Available online at www.sciencedirect.com
SCIENCE DIRECT®
Gene 329 (2004) 103–114
GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES
www.elsevier.com/locate/gene

Structural and evolutionary analysis of the transcribed sequence of *Boudicca*, a *Schistosoma mansoni* retrotransposon

Claudia S. Copeland[a,b,1], Oliver Heyers[c,1], Bernd H. Kalinna[c], Andreas Bachmair[d], Peter F. Stadler[e], Ivo L. Hofacker[f], Paul J. Brindley[a,b,*]

[a] *Department of Tropical Medicine, Tulane University Health Sciences Center, New Orleans, USA*
[b] *Interdisciplinary Program in Molecular and Cellular Biology, Tulane University Health Sciences Center, New Orleans, USA*
[c] *Department of Molecular Parasitology, Institute for Biology, Humboldt University Berlin, Berlin, Germany*
[d] *Max Planck Institute for Plant Breeding Research, Cologne, Germany*
[e] *Bioinformatics, Department of Computer Science, University of Leipzig, Leipzig, Germany*
[f] *Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Vienna, Austria*

Received 24 October 2003; received in revised form 9 December 2003; accepted 23 December 2003

Received by D. Finnegan

Abstract

Boudicca is a *gypsy*-like, long terminal repeat (LTR) retrotransposon that has colonized the genome of the human blood fluke, *Schistosoma mansoni*. Previous studies have indicated that more than 1000 copies of *Boudicca* reside within the *S. mansoni* genome, although many of them may be degenerate and inactive. Messenger RNAs transcribed from genomic copies of *Boudicca* were investigated by reverse transcription PCR. Overlapping RT-PCR products corresponding to the *gag* and *pol* polyproteins of *Boudicca*, along with relevant sequences of genomic fragments of *Boudicca*, were assembled into contigs. Consensus sequences from these contigs were used to predict the sequence and structure of transpositionally active copies of the *Boudicca* retrotransposon. They verified that *Boudicca* has a *kabuki*-like Cys–His box motif at the active site of its gag protein, a classic DTG motif as the active site of the protease domain of the pol ORF2, and indicated a contiguous integrase domain at the C-terminus of pol with strong identity to integrase from the LTR retrotransposons *CsRn1* and *kabuki*, as well as to the conserved integrase core domain, GenBank rve (pfam00665). Models of the secondary structure of the *Boudicca* transcript suggested that the first AUG was occluded by a stem loop structure, which in turn suggested a method of regulation of expression, at the level of translation, of *Boudicca* proteins. In addition, phylogenetic analysis targeting discrete domains of *Boudicca* revealed a generalized radiation in sequences among the multiple copies of *Boudicca* resident in the schistosome genome.

Keywords: Long terminal repeat (LTR); RNA secondary structure; Post-transcriptional regulation; Multiple AUG codons; *CsRn1*; *Kabuki*

1. Introduction

Uncontrolled mobilization of retrotransposons can be expected to be detrimental to their host genomes. RNA silencing systems and other mechanisms have evolved to protect the genome from deleterious movements of mobile genetic elements, among other functions (Plasterk, 2002). Further, regulation of gene expression based on mRNA secondary structure has been observed in viruses (Hohn et al., 2001) and in mobile genetic elements including the IS200 insertion sequence (Beuzon et al., 1999). Some species appear to limit accumulation of excessive numbers of mobile genetic elements (Petrov et al., 2000), resulting in smaller genomes with a higher percentage of host genes than those of related species. In turn, retrotransposons can be expected to have evolved strategies to self-limit proliferation in order to avoid overwhelming their host genome.

Schistosoma mansoni, the African blood fluke, is endemic throughout much of Africa, Latin America and the

Abbreviations: bp, base pair; EST, expressed sequence tag; IN, integrase; LTR, long terminal repeat; nt, nucleotide; ORF, open reading frame; PR, protease; RC, RT-PCR-product clone; RNase H, ribonuclease H; RT, reverse transcriptase.

* Corresponding author. Department of Tropical Medicine, SL-17, Tulane University, Health Sciences Center, 1430 Tulane Avenue, New Orleans, LA 70112, USA. Tel.: +1-504-988-4645; fax: +1-504-988-6686.
E-mail address: paul.brindley@tulane.edu (P.J. Brindley).

[1] These authors contributed equally to this work.

0378-1119/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2003.12.023

Caribbean, and is one of the three major species of the genus *Schistosoma* that parasitize humans. Schistosomiasis, the most important of the human helminthiases in terms of morbidity and mortality (Crompton, 1999), has often proved refractory to control by the more conventional public health approaches. Molecular approaches aimed at the development of new drugs, vaccines and/or diagnostics can be expected to complement more conventional interventions (Colley et al., 2001). Mobile genetic elements colonizing the chromosomes of *S. mansoni* are of interest both for their influence on the evolution and structure of the schistosome genome and for their potential in developing tools for schistosome transgenesis (Brindley et al., 2003). Recently, a new long terminal repeat (LTR) retrotransposon was characterized from the *S. mansoni* genome and designated *Boudicca* (Copeland et al., 2003). As with other retrotransposons, many copies of *Boudicca* are likely to be inactive and/or degenerate (e.g., Deininger and Batzer, 2002). Indeed, the contiguous genomic copy reported by Copeland et al. (2003) included a number of apparent mutations that may disable its capacity to retrotranspose and replicate.

Here, we investigated the sequence and structure of active forms of *Boudicca* by examination of transcripts of *Boudicca* and of sequences of genomic copies of *Boudicca* from the chromosomes of *S. mansoni.* These investigations confirmed that *Boudicca* belongs to the newly characterized *kabuki/CsRn1* assemblage of *gypsy*-like LTR retrotransposons (Abe et al., 2000; Bae et al., 2001), and that numerous sequence variants of *Boudicca* are interspersed throughout the *S. mansoni* genome. Further, they indicated that *Boudicca* may regulate its mobilization and proliferation in the schistosome genome by using alternative translation initiation sites in transcribed copies of this retrotransposon.

2. Materials and methods

2.1. PCR and RT-PCR

To obtain schistosome RNA, mixed sex adult worms of *S. mansoni*, cultured as described in Copeland et al. (2003), were homogenized in lysis buffer (RNeasy RNA extraction kit, Qiagen) using disposable micro-pestles. Residual DNA was removed by digestion with RNase-free DNAse (Promega). The schistosome RNA was precipitated to remove the DNAse, dissolved in RNase-free water and employed as the template for oligo-(dT) primed reverse transcription by reverse transcriptase (RT) (Moloney Murine Leukemia Virus H(−) Point Mutant RT) (Promega). The resulting cDNA was used as the template for PCRs designed to generate overlapping fragments spanning the protein-encoding open reading frames (ORFs) of *Boudicca*. PCR amplification was accomplished using the following thermal cycling conditions: denaturation for 4 min at 94 °C; incubation for 2 min at 94 °C, followed by the addition of 2.5 units of *Taq*-Polymerase; 30 cycles as follows: denaturation for 1 min at 94 °C, annealing for 1 min at temperature 57–65 °C (depending on specific primer pairs), extension for 2 min at 72 °C; a final extension at 72 °C for 20 min (Copeland et al., 2003). Sequences of oligonucleotide primers used for RT-PCR amplifications, regions and sizes of regions targeted for amplification, and the locations of the RT-PCR amplicons in relation to the *Boudicca* copy localized in *S. mansoni* BAC clone 53-J-5 are illustrated in Table 1 and Fig. 1.

2.2. Sequencing and analysis of RT-PCR products

PCR products were sized by electrophoresis through agarose, stained with ethidium bromide, photographed under UV light and cloned into the plasmid vector pGEM-T-Easy (Promega). The recombinant inserts were sequenced on both strands using T7- and SP6-specific primers and ABI BigDye™ Terminator chemistry (ABI, Foster City, CA) and an ABI Prism 3100 sequencer at the Center for Gene Therapy, Tulane University, after which the sequences were edited to remove vector-specific sequences using VecScreen (www.ncbi.nlm.nih.gov). Subsequently, nucleotide and/or deduced amino acid sequences from the amplicons were employed to interrogate the non-redundant database in GenBank, www.ncbi.nlm.nih.gov, using BLASTX and BLASTP.

