,913 bp) from *S. mansoni* returned 552 and 989 hits, respectively. Other comparisons with the Smα retrotransposon (M27676, 331 bp), the 18S rRNA gene of *S. mansoni* (M62652, 1,739 bp), and the cDNA encoding *S. mansoni* cathepsin D protease (U60995, 1,285 bp) returned 578, 1, and 0 hits, respectively (Table 1). Since gene copy numbers have been estimated for all these query sequences (*SR1*, 200 to 2,000 copies; *SR2*, 1,000 to 10,000 copies; Smα, 7,000 to 10,000 copies; 18S rRNA gene, 100 copies; and cathepsin D, one or a few copies [19, 20, 57, 60, 67]), and since the hit value for *Boudicca* was intermediate between those for *SR2* and Smα, we suggest that there may be as many as 1,000 to 10,000 copies of this LTR retrotransposon interspersed throughout the genome of *S. mansoni*.

Third, the copy number of *Boudicca* was also estimated by screening the BAC library of Le Paslier et al. (39) with the same ORF1-specific *Boudicca* probe used for the Southern hybridization. Approximately 5.1% of the library clones were positive,

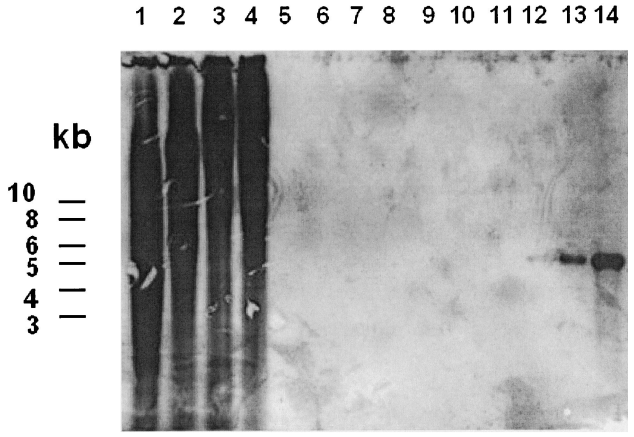


FIG. 7. Southern hybridization analysis of genomic DNA of *Schistosoma mansoni* and BAC clone 53-J-5 probed with a *Boudicca gag*-specific probe, revealing that *Boudicca* is a multicopy number element. *S. mansoni* genomic DNA was digested with *Bam*HI (lane 1), *Hin*dIII (lane 2), *Pst*I (lane 3), and *Kpn*I (lane 4), and titrations of BAC 53-J-5 DNA digested with *Hin*dIII (lanes 5 to 14, containing 3.8, 38, 380, 3800, 3.8×10^4, 3.8×10^5, 3.8×10^6, 3.8×10^7, 3.8×10^8, and 3.8×10^9 copies of *Boudicca*, respectively). Molecular size standards (in kilobase pairs) are shown at the left. Estimates of *Boudicca* copy number from densitometry scans of the X-ray film used to produce this image were calculated as follows. The number of haploid genomes contained in each digested *S. mansoni* genomic DNA lane (lanes 1, 2, 3, and 4) were calculated to be 1.1×10^8, based on the mass of genomic DNA loaded (33,000 ng/lane) and the mass of the *S. mansoni* haploid genome (2.94×10^{-4} ng) (270 Mb). Total density volume, in units of optical density per mm^2, represents the total positive signal under consideration. With this as a measure of total *Boudicca* copy number for a given area of the blot, total density volumes were obtained for each band or smear. Lanes 1, 2, 3, and 4 containing genomic DNA which represented a known number of copies of the haploid genome of *S. mansoni* were compared with the positive band in lane 14 containing *Hin*dIII-digested BAC 53-J-5, which represented 3.8×10^9 copies of *Boudicca*, according to formula given in the text. The hybridization signal shown here was obtained after a 5-h exposure. Copy number was determined from the same experiment after a 30-min exposure.

indicating a copy number of ≈1,200 copies per haploid genome (not shown). Together, these three approaches, all of which provided concordant estimates, confirmed that *Boudicca* was present at high copy number, between 1,000 and 10,000 copies per haploid genome.

