“BABES-BOLYAI” UNIVERSITY

FACULTY OF BIOLOGY AND GEOLOGY

DOCTORAL THESIS

(Summary)

Alu Transduction and Premature Polyadenylation – Insights into the Biology of SVA retrotransposons

Scientific adviser:

Prof. Dr. POPESCU Octavian

PhD student:

CHIRA Sergiu

CLUJ-NAPOCA

2011

Contents

List of abbreviations………………………………………………………………..3

Introduction…………………………………………………………………………4

Objectives……………………………………………………………………………8

Design of two vector sets for investigating the role of Alu transductions in SVA mobilization………………………………………………………………………………9

1. Construction of the H10_3 vector set………………………………………..9

2. Construction of the H10_1 vector set………………………………………11

Characterization of 3' truncated SVA elements…………………………………...16

1. The human genome reference sequence harbors 98 3' truncated SVA elements…………………………………………………………………………16

2. The SINE-R region supports the use of a broad range of alternative polyadenylation sites…………………………………………………………..19

2.1 Alternative polyadenylation sites of SVA elements in the human genome reference sequence……………………………………………...19

2.2 Poly(A) sites used for polyadenylation of 3' truncated SVA transcripts expressed in the human testis…………………………………………23

2.2.1 3' RACE analysis…………………………………………………..23

2.2.2 3' intact SVA transcripts' origin and subfamily affiliation………..25

2.2.3 Expression, characteristics, source elements and poly(A) signal utilization of 3' truncated SVA transcripts……………………………26

Conclusions………………………………………………………………………...31

References…………………………………………………………………………32

Key words: SVA retrotransposons, Alu transduction, retrotransposition reporter vectors, 3' truncated SVA elements, premature polyadenylation, SVA transcripts, human testis

List of abbreviations

A	adenine
BAC	bacterial artificial chromosome
bp	base pairs
C	cytosine
cDNA	complementary DNA
DNA	deoxyribonucleic acid
G	guanine
HERV	human endogenous retrovirus
kb	kilo bases
PCR	polymerase chain reaction
RNA	ribonucleic acid
SINE	short interspersed nuclear element
T	thymine
TPRT	target primed reverse transcription
TSD	target site duplication
U	uracil
VNTR	variable number tandem repeat

Introduction

Transposable elements, discovered in 1940, are also known as “jumping genes”, because of their capacity of moving from one genomic site to another. They represent up to 45 % of the human genome. Based on their mechanism of transposition transposable elements are divided in DNA transposons and retrotransposons (Cordaux and Batzer, 2009). The retrotransposon class of mobile elements is the most represented, comprising up to 42 % of our genome. They mobilize by a “copy and paste” mechanism using an RNA intermediate. Elements that encode factors necessary for their mobilization are called autonomous, mostly represented in our genome by the non-LTR retrotransposon L1. The non-autonomous retrotransposons, which include processed pseudogenes, Alu and SVA, do not encode any proteins and their mobilization depends on factors encoded by the autonomous retrotransposons (Cordaux and Batzer, 2009).

The youngest and still active non-autonomous retrotransposons are the SVA elements, which have expanded in the hominid primates’ genome for the past 18-25 million years. In the human genome they are represented by approximately 2800 copies, accounting for approximately 0.1% of our genome (Wang *et al*., 2005).

The SVA acronym was introduced by Shen *et al*. (1994) and is currently used (Hanks and Kazazian, 2010) to define a retrotransposon of composite structure, as SVA stands for S̲INE-V̲NTR-A̲lu. The SINE-R part is derived from HERV-K10 and its measures approximately 500 bp in length (Figure 1) (Ono *et al*., 1987). The VNTR region is highly variable among SVAs and is composed of copies of a 35-50 bp sequence (Wang *et al*., 2005). Upstream of the VNTR region is an approximately 370 bp region that contains three Alu-related sequences of 246, 54 and 25 bp in length in antisense orientation, and an undefined sequence (Shen *et al*., 1994). The full length element, which is approximately 3 kb in length, is associated at the 5’ end with a (CCCTCT) simple repeat. The poly(A) tail directly follows the putative AATAAA poly(A) signal and the whole element is enclosed by two TSDs (Figure 1) (Ostertag *et al*., 2003; Wang *et al*., 2005). Diagnostic mutations relative to the HERV_K10 sequence were used to establish a hierarchical subfamily structure. Six subfamilies have been described and named A to F. The E and F subfamilies are human specific (Wang *et al*., 2005).

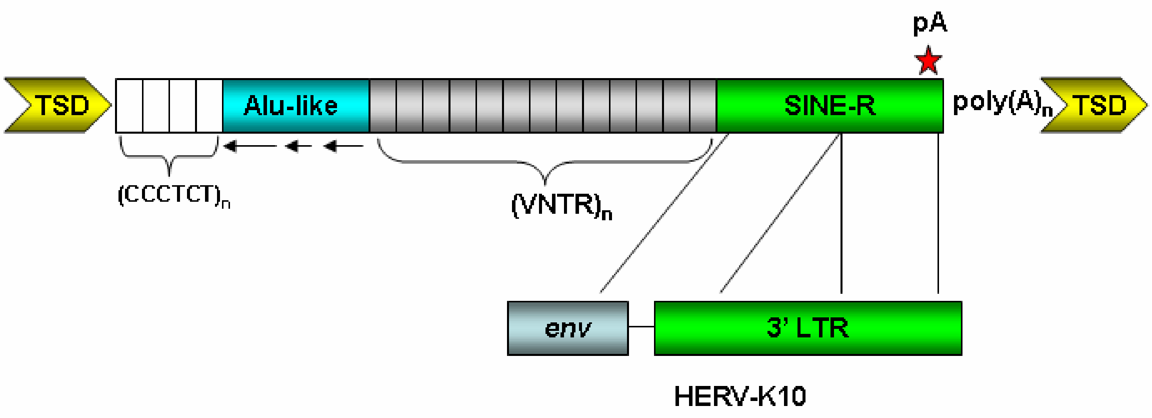


Figure 1. The composite structure of the SVA non-autonomous retrotransposon. The full length element contains a variable number of CCCTCT hexamers, which is followed by a sequence with homology to antisense Alu sequences, a VNTR region of variable length, a HERV-K10 – derived region (SINE-R) which contains the poly(A) signal (pA). The SVA element ends in a poly(A) tail and two TSDs flank the structure (adopted from Ostertag *et al*., 2003).

Efforts to identify an endogenous promoter that regulates SVA transcription lead to ambiguous results (Damert *et al*., unpublished; Hancks *et al*., 2009). However, in some cases transcription is driven by upstream cellular promoters, which results in 5' transductions (Figure 2) (Damert *et al*., 2009; Hancks *et al*., 2009). Also run-through transcription and the use of downstream poly(A) signals results in transduction of 3' flanking sequences (Figure 2) (Ostertag *et al*., 2003, Xing *et al*., 2006).

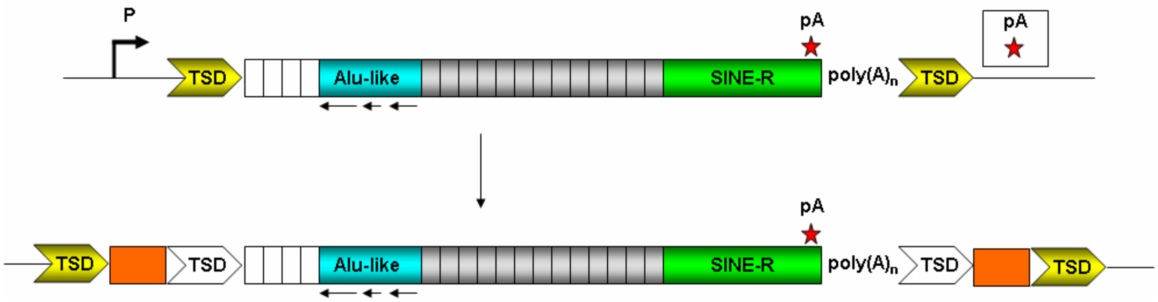


Figure 2. Acquisition of heterologous sequences by 5' and 3' transduction events. Transcription from an upstream cellular promoter (P) results in transduction of 5' heterologous sequence (orange box). Run-through transcription into the 3'-flanking genomic sequence and utilization of a downstream poly(A) signal (pA with red star, in box), results in transduction of 3' heterologous sequences (orange box).

Through a comprehensive analysis of SVAs on human chromosome 19 a novel type of SVA elements has been identified, which are characterized by 3' truncation in the SINE-R region. The elements end in a poly(A) tail and the insertions are flanked by two TSDs (Figure 3) (Damert *et al*., 2009). It has been hypothesized that such elements have arisen due to premature polyadenylation, but to date no experimental evidence exists to support such an assumption.

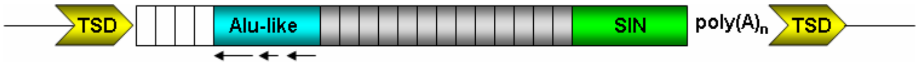


Figure 3. Structure of a 3' truncated SVA element. The element is characterized by 3' truncation in the SINE-R region (SIN) which is followed by a poly(A) tail. The entire structure is enclosed by TSDs.

As in the case of processed pseudogenes (Esnault *et al*., 2000) and Alu elements (Dewannieux *et al*., 2003), SVA elements are mobilized in *trans* by the L1 non-LTR retrotransposon encoded proteins (Hancks *et al*., 2011; Raiz *et al*., in press). It has been hypothesized that the Alu-related sequences in antisense orientation of the SVA RNA could facilitate binding to an active Alu RNA that is docked to the ribosome (Figure 4). Thus, SVA RNA could co-localize with Alu RNA in close proximity to the L1 retrotransposition proteins, facilitating its mobilization (Mills *et al*., 2007).

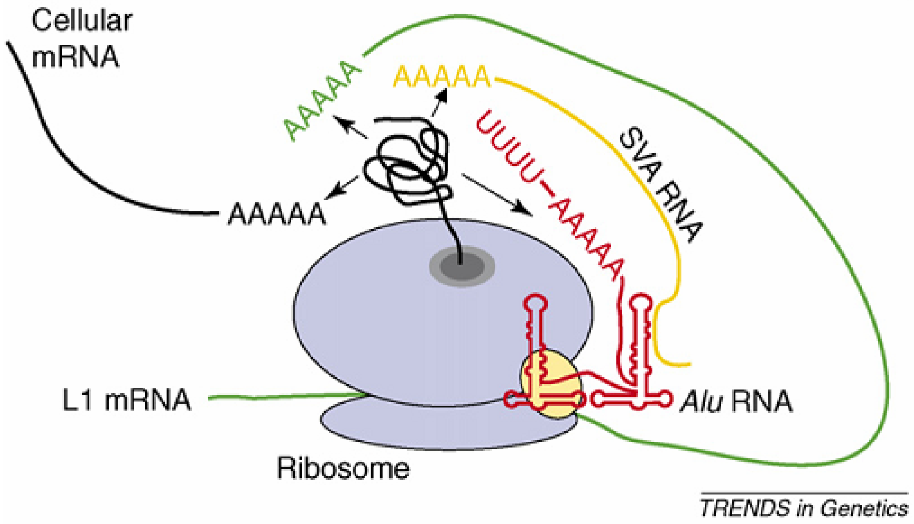


Figure 4. *Trans*-mobilization model for non-autonomous retrotransposons. The model depicts possible scenarios for L1-mediated retrotransposition in *cis* and *trans*. The translated L1 proteins (black line) interact with the RNA that encoded them (green line), resulting in a *cis*-preference of mobilization. The *trans*-mobilization of Alu involves docking of the Alu RNA (in red) to the ribosome, in close proximity to the nascent L1 proteins, thus hijacking them. Docking to the ribosome is mediated by the secondary structure of Alu RNA which facilitates binding of the left monomer by the signal recognition particle heterodimer 9/14 (SRP9/14, yellow oval). SVA might hybridize with the Alu-like region to an active Alu RNA, bringing its RNA (in orange) in proximity of the L1 proteins. Cellular mRNAs (in black) are considered to be poor substrates for mobilization, because they are not localized to the ribosome where the L1 proteins are translated. The black arrows indicate the interaction of the L1 proteins with different RNA substrates (modified from Mills *et al*., 2007).

