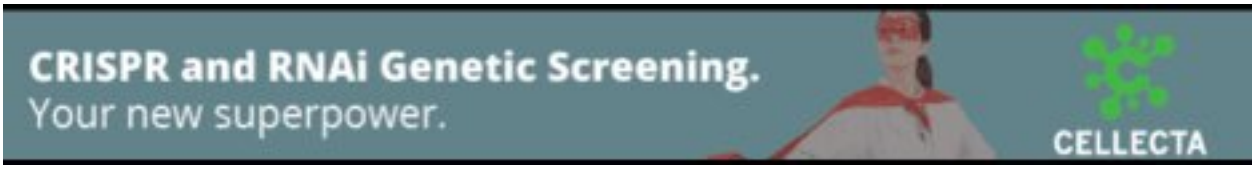
5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome

Annette Damert, Julija Raiz, Axel V. Horn, et al.

Genome Res. 2009 19: 1992-2008 originally published online August 3, 2009
Access the most recent version at doi:10.1101/gr.093435.109



Letter

5′-Transducing SVA retrotransposon groups spread efficiently throughout the human genome

Annette Damert,[1,2,6] Julija Raiz,[1,6] Axel V. Horn,[1] Johannes Löwer,[1] Hui Wang,[3] Jinchuan Xing,[4] Mark A. Batzer,[5] Roswitha Löwer,[1] and Gerald G. Schumann[1,6,7]

[1]*Fachgebiet PR2/Retroelemente, Paul-Ehrlich-Institut, D-63225 Langen, Germany;* [2]*Institute for Interdisciplinary Experimental Research, Molecular Biology Center, Babes-Bolyai-University Cluj-Napoca, RO-400271 Cluj-Napoca, Romania;* [3]*Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA;* [4]*Department of Human Genetics, Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah 84112, USA;* [5]*Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA*

SVA elements represent the youngest family of hominid non-LTR retrotransposons, which alter the human genome continuously. They stand out due to their organization as composite repetitive elements. To draw conclusions on the assembly process that led to the current organization of SVA elements and on their transcriptional regulation, we initiated our study by assessing differences in structures of the 116 SVA elements located on human chromosome 19. We classified SVA elements into seven structural variants, including novel variants like 3′-truncated elements and elements with 5′-flanking sequence transductions. We established a genome-wide inventory of 5′-transduced SVA elements encompassing ~8% of all human SVA elements. The diversity of 5′ transduction events found indicates transcriptional control of their SVA source elements by a multitude of external cellular promoters in germ cells in the course of their evolution and suggests that SVA elements might be capable of acquiring 5′ promoter sequences. Our data indicate that SVA-mediated 5′ transduction events involve alternative RNA splicing at cryptic splice sites. We analyzed one remarkably successful human-specific SVA 5′ transduction group in detail because it includes at least 32% of all SVA subfamily F members. An ancient retrotransposition event brought an SVA insertion under transcriptional control of the *MAST2* gene promoter, giving rise to the primal source element of this group. Members of this group are currently transcribed. Here we show that SVA-mediated 5′ transduction events lead to structural diversity of SVA elements and represent a novel source of genomic rearrangements contributing to genomic diversity.

[Supplemental material is available online at http://www.genome.org.]

SVA elements are nonautonomous non–long terminal repeat (non-LTR) retrotransposons that originated <25 million years ago (Mya) and represent the youngest retrotransposon family in primates. So far their copy number has increased to roughly 3000 in the human genome (Ostertag et al. 2003; Wang et al. 2005). Their ongoing activity in humans sporadically causes disease by randomly inserting into genes (for review, see Belancio et al. 2008). Considering the number of disease-causing insertions relative to their overall copy number, SVA elements are thought to represent a highly active retrotransposon family in humans (Ostertag et al. 2003; Wang et al. 2005). SVA elements stand out from the group of human non-LTR retrotransposons due to their composite structure, including modules derived from other primate repetitive elements. Starting at the 5′-end, a full-length SVA element is composed of a $(CCCTCT)_n$ hexamer repeat region; an *Alu*-like region consisting of three antisense *Alu* fragments adjacent to an additional sequence of unknown origin; a variable number of tandem repeats (VNTR) region, which is made up of copies of a 36- to 42-bp sequence or of a 49- to 51-bp sequence (Ostertag et al. 2003); and a short interspersed element of retroviral origin (SINE-R) region. The latter is derived from the 3′-end of the *env* gene and the 3′-LTR of the endogenous retrovirus HERVK-10 (Ono et al. 1987). A poly(A) tail is positioned downstream of the predicted conserved polyadenylation signal AATAAA (Ostertag et al. 2003).

The origin of SVA elements can be traced back to the beginnings of hominid primate evolution, only about 18–25 Mya. Their very young evolutionary age represents a unique opportunity to study the entire evolutionary history of a human retrotransposon. In addition, SVA elements may be valuable as markers for primate or human phylogenetic and population genetic studies, as has been the case for *Alu* elements (Bamshad et al. 2003; Watkins et al. 2003; Xing et al. 2007).

In order to draw conclusions on both the assembly process that led to the formation of SVA elements and the mechanism of their transcriptional regulation, we started out by assessing the different structural variants of SVA elements on a single chromosome. Thorough investigation of a single chromosome allows detailed structural inspection of individual elements. We chose chromosome 19 for the following reasons: (1) it is one of the best characterized single human chromosomes with regard to DNA sequence and biology (Grimwood et al. 2004), and (2) with an average density of one element per 0.62 Mb, it was reported to have the highest density of SVA insertions among human chromosomes (Wang et al. 2005). (3) Nearly 55% of this chromosome consists of repetitive elements (Grimwood et al. 2004), whereas the genome average is only 44.8% (Lander et al. 2001). This difference is due mainly to an unusually high content of short interspersed elements (SINEs) on this chromosome. Therefore, we reasoned that representatives of all human SVA variants should also be present on chromosome 19.

[6]**These authors contributed equally to this work.**
[7]**Corresponding author.**
E-mail schgr@pei.de; fax 49-6103-771280.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.093435.109.

19:1992–2008 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

The comprehensive analysis of chromosome 19 uncovered 116 SVA elements that could be assorted into seven types of structural variants. Novel variants identified were 3′-truncated elements and a significant number of SVA elements with 5′-flanking sequence transductions. Our genome-wide analysis uncovered 220 5′-transducing SVA elements that originated from 93 diverse genomic source loci. The 93 resultant 5′ transduction groups are composed of one to 84 members. The organization of these elements demonstrates that SVA elements can recruit external heterologous promoters for their own transcriptional regulation, which in turn might affect mobilization frequencies. In a number of cases, 5′-transduced sequences are constituted by spliced cellular mRNAs. Therefore, SVA-mediated 5′ transduction represents a further mechanism of retrotransposon-mediated exon shuffling. SVA-driven transduction of 5′-flanking DNA is likely to be one mechanism of genome evolution via increasing genome plasticity and facilitating new combinations of coding and regulatory sequences. Acquisition of new 5′ sequences may also be a common theme in SVA evolution, with the objective of identifying new genomic components whose incorporation increases the efficiency of SVA mobilization.

Results

Chromosome 19 harbors seven structural variants of SVA elements

A comprehensive analysis of chromosome 19 of the human genome draft sequence (hg 17; May 2004) recovered 116 insertions that are members of the SVA family of non-LTR retrotransposons accounting for roughly 179 kb of genomic sequence. Except for one SVA element, all remaining copies were characterized by hallmarks of LINE-1–mediated retrotransposition: flanking target site duplications (TSDs) of 4–18 bp (one exception, 42 bp), a variable-length poly(A) tail of 2–72 nucleotides following putative polyadenylation signals, and a consensus target sequence AA↑TTTT. Poly(A) tails were composed of patterned repeats (Szak et al. 2002) in 60 cases, while homogenous poly(A) tails were found in 56 elements. Detailed features of each SVA insertion on chromosome 19 are listed in Supplemental Table 1.

Based on their SINE-R sequences (Wang et al. 2005), we were able to allocate 105 SVA elements to subfamilies A to F. Eleven elements could not be assigned unambiguously to any SVA subfamily because 3′ truncations of the SINE-R regions were too extensive or the SINE-R sequence was missing entirely. SVA_D represents the largest subfamily, accounting for about 44.8% of SVA elements on chromosome 19. The second and third largest subfamilies B and C comprise roughly 14.7 and roughly 12%, respectively. Human-specific subfamilies E and F are represented by six (5.1%) and 13 (11.2%) members, respectively.

Due to differences in their structural organization, the 116 SVA elements can be grouped into seven types of structural variants (Fig. 1). The majority of elements are full-length (type 1, 48%) and 5′-truncated (type 2, 26%). Most of the 5′ truncation sites (19/30) are localized within VNTR sequences. Only in five and seven cases did 5′ truncation occur within the *Alu*-like and SINE-R regions, respectively. We also identified nine SVA insertions that display classical 3′ transduction events, and categorized them as type 3 elements. While these three SVA variants have been described earlier (Wang et al. 2005; Xing et al. 2006), we also found as yet unreported 3′-truncated elements (type 4) that feature the hallmarks of L1-mediated retrotransposition (Supplemental Table 1).

Type	Structural features	Copy number
1	Full-length	56
2	5' truncation	30
3	3' transduction	9
4	3' truncation	8
5	5' transduction	6
6	5' & 3' transduction	3
7	SVA2	4
	Total:	116

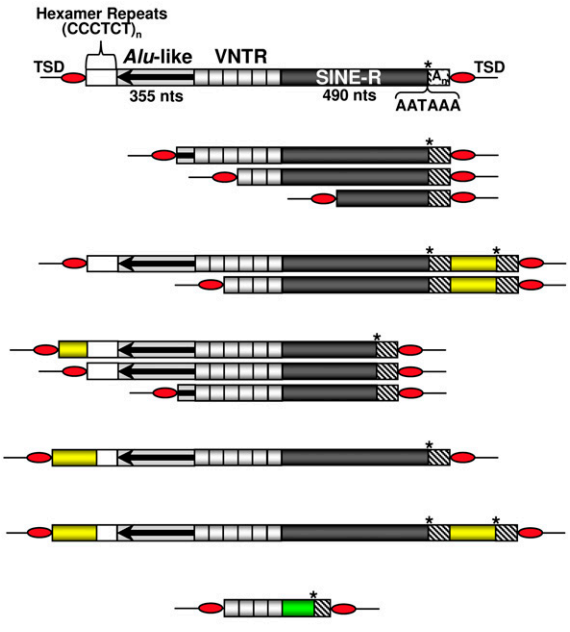


Figure 1. Structural variants of the SVA elements on human chromosome 19. Full-length SVA elements are composed of $(CCCTCT)_n$ hexamer simple repeats, an *Alu*-like region, a VNTR region, and a SINE-R sequence (type 1). Structures of 5′-truncated (type 2), 3′-transducing (type 3), 3′-truncated (type 4), 5′-transducing (type 5), 5′ as well as 3′-transducing (type 6), and SVA2 elements (type 7) are depicted. 5′- and 3′-transduced sequences are indicated as yellow boxes. Heterologous non-SVA sequences in SVA2 elements are illustrated as green box. Asterisks indicate polyadenylation signals. A_n, polyadenylate.

Two out of the eight 3′-truncated SVA elements identified are also truncated at their 5′-ends. A genome-wide analysis of 3′-truncated SVAs is the subject of a separate study (A. Damert and G.G. Schumann, in prep.). Chromosome 19 also harbors four SVA2 elements (Fig. 1, type 7) (Jurka et al. 2005; Han et al. 2007), which are composed of VNTR units fused to a DNA-sequence that has no sequence homology with any known component of SVA elements (Supplemental Fig. 1). The SVA2 elements identified display hallmarks of L1-mediated retrotransposition: They are flanked by 5- to 16-bp TSDs, and their 3′-ends are terminated by poly(A) tails of 6–14 bp.

We also located SVA elements whose 5′ TSDs adjoined to additional genomic sequences bordering on the 5′-end of the original SVA sequence (Fig. 1; type 5 and 6 variants). The structures of these variants are consistent with a previous hypothesis suggesting that at least a fraction of SVA elements may be transcribed from external promoters (Wang et al. 2005): If a cellular promoter that is localized upstream of a full-length or 5′-truncated non-LTR retrotransposon reads into the element, the consequence would be a new 5′-end for the mRNA encoding the element. Retrotransposition of such an element will result in a new genomic copy that could contain additional genomic sequences at its 5′-end that originally adjoined the 5′-end of the source element. Chromosome 19 harbors 10 SVA insertions with such 5′-transduced sequences whose lengths ranged from 24 to 849 bp (Fig. 2).

SVA-mediated 5′ transduction events are widespread and their source elements are still present in the human genome

Three out of five 5′-transducing full-length SVA elements on chromosome 19 could be traced back to their source elements (Fig. 2A). However, analysis of the genomic environment of these source elements did not uncover any putative external promoter sequences that might be responsible for the 5′ transductions. SVA H19_76 is the result of two consecutive 5′-transducing retrotransposition events with H10_10 (Fig. 2A) being the source element of H19_76. Identification of an expressed sequence tag (EST) covering the 5′ junction of H10_10 suggests that this SVA element is part of an actively transcribed region and might therefore serve currently as a source element for retrotransposition events.