***Boudicca* is transcribed in developmental stages of the schistosome.** RT-PCR targeting mRNA from the developmental stages of *S. mansoni* revealed that ORF2 sequences of *Boudicca* were expressed in all stages examined, mixed-sex adults, cercariae, and sporocysts (Fig. 8, panels A and B). This result was confirmed with a nested PCR strategy to reamplify amplicons of 447 bp from the first-round PCR with *Boudicca*-specific primers. With nested primers targeting a shorter region within the region of ORF2 targeted in the first-round PCR, the expected nested PCR product of 183 bp was obtained for cercariae, sporocysts, and mixed-sex adult stages of the schistosome, indicating that mRNAs for the entire retrotransposon were transcribed in all three of the developmental stages examined. In addition, RT-PCR with a forward primer targeting a site in ORF1 and reverse primer targeting a site in ORF2 produced a band of the expected size, 1,571 bp, indicating that mRNAs for the entire retrotransposon were transcribed in schistosomes. The control RT-PCR showed that cytochrome oxidase was also expressed in all developmental stages tested, with the expected product of 342 bp, verifying the integrity of the schistosome mRNA preparations examined in this study (Fig. 8C) (49). Omission of either the template cDNA or exogenous reverse transcriptase resulted in the absence of detectable amplicons, ruling out contamination of cDNAs with genomic DNAs (Fig. 8).

DISCUSSION

***Boudicca*, the first LTR retrotransposon to be characterized in the *Schistosoma mansoni* genome.** Blood flukes of the trematode genus *Schistosoma* have a comparatively large genome, estimated for *S. mansoni* at 270 Mbp (56). Karyotype comparisons indicate that other schistosomes have a genome of similar size and complexity (27, 28). The genome of *S. mansoni* is about one tenth the size of the human genome, almost as large as that of the fugu fish genome, and about three times the size of that of *Caenorhabditis elegans*. Previous reports have revealed that the chromosomes of *S. mansoni* have been colonized by at least two families of non-LTR retrotransposons (19, 20). Here we report the sequence and structure of *Boudicca*, apparently the first full-length LTR retrotransposon to be characterized from the genome of the African schistosome *S. mansoni*. Analysis of the *Boudicca* sequence revealed the characteristic structure of an LTR retrotransposon. Based on its size, ≈5.8 kb, and high copy number, estimated at over 1,000, *Boudicca* can be expected to have contributed substantially to the genome size of *S. mansoni*, perhaps constituting as much as 4% or more of it (5.8 kb × 1,000 = 5.8 Mb). Moreover, it will likely have influ-

TABLE 1. Estimate of gene copy number of the *Boudicca* LTR retrotransposon in the genome of *S. mansoni*[a]

Gene	GenBank accession no.	No. of nucleotides	No. of hits	Estimated copy no.	Key reference(s)
SR1	U66331	2,337[b]	552	200–2,000	19
SR2	AF025672	3,913[c]	989	1,000–10,000	20
Smα	M27676	331	578	7,000–10,000	60
18S rRNA	M62652	1,739	1	100	57
Cathepsin D cDNA	U60995	1,285	0	1	5, 67
Boudicca		5,858	743	~1,000–10,000	This study

[a] Estimate made by comparison of Blast results of the GSS sequences in the CNS and TIGR databases for five other genes for which copy numbers have been reported.

[b] Partial sequence of the *SR1* element. Results obtained in July 2002 with the TIGR (11,493 GSS) and CNS (14,791 GSS), sequences of *S. mansoni*, a total of 26,284. There is an underestimation of Smα because of a Blast score cutoff of 100 coupled to a short query sequence.

[c] The *SR2* result may be a slight underestimation due to >500 hits with a score of >100 in the CNS database.