As some SVA insertions have been previously associated with certain types of human diseases (reviewed in Belancio et al., 2008), a growing interest for these components of our genome has been observed for the past decade. However, some aspects of their biology still remained unclear and need to be addressed. Understanding the biology of SVA elements might help to evaluate the impact they had on our genome and how they can further contribute to human genome plasticity.

The purpose of this thesis is, therefore, to describe the author contribution to the research field of SVA retrotransposons.

Objectives

1. SVA elements are capable of acquiring additional heterologous sequences by 5' (Damert *et al*., 2009; Hancks *et al*., 2009) and 3' (Ostertag *et al*., 2003; Xing *et al*., 2006) transduction events. Many SVAs carry intact or partial Alu elements in their 5'/3' transduced sequences. Alu elements are known to be efficiently mobilized by interaction with the SPR9/14 heterodimer at the ribosome (Bennett *et al*., 2008). Therefore transduced intact or partial Alu elements might increase mobilization efficiency of SVA elements by providing the structural interface for interaction with the SRP9/14 heterodimer, or as suggested by Mills *et al*. (2007), by hybridizing to an active SRP9/14-bound Alu RNA. In order to provide support to such assumptions, vectors suitable for testing SVA retrotransposition in cell-based assays are required.

The first aim of this thesis is, therefore, the design of four SVA constructs: a set of two test vectors that contain an SVA element with or without a transduced Alu element in sense, and a second set of test vectors containing an SVA element with or without a transduced Alu element in antisense. These vectors can be used for in *vitro* retrotransposition assays and results obtained from these experiments could elucidate the mechanism of *trans*-mobilization of Alu-containing SVA elements.

2. Analysis of SVA elements on human chromosome 19 revealed a novel type of these elements, which is characterized by 3' truncations in the SINE-R region (Damert *et al*., 2009). It is hypothesized that these elements arise through premature polyadenylation of RNA derived from 3' intact elements. A detailed characterization of all 3' truncated SVA elements in the human genome, supplemented by detection of 3' truncated SVA transcripts in human testis, should provide evidence whether premature polyadenylation is the mechanism responsible for 3' truncation of SVA elements.

Design of two vector sets for investigating the role of Alu transduction in SVA mobilization

1. Construction of the H10_3 vector set

To investigate a possible function of transduced antisense Alu sequences in mobilization of SVA elements, SVA H10_3 (Damert *et al.*, 2009) has been selected. At its 5' end the element features an AluSg element on the minus strand and an AluSx element on the plus strand (Figure 5). The AluSg on the minus strand can underline the importance of transduced Alu elements for mobilization of SVAs in accordance to the model purposed by Mills *et al.* (2007). In addition, due to the splicing event observed between the 3' end of AluSx and the 5' end of AluSg elements in the 5' transduction of H10_3 (Damert *et al.*, 2009), experimental evidence can be provided to support the hypothesis that the Alu-like region of the SVA elements has emerged by splicing between Alu elements (Hancks *et al.*, 2009). Therefore a set of two test vectors was designed, one containing SVA H10_3 with its 5' AluSx/AluSg sequences [Alu(+)] and another one lacking these 5' flanking Alu elements [Alu(-)]. To ensure that the SVA Alu(+/-) element is co-expressed with the *neo* (neomycin) reporter cassette of the recipient vector pCEP-Neo (Raiz *et al.*, in press), the poly(A) signal of SVA H10_3 has been deleted from the final retrotransposition vectors.

The genomic locus containing SVA H10_3 with its 5' AluSg/AluSx flanking elements was PCR-amplified from BAC clone RPCIB753G12472Q (ImaGenes, Accession number AC073370) and the resulting 4 kb fragment was cloned and sequenced. Analysis of the sequencing data revealed that plasmid S1 had three modifications at the 5' end of the amplified fragment (Figure 5).

In order to obtain a mutation-free H10_3 Alu(+) retrotransposition reporter vector, the H10_3/Alu 5' fragment was re-amplified from BAC DNA, while the 3' fragment of H10_3 was re-amplified from plasmid S1. Then the two fragments were fused, and the resulting 2.6 kb fragment was cloned into pCEP-NEO downstream of the P CMV promoter and in front of the *neo* cassette, yielding the retrotransposition reporter vector pSC1 (Figure 5).

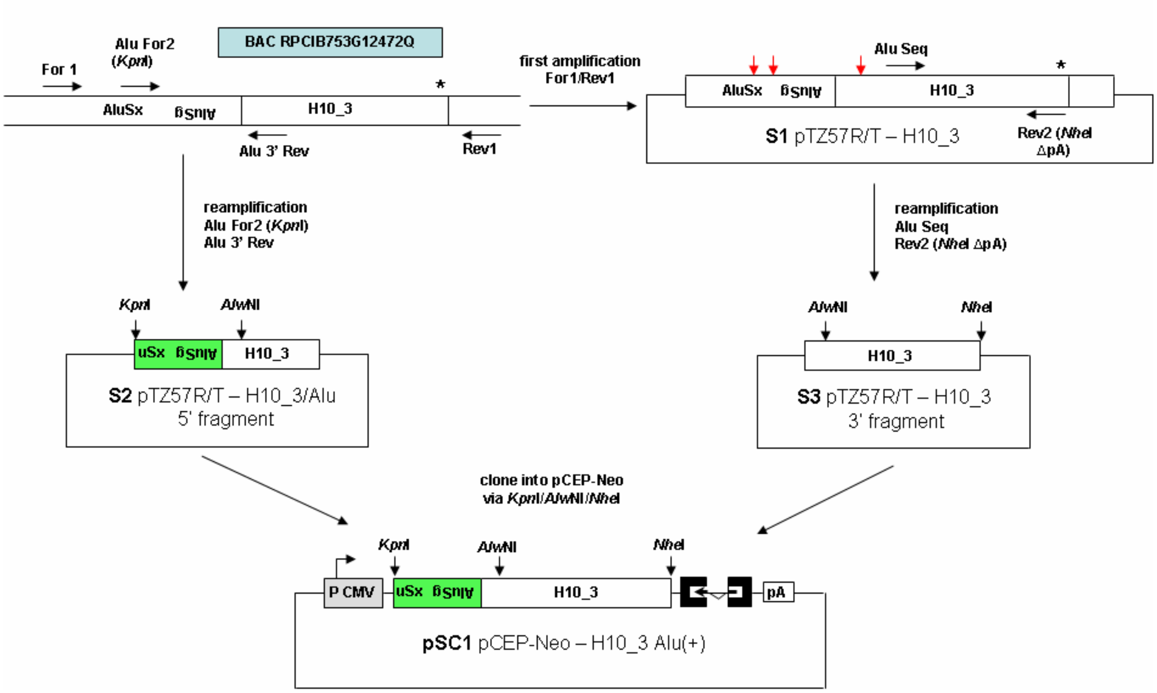


Figure 5. Construction of the SVA H10_3 Alu(+) retrotransposition reporter vector. The genomic locus containing SVA H10_3 AluSx/AluSg was amplified from BAC clone RPCIB753G12472Q with H10_3 For1 and Rev1 primers and cloned into pTZ57R/T yielding plasmid S1. Sequencing analysis indicated that S1 harbors three modifications in the cloned sequence (red arrows). A mutation-free retrotransposition reporter vector H10_3 Alu(+) was constructed by reamplification of H10_3/Alu 5' fragment from BAC clone RPCIB753G12472Q with H10_3 Alu For2 (*Kpn*I) and Alu 3'REV primers. The 3' fragment of SVA H10_3 without the polyadenylation signal (asterisk), was reamplified from plasmid S1 using Alu Seq and H10_3 Rev2 (*Nhe*I, ΔpA) primers. Both 5' and 3' fragments were cloned via *Kpn*I/*Alw*NI/*Nhe*I into the *Kpn*I/*Nhe*I pCEP-Neo vector backbone, yielding the retrotransposition vector pSC1. P CMV – cytomegalovirus immediate early enhancer/promoter; black box with a backward arrow - *neo* reporter cassette. ΔpA – deletion of the polyadenylation signal.

For generating the second retrotransposition vector, the H10_3 Alu(-) fragment was reamplified by PCR from pSC1. The resulting 1.8 kb fragment was further cloned into pCEP-Neo downstream of the P CMV promoter and in front of the *neo* cassette, yielding the retrotransposition reporter vector pSC2 (Figure 6).

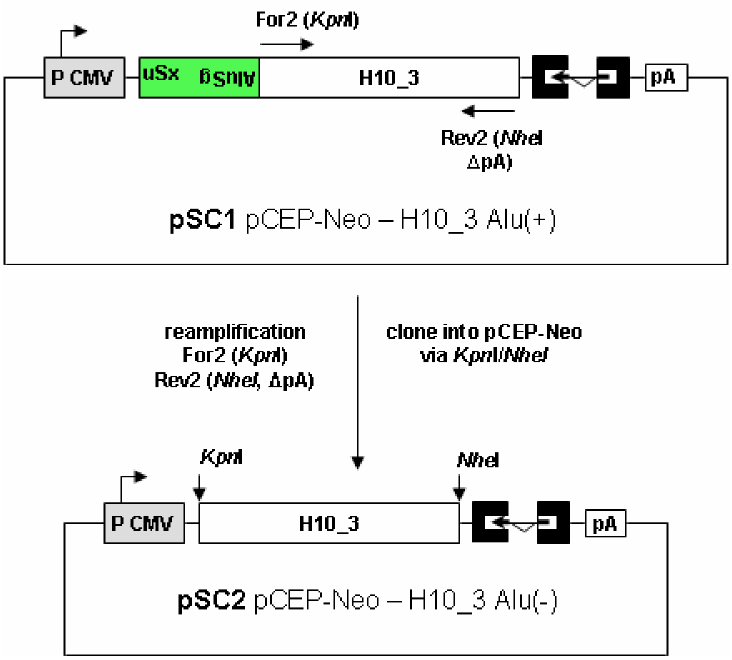


Figure 6. Construction of the SVA H10_3 Alu(-) retrotransposition reporter vector. The SVA H10_3 Alu(-) fragment was reamplified from plasmid pSC1 using H10_3 Alu For2 (*Kpn*I) and H10_3 Rev2 (*Nhe*I, ΔpA) primers and cloned into pCEP-Neo vector, yielding the retrotransposition reporter vector pSC2. P CMV – cytomegalovirus immediate early enhancer/promoter; black box with a backward arrow - *neo* reporter cassette. ΔpA – deletion of the polyadenylation signal.

2. Construction of the H10_1 vector set

H10_1 is the most active SVA element identified so far, accounting for at least 13 insertions in the human genome reference sequence (Damert *et al*., 2009; Hancks *et al*., 2009). H10_1 features a 3'-flanking AluSp element on the plus strand (Figure 7) which has been transduced to all 13 members of group 4 of SVA_F1 subfamily (Damert *et al*., 2009). This suggests that transduced Alu elements on the plus strand might enhance the mobilization efficiency of Alu-containing SVA elements. Therefore designing a retrotransposition vector set for cell-based assays, featuring SVA H10_ 1 with/without its 3'-flanking AluSp element [Alu(+/-)] may provide evidence to support this assumption.

To ensure co-expression of the H10_1 Alu(+/-) elements with the *neo* reporter cassette, the polyadenylation signals located at the 3' end of H10_1 and in the sequence downstream of the AluSp element (Figure 7) were deleted from the retrotransposition reporter vectors.

The genomic locus containing SVA H10_1 with its 3'-flanking AluSp was PCR-amplified from BAC clone RPCIB753F0114Q (ImaGenes; Accession number AL392107). The resulting 4.2 kb amplification product was subcloned, yielding plasmid S6. Sequencing data revealed that S6 had two A residues deletions in the poly(A) tail of SVA H10_1. Since the poly(A) tail will be replaced by a synthetic one (see below), such modifications will not be found in the final retrotransposition reporter vectors.