Five 5′-transducing SVAs on chromosome 19 (Fig. 2B) must have originated from the same source element because (1) junctions between the 5′-ends of the truncated *Alu*-like region and the 5′ transductions are identical, and (2) 5′-transduced sequences are overlapping and can be traced back to the same source locus. 5′ Transductions comprise 76–364 bp of exon 1 of the *MAST2* (microtubule-associated serine/threonine kinase 2; Entrez GeneID 23139) gene. SVA elements that originated from this ancestral, *MAST2* sequence-transducing source element are defined as members of the *MAST2* 5′ transduction group. 5′-Ends of the *MAST2* fragments of H19_56 and H19_108 are fused to identical *Alu*Sc sequences, indicating that both SVA insertions are derived from the same source element. Three members of the *MAST2* 5′ transduction group on chromosome 19 are characterized by identical 3′-transduced *Alu*Sp sequences in addition to their *MAST2* 5′ transductions (Fig. 2B).

Genome-wide analysis for SVA 5′ transductions

The identification of 10 5′-transducing SVA elements on chromosome 19 alone suggested that transcription from external promoters might be a more widespread phenomenon. As the mechanism of transcriptional regulation of SVA elements is unclear to date and transcription is an essential prerequisite for retrotransposition, a genome-wide analysis of SVA 5′ transductions would provide important insights into the regulation of SVA expansion. To establish a complete inventory of SVA 5′ transduction events, we screened the human genome reference sequence (hg17, May 2004) using the RepeatMasker pre-masked genome annotations. We identified 2398 SVA elements with TSDs comprising ≥6 bp, and 220 of them were found to carry additional sequences at their 5′-ends (Fig. 3), covering 49.1 kb of genomic sequences. The lengths of 5′ transductions range from 14–2161 bp, with an average of 223 bp. Almost 50% of all 5′-transduced sequences cover up to 100 bp (Fig. 4). In 31 cases, the final length of the 5′ transduction could be traced back to two consecutive transduction events. Detailed information on each SVA 5′ transduction identified is provided in Supplemental Table 2.

We were able to localize the source loci for 207 out of the 220 5′-transduced sequences. Source loci were either SVA-free or they were found to be localized upstream of the 5′ TSD of an SVA source element. Transduced sequences of 77 SVA elements could be uniquely assigned to their respective source locus (simple TD groups) (Fig. 3). Diverse SVA elements whose transductions could be mapped back to the same source locus were assigned to a single multimember transduction group.

Multimember transduction groups, sequential 5′ transduction events and the acquisition of spliced RNA sequences

Four out of the 16 multimember transduction groups (Table 1) are characterized by both 5′ and 3′ transductions originating from the same source element (5q13.1, 6q24.3, 16p12.1, and 22q13.31). We found 11 multimember transduction groups genome-wide carrying exclusively 5′ transductions. They comprise two to eight individual SVA insertions each (Table 1). The *MAST2* transduction group stands out because of its exceptionally high number of 78 elements. Transduced sequences of the majority of groups are repetitive and/or mobile DNA sequences like transposons and retroelements (Table 1).

Transduction groups 6p21.2 and ZNF 487 (Fig. 5A,B) are distinguished by a primary transduction event that was followed by a secondary transduction event in which spliced RNA sequences were acquired.

The 6p21.2 transduction group comprises six members that carry exclusively 137- to 239-bp fragments of the 6p21.2 source locus (Fig. 5A). They could be direct descendants of the source element once existing at this site. Alternatively, any given transduction group member could have served as a source element of any 6p21.2 group member whose 5′ transduction is equal in length or a 5′-truncated version of its own. Two members of this group (H3_528, H3_545) acquired additional nonhomologous sequences by secondary 5′ transduction events. H3_528 transduced an *Alu* sequence from an SVA-free source locus on chromosome X. The secondary transduction of H3_545 could be traced back to SVA H17_1694 that had retrotransposed into intron 1 of the *TOM1L2* (target of myb1-like2 [chicken]) gene. Transcription of H17_1694 from the *TOM1L2* promoter, subsequent splicing of *TOM1L2* exon 1 to a cryptic splice acceptor site upstream of the SVA integration site, and retrotransposition of the resulting processed RNA led to the formation of SVA insertion H3_545 (Fig. 5A).

The primary source locus of the ZNF487 transduction group is a fragment of the MER94 transposon on chromosome 3q25.33. The transduced MER94 sequence in an extinct secondary source element must then have provided the splice acceptor site for the

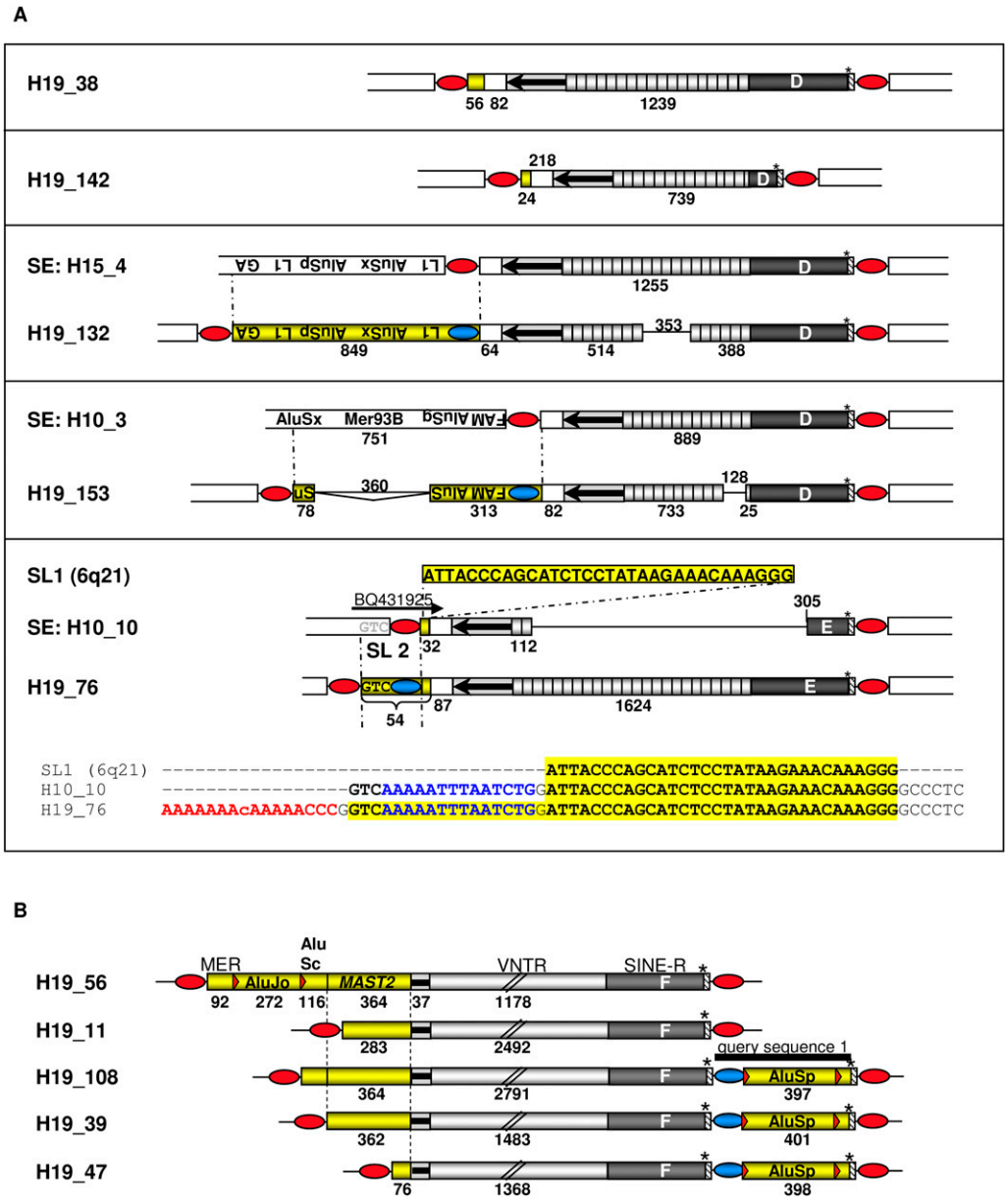


Figure 2. 5′-Transducing SVA elements on chromosome 19 and their source elements. (*A*) Full-length elements and their 5′-transduced sequences. Source elements (SEs) of SVAs H19_132, H19_153, and H19_76 could be identified and were aligned with their respective descendant SVA copy. Cryptic splice sites located within oppositely oriented *Alu* sequences (Kreahling and Graveley 2004) of the 5′-transducing pre-mRNA of source element H10_3 led to alternative splicing, which resulted in a 360-bp deletion within the 5′ transduction of H19_153. Consecutive and nested SVA-mediated 5′ transduction events generated SVA H19_76. The primary 32-bp 5′ transduction in source element H10_10 originated from the SVA-free source locus (SL) on chromosome 6q21. Transcriptional control of this source element by an external promoter is indicated by EST BQ431925 covering the junction between H10_10 SVA element and the 5′-flanking genomic sequence. Subsequent transduction of a GTC trinucleotide and of the H10_10 5′ TSD during retrotransposition resulted in the formation of the 5′-transducing H19_76 element. 5′- and 3′-transduced sequences are shown as yellow boxes. Numbers indicate lengths of VNTR regions and 5′ transductions in nucleotides. Blue ovals, TSDs of source elements; red ovals, TSDs flanking chromosome 19 SVAs; GA, low complexity GA-rich sequence; TA, $(TTAAA)_n$ repeats; MER93B, endogenous retrovirus repetitive sequence; FAM and FLAM_C, *Alu* monomer from primates (Jurka et al. 2005). (*B*) Members of the *MAST2* 5′ transduction group. Identical junctions between the 5′-ends of the truncated *Alu*-like region and the shared 5′-transduced *MAST2* sequences indicate that these SVAs originated from a common source element. MER, DNA transposon MER115; black bar, query sequence 1.

fusion to exon 1 of the *ZNF487* gene (Fig. 5B). H7_1194 is carrying the complete exon 1 sequence of the *ZNF487* gene, while the remaining three group members include only 5′-truncated versions of the exon. Because the secondary source element was under transcriptional control of the testis-specific *ZNF487* promoter, simultaneous presence of the L1-encoded protein machinery in germ cells was assured. This would have favored the amplification of this SVA transduction group in the germ line.

In multimember transduction group 17p13.3, splicing was involved in the primary transduction event. Transduced sequences of the three members of this group were mapped back to the same source locus on chromosome 17p13.3. However, examination of

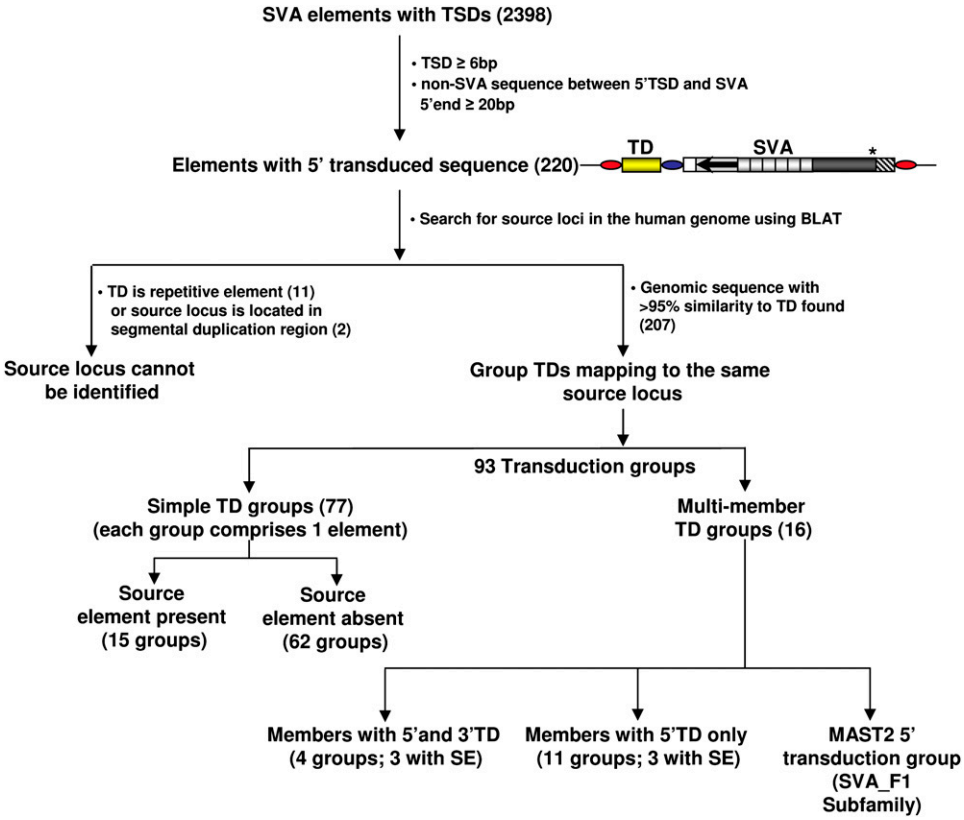


Figure 3. Schematic diagram depicting the identification of SVA 5′ transductions and their source elements. Yellow box, 5′-transduced sequence; blue oval, TSD that originally flanked the source element and is now part of the 5′ transduction; red ovals, TSDs flanking the SVA element; TD, transduction; SE, source element.

the junctions between the 5′ transduction and the 5′-end of the SVA elements and determination of their subfamily affiliation revealed that the three elements must have originated from two different source elements (Fig. 5D). 5′-Transducing SVA subfamily C members (H8_1264, H10_1542) resulted from transcription and retrotransposition of a 5′-truncated source element that is still present in the chimpanzee genome. After the human–chimpanzee divergence, an SVA_D element integrated further downstream into the same transcription unit. Cotranscription of the source locus with the SVA_D insertion, splicing to a cryptic splice acceptor site located within the SVA *Alu*-like region, and subsequent retrotransposition of the chimeric RNA led to the formation of SVA H6_1088 (Fig. 5D; Supplemental Fig. 2). *MAST2* and *RHOT1* transduction groups (Fig. 5C,E) are two additional examples for splicing of an exonic donor to an acceptor site localized within the *Alu*-like region of the SVA element. There are three genomic SVA elements that transduced spliced *RHOT1* (ras homolog gene family, member T1) sequences whose source element is still present. It has been suggested that two of these were generated by duplication and do not represent individual retrotransposition events (Hancks et al. 2009).