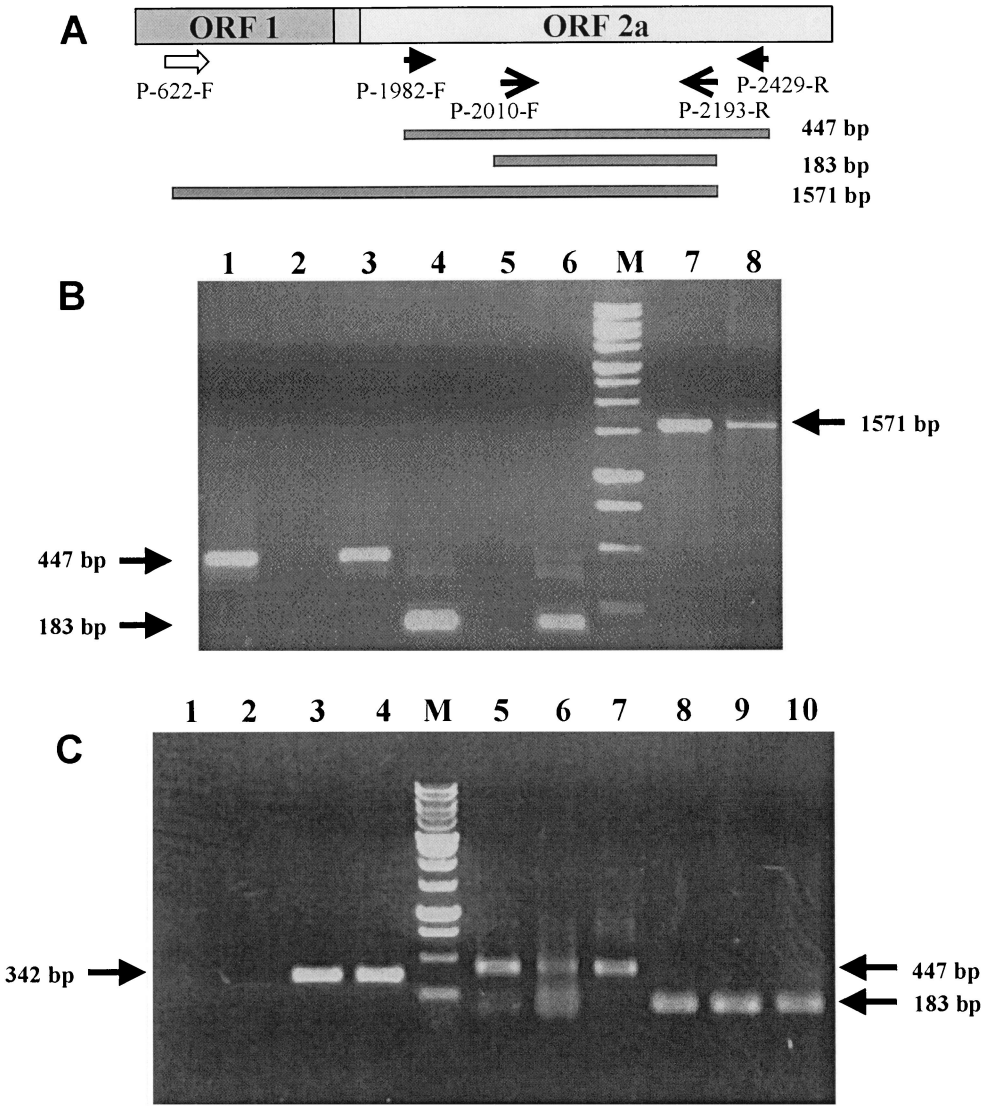


FIG. 8. RT-PCR-based detection of *Boudicca* transcripts in developmental stages of *Schistosoma mansoni*. (A) Schematic representation of RT-PCR-based detection of mRNA transcripts. The solid arrows represent primers (binding in ORF2) used in the first run of the nested PCR, and the open arrows represent primers used in the second run of the nested PCR. Primer P-622-F (white arrow), binding in ORF1, and P-2193-R span an area overlapping the first two ORFs of *Boudicca*. (B) After RNA isolation from adult worms, DNase digestion, and reverse transcription, a 447-bp fragment was amplified in the first PCR (lane 1); DNase digestion without reverse transcription revealed no contamination with genomic DNA (negative control lane 2); a corresponding fragment was amplified from genomic DNA (positive control lane 3). A 183-bp fragment was reamplified with nested primers from the 447-bp RT-PCR product (lane 4) and from the 447-bp genomic PCR product (positive control) (lane 6). Reamplification of the negative control was not possible (lane 5). A 1,571-bp fragment was amplified with primers spanning an area overlapping the first two ORFs from mRNA transcripts (lane 7) and from genomic DNA (lane 8). Lane M, size markers. (C) After RNA isolation from larval stages, DNase digestion, and reverse transcription, a 447-bp fragment was amplified in the first PCR (lane 5, sporocysts; lane 6, cercariae); a 183-bp fragment was reamplified with nested primers from the 447-bp RT-PCR product (lane 8, sporocysts; lane 9, cercariae). Corresponding fragments were amplified from genomic DNA: in the first PCR, a 447-bp fragment (lane 7), and in the nested PCR, a 183-bp fragment (lane 10). The quality of the cDNA transcripts was tested by amplification of cytochrome *c* oxidase subunit 1 (AF101196). A 342-bp fragment was amplified (lane 3, sporocysts; lane 4, cercariae). DNase digestion without transcription revealed no contamination with DNA (negative control; lane 1, sporocysts; lane 2, cercariae). Lane M, molecular size standards.

enced the evolution of the genome of this schistosome through the mutagenic action of its movement, its influence on expression of adjacent genes, and its effects on chromosomal recombination (10, 51).