In addition to this irrelevant modification, an 835 bp fragment of the VNTR region was not amenable to sequencing using VNTR-flanking primers. In order to sequence the 835 bp fragment, plasmid S6 was digested with *Alw*NI, whose recognition site is localized upstream of the VNTR region. Using Bal-31 exonuclease with increasing digestion incubation times, progressive deletions were created at the 5' end of the VNTR region. The resulting fragments were end-repaired, cloned and sequenced at their 5' ends. Using this procedure, the entire VNTR region could be sequenced. Data analysis revealed a deletion of a G residue in the VNTR region. Most likely a single nucleotide deletion does not affect the overall structure of the element, therefore plasmid S6 was used for all further cloning.

The H10_1 SVA element without its 3'-flanking AluSp was reamplified from plasmid S6 and the resulting 3.5 kb fragment was cloned, yielding plasmid S7. Sequencing and restriction analysis indicated that the amplified fragment harbored five modifications in the sequence upstream of the *Bam*HI site, in addition to an approximately 1kb deletion in the VNTR region (Figure 8).

In order to obtain a plasmid without such significant modifications in the H10_1 Alu(-) sequence, the mutation-containing fragment of plasmid S7 was replaced with the corresponding fragment of plasmid S6, yielding plasmid S8. Finally the H10_1 Alu(-) fragment was cloned into pCEP-Neo downstream of the CMV promoter and in front of the *neo* cassette, yielding the retrotransposition reporter vector pSC3 (Figure 8).

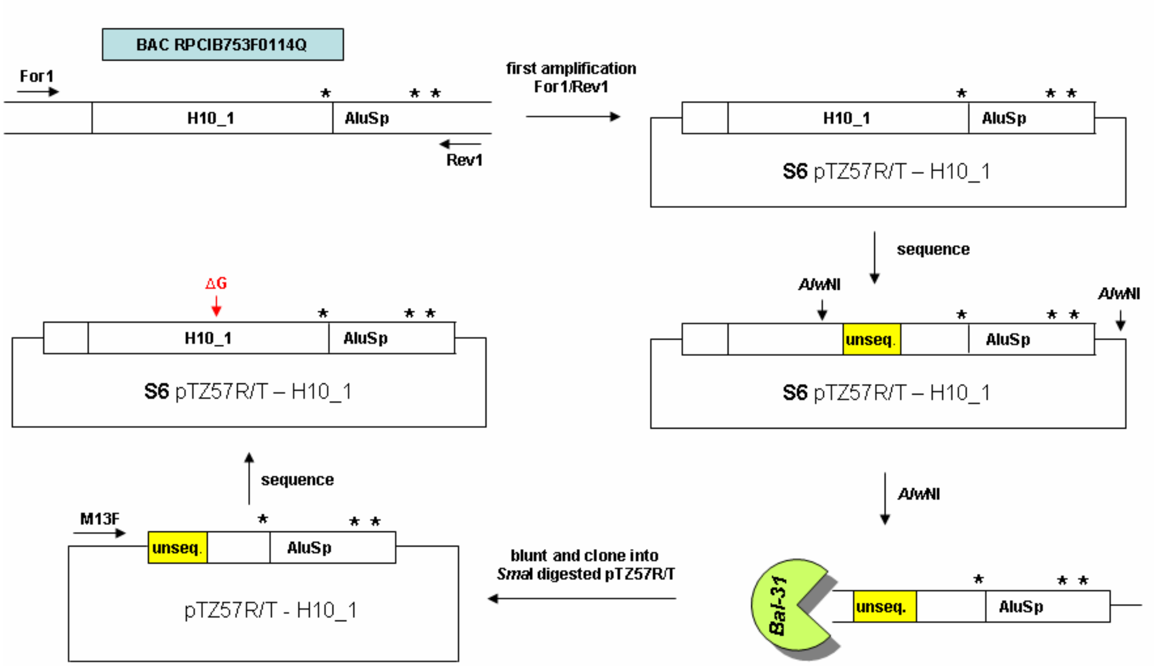


Figure 7. Amplification of the H10_1/3' AluSp genomic locus and sequencing of the VNTR region. Primers H10_1 For1 and Rev1 were used for amplification of the H10_1 with its 3' flanking AluSp element from BAC clone RPCIB753F0114Q and cloned into pTZ57R/T, yielding plasmid S6. Sequence analysis indicated that a fraction of the VNTR region remained unsequenced (yellow box). In order to sequence the VNTR region completely, plasmid S6 was digested with *Alw*NI and progressive deletions were created at the 5' end of the VNTR region using Bal-31 exonuclease. The digested fragments were further blunted and cloned into *Sma*I digested pTZ57R/T. Resulting plasmids harboring the VNTR region with different extents of deletions at the 5' end [VNTR(-)$_n$] were sequenced with M13F universal primer. Data analysis indicated the plasmid S6 had a single G residue deletion in the VNTR region (ΔG with a red arrow). Asterisks – polyadenylation signals.

For generating the H10_1 Alu(+) retrotransposition reporter vector, a synthetic oligonucleotide (Generi Biotech) substituting for the AATAAA-containing poly(A) tail in the genomic sequence was inserted in plasmid S8 downstream of the H10_1 SINE-R, yielding plasmid S10 (Figure 9). Subsequently, the 3' AluSp element was reamplified from plasmid S6 and the 389 bp amplicon was further subcloned. Sequencing analysis revealed that plasmid S11 and S12 had a two A residues deletions in the poly(A) tail and a G residue deletion at the 5' end, respectively. Such modifications can have an impact on retrotransposition, as the G deletion might affect folding of the Alu RNA in a proper conformation for interaction with the SRP9/14 heterodimer (Bennett *et al.*, 2008). Because the length of the poly(A) tail also plays an essential role for retrotransposition of the Alu elements (Roy-Engel *et al*., 2002; Dewannieux and Heidmann, 2005), the two A residues deletions could not be ignored. In order to obtain an intact AluSp element, the 5' mutation-containing fragment of plasmid S12 was replaced with the 5' mutation-free fragment of plasmid S11, yielding plasmid S13. From this plasmid, the AluSp element was cloned downstream of the synthetic poly(A) tail of H10_1 in plasmid S10. The resulting 3937 bp H10_1 Alu(+) fragment was then inserted into pCEP-Neo, downstream of the CMV promoter and in front of the *neo* reporter cassette, yielding pSC4 (Figure 9).

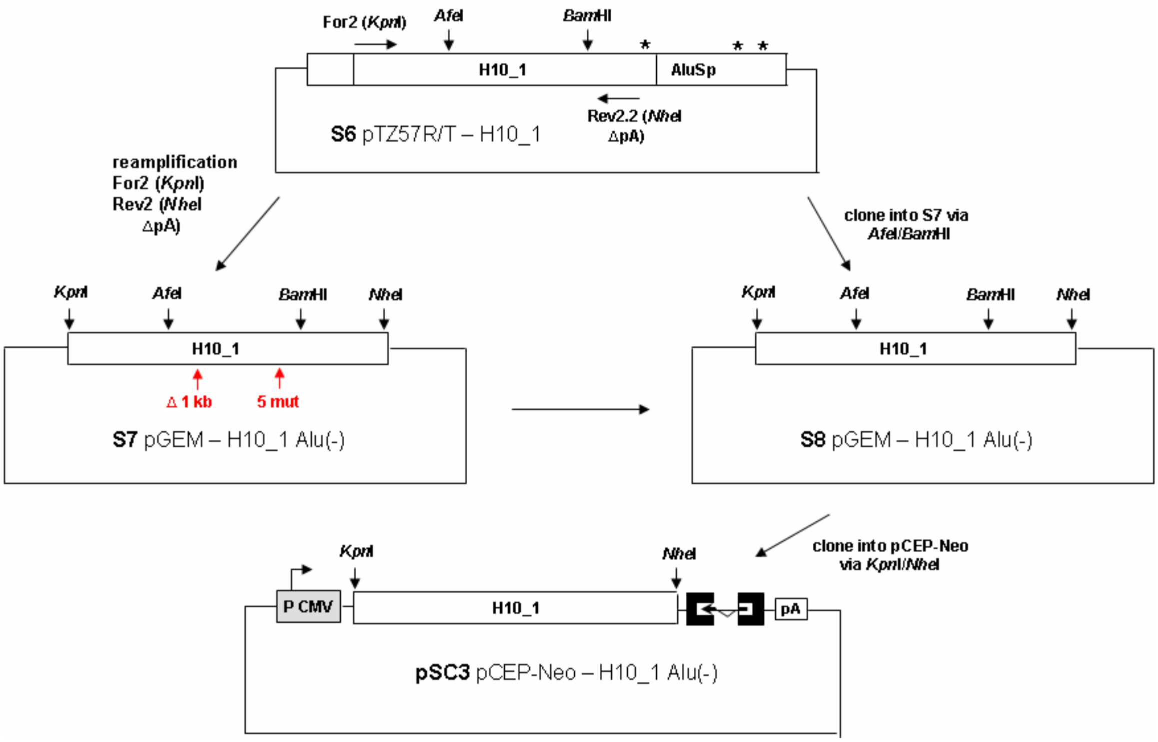


Figure 8. Construction of SVA H10_1 Alu(-) retrotransposition reporter vector. The H10_1 Alu(-) fragment lacking the polyadenylation signal (asterisk), was reamplified from plasmid S6 using H10_1 For2 (*Kpn*I) and H10_1 Rev2.2 (*Nhe*I, ΔpA) primers and cloned into pGEM, yielding plasmid S7. Sequencing and restriction analysis indicated that H10_1 Alu(-) fragment harbored five mutations at the 3' end (5 mut with a red arrow) and an approximately 1 kb internal deletion (Δ 1 kb with red arrow). In order to obtain a mutation-free plasmid, the *Afe*I/*Bam*HI fragment of plasmid S7 was replaced with the *Afe*I/*Bam*HI fragment of plasmid S6, yielding plasmid S8. The H10_1 Alu(-) retrotransposition reporter vector pSC3 was obtained by cloning of the H10_1 Alu(-) fragment from S8 into pCEP-Neo via *Kpn*I/*Nhe*I. P CMV – cytomegalovirus immediate early enhancer/promoter; black box with a backward arrow – *neo* reporter cassette. ΔpA – deletes the polyadenylation signal.