Table 1
Primers used in RT-PCR to amplify fragments of *Boudicca* transcripts

Primer name	Primer oligonucleotide	Product name(s)	Location on 53-J-5 *Boudicca* copy
P-505-F	5′-ATGACGGAACACTCACCTAAGC	no product	nt 505–2193 (gag,PR,RT)
P-622-F	5′-ATGCACGGAATCACGGAC	RC1, RC4	nt 622–2193 (gag,PR,RT)
P-2193-R	5′-GAGTGATGATGGCGGTTTTAGG	RC1, RC4	nt 622–2193 (gag,PR,RT)
P-1982-F	5′-CCCTAAAAAGGACAGCAACGATTG	RC2, RC5	nt 1982–3447 (RT,RH,IN)
P-3447-R	5′-CCTTAGATTTCTGACAGC	RC2, RC5	nt 1982–3447 (RT,RH,IN)
P-3430-F	5′-GCTGTCAGAAATCTAAGG	RC3, RC6	nt 3430–4464/28 (IN)
P-4428-R	5′-TACTCGTCTTCCAGAACG	RC3	nt 3430–4428 (IN)
P-4464-R	5′-CTACGTGCAATAGTCGTTAAGATGTTCTGG	RC6	nt 3430–4464 (IN)
P-3615-F	5′-GGACATTACTGCTGAAAC	no product	nt 3615–4882 (IN,env?)
P-4465-F	5′-GAC ACA AGC TTT ATC TCG AAT CCC	no product	nt 4465–4882 (env?)
P-4882-R	5′-ACATACAACTGATGGTCG	no product	4465–4882 (env?)

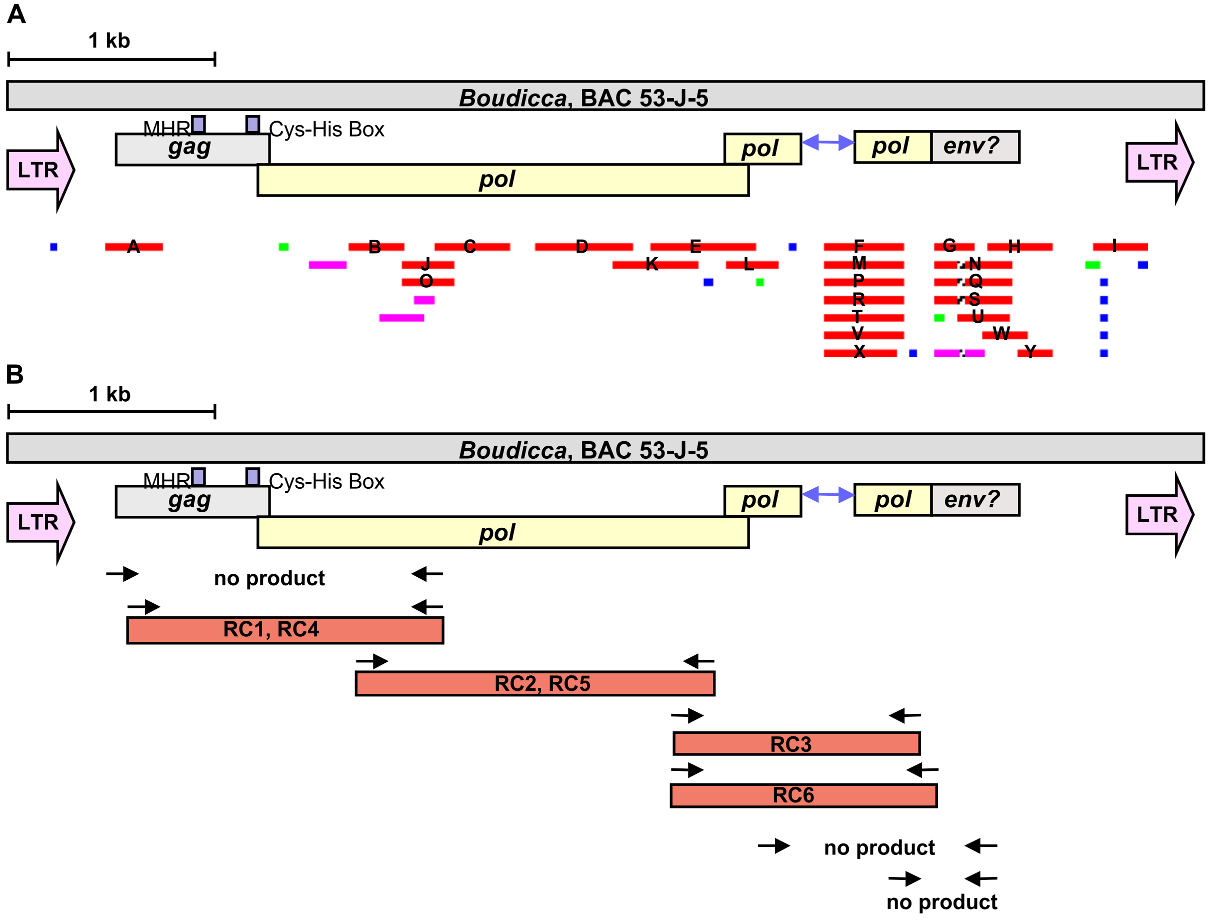


Fig. 1. *Boudicca* transcripts expressed in adult *S. mansoni*. (A) Twenty-five expressed sequence tags with identity to the region of *Boudicca* between the two LTRs are aligned with a map of the 53-J-5 copy of *Boudicca*. The accession numbers of these ESTs are: (A) AI395459, (B) BG931054, (C) BG931518, (D) AI976784, (E) AI975406, (F) BG931624, (G) L46945, (H) AI395625, (I) AI740214, (J) BG931114, (K) BG931022, (L) L81256, (M) BG931798, (N) BF440131, (O) BG931482, (P) BG931658, (Q) AI975345, (R) BG931608, (S) AW017111, (T) BG931659, (U) AA559598, (V) BG931679, (W) BF440117, (X) BG931788, (Y) BG931428. (B) Primers (arrows) and products (boxes) of RT-PCRs designed to span the ORFs of the *Boudicca* retrotransposon, aligned with the map of the 53-J-5 copy of *Boudicca*. RT-PCRs using primer sets including either the first start codon of *gag* or the unknown ORF 3′ of *integrase* were unsuccessful.

2.3. RNA secondary structure prediction, splice site prediction

RNA secondary structures were predicted using the program RNAfold of the Vienna RNA package (Hofacker et al., 1994; Hofacker, 2003) using energy parameters from Mathews et al. (1999). These predictions were undertaken on the 5858 bp *Boudicca* sequence in BAC clone 53-J-5 (Copeland et al., 2003). The positions of splice sites were predicted using the NetGene2 program (http://www.cbs.dtu.dk/services/NetGene2/), according to the algorithms of Hebsgaard et al. (1996) and Brunak et al. (1991).

2.4. Bacterial artificial chromosomes

Le Paslier et al. (2000) described the construction and partial characterization of a bacterial artificial chromosome (BAC) library of the *S. mansoni* genome. The library, constructed in the BAC plasmid vector, pBeloBac11, using genomic DNA from cercariae of a Puerto Rican strain of *S. mansoni* partially digested with *Hin*d III, consists of ∼ 21,000 clones, with an average insert size of ∼ 120 kb, providing ∼ 8-fold coverage of the *S. mansoni* genome (estimated size of haploid genome, 270 Mbp). Numerous BAC end sequences determined from randomly selected clones from this library have been deposited in the public domain, and organized into a searchable database at The Institute for Genomic Research (TIGR) (http://tigrblast.tigr.org/euk-blast/index.cgi?project=sma1). Sequences of RT-PCR products representing fragments of *Boudicca* transcripts (Section 2.2) were employed as the BLAST query to interrogate the TIGR database of *S. mansoni* gDNA BAC end sequences, and were also compared with the sequence of BAC 53-J-5, which includes the full length, contiguous copy of *Boudicca.* The GenBank accession numbers of TIGR BAC end sequences used here are listed in Table 2.