Although the *Boudicca* copy located in BAC clone 53-J-5 was clearly degenerate and inactive, it was possible to assemble a consensus sequence of a form of *Boudicca* likely to be active based on sequences of several of the large number of fragments of genomic copies of *Boudicca* that have been partially sequenced and are available in the public domain. Verification that the consensus *Boudicca* represents an ostensibly active retrotransposon may have to await the generation of the entire genome sequence of *S. mansoni*, a task that is now claiming the attention of a number of genome sequencing labs (24). Ret-

rotransposons with degraded sequences may also be capable of functioning to a limited extent, since one copy can be transcribed and reinserted with functional proteins produced by another copy by a process of transcomplementation (2, 9, 65).

Although *Ty3/gypsy*-like, *Boudicca* is closely related to *kabuki* and *CsRn1*. Phylogenetic analysis focused on the reverse transcriptase domain confirmed that *Boudicca* was a *Ty3/gypsy*-like LTR retrotransposon. Within the *Ty3/gypsy* assemblage, its closest relatives were *kabuki* from the silk moth *B. mori* and *CsRn1* from the Oriental liver fluke *C. sinensis*. In addition to the close identity to *kabuki* and *CsRn1* revealed in the phylogenetic analysis, *Boudicca*, *CsRn1*, and *kabuki* shared structural similarities in the Gag protein that differentiated this group from other *gypsy*-like elements. *Boudicca*, *CsRn1*, and *kabuki* have a CHCC zinc finger Cys-His box motif at the COOH terminus of Gag that is dissimilar to the more usual CCHC motif reported in the nucleocapsid proteins encoded by *gag* in other LTR retrotransposons and in retroviruses (4, 30, 55).

In general, the short branch lengths of the phylogenetic tree reflect the close identity among members of the *Ty3/gypsy* family, and the clear placement of *Boudicca* within this group. The fact that elements in this group bear a close relationship to each other while colonizing hosts that are phylogenetically distant suggests that these retroelements may have spread by horizontal transmission. Interestingly, this group of elements also appear to be more closely related to vertebrate retroviruses than to the other family of LTR retrotransposons, the *Ty1/copia* group. Also of interest is that *Boudicca* is more closely related to insect and liver fluke retrotransposons than to the other full-length, characterized schistosome LTR retrotransposon *Gulliver* from *Schistosoma japonicum* (36). Therefore, it is apparent that the schistosome LTR retrotransposons *Boudicca* and *Gulliver* are discrete elements that are unlikely to have evolved vertically from a common ancestor. Furthermore, examination of other reverse transcriptase-encoding sequences in the schistosome genome verified the presence of additional retrotransposons (25).

Envelope protein? Despite its phylogenetic identity as revealed by the reverse transcriptase alignments, *Boudicca* may have an important structural difference that separates it from *kabuki* and *CsRn1*, a difference that would liken it to the errantivirus group of insect LTR retrotransposons such as *gypsy*, *Tom*, and *ZAM* and to *Tas* of *Ascaris lumbricoides* (22). Specifically, *Boudicca* appears to have a third ORF that may encode an envelope protein of about 150 amino acid residues that includes several key structural motifs, a signal peptide, a transmembrane domain, and disulfide bridges, features that are characteristic of envelope proteins from other retroelements (48). The envelope protein is the key structural component that allows retroviruses and errantiviruses to function as transmissible extracellular particles because it allows them to enter new host cells via binding to membrane receptors that mediate uptake of the virus by the next host cell (32). Furthermore, through this interaction with cell surface receptors at the point of host cell entry and infection, the envelope protein confers host cell and species specificity on the retrovirus. In addition to other aspects of the molecular biology of *Boudicca*, we plan to more fully characterize this putative envelope protein in future studies.

