The *ingi* and RIME non-LTR Retrotransposons Are Not Randomly Distributed in the Genome of *Trypanosoma brucei*

Frédéric Bringaud, Nicolas Biteau,* Eduard Zuiderwijk,† Matthew Berriman,† Najib M. El-Sayed,‡§ Elodie Ghedin,‡§ Sara E. Melville,|| Neil Hall,† and Théo Baltz**

*Laboratoire de Génomique Fonctionnelle des Trypanosomatides, UMR-5162 CNRS, Université Victor Segalen Bordeaux II, Bordeaux, FRANCE; †The Wellcome Trust Sanger Institute, Hinxton, United Kingdom; ‡The Institute for Genomic Research Rockville, Maryland; §Department of Microbiology and Tropical Medicine, George Washington University, Washington, DC; and ||Molteno Institute for Parasitology, Department of Pathology, University of Cambridge, Cambridge, U.K.

The *ingi* (long and autonomous) and RIME (short and nonautonomous) non–long-terminal repeat retrotransposons are the most abundant mobile elements characterized to date in the genome of the African trypanosome *Trypanosoma brucei*. These retrotransposons were thought to be randomly distributed, but a detailed and comprehensive analysis of their genomic distribution had not been performed until now. To address this question, we analyzed the *ingi*/RIME sequences and flanking sequences from the ongoing *T. brucei* genome sequencing project (TREU927/4 strain). Among the 81 *ingi*/RIME elements analyzed, 60% are complete, and 7% of the *ingi* elements (approximately 15 copies per haploid genome) appear to encode for their own transposition. The size of the direct repeat flanking the *ingi*/RIME retrotransposons is conserved (i.e., 12-bp), and a strong 11-bp consensus pattern precedes the 5′-direct repeat. The presence of a consensus pattern upstream of the retroelements was confirmed by the analysis of the base occurrence in 294 GSS containing 5′-adjacent *ingi*/RIME sequences. The conserved sequence is present upstream of *ingi*s and RIMEs, suggesting that *ingi*-encoded enzymatic activities are used for retrotransposition of RIMEs, which are short nonautonomous retroelements. In conclusion, the *ingi* and RIME retroelements are not randomly distributed in the genome of *T. brucei* and are preceded by a conserved sequence, which may be the recognition site of the *ingi*-encoded endonuclease.

Introduction

Retrotransposons are ubiquitous mobile genetic elements, found in the genome of most organisms, which transpose through an RNA intermediate (Capy et al. 1998). The current model for transposition of non-LTR retrotransposons was developed based on the analysis of the insect R2 element (Luan et al. 1993). This model predicts that an element-encoded endonuclease creates a single-strand nick in the target DNA, generating an exposed 3′ hydroxyl that serves as a primer for reverse transcription of the element's RNA. The complementary strand of the new DNA copy of the element is thus directly synthesized onto the chromosome by the element-encoded reverse transcriptase. The second single-strand nick is created in the other strand a few base pairs downstream of the first nick, by the same element-encoded endonuclease, generating a primer for the second-strand synthesis of the retroelement. Consequently, most of the non-LTR retroelements are flanked by a direct repeat corresponding to the sequence between the two single-strand nicks generated by the element-encoded endonuclease. Most of these elements have a variable length poly(A) or A-rich 3′ tail because of the involvement of an RNA intermediate. Recently, an alternative model proposing that the non-LTR retrotransposons integrate at staggered breaks has been confirmed for the human L1 elements (Morrish et al. 2002), indicating that retrotransposition of these elements is not always endonuclease-mediated.

Non-LTR retroelements are very diverse in structure and can insert into a wide variety of different types of DNA targets. Some integrate within very specific sequences, such as rDNA genes (R2 and R4), the spliced leader RNA genes (NeSL-1, SLACS, CZAR, CRE1, and CRE2), and subtelomeric or telomeric repeats (Genie I and TRAS1) (for review see Craig [1997]). Other retroelements (Zepp, TART, and HeT-A) are restricted to the telomeric regions of chromosomes, but they do not show the extreme site-specificity (Pardue and DeBerardinis 2002). Most of the non-LTR retroelements, exemplified by the autonomous human L1 element (a long interspersed nucleotide element [LINE]), are considered to be randomly distributed in the genome. However, the observed bias in the base composition at the insertion sites of the L1 elements correlates with the relative sequence specificity of the L1-encoded endonuclease, indicating that the distribution of these retroelements is not random (Feng et al. 1996; Jurka 1997; Cost and Boeke 1998).

Trypanosomes are unicellular protists and human pathogens responsible for African sleeping sickness (*Trypanosoma brucei*) and Chagas' disease (*Trypanosoma cruzi*). Non-LTR retrotransposons constitute the most abundant mobile elements described in the genome of *T. brucei* (*ingi*, RIME, and SLACS) (Aksoy 1991). SLACS are site-specific retroelements only found in the spliced leader RNA genes (Aksoy et al. 1987), whereas *ingi*s and RIMEs have been reported as randomly distributed in the host genome (Hasan, Turner, and Cordingley 1984; Kimmel, Ole-MoiYoi, and Young 1987; Murphy et al. 1987). The *ingi* (5.25-kb) retroelement presents the characteristics of the autonomous LINE elements, whereas the RIME (0.5-kb) are short nonautonomous retroelements. The *ingi* retroelement is composed of a 4.7-kb fragment, bordered by two separate halves of the RIME retroelement called RIME-A and RIME-B for the 5′ and 3′ extremities, respectively (fig. 1). It encodes a large single protein

Key words: *ingi*, insertion site specificity, non-LTR retrotransposon, retroelement hot spot gene, RIME, *Trypanosoma brucei*.

E-mail: bringaud@u-bordeaux2.fr.

Mol. Biol. Evol. 21(3):520–528. 2004
DOI: 10.1093/molbev/msh045
Advance Access publication December 23, 2003
Molecular Biology and Evolution vol. 21 no. 3

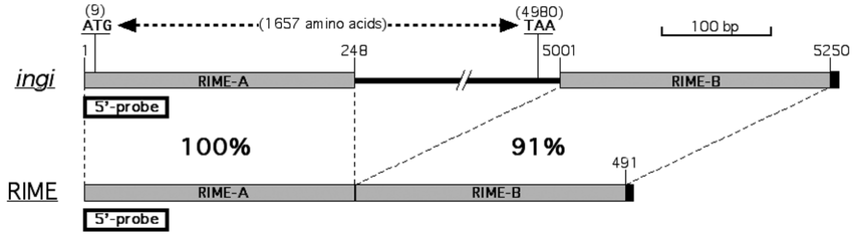


FIG. 1.—Schematic representation of the *ingi* and RIME non-LTR retrotransposons present in the *T. brucei* database. The *ingi* and RIME retroelements shown here are based on the fifth (26P8.i5) and first (26P8.r6) retrotransposons, respectively, present in the fully sequenced BAC-26P8 (GenBank accession number AC087701). RIME elements are 0.5-kb nonautonomous retroelements (Hasan, Turner, and Cordingley 1984). According to the nomenclature previously proposed, the first 248-bp of RIME are called RIME-A, and the last 250-bp are called RIME-B (Hasan, Turner, and Cordingley 1984). The "LINE-like" *ingi* is a 5.25-kb retroelement composed of an *ingi*-specific 4.75-kb DNA fragment (thick line) flanked by the RIME-A (5′ extremity) and RIME-B (3′ extremity) sequences (gray boxes) (Kimmel, Ole-MoiYoi, and Young 1987; Murphy et al. 1987). The *ingi* potentially functional retroelements contain a single long ORF (4,971-bp) from position 9 (ATG codon) to position 4,980 (TAA codon), which encodes a 1,657 aa protein. The percentage of identity between the 5′ and 3′ extremities of the *ingi* (26P8.i5) and RIME (26P8.r6) sequences is indicated, and the black box at the end of both maps represents the poly(dA) terminal sequence. The open black boxes labeled 5′ probes represent the 75-bp sequence used for Blast search.

containing a central reverse transcriptase domain, a C-terminal DNA-binding domain (Pays and Murphy 1987), and an N-terminal apurinic-apyrimidinic–like endonuclease domain (Olivares, Alonso, and Lopez 1997).

The original characterization of the *ingi* and RIME retroelements suggested they were randomly distributed in the nuclear genome of *T. brucei* (Hasan, Turner, and Cordingley 1984; Kimmel, Ole-MoiYoi, and Young 1987; Murphy et al. 1987). However, we recently characterized a large multigene family, called *RHS*, that contains a hot spot for insertion of *ingi*/RIME retrotransposons (Bringaud et al. 2002). Indeed, approximately one-third of the 280 *RHS* (pseudo)genes present in the diploid genome contain one (or more) retroelement(s) inserted at the same relative position. To study the global insertion site specificity of these retrotransposons, we have analyzed *ingi* and RIME sequences identified in the ongoing *T. brucei* genome sequencing project data: about 800 from the genome survey sequences (GSS: 90,000 sequences) and 81 from chromosome-specific sequencing reads.