Table 2
Boudicca-positive clones in the TIGR BAC end database

Domain	BAC end clones and GenBank accession numbers
gag	62-N-16 (GenBank accession BH204125), 45-H-5 (BH206669), 62-G-23 (BH211091), 46-G-15 (BH209250), 57-H-13 (BH199410), 62-C-17 (BH211047), 61-D-18 (BH203117), 54-H-1 (BH205221), 55-N-10 (BH208607), 54-H-12 (BH199857), 49-N-18 (BH203259), 61-N-7 (BH203211)
protease	59-F-1 (BH208933), 49-N-20 (BH204112), 43-P-17 (BH209123), 51-O-21 (BH210224), 54-J-21 (BH201057), 54-N-14 (BH203105), 56-C-15 (BH206783), 61-H-2 (BH211409), 49-D-16 (BH207733), 45-P-23 (BH211194), 40-H-16 (BH205011), 59-B-11 (BH203286), 50-N-14 (BH201622), 49-G-19 (BH207728), 49-L-5 (BH202898), 40-H-16 (BH205011), 50-J-14 (BH204101), 50-E-22 (BH206444), 57-A-18 (BH208309), 62-P-1 (BH200465)
reverse transcriptase	57-E-12 (BH210339), 45-M-2 (BH206603), 62-E-6 (BH199445), 59-I-14 (BH210647), 57-M-23 (BH209613), 57-P-23 (BH205772), 61-A-3 (BH210629), 62-F-10 (BH204718), 44-G-15 (BH209726), 44-C-12 (BH211335), 54-H-5 (BH206899), 42-A-9 (BH208040), 61-M-10 (BH208482), 48-E-19 (BH208015), 46-N-22 (BH208235), 54-M-22 (BH210030), 54-K-20 (BH204267), 42-K-16 (BH200935), 42-L-6 (BH200280)
RNase H	62-O-9 (BH200265), 62-F-10 (BH204718), 54-H-5 (BH206899), 48-E-19 (BH208015), 59-L-12 (BH205994), 54-M-22 (BH210030), 42-A-9 (BH208040), 60-B-10 (BH205237), 56-J-10 (BH205875), 56-O-19 (BH208145), 62-O-19 (BH211544), 61-I-9 (BH202980), 44-C-12 (BH209388), 61-M-10 (BH208482), 44-G-15 (BH209726), 44-E-1 (BH204088), 45-F-9 (BH203432), 48-I-8 (BH208424)
5′-Integrase	56-J-12 (BH206711), 44-K-10 (BH200366), 39-L-24 (BH205876), 61-M-19 (BH199499), 54-P-13 (BH207214), 46-H-4 (BH208634), 59-B-18 (BH201482), 51-O-10 (BH206019), 58-O-17 (BH206804), 58-K-19 (BH210145), 51-C-6 (BH204687), 45-H-22 (BH203839), 54-N-19 (BH204818), 55-N-14 (BH210251), 42-C-17 (BH211076), 45-F-9 (BH203432), 52-C-24 (BH201854, BH201845), 41-E-11 (BH206384), 52-H-22 (BH202114)
3′-Integrase	46-M-7 (BH208369), 57-K-8 (BH211068), 54-P-24 (BH207048), 50-L-15 (BH201909), 50-O-20 (BH200329), 45-L-19 (BH199683), 60-J-19 (BH203625), 54-I-16 (BH202872), 47-D-15 (BH210140), 56-N-2 (BH200029), 55-A-24 (BH206055), 48-A-14 (BH208816), 51-L-7 (BH203253), 59-F-13 (BH208813), 43-F-11 (BH199552), 45-N-18 (BH199365), 55-C-9 (BH202305

2.5. Phylogenetic analysis

In order to characterize the diversification and conservation of a sample of discrete genomic copies of *Boudicca*, and to compare these sequences with those of the transcribed sequences obtained by RT-PCR, phylogenetic trees were constructed based on six discrete domains of the *Boudicca* genome. The six specific domains of the 53-J-5 copy of *Boudicca* used to interrogate the TIGR database of BAC end sequences were (1) *gag*, from nt 505–1332 of the 5.8-kb copy of *Boudicca* in BAC clone 53-J-5, (2) *protease* (nt 1298–1922), (3) *reverse transcriptase* (nt 1923–2458), (4) *RNase H* (nt 2459–2919), (5) the 5′-region of integrase (5′-IN) (nt 2920–3447) and (6) the 3′-region of integrase (3′-IN) (nt 3430–4428). Alignments of the BAC end sequences located by BLAST searches and generation of bootstrapped phylogenetic trees were accomplished using CLUSTALX (Thompson et al., 1997) and Njplot Unrooted (Saitou and Nei, 1987) software, with assistance from the MacVector software suite (Accelerys, Burlington, MA). Positions with gaps were excluded and branch length ratios were preserved upon transfer into PowerPoint for display. Alignment and tree files can be found at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/Boudicca/.

2.6. Expressed sequence tag (EST) analysis

BLAST searches of the non-human, non-mouse EST database of GenBank, which included 139,064 *S. mansoni* ESTs (dbEST release 100303), were carried out to locate additional, informative sequences representing transcripts of *Boudicca*. The contiguous 5858 bp of the copy of *Boudicca* located within BAC 53-J-5 (Copeland et al., 2003) were used as the query.

2.7. GenBank entries

Nucleotide sequences reported here have been assigned accession numbers AY308018, AY308019, AY308020, AY308021, AY308022, AY308023, AY308024, AY308025 and AY308026.

3. Results

3.1. Boudicca transcripts

Previous findings indicated that the copy of *Boudicca* within BAC 53-J-5 was of full-length and included domains required by an active LTR retrotransposon. On the other hand, it exhibited several apparent mutations that may preclude independent retrotransposition (Copeland et al., 2003). These mutated regions were investigated using bioinformatic approaches and RT-PCR. Twenty-five ESTs with BLAST scores of ≥ 200 matching regions between the left- and right-hand LTRs of *Boudicca* are shown in Fig. 1A aligned on a map of the 53-J-5 copy of *Boudicca*. Whereas these 25 ESTs spanned most of the retrotransposon, they (or other currently available ESTs [not shown]) do not form a contiguous sequence of overlapping fragments. Moreover, some mutated regions and other sites of interest, including a stop codon in ORF1 at position 999 of the copy of *Boudicca* in BAC clone 53-J-5, and the clade-characteristic CHCC Cys–His box of ORF1 (Copeland et al., 2003), were not represented in these *S. mansoni* ESTs. RT-PCR amplifications were therefore performed with the aim of generating overlapping fragments spanning the protein coding regions of the *Boudicca* genome (Fig. 1B). Amplicons from the RT-PCRs were cloned and sequenced. These clones have been designated RC1 through RC6. RC1 (GenBank AY308018) and RC4 (AY308019) span *gag* and the 5′-end of *pol*, RC2 (AY308020) and RC5 (AY308021) span the central regions of *pol*, and RC3 (AY308022) and RC6 (AY308023) span the 3′-regions of *pol* (Fig. 1B). Our attempt to amplify using primers from within an uncharacterized coding region immediately downstream of *integrase*, at the position where an envelope gene could be located, based on the domain structure of *gypsy*, was not successful. (PCR from genomic DNA was successful [not shown]. These PCR products have been assigned accession numbers AY308024, AY308025 and AY308026.) The splice site detection program NetGene2 indicated that a splice site at the 5′-terminus of this region is a donor, rather than an acceptor site (not shown). Thus, this ORF appears to have been spliced in to the 53-J-5 *Boudicca* copy from a non-*Boudicca* source. Hybridization and PCR experiments using the *S. mansoni* BAC library indicate that this coding region is part of a low copy number gene not naturally associated with *Boudicca* (not shown).