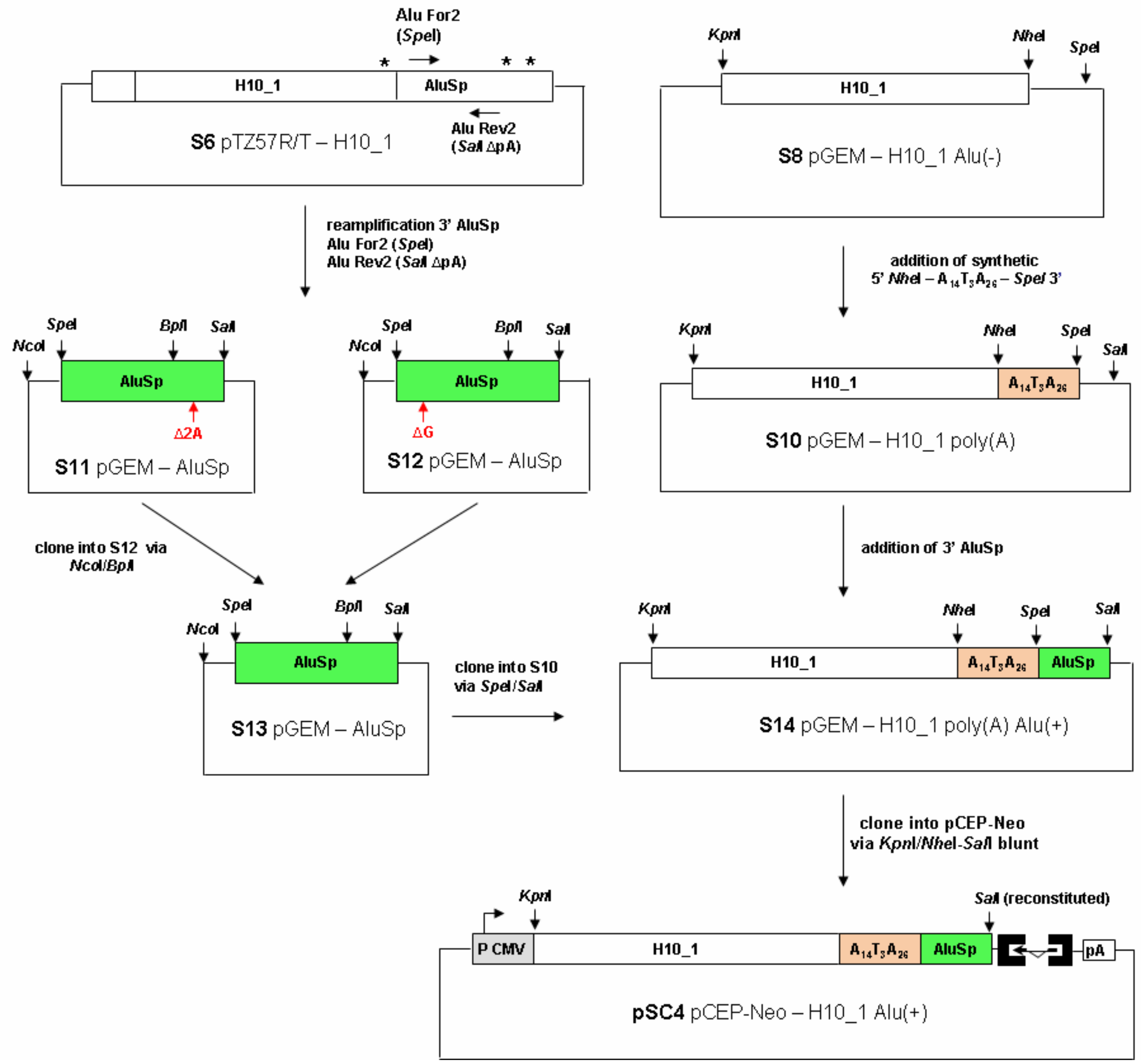


Figure 9. Construction of SVA H10_1 Alu(+) retrotransposition reporter vector. A synthetic *Nhe*I–$A_{14}T_3A_{26}$–*Spe*I oligomer substituting for the AATAAA-containing genomic poly(A) tail, was inserted via *Nhe*I/*Spe*I at the 3' end of the H10_1 of plasmid S8. The 3' AluSp element was reamplified without the downstream polyadenylation signals (asterisks) from plasmid S6 using H10_1 Alu For2 (SpeI) and H10_1 Alu Rev2 (*Sal*I, ΔpA) primers and cloned into pGEM. Sequencing analysis of the two resultant plasmids, indicated that S11 harbored a two A residues deletion in poly(A) tail of the AluSp (Δ2A with a red arrow), while S12 contained a G residue deletion at the 5' end of the AluSp element (ΔG with a red arrow). An intact AluSp was obtained by replacing the *Nco*I/*Bpl*I fragment of S12 with the mutation-free *Nco*I/*Bpl*I fragment of S11, yielding plasmid S13. The reconstituted AluSp was further cloned into plasmid S10, downstream of the synthetic poly(A) tail via *Spe*I/*Sal*I, yielding plasmid S14. The SVA H10_1 Alu(+) retrotransposition reporter vector pSC4 was constructed by cloning of the H10_1/$A_{14}T_3A_{26}$/AluSp fragment into pCEP-Neo via *Kpn*I/*Nhe*I-*Sal*I blunt. P CMV – cytomegalovirus immediate early enhancer/promoter; black box with a backward arrow - *neo* reporter cassette; ΔpA – deletion of the polyadenylation signal.

Characterization of 3’ truncated SVA elements

1. The human genome reference sequence harbors 98 3’ truncated SVA elements

The filtering method based on the end of the repeat value annotation of Repeat Masker (http://www.repeatmasker.org/cgi-bin/AnnotationRequest) had retrieved 104 SVA elements from the human genome reference sequence HG17/2004. Further analysis of their SINE-R regions revealed that 98 elements were 3’ truncated; three elements harbored SINE-R internal deletions, while the remaining three elements featured retrotransposon insertions into the SINE-R region (Figure 10).

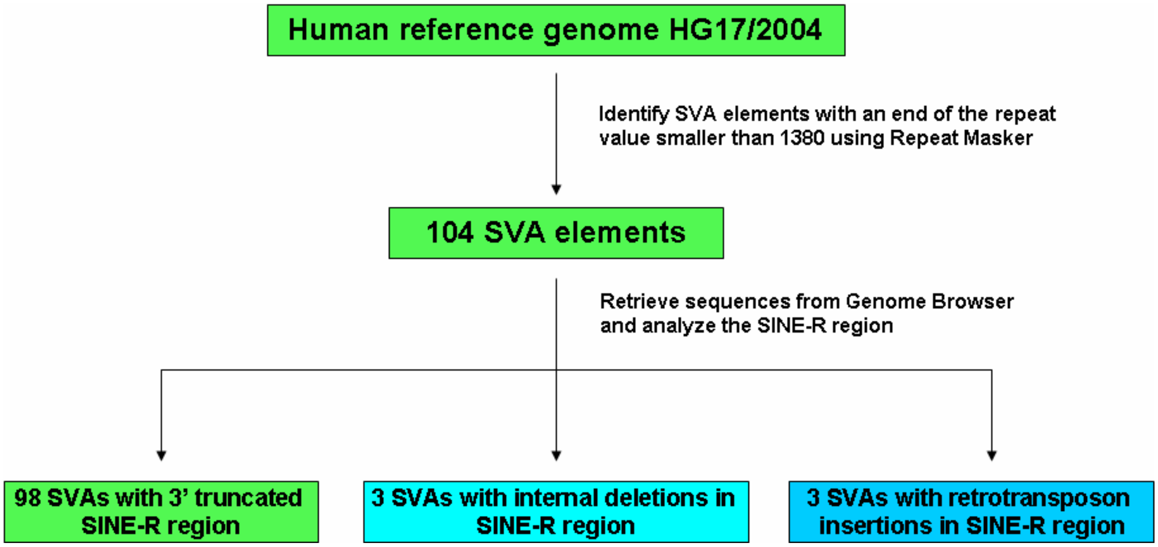

Human reference genome HG17/2004
Identify SVA elements with an end of the repeat value smaller than 1380 using Repeat Masker
104 SVA elements
Retrieve sequences from Genome Browser and analyze the SINE-R region
98 SVAs with 3’ truncated SINE-R region
3 SVAs with internal deletions in SINE-R region
3 SVAs with retrotransposon insertions in SINE-R region

Figure 10. Identification of 3’ truncated SVA elements pipeline. In the first step the reference sequence of human genome HG17/2004 was screened using RepeatMasker for SVA elements. The approximately 2700 elements identified this way (Damert *et al*., 2009), were further analyzed to detect SVAs with an end of the repeat value smaller than 1380, which corresponds to SVA elements harboring a full length SINE-R region. In the second step, based on the SINE-R analysis, the 104 SVA elements identified were further divided in SVAs with a 3’ truncated SINE-R region (98 elements), SVAs harboring internal SINE-R deletions (three elements) and SVAs harboring retrotransposon insertions into the SINE_R region (three elements).

The 98 SVA elements with a 3’ truncated SINE-R region account for 3.6% of the approximately 2700 SVA elements in the analyzed data set. Eight of them, localized on chromosome 19, have been reported previously (Damert *et al*., 2007). The SINE-R region of the 98 elements identified varies in length from 26 to 387 bp, with a mean value of 210 bp (Figure 11). This indicates that 3' truncated SVA are still capable of retrotransposition, therefore the SINE-R region is not crucial for mobilization of SVAs. However, in L1 elements premature polyadenylation which results in 3' truncated L1, transcripts was shown to attenuate L1 mobilization (Belancio *et al*., 2003).

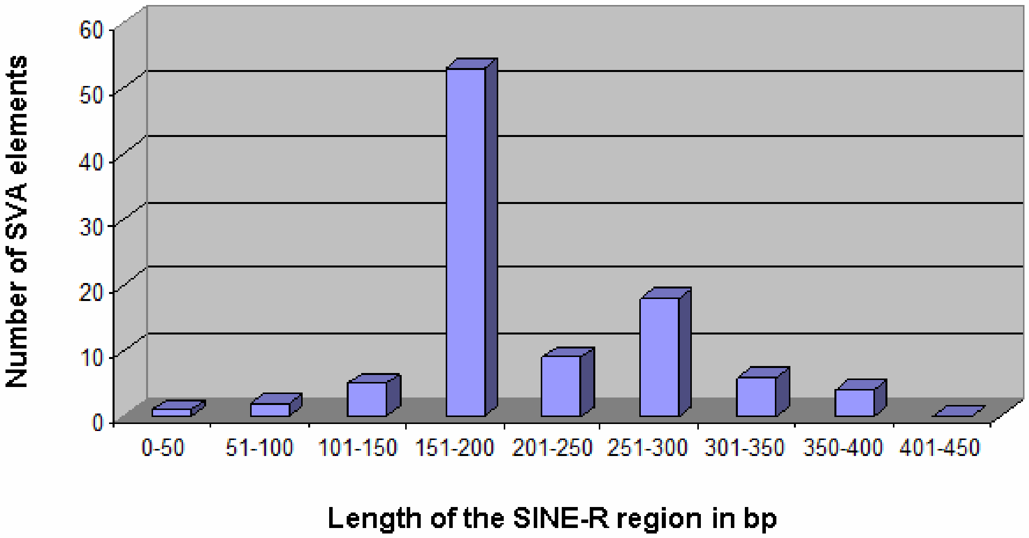


Figure 11. SINE-R length distribution in 3' truncated SVA elements. The abundance of 3' truncated SVA elements in the human genome (y axis) is shown relative to the length of their SINE-R region in 50 bp intervals (x axis).

The length of the poly(A) tails of the identified 3' truncated SVAs is in the range of 1 to 66 bp, either composed solely of A residues or displaying a patterned structure (heterogeneous). This distribution is similar to a previous report for SVA elements on chromosome 19 (Damert *et al*., 2009). In three elements, H1_69, H5_829 and H20_1877, possible poly(A) tails coincide with their A homopolymeric or A-rich TSDs. Longer poly(A) tails might have existed at the time of integration and could have undergone progressive shortening over time (Chen *et al*., 2005 and references therein).

In two elements which lack a poly(A) tail downstream of the SINE-R region, an interesting observation was made. In SVA H14_A378 and its offspring element SVA H2_413, the SINE_R region was fused to a 33 bp fragment derived from the very 3' end of an L1PA element. Splicing between the SVA and L1 RNAs is not very likely, as no functional splice sites have been identified in the SINE-R region of SVA elements. An appealing possibility is template switching during TPRT (Gogvadze and Buzdin, 2005) between the L1 RNA and the SVA RNA which lead to fusion of the L1 3' sequence to the SINE-R region of the SVA element. Therefore the 3' truncation in these two elements did not result from premature polyadenylation. A schematic representation of SINE-R-fused heterologous sequences is depicted in figure 12.

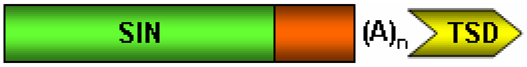


Figure 12. Schematic representation of the 3' truncated SINE-R region (SIN in green box) with fused heterologous sequences (dark orange box). $(A)_n$ – poly(A) tail.

Regarding the subfamily composition of the 3' truncated SVAs, SVA_D is the most represented, followed by SVA_B and SVA_F, which is consistent with previous reports of genome-wide analysis of SVA elements with a full length SINE-R region (Figure 13) (Wang *et al*., 2005). Subfamily A, however, is over-represented, which is in contrast to the distribution found for 3' intact elements genome-wide (Wang *et al*., 2005). An possible explanation for this difference is that SVA_A elements might provide preferential poly(A) signals for premature polyadenylation. Or it might be that elements affiliated to this subfamily might have undergone more extensive insertions/deletions/inversions than elements affiliated to younger subfamilies. For this reasoning, SVA_A should be over-represented among those 15% of SVA elements for which Wang and colleges (Wang *et al*., 2005) did not established subfamily affiliation due to their lack of an intact SINE-R region.