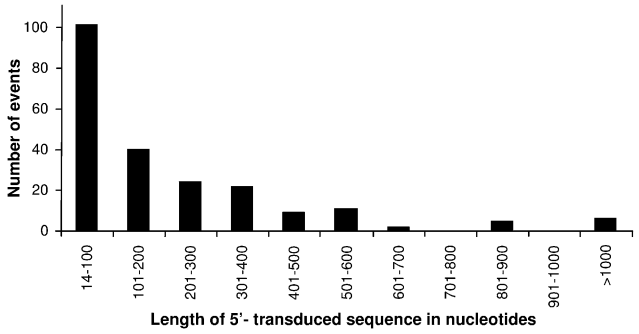


Figure 4. Length distribution of the 5′ transduction events identified. The number of SVA-mediated 5′ transduction events in the human genome in each 100-bp size interval is shown.

In total, we found 15 5′ transduction groups whose transduced sequences resulted from splicing of an exonic splice donor to a cryptic splice acceptor that is localized either upstream of an SVA 5′-end, within a primary 5′ transduction or within an SVA sequence (Table 2). Source elements of four of these 15 transduction groups are still present in the human genome. Different scenarios of SVA-mediated 5′ transduction events involving pre-mRNA splicing into SVA sequences and into sequences upstream of SVA sequences are depicted in Figure 6. In all three scenarios, the consequences of retrotransposition are 3′ truncation, 3′ modification, and translocation of host-derived RNA sequences that comprised coding sequences in several cases. Thus, SVA 5′ transduction represents an additional mechanism for exon shuffling.

The genome-wide success of the human-specific *MAST2* 5′ transduction group

The *MAST2* 5′ transduction group stands out due to its unsurpassed number of members (Fig. 5C; Supplemental Table 3). In

Table 1. Multimember 5′ transduction groups of SVA elements

5′ Transduction group	No. of members	Transduced sequence	Size range of transductions
5q13.1	3	LINE	18–26 bp
6q24.3	2	Undefined	127 bp
16p12.1	2	*Alu*	20 bp
22q13.31	2	*Alu*	88 bp
11q24.3	3	Undefined	24–30 bp
5q31.2	2	*Alu*	16–31 bp
10q26.11	2	LINE and SINE	553–593 bp
12p11.21	3	LTR9B	110–135 bp
15q21.3	4	Transposon (MER96B)	21–38 bp (–128 bp, second TD)
16p13.3	7	LTR (MER51A)	23–46 bp
GA repeat	6	Simple repeats	27–158 bp
6p21.2	8	Includes L2, spliced mRNA	137–239 bp (–438 bp)[a]
ZNF487	5	Transposon (MER94), spliced mRNA	27–32 bp (–170 bp)[a]
17p13.3	3	Spliced mRNA	220–385 bp
RHOT1	3	Spliced mRNA	491–530 bp
MAST2	78 (84)[b]	Spliced mRNA	14–364 bp (–1192 bp)[a]

[a]Secondary transductions.
[b]The number in parentheses indicates the overall number of MAST2 transduction group members, including 5′-truncated elements that lack 5′ transductions.

order to understand this exceptionally successful amplification process, we set out to investigate this group in more detail. Because a subset of the identified members of the *MAST2* 5′ transduction group shares a ~400-bp 3′ transduction including an *Alu*Sp element (Fig. 2B, query sequence 1), we used this sequence as a query to find those members of the *MAST2* - transduction subgroup that are devoid of any *MAST2* sequences as a result of 5′ truncation events. In this way we retrieved six additional SVA copies that lack any *MAST2* sequence but are clearly derived from the same 3′-transducing source element H10_1 (Fig. 7D).

Overall, we thus identified 84 SVA_F elements that are members of or derived from members of this particular *MAST2* 5′ transduction group (Fig. 7; Supplemental Table 3). It is to be expected that numerous additional genomic SVA copies originated from a *MAST2* 5′-transducing source element but cannot be identified as such because 5′ truncations might have caused the loss of any distinguishing 5′-transduced *MAST2* sequence. Thus, members of the *MAST2* 5′ transduction group represent at least 32% of the roughly 260 human-specific subfamily F members and 3% or more of all genomic SVA copies. Due to their significant number, we assigned all members of the *MAST2* 5′ transduction group to a separate subfamily termed SVA_F1. They most likely originated from one single source element that was formed and propagated after the separation of human and chimpanzee lineages about 4–6 Mya (Fig. 8).

Based on their structural organization, SVA_F1 elements can be subdivided into four major groups: Group 1 includes those elements that carry exclusively *MAST2* sequences as 5′ transduction events (Fig. 7A). In group 2, 5′ transductions contain nonhomologous sequences in addition to the *MAST2* sequence (Fig. 7B); Group 3 covers SVA elements with both *MAST2* 5′ transductions and variable 3′ transduction events (Fig. 7C). 3′-Transducing derivatives of the source element H10_1, a group 2 SVA_F1 element, constitute group 4 (Fig. 7D).

The origin of the SVA_FI subfamily

By applying an "alternative splice site prediction (ASSP) program" (Wang and Marin 2006; http://www.es.embnet.org/~mwang/assp.html), we found that in SVA_F1 elements, the splice donor site of *MAST2* exon 1 is fused precisely to a cryptic splice acceptor site in the *Alu*-like region. Therefore, it is likely that the ancestral *MAST2* SVA source element was generated by pre-mRNA splicing in *cis* or *trans* (Fig. 8A) (for *trans*-splicing, see Maniatis and Tasic 2002; Horiuchi and Aigaki 2006). Insertion polymorphism studies on a diverse human population panel suggest that this ancestral SVA_F1 source element is not present in any of the tested populations (data not shown), although sample size may be critical in the ascertainment of the element. However, the structural similarity of the *MAST2* exon 1–SVA junction to the *RHOT1* exon–SVA junction of the *RHOT1* transduction group (Fig. 5E; Table 2) whose source element is still present in the genome (Hancks et al. 2009) argues for the involvement of pre-mRNA splicing in the formation of the ancestral SVA_F1 source element. Splicing in *cis* presupposes the existence of an SVA_F element that is localized downstream of *MAST2* exon 1 within the transcriptional unit. Although such an element could not be identified in the human genome reference sequence, it might have been polymorphic for insertion presence/absence in the population at the time of initial propagation of the SVA_F1 lineage and has been lost subsequently through drift or selection. Alternatively, an SVA_F1 source element could have been generated by splicing in *trans* (Fig. 8A), as recently suggested to explain the existence of zebrafish L1–RNA fusion transcripts for which no source element could be detected (Tamura et al. 2007). Although there is also the formal possibility that retrotransposition of an SVA_F element into the exon 1–intron junction of the *MAST2* gene put the SVA element under control of the *MAST2* promoter, it is highly unlikely.

Depending on the pathway that led to the assembly of the SVA_F1 ancestral source element (Fig. 8A), it was associated either with *MAST2* exon 1 and transcribed from an external *MAST2* promoter (source element 1A) or with at least 382 bp of the 5′-transduced *MAST2* sequence and transcribed from either a nearby external promoter or from an internal promoter localized in *MAST2* exon 1 (source element 1B). As the largest *MAST2* 5′ transduction identified in the human genome encompasses 382 bp (H15_5) (Fig. 7A), a source element generated via *cis*- or *trans*-splicing must have covered at least this fraction of *MAST2* exon 1 (Fig. 8A).

The *MAST2* promoter that was most likely recruited by the ancestral SVA source element 1A for its mobilization is reported to be highly active in mammalian testicular tissue, although the promoter sequence has not been localized yet (Walden and Cowan 1993; Walden and Millette 1996). Testis-specific *MAST2* RNA increases in abundance during prepuberal testis development, peaking at the spermatid stage, and it has been suggested that the protein serine/threonine kinase activity of MAST2 plays a role in spermatid maturation (Walden and Millette 1996). Therefore, recruitment of the *MAST2* promoter would assure that the 5′-transducing SVA is coexpressed with the L1-protein machinery, which is known to be present in germ cells (Branciforte and Martin 1994; Ergun et al. 2004) and assumed to *trans*-mobilize SVA elements (Ostertag et al. 2003).

To localize any potential external or internal *MAST2* promoter that was recruited by an SVA source element, we analyzed 1000 bp of

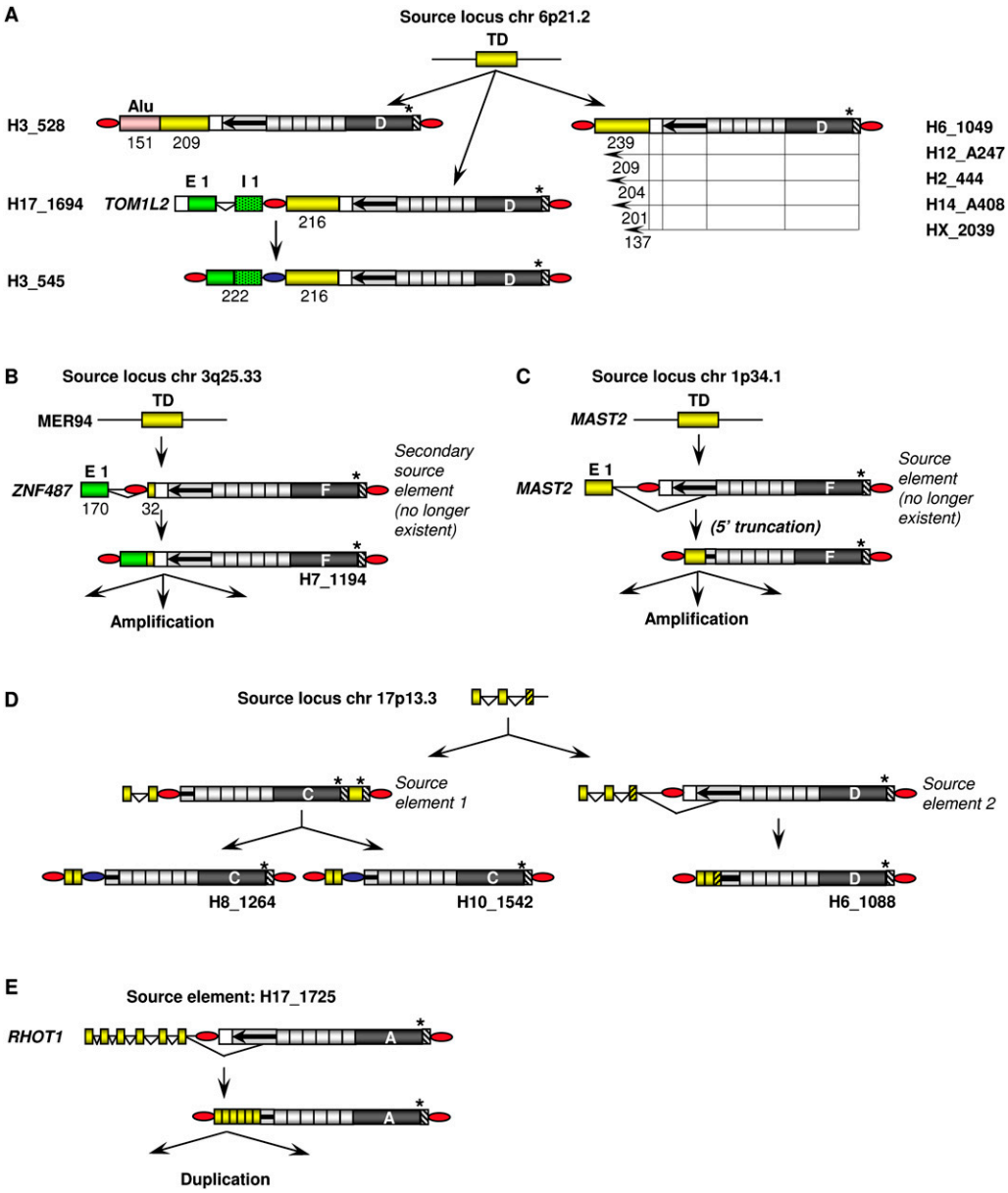


Figure 5. Multimember transduction groups with sequential 5′ transductions involving splicing of 5′-transduced RNA sequences. Transduction groups with primary transduction events followed by secondary transductions of spliced host RNAs (*A*,*B*) and those with primary transductions of spliced host RNA sequences (*C*–*E*) are shown. Splice acceptor sites can be localized either within the SVA 5′ flanking intronic sequence, within primary transductions, or within the SVA *Alu*-like region. Exons and introns are not drawn to scale. Yellow boxes, primary transductions; green and pink boxes, secondary transductions; E1, exon 1; I1, intron 1; dotted green box, intronic sequence included in 5′ transduction; red ovals, TSDs; blue ovals, TSD sequence of the source element present in the 5′-transducing SVA; TD, transduction.

genomic sequence upstream of the *MAST2* transcriptional start site as well as the *MAST2* exon 1 sequence applying different promoter prediction programs (see Methods). We found a strong external promoter candidate that was predicted to initiate transcription at position 1 of the *MAST2* gene as well as putative cryptic promoter sequences at positions 205–256 and 284–333 relative to the transcription start site (P2, P3) (Fig. 8A). The presence of such regulatory sequences in *MAST2* exon 1 is supported by analyses performed with the ESPERR regulatory potential track of the UCSC Genome Browser (King et al. 2005). These potentially regulating motifs (P2, P3) (Fig. 8A) are present in 25% of all SVA_F1 subfamily members.