Materials and Methods

Sequence Analyses

DNA and amino acid sequences were analyzed using DNA STRIDER and Artemis (The Wellcome Trust Sanger Institute, http://www.sanger.ac.uk/Software/Artemis/) programs, and database searches were performed using Blast. Multiple alignments of DNA sequences were performed using MacVector version 6.0.1 and AutoAssembler version 2.0 (PerkinElmer).

Statistical Analysis

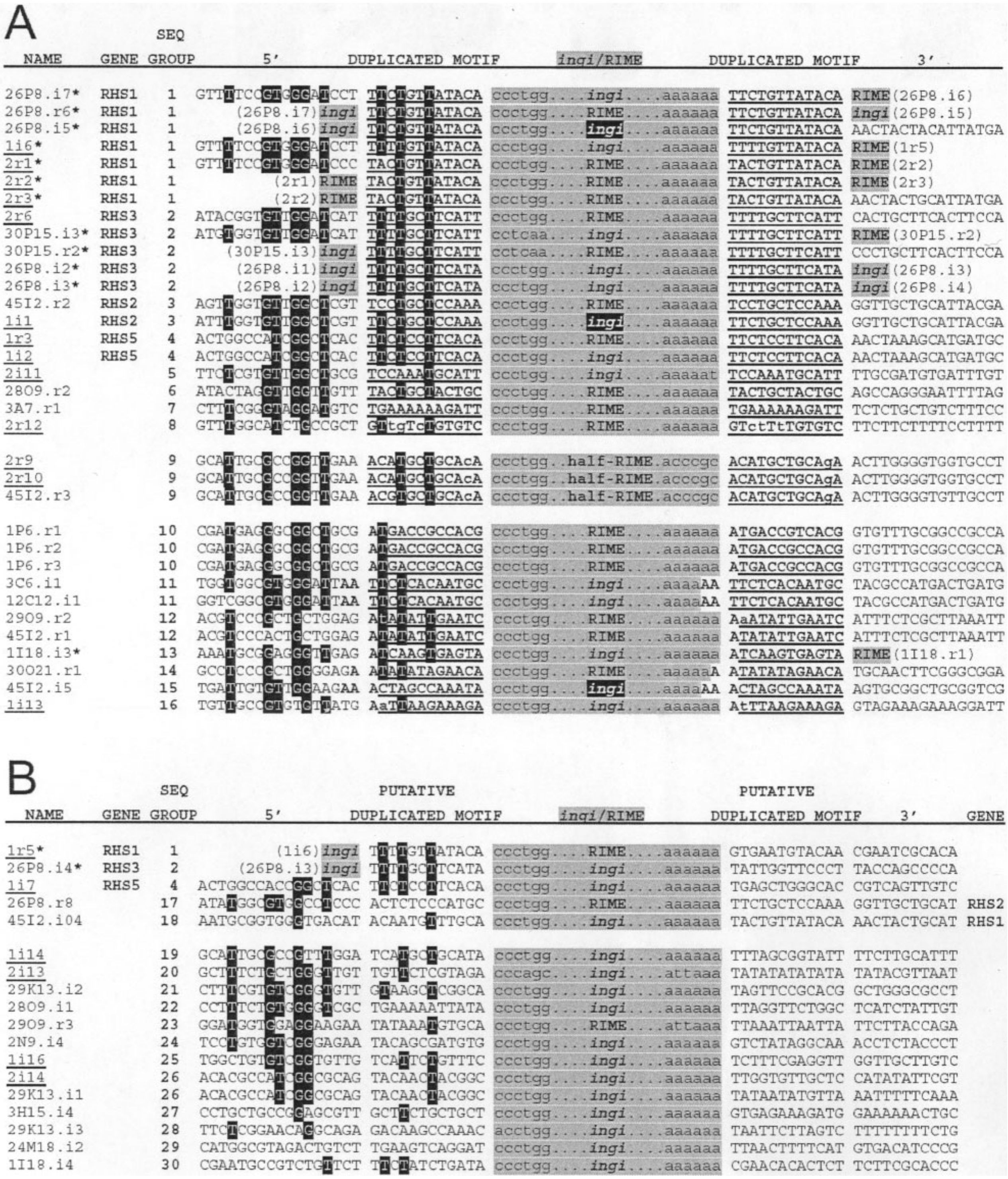
The distribution of bases in the genome survey sequence (GSS) was analyzed using the Kolmogorov-Smirnov (K-S) two-sample test. The base composition at an individual position is compared with a "background" distribution sampled further upstream from the (suspected) conserved domain. The K-S test is nonparametric and generally more powerful than parametric tests (such as a χ^2 test). It determines the probability that two observed distributions are drawn from the same parent population without making any assumption about the sampling characteristics of the distributions involved. However, the significance estimates translate directly into the more familiar χ^2 test scores (for two degrees of freedom [Siegel and Castellan Jr. 1988]) and those are the ones presented here.

Results

Analysis of *Ingi*/RIME Retroelements from Chromosome Sequencing

To analyze the *ingi* and RIME retroelements, we studied all the retroelement sequences (full-length or not) present in the *T. brucei* (TREU927/4 strain) sequence database that contains the 1.1 Mb chromosome Ia (*Chr*Ia) (the Wellcome Trust Sanger Institute) (Hall et al. 2003) and about 45 sequenced BAC clones containing genomic DNA fragments of approximately 140 kb (TIGR) (El-Sayed et al. 2000). Twelve of these BAC sequences have been assembled to generate a contig covering the 1.2 Mb chromosome II (*Chr*II) (El-Sayed et al. 2003). These contigs represent about 7.4 Mb of large, fully sequenced genomic DNA fragments, corresponding to 27.7% of the haploid genome (26.7 Mb, excluding minichromosomes). Blast searches with the *ingi* and RIME sequences identified 81 retroelement sequences within these contigs, suggesting that the nonminichromosomal haploid genome contains approximately 292 retroelements (192 *ingi*s and 100 RIMEs). This is consistent with a previous Southern blot analysis, which estimated the *ingi* copy number to be in the range of 200 per haploid genome (Murphy et al. 1987).

These 81 retroelements correspond to 46 *ingi*s (29 full length and 17 truncated), 21 RIMEs (20 full length and one truncated), three half-RIMEs (RIME-A sequences flanked by a duplicated motif [data not shown]) and 11 incomplete RIME sequences, which may correspond to either truncated *ingi* or RIME retroelements. Among them, 49 (60%) are complete RIME or *ingi* retroelements, but only three *ingi*s code for a full-length protein (1,657 amino acids), suggesting that less than 7% of the *ingi*s potentially code for their own retrotransposition.