About 55% of the 3' truncated elements, affiliated to subfamily D, E and F, are human specific insertions, indicating that more than half of the 3' truncated SVAs have expanded after the human and chimpanzee divergence.

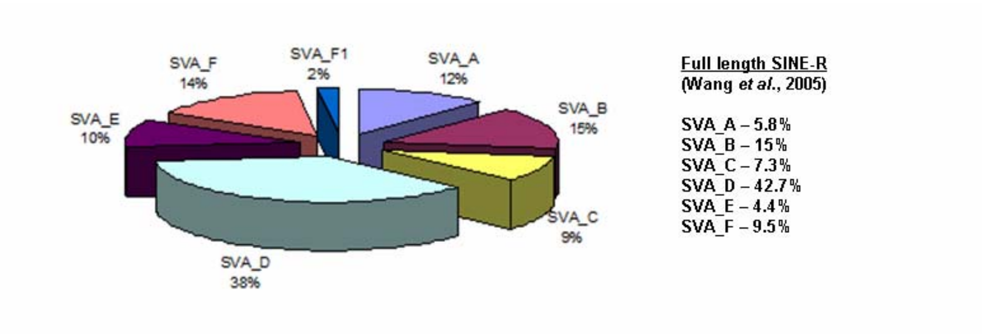


Figure 13. Subfamily distribution of 3' truncated SVA elements in the human genome. The proportion of each subfamily is shown as a percentage of the total number of 3' truncated SVAs. The subfamily distribution of SVA elements with a full length SINE-R region (Wang *et al.*, 2005) is given for comparison.

2. The SINE-R region supports the use of a broad range of alternative polyadenylation sites

2.1 Alternative polyadenylation sites of SVA elements in the human genome reference sequence

In this analysis, the two elements which contained fused heterologous sequences in the SINE-R region were excluded, because most likely these elements have not arisen by premature polyadenylation (see discussions at page 17-18). In addition, an element in which a homopolymeric C tract separates the SINE-R region and the poly(A) tail, was excluded from further analysis as well.

Based on the position of the cleavage and polyadenylation site relative to the SINE-R family consensus sequence, the remaining 95 3' truncated elements were clustered in 8 groups (Table 1) and screened for putative poly(A) signals upstream of the poly(A) site. In group 1 and group 8, no poly(A) signals were detected 40 bases upstream of the cleavage and polyadenylation site. For the rest of the groups, the majority of the identified putative poly(A) signals are single nucleotide variants of the canonical AATAAA signal or corresponding to the consensus sequence NNUANA of the human polyadenylation signals (Beaudoing *et al.*, 2000). Also two canonical poly(A) signals have been detected, AATAAA and ATTAAA, which have arisen by one base substitution of the subfamily consensus sequences (Table 1).

Regarding the distribution of the poly(A) signal relative to the cleavage and polyadenylation site, it is localized 4 to 37 bases upstream of the poly(A) site (Figure 14). In one element, the cleavage and polyadenylation site is located immediately after the poly(A) signal, as expected for elements retrotransposed by L1. Excluding this element, the distribution is similar to the general localization of the poly(A) signal, which is found 10-35 bases upstream of the cleavage and polyadenylation site (reviewed Millevoi and Vagner, 2010).

Table 1. 3' Truncation groups with corresponding putative poly(A) signals.

Group	**Range of cleavage and polyaden-ylation sites**	**Putative poly(A) signals**						
		SVA_A	**SVA_B**	**SVA_C**	**SVA_D**	**SVA_E**	**SVA_F**	**SVA_F1**
1	26						unknown (1)	
2	88-94		AAGAAA (1)	AATAGA GATAGA (1)				
3	132-135	TGTAGA AGTAGA (1)	TGTAGA AGTAGA (1)		TGTGGA GGTAGA (1) TGTAGA AGTTGA (1) TGTAGA AGTAGA (1)			
4	166-173	AAGAAA (2)	AAGAAA (8)	AAGAAA (5)	AAGAAA (17) AATAAA (1)	AAGAAA (7)	AAGAAA (6)	AAGAAA (1)
5	194-205	TCTACA (1)	TCTATA (2)		TCTGTG (1)			
6	245-264	GTTAAA (1)	GTTAAA (1)		GTTAAA (5) GTTGAA (1)	GTTAAA (1)	GTTAAA (2) ATTAAA (5)	GTTAAA (1)
7	283-302	GTTAAA (5)	GTTAAA (3)	GTTAAA (2)	GTTAAA (3)	GTTAAA (2)		
8	362-385	unknown (1)		unknown (1)	unknown (2)			

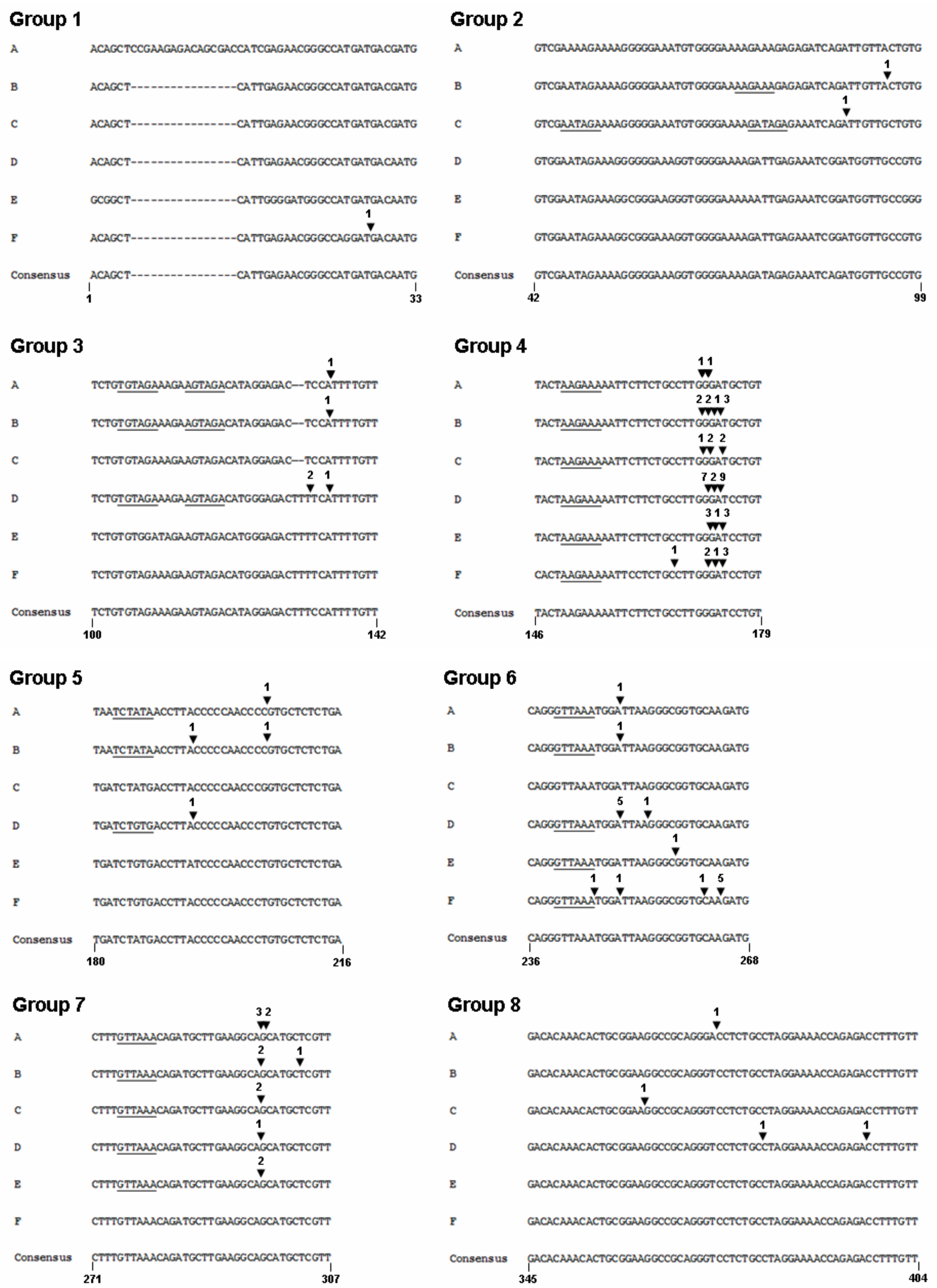
The putative poly(A) signals are indicated depending on the subfamily affiliation of each SVA element

Numbers in brackets indicate the number of SVA elements using that specific putative poly(A) signal(s)

NNNNNN – putative poly(A) signals that differ from the subfamily consensus sequence by a single base substitution

NNNNNN – putative poly(A) signals which differ from the consensus NNUANA, but are consistent with the subfamily consensus sequence

NNNNNN – putative poly(A) signals which differ by a single base substitution from the consensus poly(A) signal NNUANA and from the subfamily consensus sequence


Group 1
A ACAGCTCCGAAGAGACAGCGACCATCGAGAACGGGCCATGATGACGATG
B ACAGCT----------------CATTGAGAACGGGCCATGATGACGATG
C ACAGCT----------------CATTGAGAACGGGCCATGATGACGATG
D ACAGCT----------------CATTGAGAACGGGCCATGATGACAATG
E GCGGCT----------------CATTGGGGATGGGCCATGATGACAATG
1
F ACAGCT----------------CATTGAGAACGGGCCAGGATGACAATG
Consensus ACAGCT----------------CATTGAGAACGGGCCATGATGACAATG
1
33
Group 2
A GTCGAAAAGAAAAGGGGGAAATGTGGGGAAAAGAAAGAGAGATCAGATTGTTACTGTG
1
B GTCGAATAGAAAAGGGGGAAATGTGGGGAAAAGAAAGAGAGATCAGATTGTTACTGTG
1
C GTCGAATAGAAAAGGGGGAAATGTGGGGAAAAGATAGAGAAATCAGATTGTTGCTGTG
D GTGGAATAGAAAGGGGGGAAAGGTGGGGAAAAGATTGAGAAATCGGATGGTTGCCGTG
E GTGGAATAGAAAGGCGGGAAGGGTGGGGAAAAAATTGAGAAATCGGATGGTTGCCGGG
F GTGGAATAGAAAGGCGGGAAAGGTGGGGAAAAGATTGAGAAATCGGATGGTTGCCGTG
Consensus GTCGAATAGAAAAGGGGGAAAGGTGGGGAAAAGATAGAGAAATCAGATGGTTGCCGTG
42
99
Group 3
1
A TCTGTGTAGAAAGAAGTAGACATAGGAGAC—TCCATTTTGTT
1
B TCTGTGTAGAAAGAAGTAGACATAGGAGAC—TCCATTTTGTT
C TCTGTGTAGAAAGAAGTAGACATAGGAGAC—TCCATTTTGTT
2 1
D TCTGTGTAGAAAGAAGTAGACATGGGAGACTTTTCATTTTGTT
E TCTGTGTGGATAGAAGTAGACATGGGAGACTTTTCATTTTGTT
F TCTGTGTAGAAAGAAGTAGACATGGGAGACTTTTCATTTTGTT
Consensus TCTGTGTAGAAAGAAGTAGACATAGGAGACTTTCCATTTTGTT
100
142
Group 4
1 1
A TACTAAGAAAAATTCTTCTGCCTTGGGATGCTGT
2 2 1 3
B TACTAAGAAAAATTCTTCTGCCTTGGGATGCTGT
1 2 2
C TACTAAGAAAAATTCTTCTGCCTTGGGATGCTGT
7 2 9
D TACTAAGAAAAATTCTTCTGCCTTGGGATCCTGT
3 1 3
E TACTAAGAAAAATTCTTCTGCCTTGGGATCCTGT
1 2 1 3
F CACTAAGAAAAATTCCTCTGCCTTGGGATCCTGT
Consensus TACTAAGAAAAATTCTTCTGCCTTGGGATCCTGT
146
179
Group 5
1
A TAATCTATAACCTTACCCCCAACCCCGTGCTCTCTGA
1 1
B TAATCTATAACCTTACCCCCAACCCCGTGCTCTCTGA
C TGATCTATGACCTTACCCCCAACCCGGTGCTCTCTGA
1
D TGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGA
E TGATCTGTGACCTTATCCCCAACCCTGTGCTCTCTGA
F TGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGA
Consensus TGATCTATGACCTTACCCCCAACCCTGTGCTCTCTGA
180
216
Group 6
1
A CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
1
B CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
C CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
5 1
D CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
1
E CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
1 1 1 5
F CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
Consensus CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
236
268
Group 7
3 2
A CTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTT
2 1
B CTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTT
2
C CTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTT
1
D CTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTT
2
E CTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTT
F CTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTT
Consensus CTTTGTTAAACAGATGCTTGAAGGCAGCATGCTCGTT
271
307
Group 8
1
A GACACAAACACTGCGGAAGGCCGCAGGGACCTCTGCCTAGGAAAACCAGAGACCTTTGTT
B GACACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTT
1
C GACACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTT
1 1
D GACACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTT
E GACACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTT
F GACACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTT
Consensus GACACAAACACTGCGGAAGGCCGCAGGGTCCTCTGCCTAGGAAAACCAGAGACCTTTGTT
345
404

(legend on the next page)

Figure 14. SVA 3' truncation groups. Putative poly(A) signals are underlined in the SVA_A-F subfamily consensus sequences. For each signal the cleavage and polyadenylation site is indicated with arrow heads, while SVAs using each particular cleavage and polyadenylation sites are indicated with numbers. The sequence limits of each 3' truncation group are indicated in the consensus sequence.