***Boudicca* is actively transcribed.** Whereas the contiguous copy of *Boudicca* present in the 53-J-5 BAC clone has clearly been degraded by a number of mutations, probably to an inactive element, an RT-PCR experiment demonstrated active transcription from *Boudicca* copies. Moreover, the transcripts were detected in three developmental stages examined, adults, cercariae, and sporocysts. This suggests that *Boudicca* may be transposing within the genome of *S. mansoni* at this point in evolutionary time. Other recent reports have also suggested the activity of reverse transcriptase enzymes and/or mobilization of retrotransposable elements in schistosomes (29, 36). Given the phylogenetic proximity of *Boudicca* to the errantiviruses and the possible presence of an envelope protein, *Boudicca* may be capable of both vertical and horizontal transmission. Finally, as an endogenous retrotransposon, it is feasible that *Boudicca* or its structural components could be harnessed for use in transgenesis of schistosomes or other parasitic helminths.

ACKNOWLEDGMENTS

We thank Grit Meusel for laboratory assistance and Egbert Flory for assistance with characterization of the putative envelope protein.

This work was supported by the Deutscher Akademischer Austauschdienst in the form of a Short-Term Research Fellowship awarded to C.S.C., by research grant number KA 866/2-1 from the Deutsche Forschungsgemeinschaft to B.H.K., by the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Disease (TDR), No. 990525, to D.L.W., and by a Wellcome Trust Beowulf Genomics Project Grant to D.A.J. and A.C.I. (for the sequencing of clone 53-J-5). P.J.B. is a recipient of a Burroughs Wellcome Fund scholar award in Molecular Parasitology. The Interdisciplinary Program in Molecular and Cellular Biology and the Center for Infectious Diseases, Tulane University, both supported this study, as did the Department of Molecular Parasitology, Institute for Biology, Humboldt University Berlin.

REFERENCES

1. **Abe, H., F. Ohbayashi, T. Shimada, T. Sugasaki, S. Kawai, K. Mita, and T. Oshiki.** 2000. Molecular structure of a novel *gypsy-Ty3*-like retrotransposon (*Kabuki*) and nested retrotransposable elements on the W chromosome of the silkworm *Bombyx mori*. Mol. Gen. Gen. **263:**916–924.
2. **Ansari-Lari, M. A., and R. A. Gibbs.** 1996. Expression of human immunodeficiency virus type 1 reverse transcriptase in *trans* during virion release and after infection. J. Virol. **70:**3870–3875.
3. **Arkhipova, I., and M. Meselson.** 2000. Transposable elements in sexual and ancient asexual taxa. Proc. Natl. Acad. Sci. USA **97:**14473–14477.
4. **Bae, Y. A., S. Y. Moon, X. Kong, S. Y. Cho, and M. G. Rhyu.** 2001. *CsRn1*, a novel active retrotransposon in a parasitic trematode. Clonorchis sinensis, discloses a new phylogenetic clade of *Ty* 3/gypsy-like LTR retrotransposons. Mol. Biol. Evol. **18:**1474–1483.
5. **Becker, M. M., S. A. Harrop, J. P. Dalton, B. H. Kalinna, D. P. McManus, and P. J. Brindley.** 1995. Cloning and characterization of the *Schistosoma japonicum* aspartic protease involved in haemoglobin degradation. J. Biol. Chem. **272:**17246. (Erratum, J. Biol. Chem. **270:**24496–24501, 1997.)
6. **Bergquist, N. R.** 2002. Schistosomiasis: from risk assessment to control. Trends Parasitol. **18:**309–314.
7. **Bowen, N. J., and J. F. McDonald.** 1999. Genomic analysis of *Caenorhabditis elegans* reveals ancient families of retrovirus-like elements. Genome Res. **9:**924–935.
8. **Bowzard, J. B., J. W. Wills, and R. C. Craven.** 2001. Second-site suppressors of Rous sarcoma virus Ca mutations: evidence for interdomain interactions. J. Virol. **75:**6850–6856.
9. **Chaboissier, M. C., C. Bornecque, I. Busseau, and A. Bucheton.** 1995. A genetically tagged, defective I element can be complemented by actively transposing I factors in the germline of I-R dysgenic females in *Drosophila melanogaster*. Mol. Gen. Genet. **248:**434–438.
10. **Charlesworth, B., P. Sniegowki, and W. Stephan.** 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature **371:**215–220.
11. **Chitsulo, L., D. Engels, A. Montresor, and L. Savioli.** 2000. The global status of schistosomiasis and its control. Acta Trop. **77:**41–51.
12. **Clish, C., B. Peyton, and E. Barklis.** 1998. Solution structures of human immunodeficiency virus type 1 (HIV-1) and moloney murine leukemia virus (MoMLV) capsid protein major-homology-region peptide analogs by NMR spectroscopy. Eur. J. Biochem. **257:**69–77.
13. **Craven, R. C., Leure du A. E. Pree, R. A. Weldon, Jr., and J. W. Wills.** 1995. Genetic analysis of the Major Homology Region of the Rous sarcoma virus Gag protein. J. Virol. **69:**4213–4227.