A

NAME	GENE	SEQ GROUP	5′	DUPLICATED MOTIF	*ingi*/RIME	DUPLICATED MOTIF	3′
26P8.i7*	RHS1	1	GTTTTCCGTGGGATCCT	TTCTGTTATACA	ccctgg....*ingi*....aaaaaa	TTCTGTTATACA	RIME(26P8.i6)
26P8.r6*	RHS1	1	(26P8.i7)*ingi*	TTCTGTTATACA	ccctgg....RIME....aaaaaa	TTCTGTTATACA	*ingi*(26P8.i5)
26P8.i5*	RHS1	1	(26P8.i6)*ingi*	TTCTGTTATACA	ccctgg....*ingi*....aaaaaa	TTCTGTTATACA	AACTACTACATTATGA
1i6*	RHS1	1	GTTTTCCGTGGGATCCT	TTTTGTTATACA	ccctgg....*ingi*....aaaaaa	TTTTGTTATACA	RIME(1r5)
2r1*	RHS1	1	GTTTTCCGTGGGATCCC	TACTGTTATACA	ccctgg....RIME....aaaaaa	TACTGTTATACA	RIME(2r2)
2r2*	RHS1	1	(2r1)RIME	TACTGTTATACA	ccctgg....RIME....aaaaaa	TACTGTTATACA	RIME(2r3)
2r3*	RHS1	1	(2r2)RIME	TACTGTTATACA	ccctgg....RIME....aaaaaa	TACTGTTATACA	AACTACTGCATTATGA
2r6	RHS3	2	ATACGGTGTTGGATCAT	TTTTGCTTCATT	ccctgg....RIME....aaaaaa	TTTTGCTTCATT	CACTGCTTCACTTCCA
30P15.i3*	RHS3	2	ATGTGGTGTTGGATCAT	TTTTGCTTCATT	cctcaa....*ingi*....aaaaaa	TTTTGCTTCATT	RIME(30P15.r2)
30P15.r2*	RHS3	2	(30P15.i3)*ingi*	TTTTGCTTCATT	cctcaa....RIME....aaaaaa	TTTTGCTTCATT	CCCTGCTTCACTTCCA
26P8.i2*	RHS3	2	(26P8.i1)*ingi*	TTTTGCTTCATA	ccctgg....*ingi*....aaaaaa	TTTTGCTTCATA	*ingi*(26P8.i3)
26P8.i3*	RHS3	2	(26P8.i2)*ingi*	TTTTGCTTCATA	ccctgg....*ingi*....aaaaaa	TTTTGCTTCATA	*ingi*(26P8.i4)
45I2.r2	RHS2	3	AGTTGGTGTTGGCTCGT	TCCTGCTCCAAA	ccctgg....RIME....aaaaaa	TCCTGCTCCAAA	GGTTGCTGCATTACGA
1i1	RHS2	3	ATTTGGTGTTGGCTCGT	TTCTGCTCCAAA	ccctgg....*ingi*....aaaaaa	TTCTGCTCCAAA	GGTTGCTGCATTACGA
1r3	RHS5	4	ACTGGCCATCGGCTCAC	TTCTCCTTCACA	ccctgg....RIME....aaaaaa	TTCTCCTTCACA	AACTAAAGCATGATGC
1i2	RHS5	4	ACTGGCCATCGGCTCAC	TTCTCCTTCACA	ccctgg....*ingi*....aaaaaa	TTCTCCTTCACA	AACTAAAGCATGATGC
2i11		5	TTCTCGTGTTGGCTGCG	TCCAAATGCATT	ccctgg....*ingi*....aaaaat	TCCAAATGCATT	TTGCGATGTGATTTGT
28O9.r2		6	ATACTAGGTTGGTTGTT	TACTGCTACTGC	ccctgg....RIME....aaaaaa	TACTGCTACTGC	AGCCAGGGAATTTTAG
3A7.r1		7	CTTTCGGGTAGGATGTC	TGAAAAAAGATT	ccctgg....RIME....aaaaaa	TGAAAAAAGATT	TCTCTGCTGTCTTTCC
2r12		8	GTTTGGCATCTGCCGCT	GTtgTcTGTGTC	ccctgg....RIME....aaaaaa	GTctTtTGTGTC	TTCTTCTTTTCCTTTT
2r9		9	GCATTGCGCCGGTTGAA	ACATGCTGCAcA	ccctgg..half-RIME.acccgc	ACATGCTGCAgA	ACTTGGGGTGGTGCCT
2r10		9	GCATTGCGCCGGTTGAA	ACATGCTGCAcA	ccctgg..half-RIME.acccgc	ACATGCTGCAgA	ACTTGGGGTGGTGCCT
45I2.r3		9	GCATTGCGCCGGTTGAA	ACGTGCTGCAcA	ccctgg..half-RIME.acccgc	ACATGCTGCAgA	ACTTGGGGTGTTGCCT
1P6.r1		10	CGATGAGGGCGGCTGCG	ATGACCGCCACG	ccctgg....RIME....aaaaaa	ATGACCGTCACG	GTGTTTGCGGCCGCCA
1P6.r2		10	CGATGAGGGCGGCTGCG	ATGACCGCCACG	ccctgg....RIME....aaaaaa	ATGACCGCCACG	GTGTTTGCGGCCGCCA
1P6.r3		10	CGATGAGGGCGGCTGCG	ATGACCGCCACG	ccctgg....RIME....aaaaaa	ATGACCGCCACG	GTGTTTGCGGCCGCCA
3C6.i1		11	TGGTGGCGTGGGATTAA	TTCTCACAATGC	ccctgg....*ingi*....aaaaAA	TTCTCACAATGC	TACGCCATGACTGATG
12C12.i1		11	GGTCGGCGTGGGATTAA	TTCTCACAATGC	ccctgg....*ingi*....aaaaAA	TTCTCACAATGC	TACGCCATGACTGATG
29O9.r2		12	ACGTCCCGCTGCTGGAG	AtATATTGAATC	ccctgg....RIME....aaaaaa	AaATATTGAATC	ATTTCTCGCTTAAATT
45I2.r1		12	ACGTCCCACTGCTGGAG	ATATATTGAATC	ccctgg....RIME....aaaaaa	ATATATTGAATC	ATTTCTCGCTTAAATT
1I18.i3*		13	AAATGCGGAGGGTTGAG	ATCAAGTGAGTA	ccctgg....*ingi*....aaaaaa	ATCAAGTGAGTA	RIME(1I18.r1)
30O21.r1		14	GCCTCCCGCTGGGGAGA	ATATATAGAACA	ccctgg....RIME....aaaaaA	ATATATAGAACA	TGCAACTTCGGGCGGA
45I2.i5		15	TGATTGTGTTGGAAGAA	ACTAGCCAAATA	ccctgg....*ingi*....aaaaAA	ACTAGCCAAATA	AGTGCGGCTGCGGTCG
1i13		16	TGTTGCCGTGTGTTATG	AaTTAAGAAAGA	ccctgg....*ingi*....aaaaaa	AtTTAAGAAAGA	GTAGAAAGAAAGGATT

B

NAME	GENE	SEQ GROUP	5′	PUTATIVE DUPLICATED MOTIF	*ingi*/RIME	PUTATIVE DUPLICATED MOTIF	3′	GENE
1r5*	RHS1	1	(1i6)*ingi*	TTTTGTTATACA	ccctgg....RIME....aaaaaa	GTGAATGTACAA	CGAATCGCACA	
26P8.i4*	RHS3	2	(26P8.i3)*ingi*	TTTTGCTTCATA	ccctgg....*ingi*....aaaaaa	TATTGGTTCCCT	TACCAGCCCCA	
1i7	RHS5	4	ACTGGCCACCGGCTCAC	TTCTCCTTCACA	ccctgg....*ingi*....aaaaaa	TGAGCTGGGCAC	CGTCAGTTGTC	
26P8.r8		17	ATATGGCGTGGCCTCCC	ACTCTCCCATGC	ccctgg....RIME....aaaaaa	TTCTGCTCCAAA	GGTTGCTGCAT	RHS2
45I2.i04		18	AATGCGGTGGGTGACAT	ACAATGTTTGCA	ccctgg....*ingi*....aaaaaa	TACTGTTATACA	AACTACTGCAT	RHS1
1i14		19	GCATTGCGCCGTTTGGA	TCATGCTGCATA	ccctgg....*ingi*....aaaaaa	TTTAGCGGTATT	TTCTTGCATTT	
2i13		20	GCTTTCTGCTGGGTTGT	TGTTCTCGTAGA	cccagc....*ingi*....attaaa	TATATATATATA	TATACGTTAAT	
29K13.i2		21	CTTTCGTGTCGGGTGTT	GTAAGCTCGGCA	ccctgg....*ingi*....aaaaaa	TAGTTCCGCACG	GCTGGGCGCCT	
28O9.i1		22	CCTTTCTGTGGGGTCGC	TGAAAAATTATA	ccctgg....*ingi*....aaaaaa	TTAGGTTCTGGC	TCATCTATTGT	
29O9.r3		23	GGATGGTGGAGGAAGAA	TATAAATGTGCA	ccctgg....RIME....attaaa	TTAAATTAATTA	TTCTTACCAGA	
2N9.i4		24	TCCTGTGGTCGGGAGAA	TACAGCGATGTG	ccctgg....*ingi*....aaaaaa	GTCTATAGGCAA	ACCTCTACCCT	
1i16		25	TGGCTGTGTCGGTGTTG	TCATTCTGTTTC	ccctgg....*ingi*....aaaaaa	TCTTTCGAGGTT	GGTTGCTTGTC	
2i14		26	ACACGCCATCGGCGCAG	TACAACTACGGC	ccctgg....*ingi*....aaaaaa	TTGGTGTTGCTC	CATATATTCGT	
29K13.i1		26	ACACGCCATCGGCGCAG	TACAACTACGGC	ccctgg....*ingi*....aaaaaa	TATAATATGTTA	AATTTTTCAAA	
3H15.i4		27	CCTGCTGCCGGAGCGTT	GCTTCTGCTGCT	ccctgg....*ingi*....aaaaaa	GTGAGAAAGATG	GAAAAAACTGC	
29K13.i3		28	TTCTCGGAACAGGCAGA	GACAAGCCAAAC	acctgg....*ingi*....aaaaaa	TAATTCTTAGTC	TTTTTTTTCTG	
24M18.i2		29	CATGGCGTAGACTGTCT	TGAAGTCAGGAT	ccctgg....*ingi*....aaaaaa	TTAACTTTTCAT	GTGACATCCCG	
1I18.i4		30	CGAATGCCGTCTGTTCT	TTCTATCTGATA	ccctgg....*ingi*....aaaaaa	CGAACACACTCT	TCTTCGCACCC	