No matter which pathway led to the formation of source element 1, the ancestral element and/or its offspring spread extensively afterward, picking up additional sequences at both their 5′- and 3′-ends and dispersing them throughout the genome.

Transcriptional control by multiple external promoters causes consecutive SVA_FI-mediated 5′ transduction events

In addition to *MAST2* sequences, transductions of members of SVA_F1 5′ transduction group 2 (Fig. 7B) encompass heterologous sequences. Therefore, the (now extinct) source elements of group 2

Table 2. SVA 5′ transduction events that involved RNA splicing

Locus name	SVA subfamily	Human-specific?[a]	Source element present in human genome?	Transduced sequence from gene/RNA	No. of exons included	Splice acceptor site[b]	EST consistent with observed splice pattern
Cryptic splice site activation by SVA elements							
H1_108	D	Yes	—	*PHKA1*	6	ATTTTAAACCAC**AG**	
H2_332	D	Yes	—	AK309823	3	ATTTTAATCCTT**AG**	
H3_545	D	Yes	Yes	*TOM1L2*	1	TTTTCATTTTGC**AG**	
H7_1117	F	Yes	Yes	*WDR33*	1	GATAATTTTTAC**AG**	DB154036 DA938700
H9_1377	D	Yes	—	NA	NA	CTTTCTTTTATT**AG**	BU945341
H17_1710	E	Yes	—	*MICAL3*	1[c]	TGTCCTCTGCAC**AG**	
H22_1951	D	Yes	—	*MPPE1*	4	TTTATTTTTTTC**AG**	
H3_5	F1	Yes		*NPEPPS*	1	TCTTCCTACTGC**AG**	
H8_1264; H10_1542	C C	(Yes) —	(PanTro2)[d]	NA	NA	TTCTTCCGTTTC**AG**	BM792351 CN261643
Splicing into primary 5′ transductions							
ZNF487 group	F	Yes	—	*ZNF487*	1	NA	
Splicing into SVA sequence							
H2_413	D	Yes	Yes	AK302829	2	CCTCTGATGCCA**AG** (70)	AA974432
H6_1088	D	Yes	—	NA	NA	TTTTTTTGGTGG**AG** (208)	BU941699
RHOT1 group	A	—	Yes	*RHOT1*	6	TCAGTGTTGCCC**AG** (336)	
SVA_F1 subfamily	F1	Yes	—	*MAST2*	1	CCTACACCTCCC**AG** (386)	
H16_A580	A	—	—	*PHLPP2*[e]/AK124678	4[f]	CCTCCACCTCCC**AG** (386)	DN991859

[a]Human specificity was assessed by the presence of a gap in the Genome Browser chimpanzee net track.
[b]The position of the splice acceptor AG dinucleotide (boldface) in SVA_{rep} is given in parentheses for elements resulting from splicing into SVA sequence. For these elements, the splice acceptor sequence is taken from the respective subfamily consensus in cases where the source element is no longer present in the human genome.
[c]3′ part of the exon only.
[d]*Pan troglodytes* genome build 2006.
[e]*PHLPP2* exons 3 and 4.
[f]AK124678 exon 1 and aberrantly spliced exon 2.
NA, not annotated.

members must have been transcribed from external promoters. In the case of the most successful SVA_F1 source element H10_1, the 5′ transduction could be traced back to an annotated transcription unit (AK128272) (Fig. 8B). The AK128272 promoter which was recruited by the source element confers high level expression in testis. Members of transduction group 2 (Fig. 7B) encompass heterologous sequences that are derived from (1) a common source locus 2 (Fig. 8B) on chromosome 9p13.3 (e.g., H10_1), (2) source loci covering exclusively sequences from non-LTR retrotransposons (L1, L2, *Alu*) and endogenous retroviruses like MER21A (e.g., H6_2), (3) source loci representing short fragments of non-repetitive DNA (e.g., H8_2), or (4) those containing both non-repetitive and *Alu* sequences (e.g., H3_5). A striking feature of the subsequent transductions is the large fraction of *Alu* sequences that are found to be comobilized.

SVA_F1 elements transduce 3′-flanking sequences

The genome also harbors at least 18 SVA_F1-mediated 3′ transduction events (Fig. 7C,D). Source loci of 3′ transductions of HX_4 and H9_1 (Fig. 7C) are SVA-free. Both 3′-transducing SVA_F1 subfamily members served as source elements for H3_8 and H20_17 that carry 571 and 17 bp of secondary 3′ transduction events, respectively. The source element of H6_6 (H1_F_160) is a member of SVA_F1 transduction group 1 (Fig. 7A) that is flanked by MER57A sequences (Kohany et al. 2006). Retrotransposition of an H1_F_160 transcript generated 5′ as well as 3′ transductions in SVA H6_6.

Thirteen 3′-transducing SVA_F1 members can be assigned to one single 3′ transduction group whose members are derived from source element H10_1 and are subsumed in transduction group 4 (Fig. 7D). H10_1 is one of the most active source elements identified so far, given that it generated 13 or more copies with 3′ transductions despite the fact that it belongs to the evolutionarily younger SVA subfamily F. Transcriptional termination and polyadenylation at two alternative sites in the 3′-flanking sequence of H10_1 resulted in 3′ transductions covering 400 and 479 nucleotides of the source locus. In each case, an intact *Alu*Sp element is part of the transduced sequence.

Six members of transduction group 4 are devoid of any *MAST2* fragments or additional heterologous 5′-transduced sequences because their 5′ truncation sites are localized either within the VNTR region or at the 3′-end of the SINE-R component (Fig. 7D). The rearranged H4_2 element consists of two SVA segments of which the 5′ segment covering the VNTRs and the 5′ part of SINE-R is inverted. A similar 5′-truncated and inverted SVA structure has been shown to be the cause for a case of hereditary elliptocytosis (Ostertag et al. 2003). Analogous rearrangements have been reported for L1 elements, and it was suggested that such structures are the consequence of a twin-priming mechanism by which 5′-inverted elements are formed (Ostertag and Kazazian 2001). Clearly, members of transduction group 4 are a consequence of 3′-transducing transcription of source element H10_1 and subsequent 5′ truncation occurring during reverse transcription and/or integration.

The SVA_F1 subfamily is currently transcribed in the human genome

Because transcription of potential SVA source elements is an essential prerequisite for their retrotransposition, we investigated

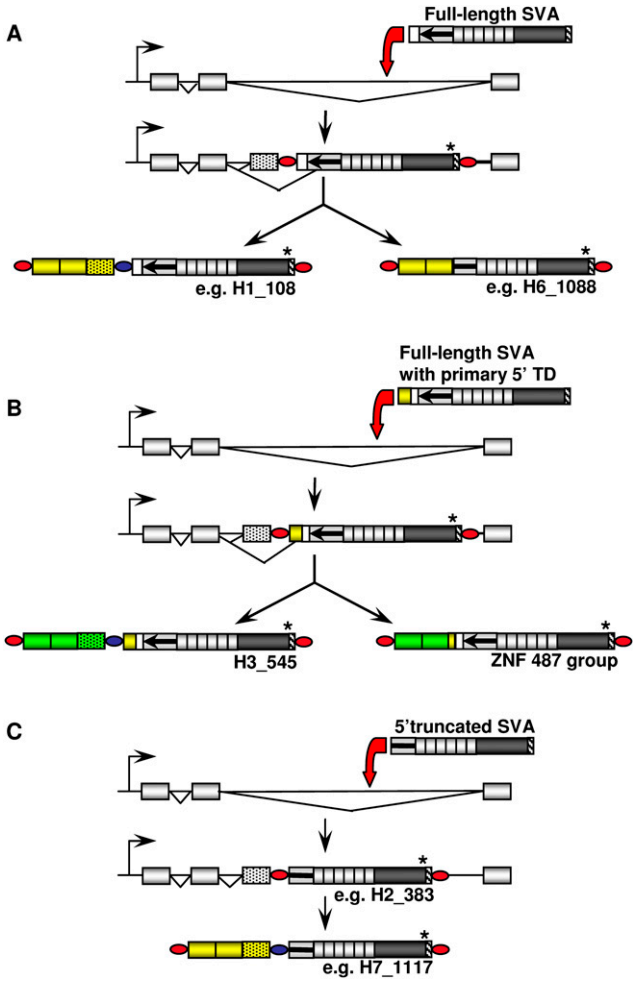


Figure 6. SVA 5′ transduction events involving pre-mRNA splicing. (*A*) After insertion of a full-length SVA element into an intron, cryptic splice acceptor sites upstream of the element can be activated. The retrotransposed copy includes both exon and intron sequences as 5′ transduction (e.g., H1_108). Alternatively, an upstream exon can be spliced to cryptic acceptor sites within the SVA sequence itself, resulting in a 5′-truncated SVA with exon sequences as 5′ transduction. (*B*) Retrotransposition of a full-length SVA with a 5′-transduced segment into intronic sequences can have two alternative outcomes: Activation of intronic splice acceptors leads to the inclusion of intron sequences (stippled box) in the secondary transduction (green boxes) (H3_545). Alternatively, the 5′-transduced sequence is fused to a splice acceptor localized in the primary transduction (e.g., ZNF487 group, cf. Fig. 5B). (*C*) After retrotransposition of 5′-truncated SVA elements into an intron (e.g., H2_383), only activation of intronic splice sites could be observed (e.g., H7_1117). A two-exon setting was chosen for the schematic representation. The actual number of exons transduced differs between elements. Yellow boxes, primary transductions; green boxes, secondary transductions; dotted green boxes, intronic sequence included in 5′ transductions; red ovals, TSDs; blue ovals, TSD sequence of the source element present in the 5′-transducing SVA; TD, transduction.

the current transcriptional activity of SVA subfamily F1. In order to identify ESTs from SVA subfamily F1 that include *MAST2* sequences, we performed a BLAST search (Altschul et al. 1990) against the human EST database using a 401-bp fragment of H10_1 (Fig. 9A) as a query. The sequence covers the 364-bp *MAST2* sequence and the adjacent truncated *Alu*-like region. ESTs matching source element H10_1 or its derivatives of SVA_F1 transduction group 4 were identified by applying query sequence 2 (Fig. 9A). ESTs aligning to the source element HX_4 and its derivative H3_8 were identified from the annotations of both loci in the UCSC Genome Browser. In total, we extracted 17 ESTs that originated from transcription of members of the SVA_F1 subfamily (Fig. 9A–C; Table 3). The identification of their most likely elements of origin suggested that at least 19 out of the 78 SVA_F1 elements are currently transcribed. However, we cannot draw any conclusion on the general expression level and/or retrotransposition rate of SVA subfamily members because (1) tissues that express SVAs or support retrotransposition might not be adequately represented in the databases, and (2) transcription of different members might be variably regulated because they are controlled by different external promoters. EST data suggest that SVA_F1 members are expressed not only in reproductive tissues, stem cells, and tumor cells but also in a variety of somatic cell types and tissues.

Two ESTs matching the SINE-R–*Alu*Sp junction specific for the H10_1 source element and its descendants of transduction group 4 argue for ongoing transcription of these elements (Fig. 9A). ESTs matching the 5′ junctions of H20_1, and the source elements H1_F160 and HX_4 (Fig. 9B,C) demonstrate current transcriptional control of these SVA elements by external promoters. We confirmed transcriptional control of H20_1 by external promoter sequences by RT-PCR on total RNA isolated from a variety of human tissues with a primer pair specifically amplifying a 397-bp fragment of the 5′ junction between H20_1 5′-end and adjacent genomic sequence (Fig. 9D). H20_1-specific RT-PCR products could be detected in spleen and testis only. The identification of three ESTs (AW627827, BI325030, BX110444) mapping the junction between the 3′-end of SINE-R and 3′ transductions of both HX_4 and its derivative H3_8 in sense orientation (Fig. 9C) is consistent with transcription of both elements. This demonstrates transcriptional read-through into genomic sequences beyond the 3′-end of the SINE-R region. Transcriptional read through into the 3′-flanking genomic sequence of HX_4 that gave rise to its derivative H3_8 is documented by EST BX110444 (Fig. 9C). This was confirmed by RT-PCRs using a primer pair specifically amplifying the 3′ junction between the 3′-end of the SINE-R region of the source element HX_4 (or its derivative H3_8) and its 3′-flanking genomic sequence, which becomes a secondary transduction in H3_8 (Fig. 9C,D). HX_4/H3_8-encoded transcripts could be identified in all tissues tested except for lung and uterus. Taken together, EST data and RT-PCR demonstrate that SVA_F1 elements are transcribed in reproductive as well as in somatic tissues. Our expression data confirm both that SVA transcription is initiated from upstream promoters and that RNA polymerase can bypass the SVA element's transcriptional termination signal.