3.2. RNA secondary structure model of Boudicca transcript suggests mechanism of regulation of transcription and translation

Although RT-PCR amplifications of the *gag* and *pol* regions of *Boudicca* were straightforward, we were not able to amplify the 5′-terminus of *Boudicca* using a primer targeting the first start codon, at nt position 505 of ORF1 of *Boudicca* (Table 1, Fig. 1B). By contrast, RT-PCR from the next in-frame start codon, at nt position 622, amplified a product of the expected size (Fig. 1B, Table 1). PCR from genomic DNA was successful in both conditions (not shown). These contrasting outcomes led us to model the RNA secondary structure of the *Boudicca* contiguous sequence since secondary structure has been reported to influence transcription of other mobile sequences (Beuzon et al., 1999). Because the precise start site of transcription within the LTR of *Boudicca* is not yet known, we modeled the complete LTR retrotransposon for this secondary structure prediction. The first 734 bp of this model is presented in Fig. 2. Models of the *Boudicca* transcript without the LTR (starting at position 329) did not change the base-pairing environment at the start of the *gag* open reading frame (not shown). Overall, the 5′ terminus of the *Boudicca* mRNA is rich in secondary structure, with three closely spaced stem-loop structures encompassing the first AUG and second AUG codons (positions 505–507 and 622–624

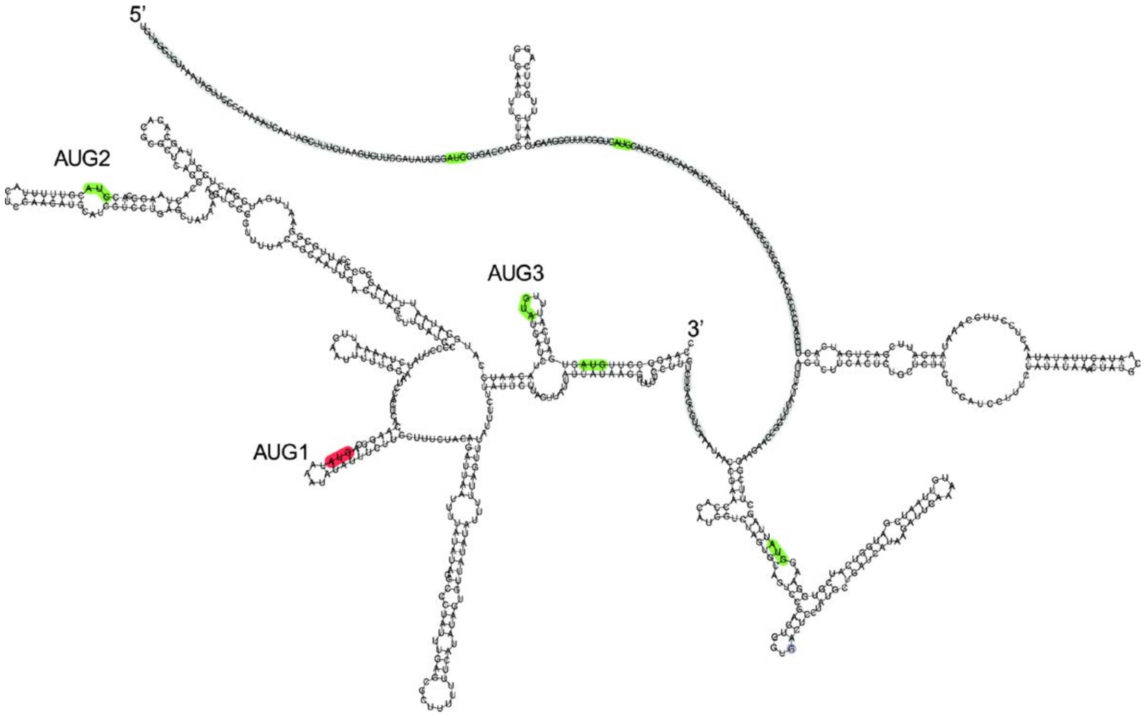


Fig. 2. Theoretical model of the RNA secondary structure of an unspliced *Boudicca* transcript showing potential stem-loop structures. Structure prediction was performed using the 5858-bp copy present in BAC 53-J-5 (a full-length contiguous copy). Only the first 734 nucleotides of the predicted structure are shown. Bases marked in gray are engaged in long range base pairs, mostly to positions close to the 3′ end. The first AUG of the ORF (marked in red) is buried in a hairpin structure that may prevent translation, whereas the AUG2 and AUG3 are situated in unpaired regions.

on the 53-J-5 sequence) (Fig. 2). Of the first three in-frame start codons of this sequence, AUG1 is the least accessible, located within the stem of a hairpin structure. The next two possible in frame start codons, AUG2 and AUG3, are located in loops rather than within stems, and presumably are more accessible. Further, there was a short hairpin loop between AUG1 and AUG2; this could result in an RT-PCR product truncated at its 3′-end, i.e. the 5′-end of *Boudicca*, through pausing of the RT. In addition, this model of the mRNA of *Boudicca* included a stem loop 24 residues upstream of the first AUG, a structural feature correlated with negative effects on translation (Gray and Hentze, 1994).

3.3. Transcribed Boudicca gag confirms a gag polyprotein of the Kabuki/CsRn1 type

The mRNA transcripts from ORF1 of *Boudicca* indicated the structure of a full length *gag* ORF, confirming that the stop codon truncating the copy in BAC clone 53-J-5 was a mutation (Fig. 3A) (Copeland et al., 2003). Amplification of the *gag* region also verified *Boudicca*'s unusual Cys–His box structure, characteristic of the *Kabuki/CsRn1* clade of *gypsy* group retrotransposons, confirming the assignment of *Boudicca* into this newly defined clade (Fig. 3B).

3.4. Analysis of BAC end and RT-PCR derived sequences

The six RT-PCR derived sequences (RC1-RC6) (Fig. 1B) were employed as queries to search the TIGR database of *S. mansoni* BAC ends. A number of sequences with strong matches to the *Boudicca*-associated transcripts were identified. These are listed in Table 2 according to the functional domains that they include: (1) *gag*, (2) *protease*, (3) *reverse transcriptase*, (4) *RNase H*, (5) 5′-*Integrase* and (6) 3′-*Integrase*. Analyses combining the genomic DNAs from BAC ends and the cDNAs from our RT-PCRs indicated DTG as the catalytic motif at the active site of *Boudicca's* protease. The 53-J-5 copy of *Boudicca* has DTD as its active site. The RT-PCR transcripts (clones RC1, RC4) of the protease region indicated that the structure of the active site of *Boudicca* is DTG. Alignment with several BAC end sequences encoding PR of *Boudicca* confirmed that the active site sequence of PR is DTG (Fig. 3C). Whereas the LTR retrotransposon *Tom* also has the DTD motif, most other LTR-retrotransposons and indeed