Approximately 50% of the 3' truncated SVA elements utilize the AAGAAA signal, while 25% utilize the GTTAAA polyadenylation signal. 20% of the 3' truncated SVAs utilize poly(A) signals which are consistent with the NNUANA consensus sequence or differ by one base from this consensus. So why is the AAGAAA poly(A) signal is mostly preferred by 3' truncated SVAs? Potential enhancer sequences might be responsible for the observed bias in poly(A) signal usage.

Analysis of the consensus sequence of group 4 revealed that the sequence upstream of the poly(A) signal contains two conserved TGTAG and TGTAC motifs (Figure 15). They correspond with the UGUAN binding consensus sequence of the cleavage factor Im (CFIm), one of the factors of the polyadenylation machinery. It has been shown that CFIm can direct polyadenylation at non-canonical poly(A) signals through interaction with cleavage and polyadenylation specificity factor (CPSF, Venkataraman *et al*., 2005).

A second potential enhancer sequence has been identified as a U-rich sequence upstream of the AAGAAA signal (Figure 15). Such sequences have been shown to be capable of directing polyadenylation at non-canonical signals through interaction with splicing factors or basal polyadenylation factors (Danckwardt *et al*., 2007). Therefore, polyadenylation using the AAGAAA signal in group 4 might also be enhanced by splicing or/and polyadenylation factors through interaction with upstream enhancer sequence elements.

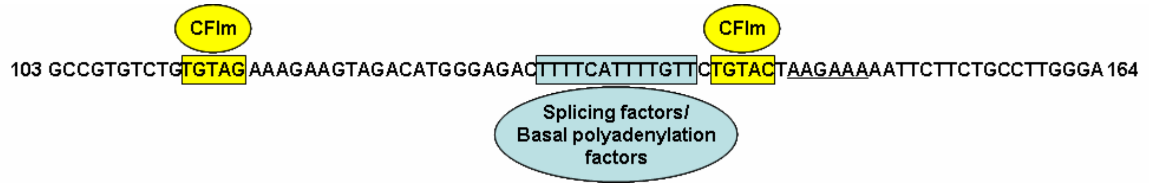

CFIm
CFIm
103 GCCGTGTCTGTGTAGAAAGAAGTAGACATGGGAGACTTTTCATTTTGTTCTGTACTAAGAAAAATTCTTCTGCCTTGGGA 164
Splicing factors/ Basal polyadenylation factors

Figure 15. Upstream sequence elements in the consensus sequence of 3'truncation group 4. The *cis*-elements are indicated in colored boxes with their corresponding *trans*-acting factors (colored ovals). CFIm – cleavage factor Im. Underlined letters – poly(A) signal.

Altogether, these results suggest that polyadenylation of SVA transcripts directed by non-canonical poly(A) signals is largely similar to alternative polyadenylation of human gene transcripts. However, for a small set of SVA elements, alternative polyadenylation seems to be driven by yet unknown signals. It might be that SVA elements are using radical different alternative poly(A) signals. This suggests that the conserved positions in the NNUANA consensus sequence (Beaudoing *et al*., 2000) might tolerate further mutations. However such poly(A) signals are most likely poor substrates for the polyadenylation machinery, as suggested by the small number of elements in which no poly(A) signals could be detected.

2.2 Poly(A) sites used for polyadenylation of 3’ truncated SVA transcripts expressed in the human testis

2.2.1 3’ RACE analysis

For investigation whether expression of 3’ truncated SVA transcripts can be detected in the human testis, a 3’ RACE procedure was used. The RACE procedure (rapid amplification of cDNA ends) is performed in two steps and involves the use of three primers, one for cDNA synthesis and two for PCR (Figure 16). The cDNA synthesis primer includes an anchor sequence at the 5’ end of an oligo(dT) sequence. This primer binds polyadenylated mRNAs which is reverse transcribed into cDNA. In the second step, specific 3’ ends are obtained from the heterogeneous cDNA pool by regular PCR using a gene-specific primer and a second primer that contains the anchor sequence alone (Borson *et al*., 1992).

For designing a SINE-R specific primer, first all 3’ truncated SVAs from human genome were sorted in 11 groups based on the end of the repeat value annotations from Repeat Masker. Group 11 accounts for only one member. A consensus sequence was inferred for each of the first 10 groups by aligning the SINE-R region of their corresponding members. Consensus sequences along with the SINE-R sequence of group 11 member were further aligned to obtain a final consensus sequence for primer design (Figure 17). The designed primer JS902+ has a T to C substitution at the 6 nt position compared to the JS902 primer previously reported (Kim *et al*., 2000).

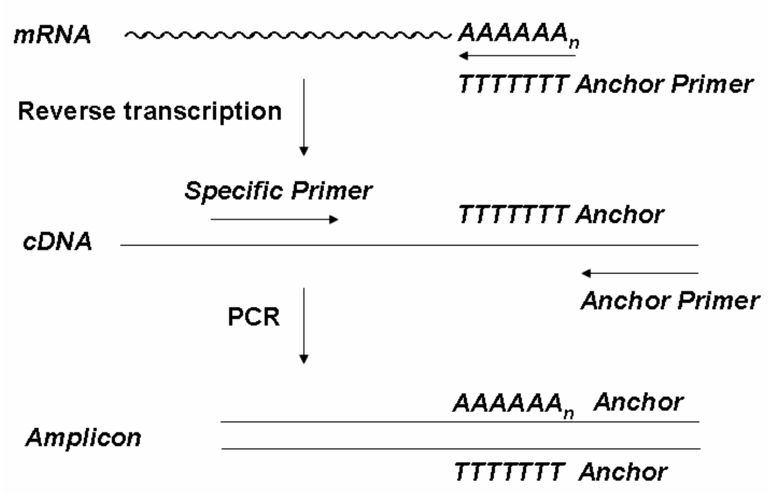
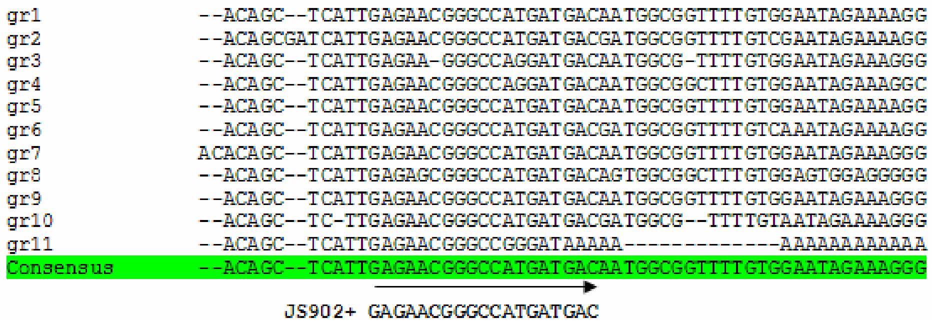


Figure 16. Principle of the 3’ RACE procedure (Rapid Amplification of cDNA Ends). In the first step an oligo(dT)-Anchor primer (TTTTTTT Anchor) is used to reverse transcribe polyadenylated mRNA (mRNA) into complementary DNA (cDNA). This cDNA is further used as template for amplification of the desired 3’ ends, which also includes the poly(A) tail (AAAAAA$_n$), using a specific primer and an Anchor primer.



```
gr1          --ACAGC--TCATTGAGAACGGGCCATGATGACAATGGCGGTTTTGTGGAATAGAAAAGG
gr2          --ACAGCGATCATTGAGAACGGGCCATGATGACGATGGCGGTTTTGTCGAATAGAAAAGG
gr3          --ACAGC--TCATTGAGAA-GGGCCAGGATGACAATGGCG-TTTTGTGGAATAGAAAGGG
gr4          --ACAGC--TCATTGAGAACGGGCCAGGATGACAATGGCGGCTTTGTGGAATAGAAAGGC
gr5          --ACAGC--TCATTGAGAACGGGCCATGATGACAATGGCGGTTTTGTGGAATAGAAAAGG
gr6          --ACAGC--TCATTGAGAACGGGCCATGATGACGATGGCGGTTTTGTCAAATAGAAAAGG
gr7          ACACAGC--TCATTGAGAACGGGCCATGATGACAATGGCGGTTTTGTGGAATAGAAAGGG
gr8          --ACAGC--TCATTGAGAGCGGGCCATGATGACAGTGGCGGCTTTGTGGAGTGGAGGGGG
gr9          --ACAGC--TCATTGAGAACGGGCCATGATGACAATGGCGGTTTTGTGGAATAGAAAGGG
gr10         --ACAGC--TC-TTGAGAACGGGCCATGATGACGATGGCG--TTTTGTAATAGAAAAGGG
gr11         --ACAGC--TCATTGAGAACGGGCCGGGATAAAAA-------------AAAAAAAAAAAA
Consensus    --ACAGC--TCATTGAGAACGGGCCATGATGACAATGGCGGTTTTGTGGAATAGAAAGGG

                            JS902+ GAGAACGGGCCATGATGAC
```

Figure 17. Schematic representation of the SINE-R specific primer used for 3’ RACE. The sequences of groups 1-10 represent the consensus sequences which were inferred by aligning the SINE-R region of each group’s member elements. Sequence of group 11 represents the SINE-R region of its one member element. Highlighted in green is the final consensus sequence inferred by aligning the consensus sequences of groups 1-11. Primer JS902+ was designed based on the consensus sequence highlighted in green.

JS902+ and Anchor primers were used for amplification of SVAs from human testis DNA. The results are depicted in figure 18. The approximately 520 bp amplicon indicates that full length SINE-R SVA sequences have been amplified from cDNA.

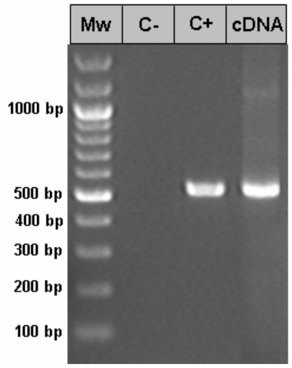


Figure 18. 3' RACE amplification products obtained from human testis cDNA. The positive control (C+) represents the full length SINE-R region of SVA H10_1 (Damert *et al.*, 2009, Hancks *et al.*, 2009). C(-) – negative control (no template); Mw - GeneRuler 100 bp.