FIG. 2.—Comparison of the 5′-adjacent and 3′-adjacent sequences flanking the *ingi*/RIME retroelements identified by *T. brucei* chromosome sequencing. The retroelements flanked by a direct repeat are shown in (*A*), and those that are not flanked by a direct repeat are shown in (*B*). The alignment of all the selected sequences was based on the retroelement sequences (gray column headed "*ingi*/RIME") from which only the first and the last 6-bp, separated by the name of the retroelement (*ingi*/RIME) or "retroelement-like" (half-RIME), are shown. The potentially functional *ingi*s, which code for a full-length protein (1,657 aa) are indicated by white characters on a black background. Retroelements from the same BAC or chromosome that are labeled by an asterisk (*) in column "NAME" are arranged head-to-tail and separated by the duplicated flanking motif. In (*A*), the direct repeat flanking the retroelements (called "DUPLICATED MOTIF") is indicated by boldfaced and capital characters, and in (*B*), the equivalent sequences of the direct repeat–less retroelements are indicated as "PUTATIVE DUPLICATED MOTIF." For the last 11 sequences of (*A*), the extent of the duplicated motif is hard to discern because of the presence of a poly(dA) sequence that always precedes the downstream duplicated motif. Additional A residues that may belong to the motif are shown (boldfaced capitals). Lowercase characters in the duplicated motifs (*A*) correspond to nonconserved residues. The 5′ and 3′ sequences that flank the duplicated motifs are indicated, and, where known, the genes to which the flanking sequences correspond are identified (column called "GENE"). In some cases, the analyzed retroelements are preceded by further retroelement

Analysis of the Duplicated Motif Flanking the *Ingi*/RIME Retroelements

The insertion of non-LTR retrotransposons, including *ingi*s or RIMEs, generates a duplication of the target sequence to form a direct repeat of a few bases flanking the inserted retroelement. Among the 52 full-length *ingi* and RIME/half-RIME elements identified by chromosome sequencing, 34 (65%) are flanked by a short duplicated motif, shown in figure 2*A* (14 *ingi*s, 17 RIMEs, and three half-RIMEs). The size of the direct repeat sequence is clearly 12-bp for the first 23 retroelements shown. However, the presence of one or more A residues at the 5′ extremity of the upstream direct repeat in the last 11 sequences (named "SEQ 10" to "SEQ 16" in figure 2*A*) prevents a precise size determination of the duplicated motif, because the poly(dA) sequence always precedes the downstream direct repeat (except for the half-RIME sequences). In these particular cases, the size of the direct repeat varies from 11- and 14-bp but may also be 12-bp (fig. 2*A*).

Analysis of the *Ingi*/RIME 5′-Adjacent Sequences in the Genome Survey Sequence Database

To analyze the *ingi*/RIME insertion sites, we also took advantage of the *T. brucei* GSS database containing 90,000 end sequences (about 1.8-fold coverage of the nonminichromosomal *T. brucei* haploid genome). The strategy consists of searching for GSS sequences containing the 5′ flanking region of *ingi*/RIME elements by Blast analyses using the first 75-bp of the 250-bp RIME-A sequence (5′ probe), which is conserved in *ingi* and RIME elements (fig. 1). After discarding all the sequences containing less than 20 nucleotides of 5′-adjacent sequence, 315 GSS were selected and analyzed after removing the *ingi*/RIME sequences. All these sequences were compared and ordered into groups of related elements. To belong to the same group, two sequences should be at least 90% identical and have the *ingi*/RIME element inserted at exactly the same position. This analysis, summarized in figure 3, indicates that among these 315 GSS sequences, 70 sequences are unique (22% of the selected GSS sequences), whereas the other 245 sequences are divided into 70 groups containing between two and 29 identical or nearly identical sequences.

The second largest group containing 21 nearly identical sequences, is composed of tandemly arranged *ingi*/RIME sequences, revealing a 12-bp sequence inserted between the poly(dA) tail of the RIME-B sequence and the 5′ extremity of the adjacent RIME-A sequence (fig. 4).

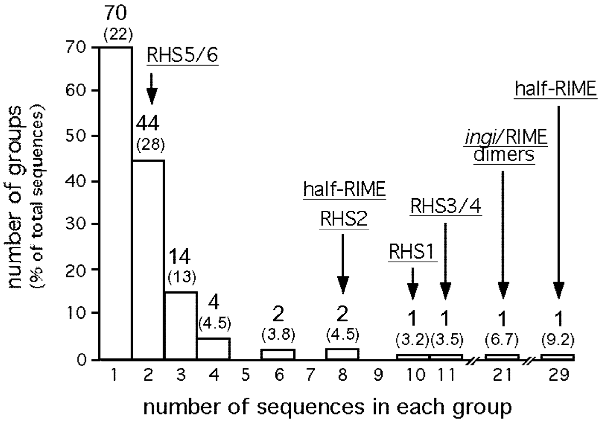


FIG. 3.—Grouping DNA sequences upstream of the *ingi*/RIME retroelements. A Blast search of the *T. brucei* GSS database (composed of about 90,000 single-pass sequences) with the 5′ probes, revealed 315 sequences containing at least 20-bp upstream of the *ingi*/RIME sequence. These sequences were aligned and *ingi*/RIME flanking sequences sharing at least 90% identity and starting at the same relative position were grouped. The resulting distribution of the number groups (Y-axis) against the number of sequences per group (X-axis) is shown. On top of each bar, the number of groups are shown with the percentage of the 315 sequences contained in each category in parentheses. The position of the groups containing the RHS sequences, some of the half-RIME sequences, and the *ingi*/RIME dimers are indicated. For example, column 8 shows that, among the 315 *ingi*/RIME 5′ flanking sequences, 16 sequences (4.5% of the total sequences) fall into two groups, and each group contains eight nearly identical sequences. One of these two groups contains the RHS2 sequence, and the other contains a half-RIME flanking sequence (RHS subfamilies were defined in Bringaud et al. [2002]).

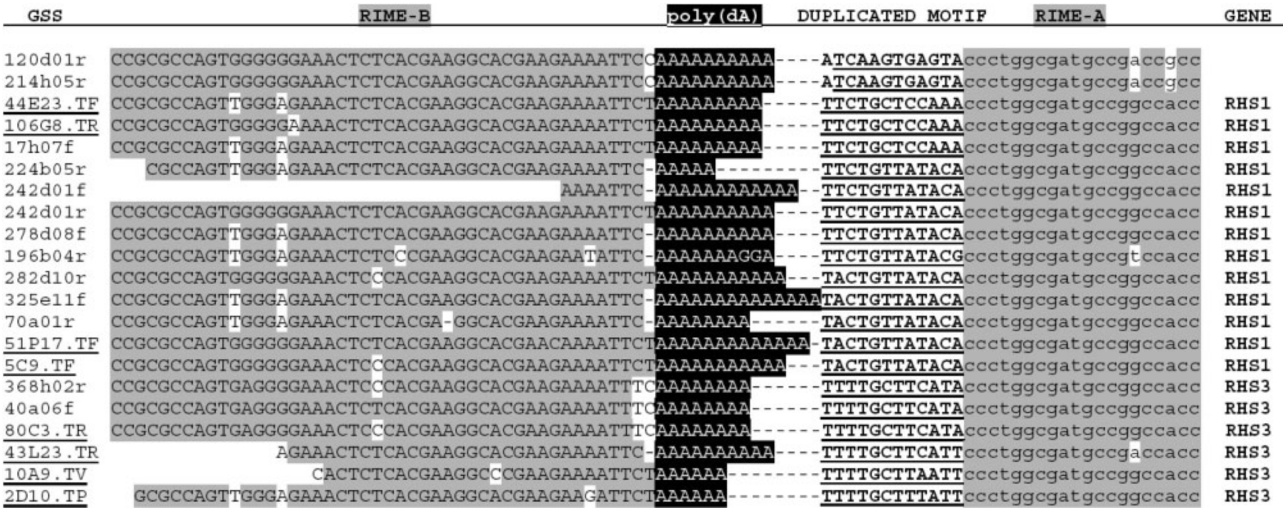
These 12-bp sequences are probably the result of multiple *ingi*/RIME insertions at the same site with, as consequence, multiple duplications of the target site, each flanked by retroelements (Bringaud et al. 2002). Interestingly, most of these 12-bp sequences (19 out of 21) are identical to the direct repeats flanking *ingi*/RIME retroelements inserted in *RHS* (pseudo)genes (figure 4 and see figure 5 in Bringaud et al. [2002]). In addition, 31 GSS sequences contain a unique retroelement sequence preceded by the 5′ extremity of a *RHS* pseudogene (fig. 3). This indicates that 15.9% of the retroelements analyzed in these 315 GSS sequences are inserted at the same relative position in the *RHS* (pseudo)genes, which confirms the presence of a hot spot for retroelement insertion in the *RHS* (pseudo)genes (Bringaud et al. 2002).