14. **Cristofari, G., D. Ficheux., and J. L. Darlix.** 2000. The gag-like protein of the yeast *Ty1* retrotransposon contains a nucleic acid chaperone domain analogous to retroviral nucleocapsid proteins. J. Biol. Chem. **275:**19210–19217.
15. **Crompton, D. W. T.** 1999. How much human helminthiasis is there in the world? J. Parasitol. **85:**397–403.
16. **Dej, K. J., T. Gerasimova, V. G. Corces, and J. D. Boeke.** 1998. A hotspot for the *Drosophila gypsy* retroelement in the *ovo* locus. Nucleic Acids Res. **26:** 4019–4025.
17. **de Jesus, A. R., A. Silva, L. B. Santana, A. Magalhaes, A. A., de Jesus, R. P. de Almeida, M. A. Rego, M. N. Burattini, E. J. Pearce, and E. M. Carvalho.** 2002. Clinical and immunological evaluation of 31 patients with acute schistosomiasis mansoni. J. Infect. Dis. **185:**98–105.
18. **Drew, A. C., and P. J. Brindley.** 1995. Female-specific sequences isolated from *Schistosoma mansoni* by representational difference analysis. Mol. Biochem. Parasitol. **71:**173–181.
19. **Drew, A. C., and P. J. Brindley.** 1997. A retrotransposon of the non-long terminal repeat class from the human blood fluke *Schistosoma mansoni*. Similarities with the chicken repeat 1-like elements from vertebrates. Mol. Biol. Evol. **14:**602–610.
20. **Drew, A. C., D. J. Minchella, L. T. King, D. Rollinson, and P. J. Brindley.** 1999. *SR2*, non-long terminal repeat retrotransposons of the RTE-1 lineage, from the human blood fluke *Schistosoma mansoni*. Mol. Biol. Evol. **16:** 1256–1269.
21. **Dupraz, P., S. Oertle, C. Meric, P. Damay, and P.-F. Spahr.** 1990. Point mutations in the proximal Cys-His box of Rous sarcoma virus nucleocapsid protein. J. Virol. **64:**4978–4987.
22. **Felder, H., A. Herzceg, Y. de Chastonay, P. Aeby, H. Tobler, and F. Muller.** 1994. *Tas*, a retrotransposon from the parasitic nemetode *Ascaris lumbricoides*. Gene **149:**219–225.
23. **Finnegan, D. J.** 1992. Transposable elements. Curr. Opin. Genet. Dev. **2:** 861–867.
24. **Forster, J. M., and D. A. J. Johnson.** 2002. Helminth genomics: from gene discovery to genome sequencing. Trends Parasitol. **18:**241–242
25. **Foulk, B. W., G. Pappas, Y. Hirai, H. Hirai, and D. L. Williams.** 2002. Adenylsuccinate lyase of *Schistosoma mansoni*: gene structure, mRNA expression, and analysis of the predicted peptide structure of a potential chemotherapeutic target. Int. J. Parasitol. **32:**1487–1495.
26. **Gad, A., E. Tanaka, K. Orii, A. Rokuhara, Z. Nooman, A. H. Serwah, M. Shoair, K. Yoshizawa, and K. Kiyosawa.** 2001. Relationship between hepatitis C virus infection and schistosomal liver disease: not simply an additive effect. J. Gastroenterol. **36:**753–758.
27. **Grossman, A. I., R. B. Short, and G. D. Cain.** 1981. Karyotype evolution and sex chromosome differentiation in schistosomes (Trematoda, Schistosomatidae). Chromosoma. **84:**413–430.
28. **Hirai, H., T. Taguchi, Sa Y. itoh, M. Kawanaka, H. Sugiyama, S. Habe, M. Okamoto, M. Hirata, M. Shimada, W. U. Tiu, K. Lai, E. S. Upatham, and T. Agatsuma.** 2000. Chromosomal differentiation of the Schistosoma japonicum complex. Int. J. Parasitol. **30:**441–452.
29. **Ivanchenko, M. G., J. P. Lerner, R. S. McCormick, A. Toumadje, B. Allen, K. Fischer, O. Hedstrom, A. Helmrich, D. W. Barnes, and C. J. Bayne.** 1999. Continuous *in vitro* propagation and differentiation of cultures of the intramolluscan stages of the human parasite *Schistosoma mansoni*. Proc. Natl. Acad. Sci. USA **96:**4965–4970.