Discussion

We set out to thoroughly analyze the structural variants of SVA elements on a single chromosome because, first, we expected to find evidence for the assembly process leading to the composite retrotransposon SVA and, second, we argued that the organization of 5′-ends of SVA insertions and their genomic environment would allow conclusions for the localization of the promoter sequences controlling SVA transcription. In this study, we found SVA elements with 3′ truncations, both 5′ and 3′ truncations, and both 5′ and 3′ transductions in combination or alone. The identification of 5′-transducing SVAs on chromosome 19 led to the first comprehensive genome-wide analysis of 5′ transduction groups of a non-LTR retrotransposon. Overall we identified 220 SVA elements with 5′ transductions comprising 8% or more of all SVA elements. Our

findings show that SVA elements can use external promoters not only for their own transcriptional regulation but also for attaching new genomic sequences to their 5′-ends by means of the retrotransposition pathway leading to structural diversity of SVA elements. Furthermore, SVA-mediated 5′ transduction represents a novel mechanism contributing to structural human genomic diversity.

SVA-mediated 5′ transduction indicates transcriptional control by external promoters

5′ Transduction events were discussed for the first time in order to explain the variability observed among 5′-end sequences in different L1 subtypes (Boeke and Pickeral 1999; Lander et al. 2001). 5′ Transduction involves retrotransposon transcription from an external promoter followed by retrotransposition of the transcript containing additional heterologous sequences at its 5′-end. Following a first description of two 5′-transducing L1 copies in the human genome (Lander et al. 2001), Symer et al. (2002) provided the first experimental evidence for retrotransposition-mediated 5′ transduction events using an L1 retrotransposition reporter assay in cell culture. More recently a disease-linked L1-mediated 5′ transduction event has been reported in mice (Chen et al. 2006).

The fact that more than 8% of human SVA elements carry 5′ transductions, together with our RT-PCR results and 5′-RACE experiments performed in the Kazazian laboratory (Hancks et al. 2009), indicates that SVA elements frequently parasitize adjacent cellular promoters. The most likely explanation for this process is the intrinsic weakness of a still elusive SVA-specific internal promoter. This reasoning is supported by previous attempts to identify any SVA-internal promoter sequence using reporter gene assays which led to ambiguous results (A Damert et al., unpubl.; Hancks et al. 2009).

The identification of 93 different primary and 26 diverse secondary 5′ transduction events (Supplemental Tables 2, 3) strongly suggests that during hominid evolution there were or still are more than 100 different cellular promoters regulating transcription of downstream SVA elements in germ cells.

It is very likely that the actual number of SVA elements that resulted from retrotransposition of an SVA source element controlled by an external promoter exceeds the number of genomic SVA elements with 5′ transductions. The reason is that those 5′-truncated SVA retrotransposition events that are derived from 5′-transducing source elements will not be identified as such if the 5′ truncation site is localized downstream the 5′-transduced sequence. This is supported by the identification of six 5′-truncated SVA_F1 subfamily members that were devoid of any *MAST2* sequence but could be unambiguously assigned to this subfamily due to a specific 3′-transduced sequence (Fig. 7D).

5′ Transductions involve a variety of genomic sequences ranging from 14–2161 bp, and it is obvious that they significantly contribute to SVA and overall genomic structural diversity. Modification of the SVA RNA secondary structure by 5′-transduced cellular sequences might result in a better target for any putative *trans*-mobilizing protein machinery. Alternatively, splicing can bring an SVA element in close vicinity of transcribed, gene-internal, host-encoded RNA polymerase II promoter sequences or transcriptional regulatory sequences thereby generating a novel SVA

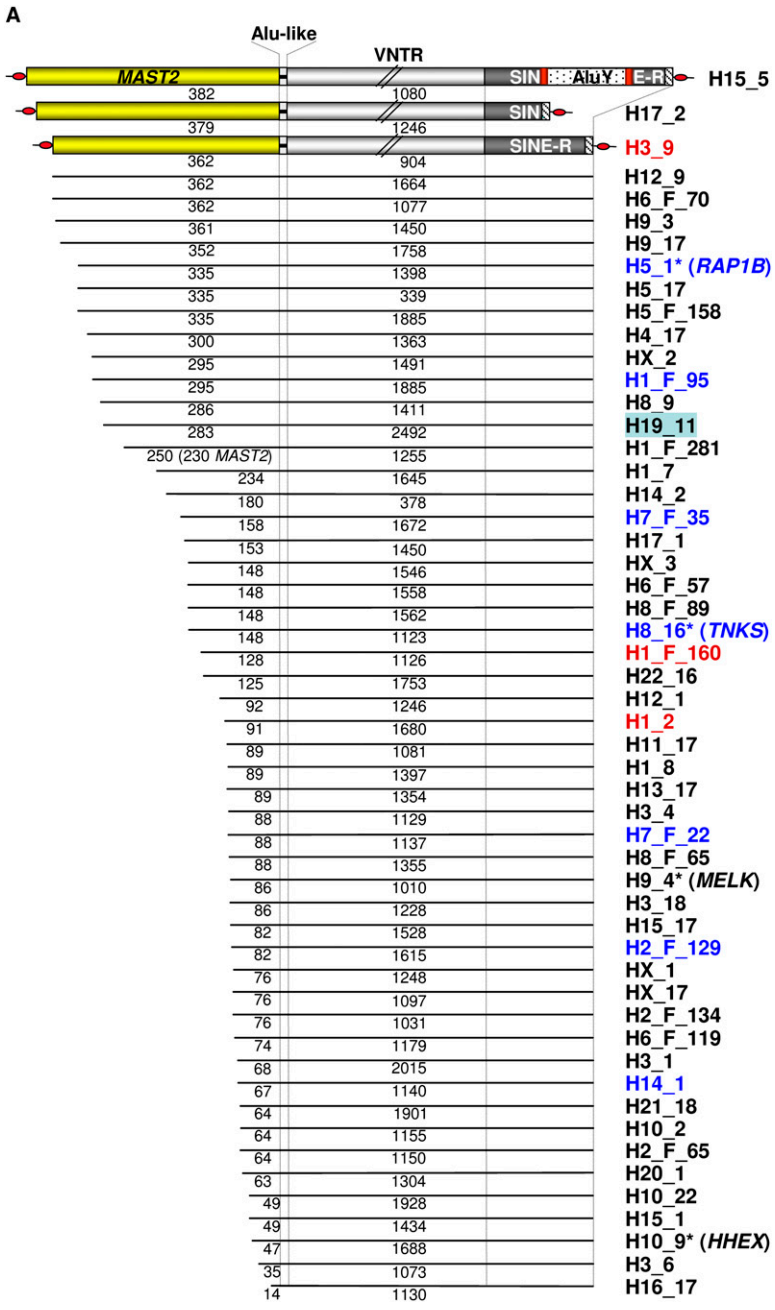


Figure 7. (Continued on next page)

subtype that might be characterized by a superior expression rate. Promoter acquisition is exemplified by the dominant A-type L1 elements in mouse which evolved from F-elements by adopting the A-type promoter (Adey et al. 1994). In both cases, 5′ transductions would be evolutionarily advantageous for the SVA element.

It is conceivable that 5′ transduction events played a role in the assembly process that led to the current organization of SVA elements. This would be supported by the structure of the "GA repeat" 5′ transduction group (Table 1; Supplemental Table 2), which is characterized by variable length additions of simple GA repeats.

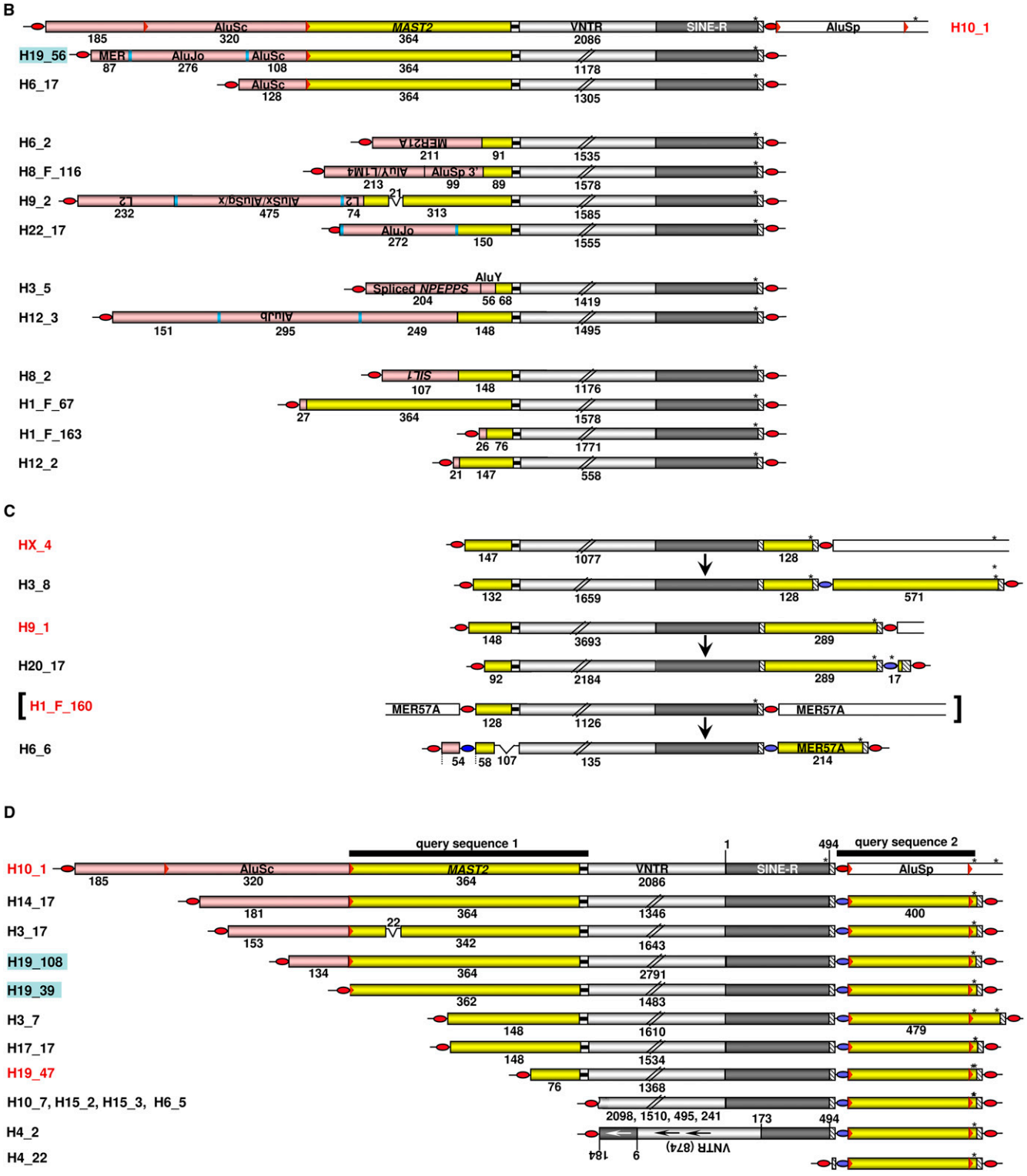


Figure 7. (Legend on next page)

The SVA hexamer repeat region might have been acquired by 5′ transduction as well. One example for the impact of structural diversity brought about by 5′ transduction events is the SVA_F1 subfamily: Transcription of a single ancestral SVA source element by the *MAST2* promoter gave rise to a whole new subfamily—with properties obviously favoring its rapid expansion in the human genome.

Exon shuffling mediated by SVA 5′ transduction

To date, SVA elements have been reported to contribute to human genome evolution through insertional mutagenesis (Kobayashi et al. 1998; Rohrer et al. 1999; Wilund et al. 2002; Ostertag et al. 2003; Conley et al. 2005; Makino et al. 2007; Takasu et al. 2007), by constituting alternative exons (Damert et al. 2004) and by exon and gene shuffling mediated by 3′ transduction (Xing et al. 2006). Similar to 3′ transduction events, the process of 5′ transduction entails the juxtaposition of two previously unlinked host genomic segments. It thus represents another mechanism of diversification of the genome via increasing genome plasticity and facilitating novel combinations of coding and regulatory sequences. Given that retrotransposition of SVA and other transduction-capable retroelements (e.g., L1) is an ongoing process in the genome, transduction events generated in this manner afford the possibility for rapid lineage-specific evolution of gene families, the proteome, and the transcriptome.