The *Ingi*/RIME Retroelements Are Preceded by a Conserved Motif

The hot spot for insertion in the *RHS* (pseudo)genes suggests that the *ingi*/RIME retroelements resemble site-

←

sequences (indicated by the words "*ingi*" or "RIME," shaded in gray and identified in parentheses). Residues within the duplicated motif and 5′ flanking sequences that match the consensus are indicated with white characters on a black background. Sequences are grouped by the upstream retroelement flanking sequences and numbered 1 to 29 in column "SEQ GROUP." Underlined names correspond to retroelements identified in the fully sequenced *Chr*Ia (the Wellcome Trust Sanger Institute—http://www.sanger.ac.uk/Projects/T_brucei/) and *Chr*II (TIGR—http://www.tigr.org/tdb/e2k1/tba1/); the first number indicates chromosome 1 versus 2, the letter and the last number indicate the retroelement (i or r for *ingi* or RIME, respectively) and its order of appearance on the chromosome. The other names (not underlined) correspond to retroelements identified in BACs of chromosomes whose sequencing is not completed (TIGR); the name of the BAC is followed by a dot, the nature of the retroelement (i or r), and the position on the BAC sequence.



```
   GSS                      RIME-B                       poly(dA)  DUPLICATED MOTIF  RIME-A          GENE

120d01r   CCGCGCCAGTGGGGGGAAACTCTCACGAAGGCACGAAGAAAATTCCAAAAAAAAAA----ATCAAGTGAGTAccctggcgatgccgaccgcc
214h05r   CCGCGCCAGTGGGGGGAAACTCTCACGAAGGCACGAAGAAAATTCCAAAAAAAAAA----ATCAAGTGAGTAccctggcgatgccgaccgcc
44E23.TF  CCGCGCCAGTTGGGAGAAACTCTCACGAAGGCACGAAGAAAATTCTAAAAAAAAA-----TTCTGCTCCAAAccctggcgatgccggccacc   RHS1
106G8.TR  CCGCGCCAGTGGGGGAAAACTCTCACGAAGGCACGAAGAAAATTCTAAAAAAAAA-----TTCTGCTCCAAAccctggcgatgccggccacc   RHS1
17h07f    CCGCGCCAGTTGGGAGAAACTCTCACGAAGGCACGAAGAAAATTCTAAAAAAAAA-----TTCTGCTCCAAAccctggcgatgccggccacc   RHS1
224b05r      CGCCAGTTGGGAGAAACTCTCACGAAGGCACGAAGAAAATTC-AAAAA---------TTCTGTTATACAccctggcgatgccggccacc   RHS1
242d01f                                         AAAATTC-AAAAAAAAAAAA--TTCTGTTATACAccctggcgatgccggccacc   RHS1
242d01r   CCGCGCCAGTGGGGGGAAACTCTCACGAAGGCACGAAGAAAATTCTAAAAAAAAAA----TTCTGTTATACAccctggcgatgccggccacc   RHS1
278d08f   CCGCGCCAGTTGGGAGAAACTCTCACGAAGGCACGAAGAAAATTC-AAAAAAAAAA----TTCTGTTATACAccctggcgatgccggccacc   RHS1
196b04r   CCGCGCCAGTTGGGAGAAACTCTCCCGAAGGCACGAAGAATATTC-AAAAAAAGGA----TTCTGTTATACGccctggcgatgccgtccacc   RHS1
282d10r   CCGCGCCAGTGGGGGGAAACTCCCACGAAGGCACGAAGAAAATTCTAAAAAAAAAAA---TACTGTTATACAccctggcgatgccggccacc   RHS1
325e11f   CCGCGCCAGTTGGGAGAAACTCTCACGAAGGCACGAAGAAAATTC-AAAAAAAAAAAAAATACTGTTATACAccctggcgatgccggccacc   RHS1
70a01r    CCGCGCCAGTTGGGAGAAACTCTCACGA-GGCACGAAGAAAATTC-AAAAAAAA------TACTGTTATACAccctggcgatgccggccacc   RHS1
51P17.TF  CCGCGCCAGTGGGGGGAAACTCTCACGAAGGCACGAACAAAATTCTAAAAAAAAAAAAA-TACTGTTATACAccctggcgatgccggccacc   RHS1
5C9.TF    CCGCGCCAGTGGGGGGAAACTCCCACGAAGGCACGAAGAAAATTCTAAAAAAAAAAA---TACTGTTATACAccctggcgatgccggccacc   RHS1
368h02r   CCGCGCCAGTGAGGGGAAACTCCCACGAAGGCACGAAGAAAATTTCAAAAAAAA------TTTTGCTTCATAccctggcgatgccggccacc   RHS3
40a06f    CCGCGCCAGTGAGGGGAAACTCTCACGAAGGCACGAAGAAAATTTCAAAAAAAA------TTTTGCTTCATAccctggcgatgccggccacc   RHS3
80C3.TR   CCGCGCCAGTGAGGGGAAACTCCCACGAAGGCACGAAGAAAATTTCAAAAAAAA------TTTTGCTTCATAccctggcgatgccggccacc   RHS3
43L23.TR                AGAAACTCTCACGAAGGCACGAAGAAAATTC-AAAAAAAAAA----TTTTGCTTCATTccctggcgatgccgaccacc   RHS3
10A9.TV                    CACTCTCACGAAGGCCCGAAGAAAATTCTAAAAAA--------TTTTGCTTAATTccctggcgatgccggccacc   RHS3
2D10.TP     GCGCCAGTTGGGAGAAACTCTCACGAAGGCACGAAGAAGATTCTAAAAAA--------TTTTGCTTTATTccctggcgatgccggccacc   RHS3
```

FIG. 4.—Comparison of all the GSS sequences containing both the 5′ and the 3′ extremities of the *ingi*/RIME retroelements separated by a short sequence. Gaps (-) were introduced to maximize the alignment of the 21 sequences, which correspond to the GSS sequences contained in column 21 of figure 3. The name of the GSS sequence determined at TIGR (underlined) or the Wellcome Trust Sanger Institute is indicated on the left side. The identical residues in the RIME-A (lowercase) and RIME-B (capital letters) sequences are shaded in gray, and the poly(dA) tail present at the 3′ extremity of the *ingi*/RIME retroelements is indicated by white characters on a black background. The sequence located between the poly(dA) tail and the RIME-A sequence, which corresponds to the duplicated motif flanking *ingi*/RIME, is underlined and bold. Where known, the name of the *RHS* gene subfamily from which this 12-bp duplicated motif is derived, is indicated on the right (RHS subfamilies were defined in Bringaud et al. [2002]).

specific non-LTR retrotransposons. However, they are not restricted to a single target site. According to the current model for retrotransposition of non-LTR retrotransposons, the process is initiated by the element-encoded endonuclease, which performs a sequence-specific or nonspecific single-strand cleavage. To determine whether conserved residues are present in the vicinity of the *ingi*/RIME insertion sites, we compared the regions flanking the full-length retroelements identified by chromosome sequencing. Only the 34 retrotransposons flanked by a direct repeat were analyzed (fig. 2*A*) because the first single-strand cleavage of the endonuclease corresponds to the 5′ extremity of the flanking duplicated motif. Based on the nature of the retroelement flanking sequences, these 34 sequences are subdivided into 16 different sequence groups (fig. 2*A*). The patterns of bases in regions flanking the retroelements are presented in figure 5*A*. Sequence conservation in the sequences adjacent to the downstream duplicated motif is limited. Indeed, among the first 20 nucleotides, only T residues at two positions (+04 and +12) are found in more than half (53%) of the sequences. For the duplicated direct repeat, greater conservation is observed. At least four out of 12 residues constituting the flanking direct repeat are moderately conserved. Upstream of the 12-bp motif shows considerable sequence conservation. For instance, in the region –15 and –34 upstream of the retroelement, specific bases occurs in three positions, with a frequency of 56% to 69%, and at six positions, with a frequency of 75% to 94%. Although the number of sequences analyzed is relatively small, this suggests that a consensus pattern is present within the first 34 bp upstream of the retroelements. Interestingly, the conserved residues are found upstream of both *ingi* and RIME retroelements (fig. 2*A*).