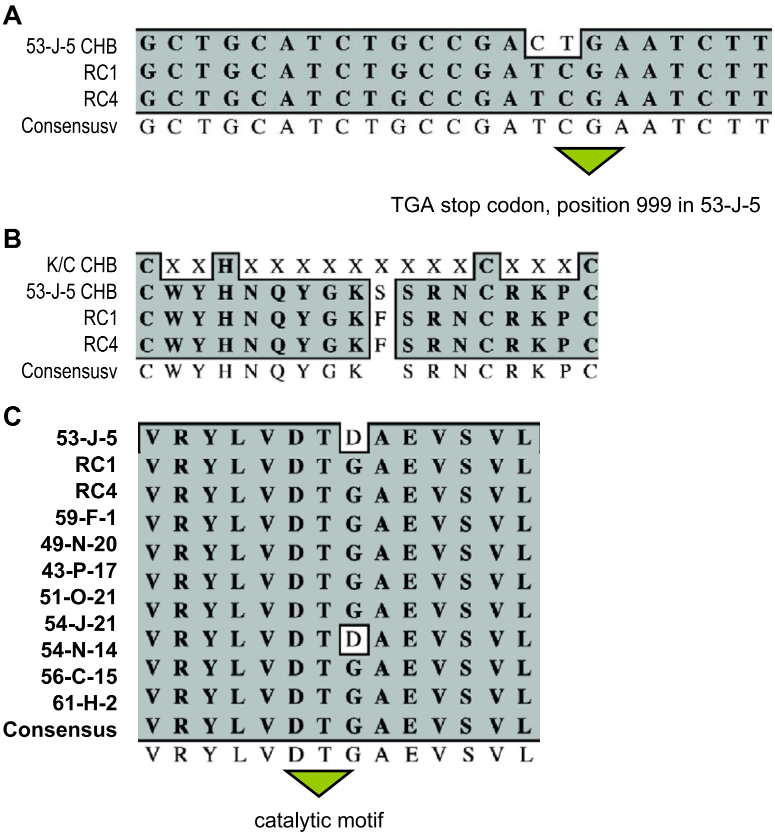

A
53-J-5 CHB G C T G C A T C T G C C G A C T G A A T C T T
RC1 G C T G C A T C T G C C G A T C G A A T C T T
RC4 G C T G C A T C T G C C G A T C G A A T C T T
Consensusv G C T G C A T C T G C C G A T C G A A T C T T
TGA stop codon, position 999 in 53-J-5
B
K/C CHB C X X H X X X X X X X X X C X X X C
53-J-5 CHB C W Y H N Q Y G K S S R N C R K P C
RC1 C W Y H N Q Y G K F S R N C R K P C
RC4 C W Y H N Q Y G K F S R N C R K P C
Consensusv C W Y H N Q Y G K S R N C R K P C
C
53-J-5 V R Y L V D T D A E V S V L
RC1 V R Y L V D T G A E V S V L
RC4 V R Y L V D T G A E V S V L
59-F-1 V R Y L V D T G A E V S V L
49-N-20 V R Y L V D T G A E V S V L
43-P-17 V R Y L V D T G A E V S V L
51-O-21 V R Y L V D T G A E V S V L
54-J-21 V R Y L V D T D A E V S V L
54-N-14 V R Y L V D T G A E V S V L
56-C-15 V R Y L V D T G A E V S V L
61-H-2 V R Y L V D T G A E V S V L
Consensus V R Y L V D T G A E V S V L
V R Y L V D T G A E V S V L
catalytic motif

Fig. 3. Alignments of sequences of key motifs of *Boudicca*. Panel A: Middle *gag* region spanning a putative TGA stop codon in the contiguous *Boudicca* copy present in *S. mansoni* genomic DNA BAC clone 53-J-5. Both RT-PCR products spanning this region exhibit the codon CGA rather than the stop codon TGA. Panel B: Cys–His Box domain showing CHCC structure, characteristic of the *Kabuki/CsRn1* clade. Panel C: Alignment of amino acid sequences of the active site of the protease domain of ORF2 of *Boudicca* shows that the active site consists of DTG rather than DTD.

most aspartic proteases have DTG as the active site motif (Dunn, 1998; Copeland et al., 2003).

We used this same method to analyze the *Boudicca* integrase sequence (Fig. 4). This region, the most degraded part of the 53-J-5 copy of *Boudicca*, encodes one of the most important proteins for retrotransposition in terms of the stable integration of transgenes (Haren et al., 1999). The RT-PCR products RC3 and RC6 (Fig. 1B) encompassed the

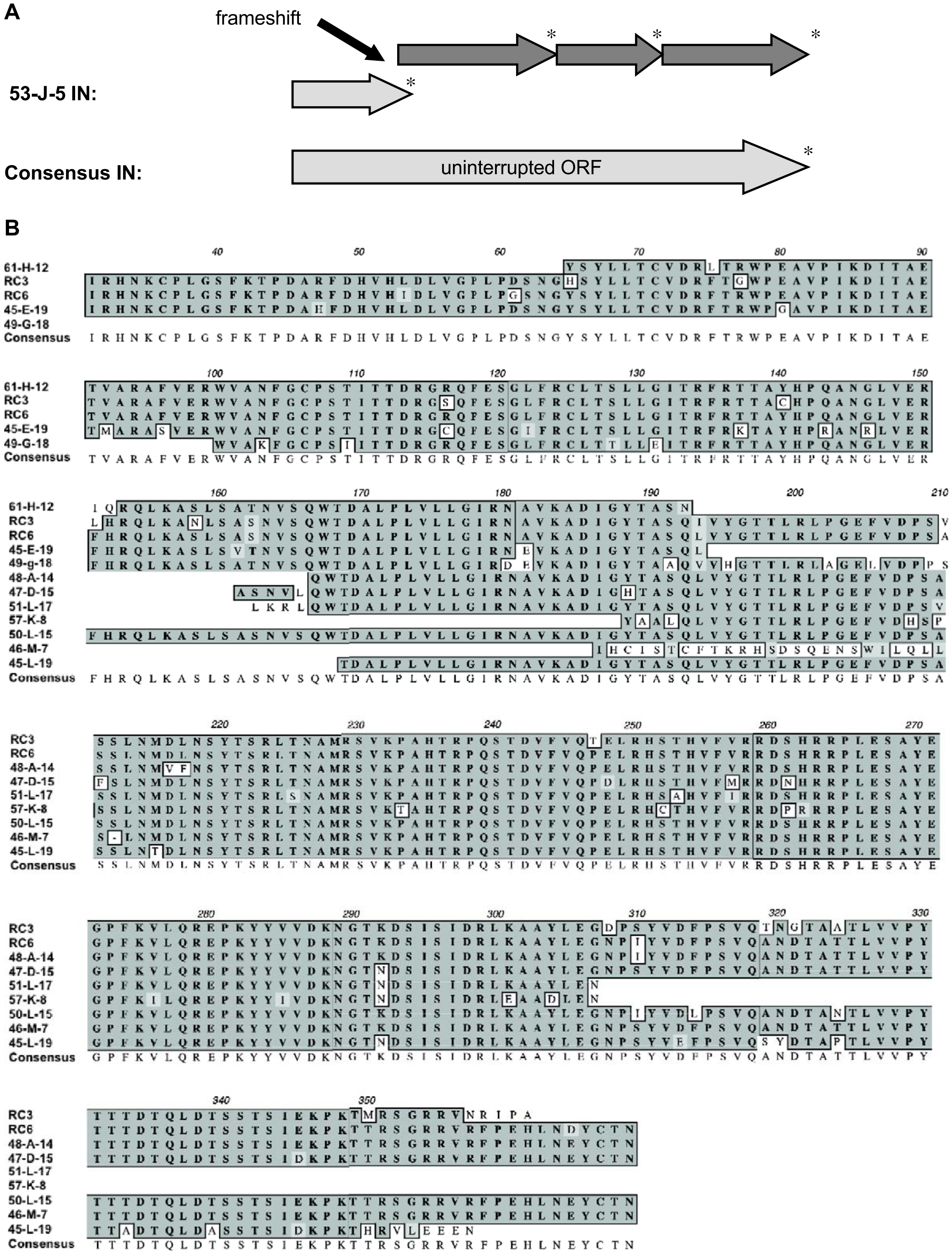
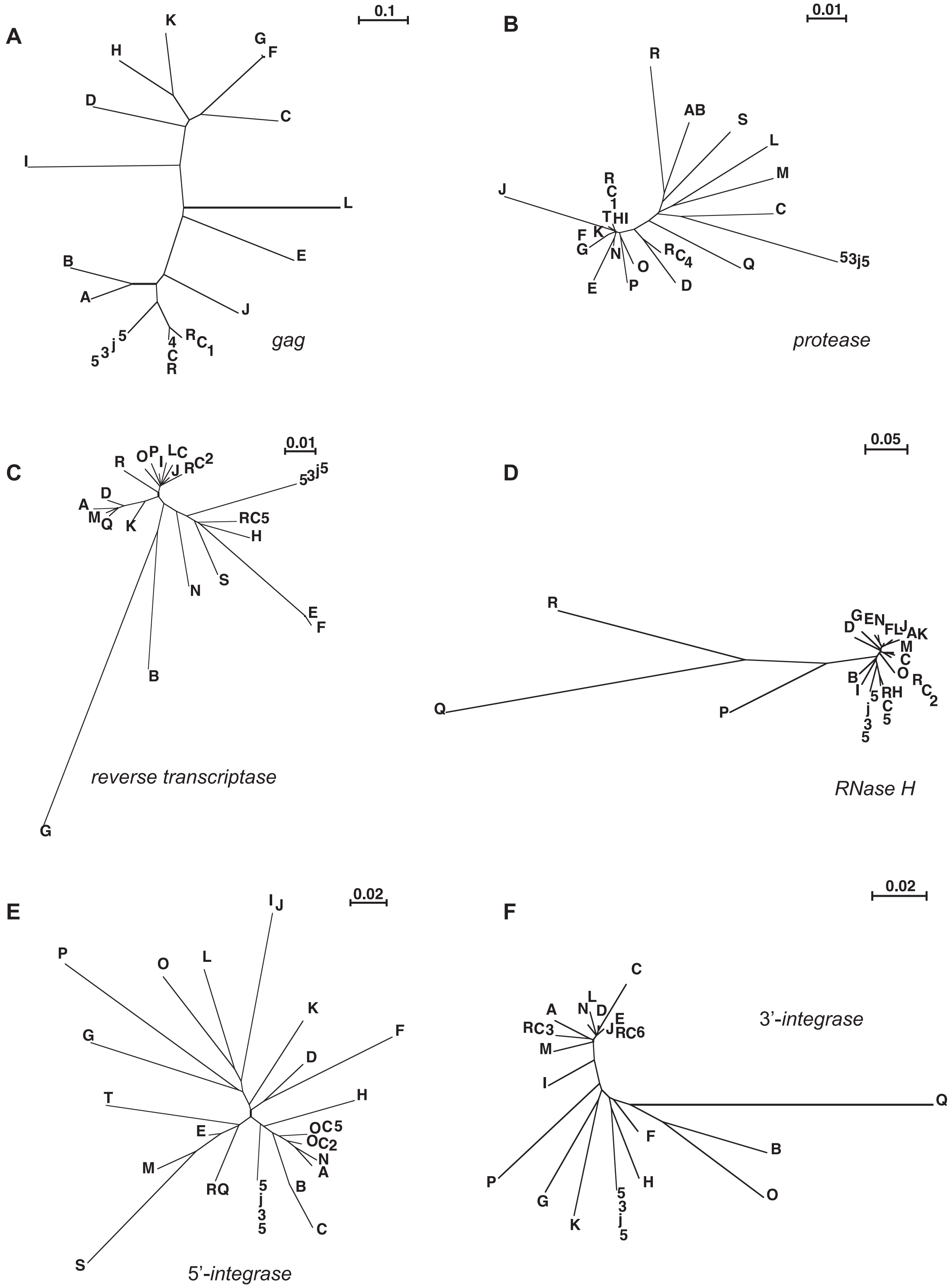