SVA-mediated 5′ transductions can be composed of genomic fragments or spliced cellular RNA sequences with functional splice acceptor sites being located either within intronic sequences or in the 5′-transducing SVA element (Fig. 6; Table 2). If RNA splicing is involved, exons of protein-coding genes or noncoding RNAs are translocated into a new regulatory and potentially also coding context. Thus, SVA-mediated 5′ transduction resembles exon shuffling by L1-mediated 3′ transduction (for review, see Boeke and Pickeral 1999). In contrast, SVA-mediated 3′ transduction has been shown to be capable of translocating entire transcriptional units, thus retaining the original regulatory context for the transduced genomic copies (Xing et al. 2006).

5′ Transduction events involving splicing at alternative/cryptic splice acceptor sites result in the modification of the 3′-end of the transduced ORF. Provided that integration of the 5′-transducing element occurs downstream of a suitable regulatory region, this could lead to expression of mutant proteins that are truncated and/or modified at their C termini. Splice acceptor sites used during SVA-mediated 5′ transduction events can be located within SVA sequences (Fig. 6A, H6_1088; Table 2), which is reminiscent of *Alu* (for review, see Kreahling and Graveley 2004) and L1 (Belancio et al. 2006; Tamura et al. 2007) elements, which have been shown to provide functional splice sites to host RNAs. However, while splicing to SVA sequences was only observed in the context of SVA-mediated exon shuffling so far, splicing of/to *Alu* (Sorek et al. 2002) and L1 sequences (Mulhardt et al. 1994; Meischl et al. 2000) was exclusively reported to lead to exonization.

The alternative mechanism resulting in the modification of the 3′-end of any SVA-transduced ORF sequence involves activation of upstream intronic cryptic splice sites (Fig. 6A, e.g., H1_108). Recent reports demonstrated activation of an upstream cryptic splice acceptor site by an *Alu* insertion into dystrophin intron 11 (Ferlini et al. 1998) and by an L1 insertion into an intron of the porcine *KPL2* gene (Sironen et al. 2006, 2007). Also in these cases, alternative splicing induced by *Alu* and L1 insertions led to exonization and not to exon shuffling as observed for SVA 5′ transductions.

Taken together, splicing into and upstream of *Alu* and L1 elements on one side and SVA elements on the other side have profoundly different consequences: *Alu* and L1 sequences are exonized within a given regulatory and coding context. In contrast, splicing of 5′ transducing SVA RNAs followed by retrotransposition leads to exon shuffling, which generates new combinations of modified coding and/or regulatory sequences at genomic integration sites.

What is the basis for the exceptional success of the SVA_F1 subfamily?

With at least 84 members, the SVA_F1 subfamily is the largest 5′ transduction group identified to date. The fact that it originated from the youngest human-specific SVA subfamily F implies that it must have reached its copy number within a relatively short period of time. There are two possible explanations for the outstanding success of the SVA_F1 subfamily:

1. SVA_F1 source elements recruited external promoters that are highly active in the germ line. The ancestral source element 1A (Fig. 8A) recruited the promoter of the *MAST2* gene whose expression peaks in germ cells at the spermatid stage in mice

Figure 7. SVA subfamily F1 can be subdivided into four transduction groups. (*A*) Transduction group 1 members carry exclusively *MAST2* sequences as 5′ transduction. They could be direct descendants of source elements 1A, 1B, 2, or 3 (Fig. 8) except for H17_2 and H15_5. The latter can only be derived from source elements 1A or B due to the extension of their transduced *MAST2* sequence. In theory, within this group any given SVA_F1 member could have served as a source element of any other SVA_F1 member whose *MAST2* sequence is equal in length or shorter than its own. Source elements and polymorphic SVA_F1 members are marked by red and blue lettering, respectively. Yellow boxes; *MAST2* sequence; red boxes, TSDs flanking *AluY* insertion; *, element was originally described by Bennett et al. 2004 and named as denoted in brackets. (*B*) Transduction group 2. In addition to *MAST2* sequences, members of this group include heterologous sequences (pink boxes) as part of their 5′ transductions. Heterologous sequences indicating consecutive transduction events are derived from source locus 2 (Fig. 8B) on chromosome 9p13.3 (H10_1, H19_56, and H6_17), from sequences of non-LTR retrotransposons or endogenous retroviruses whose source loci could be identified (H6_2, H8_F116, H9_2, H22_17, H3_5, and H12_3), or they represent nonrepetitive genomic DNA fragments ranging from 21–107 bp (H1_F_67, H12_2, H1_F_163, H8_2). TSD-flanked intact *Alu* elements in sense orientation are part of H10_1, H19_56, and H22_17. H3_5 displays secondary (*AluY*) and tertiary (*NPEPPS* exon sequences) transduction events. Numbers indicate extensions of different modules of 5′-transduced heterologous (pink boxes) and *MAST2* (yellow boxes) sequences and VNTRs. TSDs flanking *Alu*Sc and *Alu*Sp elements (red arrowheads), as well as all remaining *Alu* sequences (blue bars) are shown. Open boxes indicate genomic DNA flanking the SVA insertion. (*C*) Members of transduction group 3 include 3′ transductions. SVA_F1 members with 3′ transductions and their source elements are shown. Primary 3′ transductions of HX_4 and H9_1 originate from SVA-free source loci (data not shown). The source element of H6_6 is H1_F_160 (in brackets), a member of transduction group 1 (Fig. 7A). MER57A represents internal portions of the nonautonomous endogenous retrovirus 1 (ERV1) family. Yellow boxes depict 5′-transduced *MAST2* sequences and/or 3′ transductions, respectively. Numbers denote the length of 5′- and 3′-transduced sequences, their components, and VNTR regions in nucleotides. Red ovals, TSDs flanking SVA elements; blue ovals, TSD sequence of the respective source element that became part of 3′ or 5′ transductions. (*D*) Members of transduction group 4 are derivatives of source element H10_1. Structures of the 13 identified members of transduction group 4 and their source element H10_1 are shown. Two consecutive transcriptional termination signals (asterisks) in the 3′-flanking sequence of H10_1 resulted in 3′ transductions of 400 and 479 bp, respectively. The 3′ TSD of the source element is overlapping with the 5′ TSD of *Alu*Sp. Lengths of 5′-truncated VNTRs of H10_7, H15_2, H15_3, and H6_5 range from 241–2098 bp. H19_47 is a source element of H1_F_163 (Fig. 7B); Black bars, extension of query sequences 1 and 2.

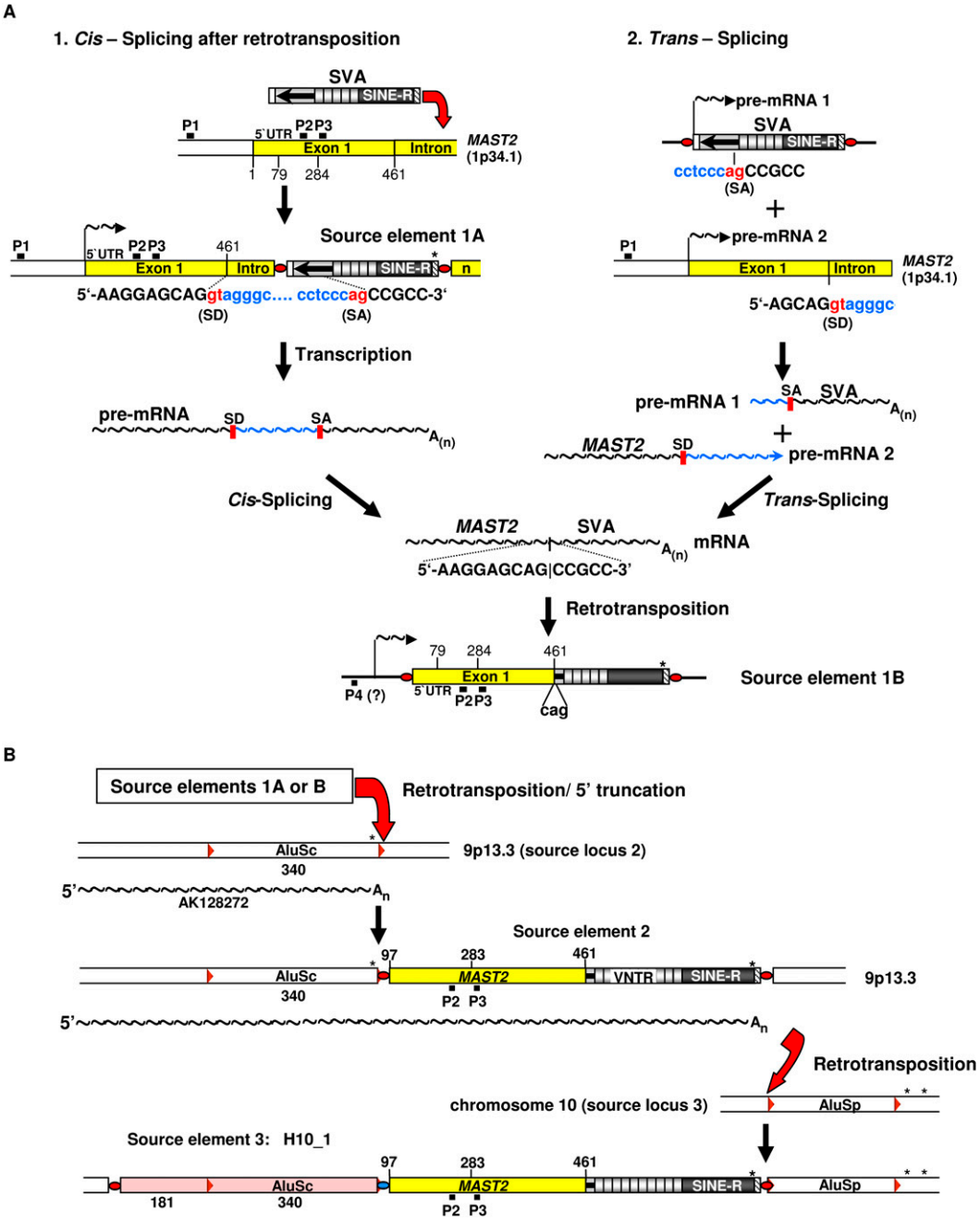


Figure 8. Evolutionary scenarios for the emergence of SVA subfamily F1. (*A*) Schematic diagram of two alternative scenarios giving rise to hypothetical precursor source elements 1A or B of the SVA_F1 subfamily, respectively. Alternative 1: *Cis*-splicing after retrotransposition; after SVA retrotransposition into the *MAST2* gene downstream of exon 1, transcription of source element 1A is controlled by the *MAST2* promoter P1 or by potential regulatory sequences in exon 1 (P2, P3) resulting in a chimeric *MAST2*-SVA pre-mRNA. *Cis*-splicing of the pre-mRNA removes the intronic sequence flanked by the splice donor (SD) site at the *MAST2* exon 1–intron junction and the cryptic splice acceptor (SA) site localized within the *Alu*-like region. Retrotransposition of the spliced, chimeric *MAST2*-SVA transcript results in the 5′-transducing source element 1B. Transcription of this element could have been controlled by internal promoter sequences located within the *MAST2* sequence (P2, P3) or by an external promoter localized upstream of the SVA_F1 insertion (P4). Nucleotide position 79 indicates the 5′-end of the *MAST2* sequence of SVA H15_5 (Fig. 7A) that harbors the most extensive *MAST2* 5′-transduction. Position 284 denotes the 5′-end of the putative cryptic promoter sequence P3; asterisks indicate transcription termination signals. Alternative 2: *Trans*-splicing; two separate pre-mRNA molecules encoding an SVA element (pre-mRNA 1) and the MAST2 protein (pre-mRNA 2), respectively, are synthesized. Pre-mRNA 2 provides the 5′ splice site (splice donor [SD]), pre-mRNA 1 harbors branchpoint sequence, poly-pyrimidine tract, and a cryptic 3′ splice site (splice acceptor [SA]). The exons of the two separate pre-mRNA molecules are joined by *trans*-splicing to create a single chimeric RNA. Retrotransposition of this RNA would result in source element 1B. The 5′-transduced *MAST2* sequence of source element 1B could either be full-length or 5′-truncated. In theory, members of transduction group 1 (Fig. 7A) could be direct descendants of source elements 1A or B. Splice donor and splice acceptor sequences are highlighted in red. P1, *MAST2* promoter; P4, hypothetical external promoter controlling transcription of source element 1B. (*B*) Evolutionary scenario giving rise to source element H10_1. Retrotransposition of source element 1A or B into the 3′ TSD of an *Alu*Sc element on chromosome 9p13.3 is setting the newly integrated 5′-truncated *MAST2*-SVA element (source element 2) under control of the AK128272 promoter. Transcriptional read-through causes a chimeric AK128272-*MAST2*-SVA transcript. 5′ Truncation in the process of retrotransposition of the chimeric RNA into the 5′ TSD of an *Alu*Sp element in source locus 3 results in the 5′-transducing source element H10_1. The SVA-free source locus 3 is still present on chromosome 10 of both rhesus macaque and chimpanzee genomes but absent from humans. H10_1 served as source element for all members of SVA_F1 transduction group 4.