To determine whether the consensus pattern upstream of the retroelements is statistically significant, this analysis should be performed on a larger set of sequences. The GSS containing sequences upstream of RIME-A are good candidates for this analysis because this set consists of 294 sequences (the 21 sequences corresponding to *ingi*/RIME dimers that do not contain the region preceding the direct repeat were not retained) (figs. 3 and 4). Un-

⟶

FIG. 5.—Base frequencies at different positions of the flanking direct repeat and adjacent regions of the *ingi*/RIME retrotransposons identified by chromosome sequencing (*A*) and GSS database analysis (*B*) and (C). In (A), the frequencies have been analyzed in the 12-bp direct repeat and the 5′-adjacent and 3′-adjacent sequences (27- and 16-bp, respectively) of 32 full-length *ingi*s and RIMEs flanked by a direct repeat (fig. 2*A*). The *ingi*/RIME and downstream 12-bp direct repeats have not been analyzed. In (B) and (C), the region upstream of *ingi*/RIME retroelements identified in 294 GSS sequences (*B*) or in a smaller set of 139 GSS sequences, using only one sequence per group of nearly identical GSS sequences as defined in figure 3 (*C*), have been compared. The first column (called "pos") indicates the nucleotide position: for the 5′ flanking region of both panels (from position −01 to −40) the numbering starts before the 5′ extremity of the retroelement; for the 3′ flanking region of (*A*) (from position +01 to +16), the numbering starts after the 3′ extremity of the downstream 12-bp direct repeat; and for the *ingi*/RIME sequences of (*B*) and (*C*) (from position +01 to +20) the numbering starts from the 5′ extremity of the retroelements. The values in columns "T," "C," "A," and "G" represent the percentage of the T, C, A, and G residues, respectively, at individual positions. Values superior to 50% are indicated: 50% to 60% (underlined), 60% to 70% (underlined and boldfaced), 70% to 80% (underlined and gray shaded), 80% to 90% (underlined, boldfaced and gray shaded), and 90% to 100% (white characters on a black background). The last column (named "cons") shows the conserved residues. An arrow in the right margin indicates the position of the first single-strand cleavage.

A

Pos	T	C	A	G	cons	
-40	13	25	31	31		
-39	37	25	13	25		
-38	37	13	31	19		
-37	19	13	19	49		
-36	0	38	31	31		
-35	31	25	31	13		
-34	0	0	88	12	A	
-33	12	19	44	25		
-32	19	19	25	37		
-31	69	6	6	19	T	
-30	6	49	14	31		
-29	25	13	37	25		
-28	38	25	6	31		
-27	38	12	37	13		
-26	81	13	0	6	T	
-25	26	25	0	49		
-24	0	38	13	49		
-23	26	49	0	25		
-22	0	0	12	88	G	
-21	69	19	6	6	T	
-20	44	25	6	25		
-19	12	0	0	88	G	
-18	0	6	0	94	G	
-17	31	31	31	7		
-16	75	6	6	13	T	
-15	6	25	13	56	G	
-14	19	25	44	22		
-13	31	13	25	31		
-12	50	0	43	7	T	
-11	63	25	6	6	T	12
-10	25	44	25	6		bp
-09	63	0	31	6	T	
-08	6	19	37	38		R
-07	19	49	25	7		E
-06	63	13	12	12	T	P
-05	13	12	37	38		E
-04	12	44	38	6		A
-03	13	0	75	12	A	T
-02	44	31	6	19		
-01	19	25	49	7		

←

***ingi* / RIME**

12 bp repeat

Pos	T	C	A	G	cons
+01	33	7	40	20	
+02	33	13	27	27	
+03	33	46	7	14	
+04	53	20	7	20	T
+05	20	20	33	27	
+06	13	40	20	27	
+07	40	13	20	27	
+08	33	13	7	47	
+09	27	40	13	20	
+10	20	0	47	33	
+11	33	40	13	14	
+12	53	13	7	27	T
+13	33	7	27	33	
+14	46	27	20	7	
+15	27	33	7	33	
+16	33	13	33	20	

B

Pos	T	C	A	G	cons	
-40	30	24	28	18		
-39	29	32	18	21		
-38	39	23	21	17		
-37	30	14	28	28		
-36	25	27	21	27		
-35	25	22	31	22		
-34	17	10	56	17	A	
-33	26	22	38	14		
-32	21	14	23	42		
-31	39	21	25	15		
-30	31	32	13	24		
-29	22	23	24	31		
-28	44	26	15	15		
-27	39	13	36	12		
-26	64	19	11	6	T	
-25	41	15	5	38		
-24	13	28	13	46		
-23	29	39	15	17		
-22	13	6	13	68	G	
-21	55	23	8	14	T	
-20	35	29	11	25		
-19	16	4	9	71	G	
-18	17	6	7	70	G	
-17	39	23	22	16		
-16	64	7	14	15	T	
-15	20	23	19	38		
-14	24	22	35	19		
-13	34	21	27	18		
-12	42	2	45	11		
-11	47	22	22	9		12
-10	27	32	24	17		bp
-09	62	10	21	7	T	
-08	21	14	26	39		R
-07	24	44	16	16		E
-06	68	10	12	12	T	P
-05	23	21	21	15		E
-04	35	30	27	8		A
-03	18	4	53	25	A	T
-02	32	42	14	12		
-01	27	17	46	10		
+01	2	96	2	0	C	
+02	1	98	1	0	C	
+03	1	96	2	1	C	
+04	94	4	1	1	T	*i*
+05	0	2	5	93	G	*n*
+06	1	3	5	91	G	*g*
+07	20	78	2	0	C	*i*
+08	3	0	4	93	G	/
+09	0	1	96	3	A	R
+10	92	8	0	0	T	I
+11	3	0	2	95	G	M
+12	1	98	0	1	C	E
+13	13	86	1	0	C	
+14	3	0	2	95	G	
+15	1	0	0	99	G	
+16	3	97	0	0	C	
+17	6	93	1	0	C	
+18	2	4	93	1	A	
+19	1	99	0	0	C	
+20	1	83	16	0	C	

←

C

Pos	T	C	A	G	cons	
-40	34	18	26	22		
-39	28	26	19	27		
-38	26	26	24	24		
-37	37	12	23	28		
-36	29	22	25	24		
-35	27	15	35	23		
-34	17	10	50	23	A	
-33	23	25	34	18		
-32	23	15	18	44		
-31	43	19	20	18		
-30	34	30	16	20		
-29	23	23	31	23		
-28	45	21	16	18		
-27	41	17	28	14		
-26	56	24	12	8	T	
-25	35	29	9	37		
-24	14	32	14	40		
-23	34	32	14	20		
-22	11	6	18	65	G	
-21	59	18	9	14	T	
-20	35	26	12	27		
-19	18	4	10	68	G	
-18	17	6	12	65	G	
-17	32	25	22	21		
-16	57	8	19	16	T	
-15	25	19	22	34		
-14	31	20	28	21		
-13	37	21	22	20		
-12	43	4	41	12		
-11	51	14	23	11	T	12
-10	28	34	23	15		bp
-09	52	11	26	11	T	
-08	24	20	29	27		R
-07	30	34	18	18		E
-06	59	14	14	13	T	P
-05	25	17	23	35		E
-04	38	18	33	11		A
-03	25	6	37	32		T
-02	32	38	14	18		
-01	34	23	34	10		
+01	2	96	2	0	C	
+02	1	98	1	0	C	
+03	1	96	2	1	C	
+04	94	4	1	1	T	*i*
+05	1	1	5	93	G	*n*
+06	2	2	4	92	G	*g*
+07	11	87	2	0	C	*i*
+08	1	0	3	96	G	/
+09	0	1	97	2	A	R
+10	95	4	1	0	T	I
+11	6	1	0	94	G	M
+12	1	97	1	1	C	E
+13	6	93	1	0	C	
+14	4	1	1	94	G	
+15	2	1	1	96	G	
+16	3	96	1	0	C	
+17	4	95	1	0	C	
+18	1	1	97	1	A	
+19	1	98	0	1	C	
+20	2	90	8	0	C	