A
frameshift
53-J-5 IN:
Consensus IN:
uninterrupted ORF
*
B
61-H-12
RC3
RC6
45-E-19
49-G-18
49-g-18
48-A-14
47-D-15
51-L-17
57-K-8
50-L-15
46-M-7
45-L-19
Consensus
I R H N K C P L G S F K T P D A R F D H V H L D L V G P L P D S N G Y S Y L L T C V D R F T R W P E A V P I K D I T A E
T V A R A F V E R W V A N F G C P S T I T T D R G R Q F E S G L F R C L T S L L G I T R F R T T A Y H P Q A N G L V E R
F H R Q L K A S L S A S N V S Q W T D A L P L V L L G I R N A V K A D I G Y T A S Q L V Y G T T L R L P G E F V D P S A
S S L N M D L N S Y T S R L T N A M R S V K P A H T R P Q S T D V F V Q P E L R H S T H V F V R R D S H R R P L E S A Y E
G P F K V L Q R E P K Y Y V V D K N G T K D S I S I D R L K A A Y L E G N P S Y V D F P S V Q A N D T A T T L V V P Y
T T T D T Q L D T S S T S I E K P K T T R S G R R V R F P E H L N E Y C T N

Fig. 4. Amino acid sequence and ORF structure of the integrase domain of *Boudicca*. Panel A: Comparison of the open reading frame structures of the 53-J-5 integrase domain with the consensus integrase sequence assembled from RT-PCR transcripts and BAC ends. Asterisks denote stop codons. Panel B: Amino acid alignments assembled to form a probable consensus of the entire integrase polypeptide.

entire integrase gene region of *pol*. Using the RC3 and RC6 PCR products as a framework, overlapping BAC ends were combined to assemble a consensus of the entire integrase domain. This consensus sequence forms a single, uninterrupted open reading frame (Fig. 4A), and includes the conserved integrase catalytic and zinc-finger domains (GenBank conserved domains pfam00665 and pfam02022). It also includes a COOH-terminal non-specific DNA binding domain with homology to *CsRn1* (AAK07486), *Kabuki* (BAA92690) and, to a lesser degree, *412* (P10394), *Osvaldo*



(CAB39733) and other LTR-retrotransposons. Amino acid alignments of this domain are presented in Fig. 4B.

3.5. Variability of Boudicca copies within the S. mansoni genome

Comparison of the mRNA transcripts in clones RC1-RC6 and copies of *Boudicca* from BAC clones shows differing patterns of conservation and divergence for six discrete domains of the *Boudicca* genome (Table 2, Fig. 5). RT and protease were the most highly conserved (note scale bars). The *gag* region, on the other hand, showed the most divergence. The placement of the transcript sequences within the tree of BAC genomic sequences indicates the diversity of specific regions of active copies. The 53-J-5 copy of *Boudicca* may be more, rather than less, recent, when compared with others of the 1000–10,000 copies, based on its identity to the transcribed copies. Close clustering of the 53-J-5 copy with the transcript sequences was especially evident in the least conserved regions, *gag* and RNase H.

4. Discussion

4.1. Sequences of the transcribed copies of Boudicca

The RT-PCR results presented here indicated that *Boudicca* is an actively transcribed LTR retrotransposon. By contrast, the RT-PCR results do not support the notion that *Boudicca* is actively replicated as an enveloped retrovirus-like retrotransposon. Although GenBank dbEST included fragments corresponding to this locus, we were unable to amplify mRNA using primers encompassing the *integrase* and part of the unknown coding region or from primers targeting only the unknown coding region downstream of integrase. Failure to amplify the third ORF using RT-PCR identified in the 53-J-5 copy of *Boudicca*, information from screening of the *S. mansoni* BAC library, and splice site analysis (not shown) suggested that this gene is of non-*Boudicca* origin, and may have been annexed through splicing of a gene fragment into this copy of *Boudicca*.

4.2. RNA secondary structure and reverse transcription

The region of the *Boudicca* mRNA encompassing the translation initiation site is rich in secondary structure, with three closely spaced stem-loop structures encompassing the first and second AUG codons (positions 505–507 and 622–624 on the 53-J-5 sequence) (Fig. 2). The hairpin between the two start codons, in particular, could result in an RT-PCR product truncated at its 3′ end (*Boudicca*'s 5′ end) if it induced pausing by the RT. Indeed, the fact that reverse transcription proceeded successfully to AUG2 (past AUG3, position 709–711 on the 53-J-5 sequence) indicated that the hairpin at AUG3 (position 704–719 on the 53-J-5 sequence) did not significantly impede the progress of the RT. Perhaps the proximity of the large stem at nt 550–693 on the 53-J-5 sequence to this smaller stem interferes with the processivity of the RT. Alternatively, each of the several closely spaced hairpins might interfere with the processivity of RT, such that a gradient of transcripts is produced, with shorter transcripts predominating. In this case, it would be expected that a full length transcript including AUG1 would result only occasionally. Since S. *mansoni* ESTs span this region (e.g., AI395459 and AI067132), some transcripts that include AUG1 are produced. RNA secondary structure can adversely affect reverse transcription, based on the ability of the specific RT to denature the secondary structure of the template RNA (Brooks et al., 1995). With HIV, for example, RT pause sites have been correlated with predicted hairpin structures of the RNA template (Suo and Johnson, 1997).