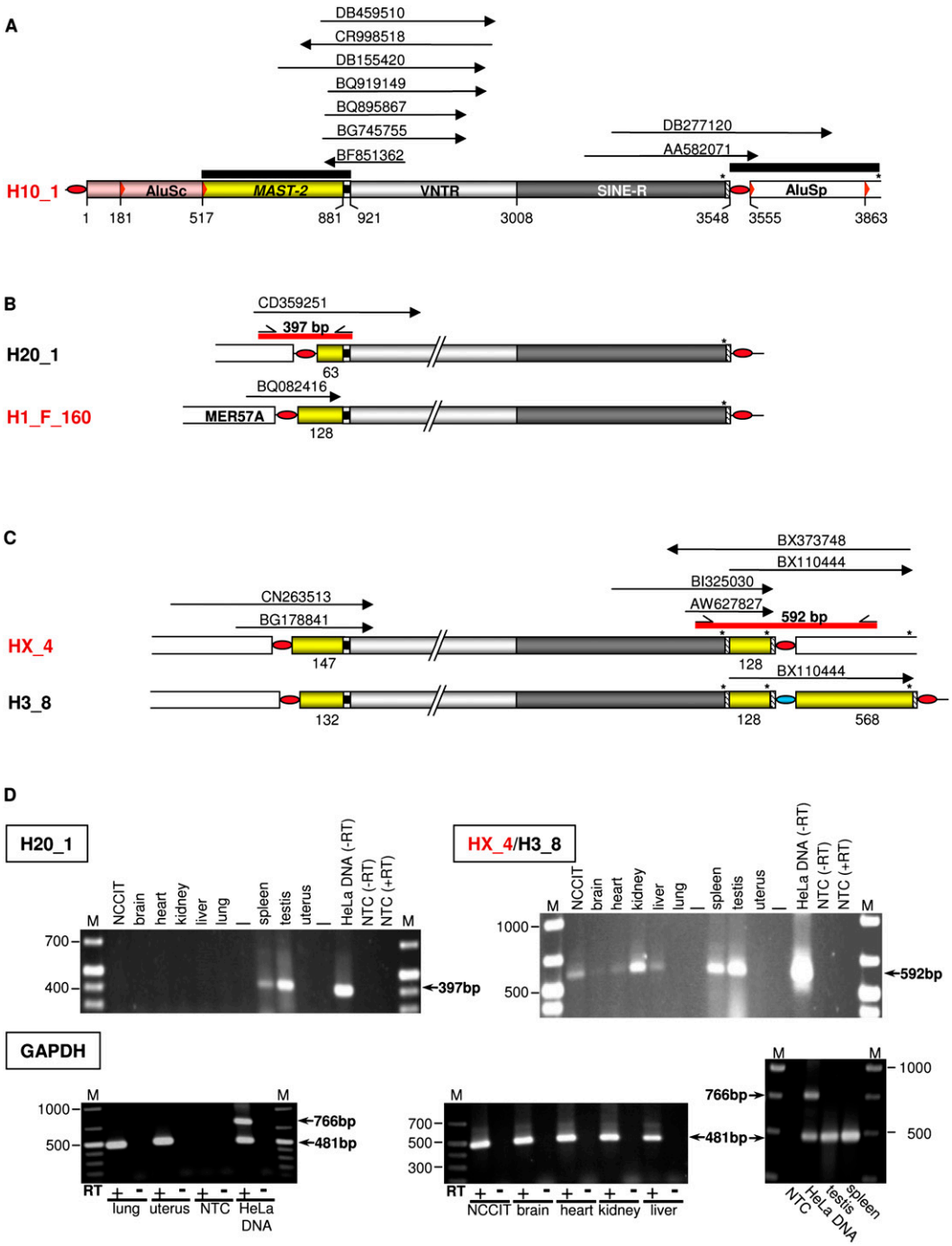


Figure 9. Ongoing transcription of SVA_F1 subfamily members is demonstrated by ESTs and RT-PCR analyses. (*A*) ESTs (arrows) aligning specifically to the source element H10_1 and its descendants (DB277120, AA582071) and to *MAST2*–SVA junctions of a subset of SVA_F1 subfamily members. Extensions of the queries used to identify ESTs (black bars) are shown. (*B*) Transcriptional control of transduction group 1 members H20_1 and H1_F_160 is effected by external promoters. Identified ESTs match 5′ junctions of the SVA elements with primary 5′ transductions. MER57A is an internal portion of the nonautonomous retrovirus 1 (ERV1). (*C*) ESTs indicating ongoing transcription of source element HX_4. Those ESTs that correspond to the 3′-end (BI325030, AW627827) and the 3′ junction (BX110444) of HX_4 are also consistent with transcription of the HX_4 derivative H3_8 (Fig. 7C). 3′-Ends of sense ESTs BI325030 and BX110444 as well as the 5′-end of antisense EST BX373748 are directly at the polyadenylation sites of HX_4 and/or H3_8. SVA copies highlighted in red were verified to be source elements. (*D*) RT-PCR analyses confirm transcription of distinct SVA_F1 subfamily members. Agarose gel electrophoresis of H20_1 and HX_4/H3_8-specific RT-PCR products derived from total RNA from human tissues and NCCIT cells. Primer binding sites and extension of the RT-PCR products are indicated as short arrows and red bars, respectively, in *B* and *C*. To control for both successful cDNA synthesis and DNA-free RNA preparations, PCRs with *GAPDH*-specific, intron-spanning primers (see Methods) were performed on RNA preparations from each tissue before (−RT) and after (+RT) cDNA synthesis. PCR on genomic HeLa DNA served as a positive control. The 766- and 481-bp products originate from the unspliced *GAPDH* gene and a processed *GAPDH* pseudogene, respectively (Harper et al. 2003). Sizes of obtained PCR products are indicated and correspond to expected fragment sizes. RT, reverse transcriptase; M, size marker; NTC, no template control.

Table 3. Human ESTs corresponding to members of the SVA_F1 subfamily

ESTs[a]	Tissue	Length in bp	High-quality sequence range (as indicated by NCBI)	Percent identity to best match SVA	Best match SVA_F1 (Fig. 7A–D)	Best match SVA aligns to EST position *x* to *y*	EST aligns to H10_1 ref. sequence position *x* to *y*
Spanning *MAST2*–*Alu*-like–VNTR regions							
DB155420	Thymus	571	Not specified	99% (567/571)	HX_4 (C)	1-571	737-1266
BG745755	Spleen	665	1–657	97% (649/663)	H2_F_65(A)	1-663	821-977
DB459510	Testis	483	Not specified	100% (482/482)	H13_17 (A)	2-483	824-1305
				99% (480/482)	H1_F_67(B)	2-483	
				99% (480/482)	H22_17 (B)	2-483	
BQ895867	Carcinoma cell line	1316	26–224	93% (338/361)	H6_2 (B)	1-361	851-1207
BF851362 (antisense)	Lung	384	1–342	96% (338/361)	H10_2 (A)	384-11	838-1092
BQ919149	Prostate carcinoma cell line	1144	1–384	95% (394/414)	H14_1 (A)	1-411	855-1261
CR998518 (antisense)	T lymphocytes	768	Not specified	98% (629/636)	H9_2 (B)	636-1	817-1452
				99% (616/618)	H1_F_67(B)	636-19	
				99% (616/618)	H15_17 (B)	636-19	
				99% (616/618)	H17_1 (A)	636-19	
				99% (617/618)	H1_F_160 (A)	636-19	
				99% (628/629)	H6_F_70 (A)	636-8	
				98% (626/633)	H9_3 (A)	636-4	
Spanning 5′-flanking/5′-transduced sequences							
CD359251	Testis	868	12–189	96% (602/627)	H20_1 (A)	19-665[b]	817-1134
CN263513	ES cells	585	Not specified	99% (569/573)	HX_4 (C)	13-585[b]	737-1082
BG178841	Adenocarcinoma cell line	971	1–374	97% (404/413)	HX_4 (C)	1-410[b]	737-1075
BQ082416	Stomach, ascites	271	1–271	100%(271/271)	H1_F_160 (A)	1-271[b]	757-860
Spanning 3′-flanking/3′-transduced sequences							
DB277120	Uterus	535	Not specified	97% (528/539)	H10_1 (B,D)	1-535	3262-3798
				97% (528/539)	H19_108 (D)	1-535	
AA582071	Gastric tumor	395	1–380	95% (389/407)	H10_1 (B,D)	3-395	3178-3584
				97% (390/4402)	H19_39 (D)	3-395	
				95% (389/407)	H19_108 (D)	3-395	
BI325030	Pancreatic island	437	1–411	100% (437/437)	HX_4 (C)	1-437	1485-1921 (HX_4)
AW627827	Genitourinary tract, tumor	228	1–227	98% (223/227)	HX_4 (C)	3-228	1696-1921 (HX_4)
BX373748 (antisense)	Placenta	921	Not specified	89% (267/297)	HX_4 (C)	846-539	1653-1933 (HX_4)
				98% (659/668)	H3_8 (C)	662-1	2154-3089 (H3_8)
BX110444	Senescent fibroblast	668	Not specified	99% (131/132)	HX_4 (C)	1-132	1815-1933 (HX_4)
				99% (645/649)	H3_8 (C)	20-668	2382-3049 (H3_8)

[a]BF851362, CR998518, and BX373748 are directed in antisense orientation.
[b]5′-Ends of ESTs CD359251, CN263513, BG178841, and BQ082416 are localized in the 5′-flanking genomic regions of SVAs H20_1, HX_4, and H1_F_160, respectively (Fig. 7B,C).

(Walden and Cowan 1993). Transcriptional control of a potential source element by the *MAST2* promoter assures that the 5′-transducing SVA element is coexpressed with the L1 protein machinery, which is known to be present in germ cells (Branciforte and Martin 1994; Ergun et al. 2004) and suspected to *trans*-mobilize SVA elements (Ostertag et al. 2003). Additional heterologous sequences that are part of SVA_F1 5′ transductions (Fig. 7B–D) indicate that their source elements were subject to transcriptional control by external promoters other than the *MAST2* promoter. Identification of SVA_F1-specific ESTs and RT-PCR products extending into genomic flanks upstream of specific SVA_F1 elements indicate ongoing transcriptional regulation by external promoters in both reproductive and somatic tissues (Fig. 9).

2. Newly acquired 5′- and 3′-transduced sequences made SVA_F1 subfamily members a better target for *trans*-mobilization by the L1 protein machinery. A striking feature of numerous SVA_F1 secondary transductions is the presence of intact *Alu* elements (Fig. 7B,D). The most successful source element H10_1, in particular, can give rise to transcripts containing intact *Alu*Sc and *Alu*Sp RNAs at their 5′- and 3′-ends, respectively (Fig. 7D). The same mechanism hypothesized recently to be responsible for the preferential *trans*-mobilization of *Alu* elements by the L1-encoded protein machinery (Boeke 1997) could thus explain the favored mobilization of SVA RNAs harboring intact *Alu* elements in their transduced sequences. In this model, the *Alu* sequence is docked on ribosomes via the SRP9/14 complex and captures the L1 ORF2 protein as it is translated from an active L1 element mRNA. By capturing ORF2p at the ribosome, *Alu* and, possibly, *Alu*-containing SVA RNAs can efficiently substitute its RNA for the normal L1 mRNA during the process of target-site primed reverse transcription (TPRT). Functionally intact *Alu* core elements were recently identified from four of the six *Alu*S subfamilies, and binding of the *Alu* RNA to SRP9/14 was shown to be essential for *Alu* retrotransposition activity (Bennett et al. 2008). Provided that sequences flanking *Alu* elements in SVA transcripts allow formation of the three-dimensional structure required for SRP9/14 interaction (Weichenrieder et al. 2000), *Alu* sequences could mediate docking of the respective RNA to the ribosome via SRP9/14 and thus facilitate efficient capture of ORF2 proteins (Boeke 1997; Dewannieux et al. 2003; Mills et al. 2007).

In a different scenario, 5′-transduced *Alu* sequences in antisense orientation (Fig. 7B) might substitute for the truncated *Alu*-like region in SVA_F1 subfamily members. The SRP9/14 model of *Alu trans*-mobilization suggests that hybridization of the SVA *Alu*-like region that is complementary to intact ribosome-bound *Alu* elements could bring SVA transcripts into close proximity to the nascent L1 ORF2 polypeptide chain and may, in this way, enhance the probability of ORF2p capture (Mills et al. 2007).

The presence of both sense and antisense *Alu* sequences on the same transducing RNA (Fig. 7B) might increase the *trans*-mobilization rate of such an SVA mRNA even more. Clearly, all three of these scenarios remain to be tested experimentally.

Methods

Identification of SVA elements on human chromosome 19 and genome-wide screening for 5′-transducing SVA elements

To identify the complete set of SVA elements on chromosome 19 and 5′-transducing SVA elements genome-wide, we started out by using the RepeatMasker premasked genome section at http://www.repeatmasker.org/cgi-bin/AnnotationRequest (hg 17; May 2004; repeat class: other). Each of the identified SVA sequences was extracted and manually inspected for correct annotation. In order to identify TSD sequences, the BLAST-2–sequences program (bl2seq; http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi; Tatusova and Madden 1999) was applied to align 500-bp 5′- and 3′-flanking sequences of each identified SVA against each other. In case TSDs could not be identified in this way, SVA insertions and their flanking sequences were tested for the presence of TSDs individually. In principle, the minimum number of nucleotides required for TSDs to be considered as such was six.

