Exon-trapping mediated by the human retrotransposon SVA


Although most human retrotransposons are inactive, both inactive and active retrotransposons drive genome evolution and may influence transcription through various mechanisms. In humans, three retrotransposon families are still active, but one of these, SVA, remains mysterious. Here we report the identification of a new subfamily of SVA, which apparently formed after an alternative splicing event where the first exon of the MAST2 gene spliced into an intronic SVA and subsequently retrotransposed. Additional examples of SVA retrotransposing upstream exons due to splicing into SVA were also identified in other primate genomes. After molecular and computational experiments, we found a number of functional SVA transcripts within many different transcribed SVAs across the human and chimpanzee genomes. Using a minigene splicing construct containing an SVA, we observed splicing in cell culture, along with SVA exonization events that introduced premature termination codons (PTCs). These data imply that an SVA residing within an intron in the same orientation as the gene may alter normal gene transcription either by gene-trapping or by introducing PTCs through exonization, possibly creating differences within and across species.

[Supplemental material is available online at http://www.genome.org. The 5' RACE sequence data from this study have been submitted to dbEST (http://www.ncbi.nlm.nih.gov/dbEST) under accession nos. GR564526-GR564716.]

Most eukaryotic genomes harbor retrotransposable elements (Malik et al. 1999). About 35% of the human genome is derived from retrotransposed sequences such as LINE-1, Alu, SVA, endogenous retroviruses, and processed pseudogenes (Lander et al. 2001). Although most human retroelement copies are no longer mobile, both active and inactive human elements have been shown to drive genome evolution and influence gene expression (Moran et al. 1999; Han et al. 2004; for reviews, see Belancio et al. 2005; Goodier and Kazazian 2008).

SVA RNAs are hominin-specific noncoding RNAs, which vary in size from 700–4000 bp, and are likely mobilized by the human LINE-1 in trans (Ono et al. 1987; Shen et al. 1994; Ostertag et al. 2003; Wang et al. 2005), similar to the human Alu (Dewannieux et al. 2003). There are roughly 2700 SVA copies in the human genome (Wang et al. 2005). A canonical full-length SVA (Fig. 1A) contains a number of sequence features proceeding from its 5' end: (1) a CCGTCTT hexameric repeat, ranging in repeat number from a few to as many as 71; (2) a sequence that shares homology with two antisense Alu fragments; (3) a variable number of tandem repeat sequence (VNTR); and (4) a sequence derived from the ENV gene and rightLTR of an extinct HERV-K, hereafter referred to as SINE-R. SVAs typically terminate at their own polyA signal, with genomic insertions usually containing a number of adenines at the 3' end. A target site duplication common to other LINE-1-driven retroelements (6–20 bp) flanks the inserted element (Ostertag et al. 2003).

Little is known about the biology of SVA apart from its structure. SVAs are currently active in the human genome, as indicated by the identification of de novo SVA insertions associated with disease (Hassoun et al. 1994; Rohrer et al. 1999; Makino et al. 2007). SVA disease insertions are associated with exon-skipping (Hassoun et al. 1994; Rohrer et al. 1999), deletion of genomic DNA (Takasu et al. 2007), and reduced or absent mRNA expression (Kobayashi et al. 1998; Wihund et al. 2002; Makino et al. 2007). In a manner similar to L1, SVAs have been shown to transduce 3' flanking sequences to new genomic locations (Moran et al. 1999; Ostertag et al. 2003; Xing et al. 2006). SVAs are thought to be highly active due to the ratio of disease insertions to genomic copies. Furthermore, the high levels of insertion polymorphism of the human-specific subfamilies, E and F (Bennett et al. 2004; Wang et al. 2005), support the notion that SVAs are evolutionarily young and relatively active in the human population.