4.3. Multiple in-frame start codons, accompanying mRNA secondary structure and regulation of copy number through translation

Retrotransposons can be thought of as genetic parasites which have coevolved with their hosts. As such, they have developed mechanisms to limit their adverse impact on the survival of the host genome, including methods to limit copy number (see Deininger and Batzer, 2002; Carr et al., 2000). One possible method of copy number limitation is the interference with reverse transcription hypothesized to explain the truncated RT-PCR results discussed above.

Fig. 5. Phylogenetic relationships among transcripts and genomic copies of the *Boudicca* retrotransposon. Panel A: *gag*. Panel B: *protease*. Panel C: *reverse transcriptase*. Panel D: *RNase H*. Panel E: Integrase 5′ region. Panel F: Integrase 3′ region. Note scale bars; scales differ by up to a factor of 10 for each panel. Bootstrap values in general were low in regions of tight clustering and higher for longer branch lengths. These bootstrap values are available as supplementary data at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/03-023/. BAC ends containing the *Boudicca* fragments used in each panel were as follows: Panel A: A, 62-N-16; B, 45-H-5; C, 46-G-15; D, 55-N-10; E, 61-N-7; F, 61-D-18; G, 62-C-17; H, 54-H-1; I, 57-H-13; J, 62-G-23; K, 54-H-12; L, 49-N-18. Panel B: A, 59-B-11; B, 59-F-1; C, 43-P-17; D, 49-D-16; E, 40-H-16; F, 56-C-15; G, 45-P-23; H, 49-N-20; I, 61-H-2; J, 54-N-14; K, 49-G-19; L, 54-J-21; M, 49-L-5; N, 40-H-19; O, 50-J-14; P, 50-E-22; Q, 57-A-18; R, 62-P-1; S, 51-O-21; T, 50-N-14. Panel C: A, 57-E-12; B, 45-M-2; C, 62-E-6; D, 59-I-14; E, 57-M-23; F, 57-P-23; G, 61-A-3; H, 62-F-10; I, 44-G-15; J, 44-C-12; K, 54-H-5; L, 42-A-9; M, 61-M-10; N, 48-E-19; O, 46-N-22; P, 54-M-22, Q, 54-K-20; R, 42-K-16; S, 42-L-6. Panel D: A, 62-O-9; B, 62-F-10; C, 54-H-5; D, 48-E-19; E, 59-L-12; F, 54-M-22; G, 42-A-9; H, 60-B-10; I, 56-J-10; J, 56-O-19; K, 62-O-19; L, 61-I-19; M, 44-C-12; N, 61-M-10; O, 44-G-15; P, 44-E-1; Q, 45-F-9; R, 48-I-8. Panel E: A, 56-J-12; B, 44-K-10; C, 39-L-24; D, 61-M-19; E, 54-P-13; F, 46-H-4; G, 59-B-18; H, 51-O-10; I, 58-O-17; J, 58-K-19; K, 51-C-6; L, 45-H-22; M, 54-N-19; N, 55-N-14; O, 42-C-17; P, 45-F-9; Q, 52-C-24 (rev); R, 52-C-24 (fwd); S, 41-E-11; T, 52-H-22. Panel F: A, 46-M-7; B, 57-K-8; C, 54-P-24; D, 50-L-15; E, 50-O-20; F, 45-L-19; G, 60-J-19; H, 54-I-6; I, 47-D-15; J, 56-N-2; K, 55-A-24; L, 48-A-14; M, 51-L-17; N, 59-F-13; O, 43-F-11; P, 45-N-18; Q, 55-C-9.

Since reverse transcription is an intrinsic part of retrotransposition, premature disengagement of RT would result in the insertion of truncated copies of the element into the genome. Another potential function of this RNA secondary structure is at the level of translation. One possible method of copy number limitation could be through the use of alternative in-frame start codons such that the fully functional protein is produced only occasionally. The secondary structure model, in tandem with the findings of the RT-PCRs, presents the possibility that *Boudicca* may regulate its copy number through alternative translation initiation sites. Eukaryotic and prokaryotic genes generally lack secondary structure at the start AUG (Ganoza and Louis, 1994). The first AUG of *Boudicca*, however, is located within the helix stem of a stem-loop structure, in contrast to the positions of the second and third AUG within loops. Ribosome binding can be reduced or even eliminated when a start codon is deeply embedded within secondary structure (Yun et al., 1996; Satchidanandam and Shivashankar, 1997). Comparison of clones with equal overall stability of RNA secondary structure has shown that an AUG made inaccessible within a stem structure is correlated with highly attenuated expression levels (Wang et al., 1995). In addition, RNA secondary structure upstream of start codons can impede translation (Gray and Hentze, 1994); the stem-loop structure found 5′ of the first AUG could have this function.

In terms of the issue of accessibility, the RNA secondary structure of *Boudicca* appears to favor binding of the ribosome to AUG2 or AUG3 over AUG1 (Fig. 2). These start codons are located in the more easily accessible loops, rather than the stems, of their associated stem-loop structures. The resulting truncated polypeptide could in turn contribute to element down-regulation if this shorter polypeptide is non-functional or subfunctional compared with the full-length version. Alternatively, the more accessible AUG2 or AUG3 may be the optimal sites for translation initiation. In either case, translation efficiency and/or choice of translation initiation site may be governed by RNA secondary structure, and may serve as a regulation point for element activity. Without experimental data to support this hypothesis on secondary structure/function, our interpretation remains speculative at this point. (Future analysis by 5′ RACE and other techniques can be expected to determine the precise start sites of *Boudicca* transcripts.) Nonetheless, examples of control of expression by RNA secondary structure working together with other regulatory systems are known. For example, in plant pararetroviruses, the RNA secondary structure induces ribosomes to “shunt”, or skip over, entire large stem-loop structures to target specific downstream AUGs, a process thought to be controlled by specific factors in the initiation complex (Hohn et al., 2001). A critical role for host factors in the process of regulation is consistent with the notion of a presumed host influence on element activity (Plasterk, 2002).

4.4. Unusual gag structure confirmed, copy-specific mutations resolved

In addition to illustrating the general structure of the transcript of *Boudicca*, the RT-PCR products also confirmed more specific primary sequence level motifs. The sequences of the individual RT-PCR products spanning the gag polyprotein confirmed both a continuous open reading frame for gag and the unique CHCC Cys–His Box characteristic of the *Kabuki*/*CsRn1* clade, of which *Boudicca* is one of the few known members. On the other hand, the unusual DTD protease motif seen in 53-J-5′s *Boudicca* and in another *gypsy* group retrotransposon, *Tom* (CAA80824), was refuted in favor of the more common DTG motif. The most obviously degraded individual gene in *Boudicca* was that of integrase. Stop mutations and a frameshift mutation divided the integrase of the copy of *Boudicca* in BAC 53-J-5 into a series of small open reading frames. If *Boudicca* were to be used as a vector for transgenesis of *S. mansoni*, a continuous, functional integrase protein would be crucial, as this is the protein which mediates the stable integration of genetic material into the *S. mansoni* genome. The consensus integrase sequence formed by combining the *Boudicca* RT-PCR products with fragments of integrase obtained from BAC ends from the TIGR database has as its structure a single open reading frame. Moreover, the amino acid sequence is, for the overwhelming majority of the protein, clear, with few individual amino acids that are ambiguous. This transcript and multi-copy genomic based consensus could form a valuable tool as the first step of the process of back mutation which would be necessary if *Boudicca* were to be used as a transgenesis vector.

4.5. Evolution/phylogeny of the population of Boudicca copies

By placing the sequences of RT-PCR transcripts of specific regions within phylogenetic trees of corresponding BAC end fragments, it was possible to assign the 53-J-5 copy of *Boudicca* within the distribution of various *Boudicca* copies. This analysis also served to elucidate the level of divergence tolerated for each region and other patterns of divergence and conservation among copies of *Boudicca*. The well-characterized genomic copy of *Boudicca* in BAC clone 53-J-5 falls within the conserved cluster of *Boudicca* copies based on the RT-PCR products. This suggests that the 53-J-5 sequence represents a copy of *Boudicca* relatively close to the putative “active”, or non-mutated, copy. BLAST searches of the TIGR BAC end database using the RT-PCR products as queries revealed a general pattern of a cluster of BAC end sequences similar to the RT-PCR sequences, with other BAC ends lying farther out in the tree (Fig. 5, Table 2). The two RT-PCR sequences from each area showed a tendency to cluster together.