2.2.2 3' intact SVA transcripts' origin and subfamily affiliation

Cloning and sequencing of the approximately 520 bp cDNA amplification product yielded 14 full-length SINE-R SVA sequences which were mapped to ten genomic SVA-containing loci (Table 2). Seven out of these 10 SVAs were SVA_A, B, C subfamily members and while the remaining three SVAs are affiliated to the SVA_D subfamily. No SVA_E or SVA_F subfamily sequences have been identified. This does not exclude the possibility that such elements are expressed in the human testis. SVA_E and SVA_F elements comprise only 4.4% and 9.5% of the total number of SVA elements in the human genome, respectively (Wang *et al.*, 2005). Their representation relative to the other subfamilies might therefore not have been sufficient to be detected in the relatively small number of analyzed sequences.

Table 2. 3' intact SINE_R SVA transcripts expressed in human testis

Sequence name	Source element	Subfamily affiliation
IB1L1	H2_361	SVA_A
HB1L3	H5_858	SVA_A
IB1L3		
HB1L1	H13_A287	SVA_A
S11	H21_1992	SVA_A
HB1L4		
S3	H6_978	SVA_B
S2	H7_1225	SVA_B
S5	H7_1211	SVA_C
S16		
IB1L5		
S7	H1_179	SVA_D
S14	H5_873	SVA_D
S15	H5_859	SVA_D

2.2.3 Expression, characteristics, source elements and poly(A) signal utilization of 3’ truncated SVA transcripts

Because no visible amplicons corresponding to 3’ truncated SVA elements were detected by electrophoresis (Figure 17), a “shotgun” procedure was used. Gel pieces corresponding to the 400 bp, 300 bp and 200 bp bands of the DNA ladder (Mw) were cut out, purified, and the purified products were cloned and sequenced. From the total number of 21 SVA sequences obtained (Table 3), one SINE-R sequence with a length of 415 bp harbored an 84 bp internal deletion. The remaining 20 SVA sequences were 3’ truncated with a SINE-R length ranging from 73 to 261 bp, with an average value of 170 bp. One of them displays a *de novo* 3’ transduction, whereas two transcripts carry 3’ L1PA fusions (see discussions at page 17-18).

Table 3. SINE_R truncated SVA transcripts expressed in human testis

Sequence name	**SINE-R length (bp)**	**Characteristics**	**Source element 3’ truncated/intact**		**Subfamily affiliation**	**Putative poly(A) signal**
IB3L4	133	3’ truncated	H2_434	3’ intact	SVA_B	TGTAGA/AGTAGA
IB3L10	83		H2_418		SVA_C	AATACA/GATAGA
IB4L1	171		H19_105		SVA_C	AAGAAA
IB3L6	171		H2_456		SVA_D	AAGAAA
IB2L10	171		H19_17		SVA_D	AAGAAA
IB2L1	73		H19_70		SVA_D	AATACA/GATAGA
IB4L10	261					GTTAAA
IB3L3	220	3’ truncated	H7_1026	3’ truncated	SVA_A	TCTACA
IB3L5	149		H12_A190		SVA_A	TGTAGA/AGTAGA
IB3L7	149					
IB3L9	149					
IB2L9	253		H11_A52		SVA_D	GTTGAA
IB3L1	253					
IB4L5	171		H14_A405		SVA_D	AAGAAA
IB4L7	171		H16_A573		SVA_D	AAGAAA
IB3L2	170	3’ truncated	UD	UD	UD	AAGAAA
IB3L8	173					
IB2L6	161	3’ truncated; 3’ transduction	H14_A413	3’ truncated	SVA_C	3’ transduction-derived AAGAAA
IB2L2	161	3’ truncated; 3’ fusion to L1	H14_A378	3’ truncated; 3’ fusion to L1	SVA_D	L1-derived AATAAA
IB2L4	161					
IB4L3	415	internal SINE-R deletion	H20_1873	internal SINE-R deletion; 3’ transduction	SVA_D	AATAAA

UD – undetermined

18 out of 20 3’ truncated SVA retrieved sequences were mapped to 13 genomic SVAs, while the remaining two sequences could not be traced to a single genomic SVA source element (Table 3). Analysis of the loci of origin in the genome revealed that six out of 13 elements have a full length SINE-R region. The remaining seven genomic SVA copies are 3’ truncated in their SINE-R region. Regarding the subfamily affiliation of the 13 genomic SVA copies, seven were SVA_D, three were SVA_C, two were SVA_A and one was an SVA_B element. As in the case of 3’ intact transcripts, no SVAs affiliated to E and F subfamilies have been identified.

In order to assess, which poly(A) signals drive polyadenylation at alternative poly(A) sites in the SINE-R region of 3’ truncated SVA transcripts expressed in human testis, the three transcripts displaying 3’ transductions and fusions (Table 3) were excluded from this analysis. Most likely the poly(A) signal of the transductions/fusions was used for polyadenylation of these SVA transcripts.

For the remaining 17 cloned 3’ truncated SVA transcripts, the identified putative poly(A) signals were categorized in the same manner as for 3’ truncation groups characterized in the human genome. They are mainly represented by hexamer signals that are also potentially used by the 3’ truncated SVA found in the human genome reference sequence (Table 4). However new putative poly(A) signals could be identified as well. One such variant is the AATACA signal which is potentially used by an SVA_C element of group 5. This signal has arisen by substitution of the G residue of the AATAGA hexamer found in the subfamily C consensus sequence. The second new putative poly(A) signal, GGTGAA, has a G substitution at the fourth position of its counterpart GGTAAA signal.

Polyadenylation driven by these putative poly(A) signals (Table 4) is largely the same as in the case of 3’ truncated SVA from the human genome reference sequence. The positioning of the cleavage and polyadenylation sites relative to the poly(A) signals is the same for a given poly(A) signal. Only in three cases, the cleavage and polyadenylation site was positioned different from the one observed for the same poly(A) signal in the 3’ truncated SVAs from the human genome reference sequence (Figure 19). However, for both 3’ truncated SVAs from human genome reference sequence and expressed in human testis, the cleavage and polyadenylation sites, and therefore the poly(A) tail is found at least 4 bases downstream of the putative non-canonical poly(A) signal. This is in contrast to L1 elements, in which the canonical AATAAA poly(A) signal is immediately followed by the poly(A) tail (Belancio *et al.*, 2007). Therefore non-canonical poly(A) signals might represent substrates for different cellular polyadenylation factors than the ones used by canonical poly(A) signals. This might promote polyadenylation by a different mechanism than in the case of canonical poly(A) signals.

Table 4. Putative poly(A) signals used for polyadenylation of 3' truncated SVA transcripts

Group	**Putative poly(A) signals**			
	SVA_A	**SVA_B**	**SVA_C**	**SVA_D**
2			AATACA GATAGA (1)	AATAGA GATTGA (1)
3	TGTAGA AGTAGA (3)	TGTAGA AGTAGA (1)		
4			AAGAAA (2)	AAGAAA (5)
5	TCTACA (1)			
6				GTTAAA (1)
				GTTGAA (2)

The putative poly(A) signals are indicated depending on the subfamily affiliation of each SVA transcript

Numbers in brackets indicate the number of SVA transcripts using that specific putative poly(A) signal

NNNNNN – putative poly(A) signals that differs from the subfamily consensus sequence by a single base substitution

NNNNNN – putative poly(A) signals which differ from the consensus poly(A) signal NNUANA, but are consistent with the subfamily consensus sequence

NNNNNN – putative poly(A) signals which differ by a single base substitution from the consensus poly(A) signal NNUANA and from the subfamily consensus sequence

The poly(A) tail lengths for the cloned 3' truncated SVA transcripts expressed in the human testis, ranged from 13 to 126 bases (Table 5). This high variation can be attributed to internal priming during reverse transcription. No correlation could be established between the length of the poly(A) tail and the length of the SINE-R region.