←

fortunately, the site of the first single-strand cleavage is unknown for all these sequences because the sequence downstream of the retroelements encoded by these GSS sequences is unknown. However, we have shown that most, if not all, of the direct repeats flanking the *ingi*/RIME retroelements are the same size (i.e., 12-bp) (see fig. 2*A*). Consequently, if we assume that cleavage occurs 12-bp upstream of the *ingi*/RIME sequence, the analysis of the base occurrence in the vicinity of the first single-strand nick can be performed on this set of 294 GSS sequences. Because this set of GSS sequences contains a lot of repeated sequences (including 31 RHS sequences), which may skew the statistical analysis, we also considered a smaller set of 139 GSS sequences, which contains only one sequence per group of nearly identical sequences (fig. 3). The base occurrence analysis of these two sets of 294 sequences (fig. 5*B*) and 139 sequences (fig. 5*C*), confirms the presence of a consensus pattern upstream of the retroelements. This conclusion is confirmed by the K-S test performed on the set of 139 sequences (fig. 6). It is noteworthy that the conserved G nucleotides located upstream (positions –26, –22, –19, and –18) and within (positions +5, +6, +8, +11, +14, and +15) the retroelement sequence, show comparable χ^2 values, which further confirms the high level of conservation observed for residues located upstream of the *ingi*/RIME insertion sites. In conclusion, these statistical analyses demonstrate the presence of a conserved sequence (–34 **A**xxxxxxx**Ttg**x**TG**x**GG**x**T**xxx↑**tT**x**T**x**T** –6) upstream of the *ingi*/RIME retroelements, with an 11-bp core consensus sequence (underlined residues) located 4- to 14-bp upstream of the first single-strand cleavage (arrow).

Discussion

The *ingi* and RIME non-LTR retrotransposons, which were believed to be randomly distributed, are the most abundant mobile elements characterized so far in the genome of *Trypanosoma brucei* (Hasan, Turner, and Cordingley 1984; Kimmel, Ole-MoiYoi, and Young 1987; Murphy et al. 1987). Our analysis of 81 *ingi*/RIME retroelements and approximately 800 GSS obtained from the *T. brucei* genome sequencing project has revealed that (1) the size of the direct repeat flanking the *ingi*/RIME retrotransposons is conserved at 12-bp; (2) these retroelements are not randomly distributed in the genome; and (3) they are preceded by a highly conserved consensus pattern (–34 **A**xxxxxxx**Ttg**x**TG**x**GG**x**T**xxx↑**tT**x**T**x**T** –6), with a core consensus sequence located 4- to 14-bp upstream of the first single-strand nick.

The presence of a consensus pattern in the vicinity of the *ingi*/RIME retroelement insertion sites suggests that retrotransposition is mediated by protein binding to a conserved motif. Because the current model for retrotransposition predicts that the first step is mediated by a retroelement-encoded endonuclease, which determines the sequence specificity of site-specific retroelements (Feng, Schumann, and Boeke 1998; Yang, Malik, and Eickbush 1999; Christensen, Pont-Kingdon, and Carroll 2000; Anzai, Takahashi, and Fujiwara 2001), it is tempting to assume that the observed consensus is the DNA binding site of the *ingi*-encoded endonuclease. The *T. brucei ingi* belong to the group of non-LTR retrotransposons, which encode an apurinic-apyrimidinic (AP)-like endonuclease domain related to the *Escherichia coli* exonuclease III (Feng et al. 1996; Olivares, Alonso, and Lopez 1997; Malik, Burke, and Eickbush 1999). The AP-endonucleases recognize modified purine and pyrimidine residues in the DNA, as observed for the AP-like endonuclease domain encoded by the *Trypanosoma cruzi* L1Tc retrotransposon (Olivares, Alonso, and Lopez 1997). However, other analyzed AP-like endonuclease domains from non-LTR retroelements present a strong bias for insertion (human L1 [Feng et al. 1996; Cost and Boeke 1998]) or an absolute target site specificity (TRAS1 [Anzai, Takahashi, and Fujiwara 2001] and Tx1L [Christensen, Pont-Kingdon, and Carroll 2000]) not related to apurinic-apyrimidinic sequences. This indicates that the retroelement-encoded AP-like endonucleases have a wide range of site specificity, which is in agreement with the relative site specificity proposed for the *ingi*-encoded AP-like endonuclease. However, we cannot rule out that another domain of the long *ingi*-encoded protein (1,657 amino acids) is responsible for this relative site specificity. Indeed, McClure, Donaldson, and Corro (2002) recently found that the C-terminal half of the protein encoded by the *ingi* element contains two additional endonuclease signatures related to retrovirus encoded integrase. In addition, this part of the *ingi*-encoded protein contains a large DNA-binding domain with five zinc-binding motifs, which may play a role in recognition of the target site (Pays and Murphy 1987). Functional expression of the different DNA-binding and endonuclease-like domains encoded by the *T. brucei ingi* will help to confirm its relative site specificity and to help to characterize any domain that are involved in target-site recognition.

The *T. brucei* chromosome sequencing project has revealed 34 *ingi*/RIME retroelements flanked by a direct repeat to date. For 23 of them, the size of the direct repeat is 12-bp. The size of the duplicated motif flanking the 11 other retroelements could also be 12-bp. However, the exact size cannot be determined because of the presence of A residues at the 5′ extremity of the direct repeats located upstream of the elements (fig. 2*A*). As far as we know, the size conservation of the flanking duplicated motif is unique to trypanosome retroelements. Indeed, all of the other non–site-specific non-LTR retrotransposons characterized so far have polymorphic flanking direct repeats, as exemplified by human L1, Alu, and ID elements, whose sizes range between 4- and 26-bp (Jurka 1997). The size of the flanking direct repeat primarily depends on the position of the second-strand cleavage. The mechanism of the second-strand cleavage at the downstream site is poorly understood. However, it is commonly accepted that the element-encoded endonuclease is responsible for the first and second single-strand nick of the target DNA. Thus, it is tempting to propose that the conservation of the size of the direct repeat, resulting from the retrotransposition of the *T. brucei* retroelements, occurs because of mechanistic properties of the *ingi*-encoded AP-like endonuclease.

About one-third of the full-length *ingi* and RIME/half-RIME elements identified by chromosome sequencing (18 out of 52 elements) are not flanked by a direct repeat

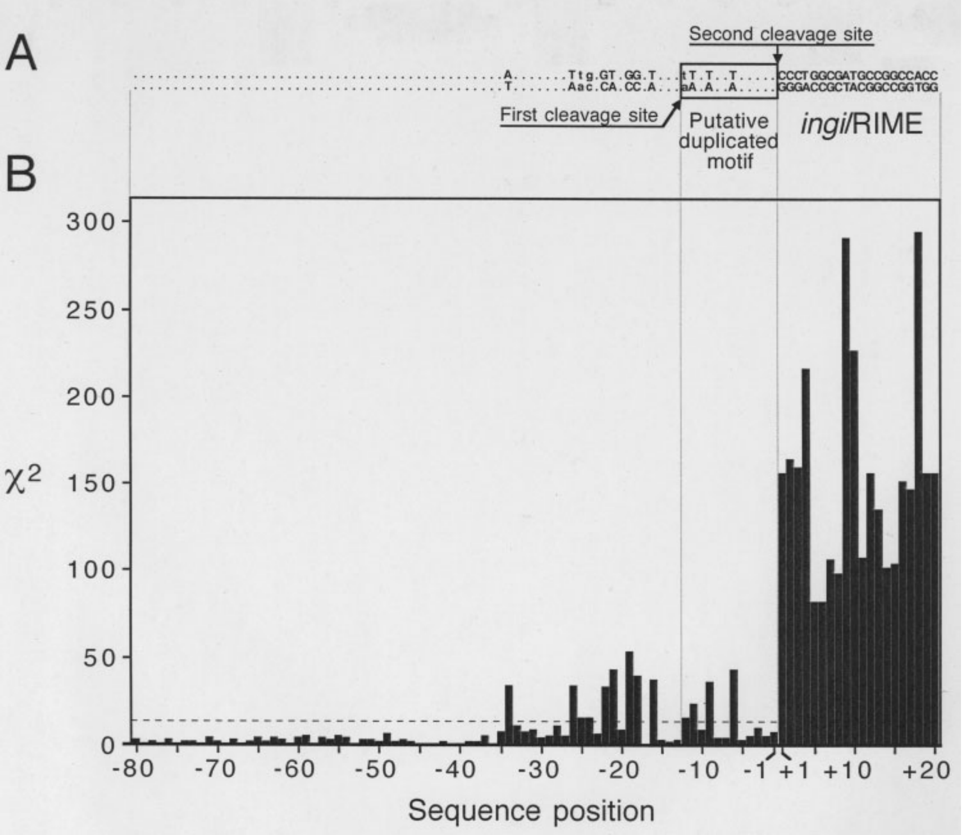


FIG. 6.—The χ^2 values for individual positions of the putative duplicated motif (flanking direct repeat) and adjacent regions. (*A*) Nucleotide sequence of residues presenting a nonrandom distribution, as determined in (*B*). The putative 12-bp duplicated motif preceding the retroelement sequence (called "*ingi*/RIME") is boxed, and the position of the first and second cleavages, probably performed by an *ingi*-encoded endonuclease, are indicated. (*B*) The χ^2 values were calculated as described in *Materials and Methods* from the set of 139 GSS sequences analyzed in figure 5*C*. The base composition upstream from position –40 was used to determine the background base distribution (the same result was obtained using the chromosome I sequence as reference) χ^2 values above the broken horizontal line (13.8) correspond to significance levels of $P < 0.001$ for two degrees of freedom. The discrete χ^2 values were represented as vertical bars. Positions –34, –26, –22, –21, –19, –18, –16, –11, –9, –6, and +1 to +20 have very high χ^2 values, indicating a nonrandom distribution of nucleotides at these positions. These nucleotides are indicated by capital letters in (*A*). The three nucleotide positions (–25, –24, and –12), which present χ^2 values close to 13.8 ($P < 0.001$ for two degrees of freedom), are indicated by small case characters in (*A*).