In general, certain domains of LTR retrotransposons exhibit a greater degree of conservation than others (Li et al., 1995). Through the use of the RT-PCR products and BAC end sequences, we have addressed the question of whether this is also true for *Boudicca*. *Reverse transcriptase* and *protease* showed the highest degree of conservation, while *gag* was the least conserved. Certain regions showed qualitatively different patterns from others, as well. Some, such as the 5′ end of *integrase*, showed a pattern of general radiation. This pattern may reflect a gene with a relatively high tolerance for mutation and reduced sequence stringency for functionality of the gene. The other pattern, exemplified by *RNase H*, was that of a tight cluster of closely related sequences, with other copies lying much farther out. This pattern could reflect a region with selection pressure for a conserved sequence, with the outliers corresponding to inactive, nonfunctional copies of the element. Finally, since the copy of *Boudicca* in BAC clone 53-J-5 is both contiguous and well characterized, and since it appears to be close in sequence to a recently active copy, it may be feasible to reconstruct an active copy of this retrotransposon, as has been accomplished with some other mobile genetic elements (e.g., Ivics et al., 1997).

Acknowledgements

Preliminary sequence data for the *S. mansoni* BAC inserts analyzed here (listed in Table 2) were obtained from The Institute for Genomic Research website (http://www.tigr.org/tdb/e2k1/sma1/genome.shtml). Sequencing of these BAC inserts is supported by an award from the National Institute of Allergy and Infectious Diseases, National Institutes of Health. This study was supported by grant number KA 866/2-1 of the Deutsche Forschungsgemeinschaft (DFG) to BHK, by an Infrastructure Grant no. ID-IA-0037-02 from the Ellison Medical Foundation to PJB and others, by a DFG bioinformatics initiative BIZ-6/1-2 to PFS, and by Austrian FWF Project number P15893 to ILH and PFS. PJB is a recipient of a Burroughs Wellcome Fund Scholar Award in Molecular Parasitology. Both The Interdisciplinary Program in Molecular and Cellular Biology and the Center for Infectious Diseases, Tulane University supported this study, as did the Department of Molecular Parasitology, Institute for Biology, Humboldt University Berlin.

References

Abe, H., Ohbayashi, F., Shimada, T., Sugasaki, T., Kawai, S., Mita, K., Oshiki, T., 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. Genet. Genomics 263, 916–924.

Bae, Y.A., Moon, S.Y., Kong, Y., Cho, S.Y., Rhyu, M.G., 2001. CsRn1, a novel active retrotransposon in a parasitic trematode, *Clonorchis sinensis*, discloses a new phylogenetic clade of *Ty3/gypsy*-like LTR retrotransposons. Mol. Biol. Evol. 18, 1474–1483.

Beuzon, C.R., Marques, S., Casadesus, J., 1999. Repression of IS200 transposase synthesis by RNA secondary structures. Nucleic Acids Res. 27, 3690–3695.

Brindley, P.J., Laha, T., McManus, D.P., Loukas, A., 2003. Mobile genetic elements colonizing the genomes of metazoan parasites. Trends Parasitol. 19, 79–87.

Brooks, E.M., Sheflin, L.G., Spaulding, S.W., 1995. Secondary structure in the 3′ UTR of EGF and the choice of reverse transcriptases affect the detection of message diversity by RT-PCR. BioTechniques 19, 806–815.

Brunak, S., Engelbrecht, J., Knudsen, S., 1991. Prediction of human mRNA donor and acceptor sites from the DNA sequence. J. Mol. Biol. 220, 49–65.

Carr, M., Soloway, J.R., Robinson, T.E., Brookfield, J.F.Y., 2000. Mechanisms regulating the copy numbers of six LTR retrotransposons in the genome of *Drosophila melanogaster*. Chromosoma 110, 511–518.

Colley, D.G., LoVerde, P.T., Savioli, L., 2001. Medical helminthology in the 21st century. Science 293, 1437–1438.

Copeland, C.S., Brindley, P.J., Heyers, O., Michael, S.F., Johnston, D.A., Williams, D.J., Ivens, A., Kalinna, B.H., 2003. Boudicca, a retrovirus-like, LTR retrotransposon from the genome of the human blood fluke, *Schistosoma mansoni*. J. Virol. 77, 6153–6166.

Crompton, D.W., 1999. How much human helminthiasis is there in the world? J. Parasitol. 85, 397–403.

Deininger, P.L., Batzer, M.A., 2002. Mammalian retroelements. Genome Res. 12, 1455–1465.

Dunn, B.M., 1998. Retropepsin. In: Barrett, A.J., Rawlings, N.D., Woessner, J.F. (Eds.), Handbook of Proteolytic Enzymes. Academic Press, Oxford, pp. 919–928.

Ganoza, M.C., Louis, B.G., 1994. Potential secondary structure at the translational start domain of eukaryotic and prokaryotic mRNAs. Biochimie 76, 428–439.

Gray, N.K., Hentze, M.W., 1994. Regulation of protein synthesis by mRNA structure. Mol. Biol. Rep. 19, 195–200.

Haren, L., Ton-Toang, B., Chandler, M., 1999. Integrating DNA: transposases and retroviral integrases. Annu. Rev. Microbiol. 53, 245–281.

Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., Brunak, S., 1996. Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. Nucleic Acids Res. 24, 3439–3452.

Hohn, T., Corsten, S., Dominguez, D., Fütterer, J., Kirk, D., Hemmings-Mieszczak, M., Pooggin, M., Schärer-Hernandez, N., Ryabova, L., 2001. Shunting is a translation strategy used by plant pararetroviruses (Caulimoviridae). Micron 32, 51–57.

Hofacker, I.L., 2003. The Vienna RNA secondary structure server. Nucleic Acids Res. 31, 3429–3431.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures (The Vienna RNA Package). Monatsh. Chem. 125, 167–188.

Ivics, Z., Hackett, P.B., Plasterk, R.H., Izsvak, Z., 1997. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91, 501–510.

Le Paslier, M.C., Pierce, R.J., Merlin, F., Hirai, H., Wu, W., Williams, D.L., Johnston, D., LoVerde, P.T., Le Paslier, D., 2000. Construction and characterization of a *Schistosoma mansoni* bacterial artificial chromosome library. Genomics 65, 87–94.

Li, M.D., Bronson, D.L., Lemke, T.D., Faras, A.J., 1995. Phylogenetic analyses of 55 retroelements on the basis of the nucleotide and product amino acid sequences of the *pol* gene. Mol. Biol. Evol. 12, 657–670.

Mathews, D.H., Sabina, J., Zucker, M., Turner, D.H., 1999. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. J. Mol. Biol. 288, 911–940.

Petrov, D.A., Sangster, T.A., Johnston, J.S., Harl, D.L., Shaw, K.L., 2000. Evidence for DNA loss as a determinant of genome size. Science 287, 1060–1062.

Plasterk, R.H.A., 2002. RNA silencing: the genome's immune system. Science 296, 1263–1265.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Satchidanandam, V., Shivashankar, Y., 1997. Availability of a second upstream AUG can completely overcome inhibition of protein synthesis initiation engendered by mRNA secondary structure encompassing the start codon. Gene 196, 231–237.

Suo, Z., Johnson, K.A., 1997. RNA secondary structure switching during DNA synthesis catalyzed by HIV-1 reverse transcriptase. Biochemistry 36, 14778–14785.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeannmougin, F., Higgins, D.G., 1997. The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882.

Wang, G., Liu, N., Yang, K., 1995. High-level expression of prochymosin in *Escherichia coli*: effect of the secondary structure of the ribosome binding site. Protein Expr. Purif. 6, 284–290.

Yun, D.F., Laz, T.M., Clements, J.M., Sherman, F., 1996. mRNA sequences influencing translation and the selection of AUG initiator codons in the yeast *Saccharomyces cerevisiae*. Mol. Microbiol. 19, 1225–1239.