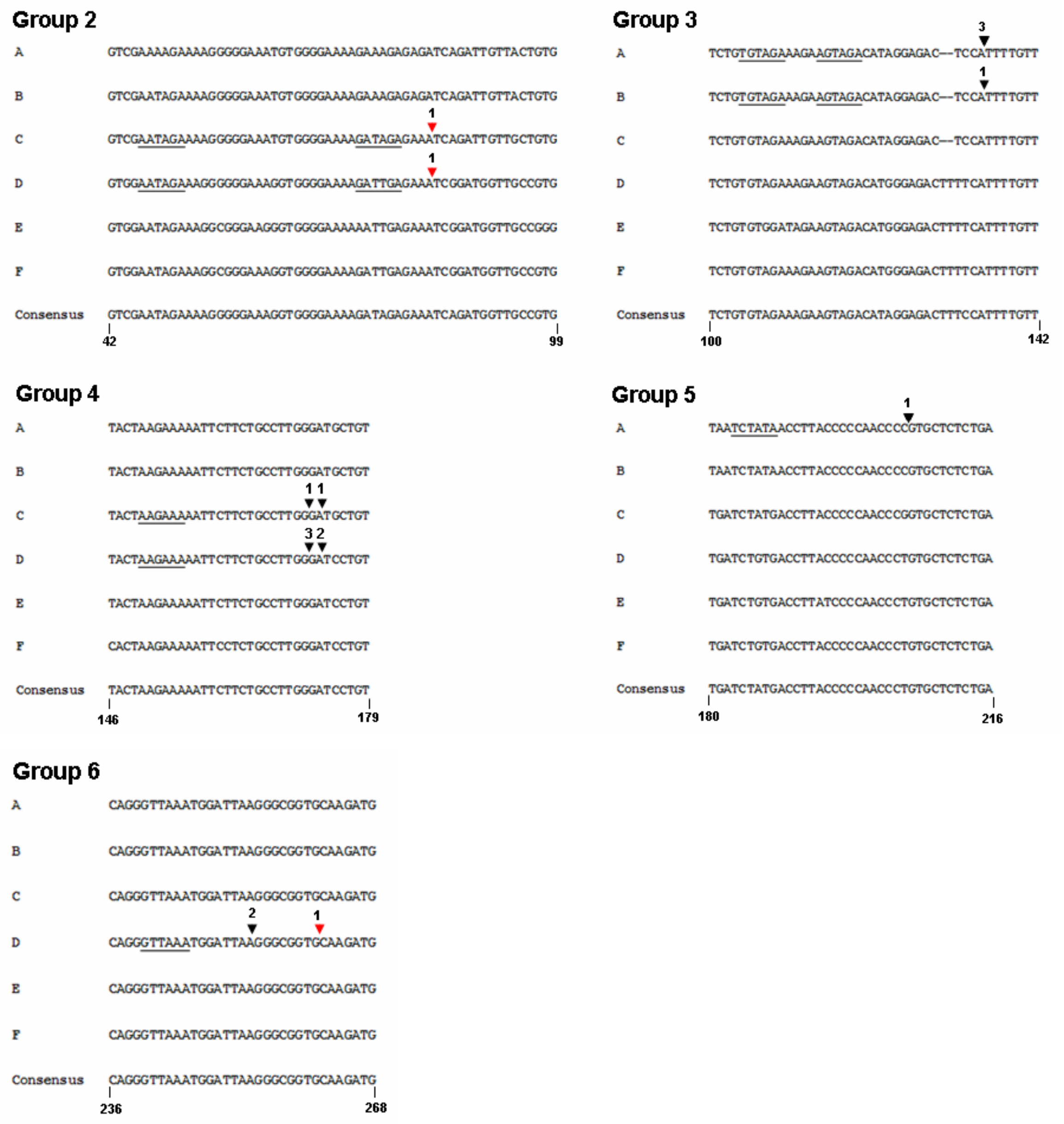

Group 2
A GTCGAAAAGAAAAGGGGGAAATGTGGGGAAAAGAAAGAGAGATCAGATTGTTACTGTG
B GTCGAATAGAAAAGGGGGAAATGTGGGGAAAAGAAAGAGAGATCAGATTGTTACTGTG
1
C GTCGAATAGAAAAGGGGGAAATGTGGGGAAAAGATAGAGAAATCAGATTGTTGCTGTG
1
D GTGGAATAGAAAGGGGGGAAAGGTGGGGAAAAGATTGAGAAATCGGATGGTTGCCGTG
E GTGGAATAGAAAGGCGGGAAGGGTGGGGAAAAAATTGAGAAATCGGATGGTTGCCGGG
F GTGGAATAGAAAGGCGGGAAAGGTGGGGAAAAGATTGAGAAATCGGATGGTTGCCGTG
Consensus GTCGAATAGAAAAGGGGGAAAGGTGGGGAAAAGATAGAGAAATCAGATGGTTGCCGTG
42
99
Group 3
3
A TCTGTGTAGAAAGAAGTAGACATAGGAGAC--TCCATTTTGTT
1
B TCTGTGTAGAAAGAAGTAGACATAGGAGAC--TCCATTTTGTT
C TCTGTGTAGAAAGAAGTAGACATAGGAGAC--TCCATTTTGTT
D TCTGTGTAGAAAGAAGTAGACATGGGAGACTTTTCATTTTGTT
E TCTGTGTGGATAGAAGTAGACATGGGAGACTTTTCATTTTGTT
F TCTGTGTAGAAAGAAGTAGACATGGGAGACTTTTCATTTTGTT
Consensus TCTGTGTAGAAAGAAGTAGACATAGGAGACTTTCCATTTTGTT
100
142
Group 4
A TACTAAGAAAAATTCTTCTGCCTTGGGATGCTGT
B TACTAAGAAAAATTCTTCTGCCTTGGGATGCTGT
1 1
C TACTAAGAAAAATTCTTCTGCCTTGGGATGCTGT
3 2
D TACTAAGAAAAATTCTTCTGCCTTGGGATCCTGT
E TACTAAGAAAAATTCTTCTGCCTTGGGATCCTGT
F CACTAAGAAAAATTCCTCTGCCTTGGGATCCTGT
Consensus TACTAAGAAAAATTCTTCTGCCTTGGGATCCTGT
146
179
Group 5
1
A TAATCTATAACCTTACCCCCAACCCCGTGCTCTCTGA
B TAATCTATAACCTTACCCCCAACCCCGTGCTCTCTGA
C TGATCTATGACCTTACCCCCAACCCGGTGCTCTCTGA
D TGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGA
E TGATCTGTGACCTTATCCCCAACCCTGTGCTCTCTGA
F TGATCTGTGACCTTACCCCCAACCCTGTGCTCTCTGA
Consensus TGATCTATGACCTTACCCCCAACCCTGTGCTCTCTGA
180
216
Group 6
A CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
B CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
C CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
2 1
D CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
E CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
F CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
Consensus CAGGGTTAAATGGATTAAGGGCGGTGCAAGATG
236
268

Figure 19. Alternative cleavage and polyadenylation sites driven by putative poly(A) signals in SVA transcripts expressed in human testis. The putative poly(A) signals (underlined) and their corresponding cleavage and polyadenylation sites (arrow heads) are indicated in accordance with the 3' truncation groups characterized for SVA elements from the human genome (see figure 11). The red arrow heads indicate the cleavage and polyadenylation sites that differ from the ones observed in the 3' truncation groups. Numbers indicate the number of SVA transcripts using that particular cleavage and polyadenylation site. The sequence limits are indicated in the consensus sequence.

Also the length of the poly(A) tail seems to be independent of the type of poly(A) signal or the subfamily affiliation of the element which generated the transcript. In four

SVA transcripts, the poly(A) tail lengths were in range of 81 to 126 bases, which exceeds the ones observed in genomic SVA 3' truncated elements, in which the largest poly(A) tail observed was 66 bases in length. It could be that these genomic SVA copies might have had larger poly(A) tails at the time of integration and suffered progressive shortening over time (Chen *et al*., 2005 and reference therein).

Table 5. **Poly(A) length in 3' truncated SVA cloned transcripts**

Sequence name	SINE-R length (bp)	Putative poly(A) signal	Poly(A) tail length (bp)	Source element	Subfamily affiliation
IB2L1	73	AATAGA/GATTGA	123	H19_70	SVA_D
IB3L10	83	AATACA/GATAGA	100	H2_418	SVA_C
IB3L4	133	TGTAGA/AGTAGA	81	H2_434	SVA_B
IB3L5	149	TGTAGA/AGTAGA	13	H12_A190	SVA_A
IB3L7	149	TGTAGA/AGTAGA	18		
IB3L9	149	TGTAGA/AGTAGA	14		
IB2L10	171	AAGAAA	126	H19_17	SVA_D
IB3L2	170	AAGAAA	16	undetermined	undetermined
IB3L6	171	AAGAAA	17	H2_456	SVA_D
IB4L1	171	AAGAAA	15	H19_105	SVA_C
IB4L5	171	AAGAAA	16	H14_A405	SVA_D
IB4L7	171	AAGAAA	25	H16_A573	SVA_D
IB3L8	173	AAGAAA	13	undetermined	undetermined
IB3L3	220	TCTACA	16	H7_1026	SVA_A
IB4L10	261	GTTAAA	16	H19_70	SVA_D
IB2L9	253	GTTGAA	14	H11_A52	SVA_D
IB3L1	253	GTTGAA	14		

Conclusions

1. The retrotransposition efficiency of SVA elements has been hypothesized to be enhanced by transduced Alu elements (Damert *et al*., 2009). Such an assumption requires experimental validation. This has been partly addressed by construction of two vector sets for cell-based retrotransposition assays. One vector set contains an SVA element with or without a 3' transduced Alu element on the plus strand. The second vector set contains an SVA element with or without a 5' transduced Alu element on the minus strand. Therefore, testing of these vector sets in cell-based assays should provide evidence that transduced Alu elements indeed enhance the retrotransposition efficiency of SVAs and should help to elucidate which orientation of the Alu element is more relevant for L1-mediated mobilization of SVA elements.

2. A complete inventory of annotated 3' truncated SVA in the human genome reference sequence has been established. The majority of these elements have a poly(A) tail at their 3' end, indicating that premature polyadenylation might be the mechanism responsible for the 3' truncation events. Sequence analysis indicated that polyadenylation mostly occurs at preferred poly(A) sites and potential upstream enhancer sequences might be responsible for this bias.

The genome-wide screening for 3' truncated SVAs had revealed novel structural variants which are characterized by internal deletions in the SINE-R region and by retrotransposon insertions within the SINE-R region. The latter ones indicate that SVA elements contain target sites for L1 endonuclease within the SINE-R region.

Identification of polyadenylated 3' truncated SVA transcripts expressed in human testis provides evidence that premature polyadenylation is indeed the mechanism responsible for the 3' truncation events observed in genomic SVAs carrying poly(A) tails. The poly(A) sites used by the 3' SVA expressed transcripts in the human testis resemble the ones observed in the genomic 3' truncated SVAs.

Finally, experimental evidence has been provided to support of transcription of the older SVA A, B and C subfamilies in human testis, indicating that they are still capable of impacting our genome.

References

Beaudoing, E., S. Freier, J.R. Wyatt, J.M. Claverie, and D. Gautheret. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Research* **10:** 1001-1010.

Belancio, V.P. and P. Deininger. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nature Genetics* **35:** 363-366.

Belancio, V.P., D.J. Hedges, and P. Deininger. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Research* **18:** 343-358.

Belancio, V.P, M. Whelton, and P. Deininger. 2007. Requirements for polyadenylation at the 3' end of LINE-1 elements. *Gene* **390:** 98-107.

Bennett, E.A., H. Keller, R.E. Mills, S. Schmidt, J.V. Moran, O. Weichenrieder, and S.E. Devine. 2008. Active Alu retrotransposons in the human genome. *Genome Research* **18:** 1875-1883.

Borson, N.D., W.L. Salo, and L.R. Drewes. 1992. A lock-docking oligo(dT) primer for 5' and 3' RACE PCR. *PCR Methods and Applications* **2:** 144-148.

Chen, J.-M., P.D. Stenson, D.N. Cooper, and C. Ferec. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Human Genetics* **117:** 411-427.

Cordaux, R., and M.A. Batzer. 2009. The impact of retrotransposons on human genome evolution. *Nature Reviews. Genetics* **10**: 691-703

Damert, A., J. Raiz, A.V. Horn, J. Lower, H. Wang, J. Xing, M.A. Batzer, R. Lower, and G.G. Schumann. 2009. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Research* **19:** 1992-2008.

Danckwardt, S., I. Kaufmann, M. Gentzel, K.U. Foerstner, A.-S. Gantzert, N.H. Gehring, G. Neu-Yilik, P. Bork, W. Keller, M. Wilm et al. 2007. Splicing factors stimulate polyadenylation via USEs at non-canonical 3' end formation signals. *The EMBO Journal* **26:** 2658-2669.

Dewannieux, M., C.C. Esnault, and T. Heidmann. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics* **35:** 41-48.

Dewannieux, M. and T. Heidmann. 2005. Role of poly(A) tail length in Alu retrotransposition. *Genomics* **86:** 378-381.

Esnault, C., J. Maestre, and T. Heidmann. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nature Genetics* **24:** 363-367.

Gogvadze, E.V., and A.A. Buzdin. 2005. New mechanism of retrogene formation in mammalian genomes: in vivo recombination during RNA reverse transcription. *Molekuliarnaia Biologiia* **39**: 364-373.

Hancks, D.C., A.D. Ewing, J.E. Chen, K. Tokunaga, and H.H. Kazazian, Jr. 2009. Exon-trapping mediated by the human retrotransposon SVA. *Genome Research* **19:** 1983-1991.

Hancks, D.C. and H.H. Kazazian, Jr. 2010. SVA retrotransposons: Evolution and genetic instability. *Seminars in Cancer Biology* **20:** 234-245.

Hancks, D.C, J.L. Goodier, P.K Mandal, L.E. Cheung, and H.H. Kazazian, Jr. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Human Molecular Genetics* **20**: 3386-3400

Kim, H.S., B.H. Hyun, J.Y. Choi, and T.J. Crow. 2000. Phylogenetic analysis of a retroposon family as represented on the human X chromosome. *Genes & Genetic Systems* **75:** 197-202.

Millevoi, S. and S.p. Vagner. 2010. Molecular mechanisms of eukaryotic pre-mRNA 3' end processing regulation. *Nucleic Acids Research* **38:** 2757-2774.

Mills, R.E., E.A. Bennett, R.C. Iskow, and S.E. Devine. 2007. Which transposable elements are active in the human genome? *Trends in Genetics: TIG* **23:** 183-191.

Ono, M., M. Kawakami, and T. Takezawa. 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Research* **15:** 8725-8737.

Ostertag, E.M., J.L. Goodier, Y. Zhang, and H.H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *American Journal of Human Genetics* **73:** 1444-1451.

Raiz J., A. Damert, S. Chira, U. Held, S. Klawitter, M. Hamdorf, J. Löwer, W.H. Strätling, R. Löwer, G.G Schumann. In press. The non-autonomous retrotransposon SVA is *trans*-mobilized by the human LINE-1 protein machinery. *Nucleic Acid Research.*

Roy-Engel, A.M., A.-H. Salem, O.O. Oyeniran, L. Deininger, D.J. Hedges, G.E. Kilroy, M.A. Batzer, and P.L. Deininger. 2002. Active Alu element "A-tails": size does matter. *Genome Research* **12:** 1333-1344.

Shen, L., L.C. Wu, S. Sanlioglu, R. Chen, A.R. Mendoza, A.W. Dangel, M.C. Carroll, W.B. Zipf, and C.Y. Yu. 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *The Journal of Biological Chemistry* **269:** 8466-8476.

Venkataraman, K., K.M. Brown, and G.M. Gilmartin. 2005. Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition. *Genes & Development* **19:** 1315-1327.

Wang, H., J. Xing, D. Grover, D.J. Hedges, K. Han, J.A. Walker, and M.A. Batzer. 2005. SVA elements: a hominid-specific retroposon family. *Journal of Molecular Biology* **354:** 994-1007.

Xing, J., H. Wang, V.P. Belancio, R. Cordaux, P.L. Deininger, and M.A. Batzer. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proceedings of the National Academy of Sciences of the United States of America* **103:** 17608-17613.