In case of the genome-wide survey for 5′ transductions, only elements containing at least a fraction of the SINE-R region were analyzed. Thus, SVA2 elements and fragments lacking a poly(A) tail were excluded from the analysis. An element was considered candidate for containing a 5′ transduction if it had (1) unambiguous TSDs (≥6 bp) and (2) a ≥20-bp sequence between the TSD and the 5′-end of the SVA sequence. These criteria were relaxed in those few cases where an unambiguous assignment to a particular transduction group was possible based on the respective transduced sequence. Multiple sequence alignments were calculated using the ClustalW (2.0) program (Labarga et al. 2007).

To identify source loci of transduced sequences, we searched the human genome by using BLAT. Repetitive sequence annotations were obtained from the UCSC Genome Browser. The corresponding descriptions were retrieved from Repbase Update at http://www.girinst.org/server/RepBase/index.php.

Prediction of potential promoter sequences

In order to identify sequences that had the potential to serve as promoters controlling either transcription of the *MAST2* gene or expression of members of the SVA_F1 subfamily, we used a specific query sequence. The query encompasses the 461 bp of *MAST2* exon 1 and 1000 bp of genomic 5′-flanking sequence. Genomic sequences were analyzed applying promoter prediction programs NNPP2.2 (Reese 2001), TSSW (Solovyev and Shahmuradov 2003), and TSSG (Prestridge 1995; http://www.softberry.com/berry.phtml?topic=tssg&group=help&subgroup=promoter). Obtained results were scrutinized by applying the ESPERR regulatory track of the UCSC Genome Browser (King et al. 2005). Data for this track are obtained by comparing frequencies of short alignment patterns between known regulatory elements and neutral DNA. Score values above 0.1 indicate significant resemblance to alignment patterns typical of regulatory elements in the training set.

Identification of ESTs matching members of SVA subfamily F1

To identify ESTs originating from members of the SVA_F1 subfamily, we searched the human division of dbEST, a descriptive catalogue of ESTs in GenBank, by using the BLASTN program (BLASTN at http://blast.ncbi.nlm.nih.gov/Blast.cgi; default settings except for: word length 16, no filter) with a 401-bp DNA fragment of SVA element H19_56, which is also part of the source element H10_1. The fragment comprises 364 bp of *MAST2* exon 1 and 37 bp of the 5′-truncated *Alu*-like region specific for the *MAST2* 5′ transduction groups (query sequence 1). ESTs specific for members of SVA_F1 transduction group 4 were searched for by using the 394-bp 3′ transduction specific for this transduction group (Fig. 7D). The sequence is flanked by the H10_1-terminating poly(A) sequence and the first transcriptional termination signal following the *Alu*Sp 3′ TSD (query sequence 2). ESTs matching the source element HX_4 and its derivative H3_8 were identified from the annotation of the respective loci in the UCSC Genome Browser.

RT-PCR analyses

Total RNA from different human tissues was purchased from BioCat (brain, heart, liver, kidney) or Agilent/Stratagene (uterus, lung). Total RNA was extracted from human teratocarcinoma cell line NCCIT (ATCC CRL-2073) using TRIzol reagent (Invitrogen). First-strand synthesis was performed using random hexamers (Fermentas) and SuperScript III RT (Invitrogen) according to the manufacturer's instructions. cDNAs synthesized from testis and spleen RNAs were purchased from BioCat. For cDNA amplification AmpliTaq DNA polymerase (Applied Biosystems) was used. PCR cycling conditions were as follows: 5 min at 94°C, 30 sec at 94°C, 30 sec at primer-specific annealing temperature, 30 sec at 72°C, 7 min at 72°C; 40 cycles. HX_4/ H3_8 cDNA was amplified using primers SVA-RT-FW2 5′-TGCCTAGGAAAACCAGAGAC-3′ and HX4-REV 5′-CTCTAATAGTGGGTCACCAATGCC-3′ (62°C). For H20_1, the primer combination 20_1 FW 5′-CCTGTCATTGA TCTTGACTTACC-3′ and *Alu* 3′REV 5′-CGGGCAGAGG CTGC AAT-3′ (63°C) was used.

Control PCRs to exclude genomic DNA contamination and to test for successful cDNA synthesis were performed with GAPDH intron-spanning primers hGAPDH-for 5′-CCATGAGAAGTATG ACAACAGC-3′ (Exon 6) and hGAPDH-rev 5′-GTCAAAGGTGG AGGAGTGG-3′ (Exon 8).

Acknowledgments

We thank Dustin Hancks, Haig Kazazian, and Sandy Martin for helpful discussions. We thank Matthias Hamdorf for his indispensable contribution to artwork generation. This research was supported by grant DA 545/2-1 of the *Deutsche Forschungsgemeinschaft* (A.D., G.G.S), National Science Foundation grant BCS-0218338 (M.A.B.), and National Institutes of Health RO1 GM59290 (M.A.B.).

References

Adey NB, Schichman SA, Graham DK, Peterson SN, Edgell MH, Hutchison CA III. 1994. Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Mol Biol Evol* **11:** 778–789.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403–410.
Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. 2003. Human population genetic structure and inference of group membership. *Am J Hum Genet* **72:** 578–589.
Belancio VP, Hedges DJ, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res* **34:** 1512–1521.
Belancio VP, Hedges DJ, Deininger P. 2008. Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Res* **18:** 343–358.
Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168:** 933–951.
Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE. 2008. Active *Alu* retrotransposons in the human genome. *Genome Res* **18:** 1875–1883.
Boeke JD. 1997. LINEs and *Alu*s—the polyA connection. *Nat Genet* **16:** 6–7.
Boeke JD, Pickeral OK. 1999. Retroshuffling the genomic deck. *Nature* **398:** 108–109, 111.
Branciforte D, Martin SL. 1994. Developmental and cell type specificity of LINE-1 expression in mouse testis: Implications for transposition. *Mol Cell Biol* **14:** 2584–2592.
Chen J, Rattner A, Nathans J. 2006. Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: Lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements. *Hum Mol Genet* **15:** 2146–2156.
Conley ME, Partain JD, Norland SM, Shurtleff SA, Kazazian HH Jr. 2005. Two independent retrotransposon insertions at the same site within the coding region of BTK. *Hum Mutat* **25:** 324–325.
Damert A, Lower J, Lower R. 2004. Leptin receptor isoform 219.1: An example of protein evolution by LINE-1-mediated human-specific retrotransposition of a coding SVA element. *Mol Biol Evol* **21:** 647–651.
Dewannieux M, Esnault C, Heidmann T. 2003. LINE-mediated retrotransposition of marked *Alu* sequences. *Nat Genet* **35:** 41–48.
Ergun S, Buschmann C, Heukeshoven J, Dammann K, Schnieders F, Lauke H, Chalajour F, Kilic N, Stratling WH, Schumann GG. 2004. Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. *J Biol Chem* **279:** 27753–27763.
Ferlini A, Galie N, Merlini L, Sewry C, Branzi A, Muntoni F. 1998. A novel *Alu*-like element rearranged in the dystrophin gene causes a splicing mutation in a family with X-linked dilated cardiomyopathy. *Am J Hum Genet* **63:** 436–446.
Grimwood J, Gordon LA, Olsen A, Terry A, Schmutz J, Lamerdin J, Hellsten U, Goodstein D, Couronne O, Tran-Gyamfi M, et al. 2004. The DNA sequence and biology of human chromosome 19. *Nature* **428:** 529–535.
Han K, Konkel MK, Xing J, Wang H, Lee J, Meyer TJ, Huang CT, Sandifer E, Hebert K, Barnes EW, et al. 2007. Mobile DNA in Old World monkeys: A glimpse through the rhesus macaque genome. *Science* **316:** 238–240.
Hancks DC, Ewing AD, Tokunaga K, Kazazian HH. 2009. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res* (this issue). doi: 10.1101/gr.093153.109.
Harper LV, Hilton AC, Jones AF. 2003. RT-PCR for the pseudogene-free amplification of the glyceraldehyde-3-phosphate dehydrogenase gene (*gapd*). *Mol Cell Probes* **17:** 261–265.
Horiuchi T, Aigaki T. 2006. Alternative *trans*-splicing: A novel mode of pre-mRNA processing. *Biol Cell* **98:** 135–140.
Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462–467.
King DC, Taylor J, Elnitski L, Chiaromonte F, Miller W, Hardison RC. 2005. Evaluation of regulatory potential and conservation scores for detecting *cis*-regulatory modules in aligned mammalian genome sequences. *Genome Res* **15:** 1051–1060.
Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M, et al. 1998. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* **394:** 388–392.
Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7:** 474. doi: 10.1186/1471-2105-7-474.
Kreahling J, Graveley BR. 2004. The origins and implications of *Alu*ternative splicing. *Trends Genet* **20:** 1–4.
Labarga A, Valentin F, Anderson M, Lopez R. 2007. Web services at the European Bioinformatics Institute. *Nucleic Acids Res* **35:** W6–W11.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.
Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S, Matsumoto S, Tabuena MD, Maranon E, Dantes M, et al. 2007. Reduced neuron-specific expression of the *TAF1* gene is associated with X-linked dystonia-parkinsonism. *Am J Hum Genet* **80:** 393–406.
Maniatis T, Tasic B. 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* **418:** 236–243.
Meischl C, Boer M, Ahlin A, Roos D. 2000. A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. *Eur J Hum Genet* **8:** 697–703.
Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends Genet* **23:** 183–191.
Mulhardt C, Fischer M, Gass P, Simon-Chazottes D, Guenet JL, Kuhse J, Betz H, Becker CM. 1994. The spastic mouse: Aberrant splicing of glycine receptor beta subunit mRNA caused by intronic insertion of L1 element. *Neuron* **13:** 1003–1015.
Ono M, Kawakami M, Takezawa T. 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res* **15:** 8725–8737.
Ostertag EM, Kazazian HH Jr. 2001. Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* **11:** 2059–2065.
Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73:** 1444–1451.
Prestridge DS. 1995. Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* **249:** 923–932.
Reese MG. 2001. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput Chem* **26:** 51–56.
Rohrer J, Minegishi Y, Richter D, Eguiguren J, Conley ME. 1999. Unusual mutations in Btk: An insertion, a duplication, an inversion, and four large deletions. *Clin Immunol* **90:** 28–37.
Sironen A, Thomsen B, Andersson M, Ahola V, Vilkki J. 2006. An intronic insertion in KPL2 results in aberrant splicing and causes the immotile short-tail sperm defect in the pig. *Proc Natl Acad Sci* **103:** 5006–5011.
Sironen A, Vilkki J, Bendixen C, Thomsen B. 2007. Infertile Finnish Yorkshire boars carry a full-length LINE-1 retrotransposon within the *KPL2* gene. *Mol Genet Genomics* **278:** 385–391.
Solovyev VV, Shahmuradov IA. 2003. PromH: Promoters identification using orthologous genomic sequences. *Nucleic Acids Res* **31:** 3540–3545.
Sorek R, Ast G, Graur D. 2002. Alu-containing exons are alternatively spliced. *Genome Res* **12:** 1060–1067.
Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327–338.
Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD. 2002. Molecular archeology of L1 insertions in the human genome. *Genome Biol* **3:** research0052. doi: 10.1186/gb-2002-3-10-research0052.
Takasu M, Hayashi R, Maruya E, Ota M, Imura K, Kougo K, Kobayashi C, Saji H, Ishikawa Y, Asai T, et al. 2007. Deletion of entire *HLA-A* gene accompanied by an insertion of a retrotransposon. *Tissue Antigens* **70:** 144–150.
Tamura M, Kajikawa M, Okada N. 2007. Functional splice sites in a zebrafish LINE and their influence on zebrafish gene expression. *Gene* **390:** 221–231.
Tatusova TA, Madden TL. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett* **174:** 247–250.
Walden PD, Cowan NJ. 1993. A novel 205-kilodalton testis-specific serine/threonine protein kinase associated with microtubules of the spermatid manchette. *Mol Cell Biol* **13:** 7625–7635.
Walden PD, Millette CF. 1996. Increased activity associated with the MAST205 protein kinase complex during mammalian spermiogenesis. *Biol Reprod* **55:** 1039–1044.
Wang M, Marin A. 2006. Characterization and prediction of alternative splice sites. *Gene* **366:** 219–227.
Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA. 2005. SVA elements: A hominid-specific retroposon family. *J Mol Biol* **354:** 994–1007.
Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, Carroll ML, Nguyen SV, Walker JA, Prasad BV, et al. 2003. Genetic variation among world populations: Inferences from 100 *Alu* insertion polymorphisms. *Genome Res* **13:** 1607–1618.
Weichenrieder O, Wild K, Strub K, Cusack S. 2000. Structure and assembly of the *Alu* domain of the mammalian signal recognition particle. *Nature* **408:** 167–173.
Wilund KR, Yi M, Campagna F, Arca M, Zuliani G, Fellin R, Ho YK, Garcia JV, Hobbs HH, Cohen JC. 2002. Molecular mechanisms of autosomal recessive hypercholesterolemia. *Hum Mol Genet* **11:** 3019–3030.
Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci* **103:** 17608–17613.
Xing J, Witherspoon DJ, Ray DA, Batzer MA, Jorde LB. 2007. Mobile DNA elements in primate and human evolution. *Am J Phys Anthropol* **135** (Suppl. 45)**:** 2–19.

Received March 4, 2009; accepted in revised form July 24, 2009.