30. **Katz, R. A., and A. M. Skalka.** 1994. The retroviral enzymes. Annu. Rev. Biochem. **63:**133–173.
31. **Kawanaka, M., S. Hayashi, and H. Ohtomo,** 1983. A minimum essential medium for cultivation of Schistosoma japonicum eggs. J. Parasitol. **69:**991–992.
32. **Kim, A., C. Terzian, P. Santamaria, A. Pelisson, N. Prud'homme, and A. Bucheton.** 1994. Retroviruses in invertebrates: the *gypsy* retrotransposon is apparently an infectious retrovirus of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **91:**1285–1289.
33. **Krogh, A., B. Larsson, G. von Heijne, and E. Sonnhammer.** 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. **305:**567–580.
34. **Laha, T., P. J. Brindley, M. J. Smout, C. K. Verity, D. P. McManus, and A. Loukas.** 2002. Reverse transcriptase activity and UTR sharing of a new *RTE*-like, non-LTR retrotransposon from the human blood fluke. Schistosoma japonicum. Int. J. Parasitol. **32:**1163–1174.
35. **Laha, T., P. J. Brindley, C. K. Verity, D. P. McManus, and A. Loukas.** 2002. *pido*, a non-long terminal repeat retrotransposon of the chicken repeat 1 family from the genome of the. Oriental blood fluke *Schistosoma japonicum.* Gene **284:**149–159.
36. **Laha, T., A. Loukas, C. K. Verity, D. P. McManus, and P. J. Brindley.** 2001. *Gulliver*, a long terminal repeat retrotransposon from the genome of the oriental blood fluke *Schistosoma japonicum*. Gene **264:**59–68.
37. **Lambertucci, J. R.** 1993. *Schistosoma mansoni*: pathological and clinical aspects, p. 195–235. *In* P. Jordan, G. Webbe, and R. F. Sturrock (ed.), Human schistosomiasis. CAB International, Wallingford, United Kingdom.
38. **Leblanc, P., S. Desset, B. Dastugue, and C. Vaury.** 1997. Invertebrate retroviruses: *ZAM*, a new candidate in *D. melanogaster*. EMBO J. **16:**7521–7531.
39. **Le Paslier, M. C., R. J. Pierce, F. Merlin, H. Hirai, W. Wu, D. L. Williams, D. Johnston, P. T. LoVerde, and Le D. Paslier.** 2000. Construction and characterization of a *Schistosoma mansoni* bacterial artificial chromosome library. Genomics **65:**87–94.
40. **Lerat, E., and P. Capy.** 1998. Retrotransposons and retroviruses: analysis of the envelope gene. Mol. Biol. Evol. **16:**1198–1207.
41. **Malik, H. S., and T. H. Eickbush.** 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res. **11:**1187–1197.
42. **Malik, H. S., S. Henikoff, and T. H. Eickbush.** 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. Genome Res. **10:**1307–1318.
43. **Malik, H. S., W. D. Burke, and T. H. Eickbush.** 1999. The age and evolution of non-LTR retrotransposable elements. Mol. Biol. Evol. **16:**793–805.
44. **Meric, C., and S. P. Goff.** 1989. Characterization of Moloney Murine Leukemia Virus mutants with single amino acid substitutions in the Cys-His box of the nucleocapsid protein. J. Virol. **63:**1558–1568.
45. **Miller, K., C. Lynch, J. Martin, E. Herniou, and M. Tristem.** 1999. Identification of multiple *gypsy* LTR-retrotransposon lineages in vertebrate genomes. J. Mol. Evol. **49:**358–366.
46. **Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne.** 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10:**1–6.
47. **Page, R. D. M.** 1996. Treeview: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12:**357–358.
48. **Pantazidis, A., M. Labrador, and A. Fontdevila.** 1999. The retrotransposon *Osvaldo* from *Drosophila buzzatii* displays all structural features of a functional retrovirus. Mol. Biol. Evol. **16:**909–921.