(fig. 2*B*). Two lines of evidence suggest that the loss of the duplicated flanking motif may be the consequence of homologous recombination between retroelements, which would generate chimeric retrotransposons flanked by unrelated sequences. First, one flanking extremity in five out of these 18 retroelements corresponds to a *RHS* pseudogene, whereas the sequence of the other flanking extremity is not related to the *RHS* multigene family (first five sequences in fig. 2*B*). Second, among the 29 full-length *ingi*s, 52% (15 retroelements) are not flanked by a short duplicated motif, compared with only 15% of the full-length RIMEs (3 out of 20). Therefore the loss of the duplicated flanking motif is about three times more common in *ingi*s (5.25 kb), as compared with RIMEs (0.5 kb) (fig. 2), commensurate with the expected increase of homologous recombination frequency caused by increased size of the homologous sequences. However, we cannot rule out the possibility that endonuclease-independent retrotransposition may generate *ingi*/RIME retroelements lacking the flanking direct repeats. Indeed, endonuclease-independent retrotransposition of L1 elements, recently observed in mammalian cells to repair double-stranded DNA breaks, is not associated with the duplication of the target DNA (Morrish et al. 2002). Interestingly, some of the direct repeat–less *ingi*/RIME elements do not contain the consensus pattern (fig. 2*B*) proposed to be the endonuclease-binding site, suggesting that retrotransposition of these few elements was an endonuclease-independent process.

In conclusion, the autonomous *ingi* and nonautonomous RIME *T. brucei* non-LTR retrotransposons present some characteristics that are not (or rarely) observed in other non-site-specific elements encoding an AP-like endonuclease domain: (1) very few 5′ truncations, (2) conservation of the direct repeat size (12-bp), (3) a strong consensus pattern 4- to 14-bp upstream of the direct repeat, and (4) a tendency to form head-to-tail retroelement clusters. These features are probably the consequence of enzymatic activities encoded by the *ingi* element, which are involved in retrotransposition of *ingi*s and probably RIMEs. Functional analyses of the *ingi*-encoded enzymatic activities, such as the AP-like endonuclease and the reverse transcriptase, will help to explain the molecular mechanisms leading to these particular features observed in the *T. brucei* retroelements described here.

Acknowledgments

We are grateful to M. C. Lopez for critical reading of the manuscript. This work was supported by the CNRS, the Conseil Régional d'Aquitaine, the GDR/CNRS-Parasitologie, and the Ministère de l'Education Nationale de la Recherche et de la Technologie.

Literature Cited

Aksoy, S. 1991. Site-specific retrotransposons of the trypanosomatid protozoa. Parasitol. Today **7**:281–285.

Aksoy, S., T. M. Lalor, J. Martin, L. H. Van der Ploeg, and F. F. Richards. 1987. Multiple copies of a retroposon interrupt spliced leader RNA genes in the African trypanosome, *Trypanosoma gambiense*. EMBO J. **6**:3819–3826.

Anzai, T., H. Takahashi, and H. Fujiwara. 2001. Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)(n) by endonuclease of non-long terminal repeat retrotransposon TRAS1. Mol. Cell. Biol. **21**:100–108.

Bringaud, F., N. Biteau, S. E. Melville, S. Hez, N. M. El-Sayed, V. Leech, M. Berriman, N. Hall, J. E. Donelson, and T. Baltz. 2002. A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei*. Eukaryotic Cell **1**:137–151.

Capy, P., C. Bazin, D. Higuet, and T. Langin. 1998. Dynamics and evolution of transposable elements. Landes Bioscience, Austin, Tex.

Christensen, S., G. Pont-Kingdon, and D. Carroll. 2000. Target specificity of the endonuclease from the *Xenopus laevis* non-long terminal repeat retrotransposon, Tx1L. Mol. Cell. Biol. **20**:1219–1226.

Cost, G. J., and J. D. Boeke. 1998. Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry **37**:18081–18093.

Craig, N. L. 1997. Target site selection in transposition. Annu. Rev. Biochem. **66**:437–474.

El-Sayed, N. M. A., G. Ghedin, J. Song et al. (41 co-authors). 2003. The sequence and analysis of *Trypanosoma brucei* chromosome II. Nucleic Acids Res. **31**:4856–4863.

El-Sayed, N. M., P. Hegde, J. Quackenbush, S. E. Melville, and J. E. Donelson. 2000. The African trypanosome genome. Int. J. Parasitol. **30**:329–345.

Feng, Q., G. Schumann, and J. D. Boeke. 1998. Retrotransposon R1Bm endonuclease cleaves the target sequence. Proc. Natl. Acad. Sci. USA **95**:2083–2088.

Feng, Q., J. V. Moran, H. H. Kazazian, and J. D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell **87**:905–916.

Hall, N., M. Berriman, N. J. Lennard, et al. (40 co-authors). 2003. The DNA sequence of chromosome I of an African trypanosome: gene content, chromosome organisation, recombination and polymorphism. Nucleic Acids Res. **31**:4864–4873.

Hasan, G., M. J. Turner, and J. S. Cordingley. 1984. Complete nucleotide sequence of an unusual mobile element from *Trypanosoma brucei*. Cell **37**:333–341.

Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA **94**:1872–1877.

Kimmel, B. E., O. K. Ole-MoiYoi, and J. R. Young. 1987. *Ingi*, a 5.2-kb dispersed sequence element from *Trypanosoma brucei* that carries half of a smaller mobile element at either end and has homology with mammalian LINEs. Mol. Cell. Biol. **7**:1465–1475.

Luan, D. D., M. H. Korman, J. L. Jakubczak, and T. H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell **72**:595–605.

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.

McClure, M. A., E. Donaldson, and S. Corro. 2002. Potential multiple endonuclease functions and a ribonuclease H encoded in retroposon genomes. Virology **296**:147–158.

Morrish, T. A., N. Gilbert, J. S. Myers, B. J. Vincent, T. D. Stamato, G. E. Taccioli, M. A. Batzer, and J. V. Moran. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat. Genet. **31**:159–165.

Murphy, N. B., A. Pays, P. Tebabi, H. Coquelet, M. Guyaux, M. Steinert, and E. Pays. 1987. *Trypanosoma brucei* repeated element with unusual structural and transcriptional properties. J. Mol. Biol. **195**:855–871.

Olivares, M., C. Alonso, and M. C. Lopez. 1997. The open reading frame 1 of the L1Tc retrotransposon of *Trypanosoma cruzi* codes for a protein with apurinic-apyrimidinic nuclease activity. J. Biol. Chem. **272**:25224–25228.

Pardue, M. L., and R. J. DeBerardinis. 2002. Telomeres and transposable elements. Pp. 870–887 *in* N. L. Craig, R. Craigie, M. Gellert, and A. M. Lambowitz, eds. Mobile DNA II. ASM Press, Washington.

Pays, E., and N. B. Murphy. 1987. DNA-binding fingers encoded by a trypanosome retroposon. J. Mol. Biol. **197**: 147–148.

Siegel, S., and N. J. Castellan Jr. 1988. Nonparametric statistics for the behavioural sciences. Pp. 144–151 *in* S. Siegel, S. Sidney, and N. J. Castellan. The Kolmogorov-Smirnov two-sample test. McGraw Hill Education, ISE Editions.

Yang, J., H. S. Malik, and T. H. Eickbush. 1999. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc. Natl. Acad. Sci. USA **96**:7847–7852.

Pierre Capy, Associate Editor

Accepted October 22, 2003