49. **Pereira, C., P. G. Fallon, J. Cornette, A. Capron, M. J. Doenhoff, and R. J. Pierce.** 1998. Alterations in cytochrome-c oxidase expression between praziquantel-resistant and susceptible strains of *Schistosoma mansoni*. Parasitology. **117:**63–73.
50. **Perriere, G., and M. Gouy.** 1996. WWW-Query: an on-line retrieval system for biological sequence banks. Biochimie **78:**364–369.
51. **Petrov, D. A., T. A. Sangster, J. S. Johnston, D. L. Hartl, and K. L. Shaw.** 2000. Evidence for DNA loss as a determinant of genome size. Science **287:** 1060–1062.
52. **Provitera, P., A. Goff, A. Harenberg, F. Bouamr, C. Carter, and S. Scarlata.** 2001. Role of the Major Homology Region in assembly of HIV-1 Gag. Biochemistry **40:**5565–5572.
53. **Robertson, H. M.** 1997. Multiple *mariner* transposons in flatworms and hydras are related to those of insects. J. Hered. **88:**195–201.
54. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
55. **Schüler, W., C. Z. Dong, K. Wecker, and B. P. Roques.** 1999. NMR structure of the complex between the zinc finger protein NCp10 of Moloney murine leukemia virus and the single-stranded pentanucleotide d(ACGCC): comparison with HIV-NCp7 complexes. Biochemistry **38:**12984–12994.
56. **Simpson, A. J. G., A. Sher, and T. F. McCutchan.** 1982. The genome of *Schistosoma mansoni*: isolation of DNA, its size, bases and repetitive sequences. Mol. Biochem. Parasitol. **6:**125–137.
57. **Simpson, A. J. G., J. B. Dame, F. A. Lewis, and T. F. McCutchan.** 1984. The arrangement of the ribosomal RNA genes in *Schistosoma mansoni*. Identification of polymorphic structural variants. Eur. J. Biochem. **139:**41–45.
58. **Smit, A. F.** 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. Nucleic Acids Res. **21:**1863–1872.
59. **Southern, E. M.** 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. **98:**503–517.
60. **Spotila, L. D., H. Hirai, D. M. Rekosh, and P. T. LoVerde.** 1989. A retroposon-like short repetitive DNA element in the genome of the human blood fluke *Schistosoma mansoni*. Chromosoma **97:**421–428.
61. **Swofford, D. L.** 1998. PAUP*. phylogenetic analysis with parsimony (and other methods). Sinauer Associates, Sunderland, Mass.
62. **Tanda, S., J. L. Mullor, and V. G. Corces.** 1994. The *Drosophila tom* retrotransposon encodes an envelope protein. Mol. Cell. Biol. **14:**5392–5401.
63. **Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeannmougin, and D. G. Higgins.** 1997. The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **24:**4876–4882.
64. **Verity, C. K., D. P. McManus, and P. J. Brindley.** 1999. Developmental expression of cathepsin D aspartic protease in *Schistosoma japonicum.* Int. J. Parasitol. **29:**1819–1824.
65. **Wei, W., N. Gilbert, S. L. Ooi, J. F Lawler,., E. M. Ostertag, H. H. Kazazian, J. D. Boeke, and J. V. Moran.** 2001. Human L1 retrotransposition: *cis* preference versus *trans* complementation. Mol. Cell. Biol. **21:**1429–1439.
66. **Weiss, R. A.** 1981. RNA tumor viruses: molecular biology of tumor viruses, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.
67. **Wong, J. Y. M., S. A. Harrop, S. R. Day, and P. J. Brindley.** 1997. Schistosomes express two forms of cathepsin D. Biochim. Biophys. Acta **1338:**156–160.
68. **Xiong, Y., and T. Eickbush.** 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9:**3353–3362.
69. **Zhou, X., M. Chen, D. McManus, and R. Bergquist.** 2002. Schistosomiasis control in the 21st century. Proceedings of the International Symposium on Schistosomiasis, Shanghai, July 4–6, 2001. Acta Trop. **82:**95–114.