The mechanism of SVA transcription and the location of its promoter are unknown and are critical to our understanding of SVA retrotransposition. To date, experiments to characterize the SVA promoter have led to ambiguous results. Recently, we reexamined an SVA insertion on CH6 associated with a genomic deletion including the entire HLA-A gene that resulted in leukemia in three Japanese individuals (Takasu et al. 2007). The HLA-A insertion led to several interesting findings, including the identification of a new SVA subfamily formed by alternative splicing from the first exon of the MAST2 gene on CH1 into an SVA, and the identification of a SVA master element on CH10. We have also identified numerous functional 3' splice sites (SSs) within SVA while analyzing human and chimpanzee SVA 5' transcriptional start sites (TSSs), and three more examples where splicing into the SVA followed by retrotransposition led to exon shuffling. Using a minigene construct containing an intronic SVA, we are able to show that splicing into the SVA is not rare and that this splicing results in both exonization and SVA gene-trapping. These data suggest that splicing into SVA elements enables their expression and can allow for adaptive evolution at the cost of altering the transcriptome in both humans and great apes.
Using the SVA site identified an SVA insertion, hereafter referred to as SVA in one individual from each of the families. Analysis of the deletion progenitor revealed several interesting details.

Chromosome 3p21.31 as the likely progenitor of the SVAHLA-A in-nucleotides (nt) to the 5' of the SVA (M2) gene on chromosome 1. Further analysis showed that 262 nt native splicing event. The SVA splice site relative to SVARep is aligned to the known splice site sequence genome. Note the loss of the entire CCCTCT hexamer and most of the locus, sometime after the human–chimp split, containing a full-length SVA within intron 1 (SVAF1 founder insertion (1984 Genome Research). The number of nucleotides derived from MAST2 directly upstream of SVA varied from 35–382 (Supplemental Table 1). Given that the MAST2 5' UTR and first exon combined is 460 nt, no SVAf1 present in the human reference genome contains the entire 5' UTR-first exon.

The MAST2 sequence abutting the genomic copies of SVA in SVAF1 elements terminated directly at the 5' SS of the MAST2 first exon. However, there is no SVA present in the reference MAST2 intron sequence. It is likely that an SVA with an allele frequency <1 resided in intron 1 of MAST2 (Fig. 1B) in the individual in which the first M2-SVA splicing and subsequent retrotransposition event occurred. We generated a consensus sequence from the SVAf1 in the human genome and aligned it to the SVA present in Repbase (Jurka 2000; Jurka et al. 2005), henceforth called SVArep to determine whether the site where MAST2 and SVA intersect would have provided a suitable 3' SS in a consensus SVAF1. The 3' SS consensus sequence is YYYYYYYYYNN CAG/G, where Y represents a pyrimidine and N is any nucleotide (Wang and Cooper 2007). The SVA portion of SVAF1 aligns to SVArep beginning at position 388 of SVArep which is located in the 3' region of the Alu-like fragment, 35 bp upstream of the VNTR. The sequence upstream of position 388 in SVArep is CCTCCACCTCC CAG (YYYY-YYYYNCAG), a close match to the 3' SS consensus sequence.

An SVA master element on CH10
The SVA on CH3, the progenitor to SVAFch3-A (Fig. 2A), lacks a target-site duplication (TSD) directly flanking the SVAf1 (Fig. 2B). Given that retrotransposons are able to transduce sequences 5' and 3' of their location in the genome (Moran et al. 1999; Xing et al. 2006; Goodier and Kazazian 2008), we searched for a TSD further upstream and downstream from the CH3 SVAf1. We identified a 15-nt TSD, with the 5' duplication directly in front of a truncated AluS (153 nt) and the 3' duplication following a polyA tail downstream from a non-RepeatMasker annotated sequence. The entire insertion between the AluS and the terminal 3' polyA tail nearest the 3' TSD consisted of (1) the AluS, followed by (2) a SVAf1, (3) an AluSp, and then (4) a 3' transduction of 82 nt 3' to the AluSp. When using BLAT (Kent 2002) to identify the source locus for the CH3 3' transduction, 13 hits in addition to the CH3 query sequence were obtained (Supplemental Table 2). The source locus was identified on CH10 (Fig. 2C) due to the absence of a polyA tail 3' of the transduced sequence present on CH3. Interestingly, the SVA on CH10 was flanked by a 5' AluS (320 nt) and a 3' AluSp (299 nt). Overall, 13 SVAF1 insertions contained the 3' transduction from chromosome 10 (Supplemental Table 2), and all 13 had the AluSp and were variably truncated with three containing the AluS, four containing some portion of the MAST2 sequence and no AluS, five truncated in the VNTR, and one truncated in the SVA polyA tail. Furthermore, one of the SVAs, which

### Results

**Identification of MAST2-SVA**

In 2007, Takasu et al. described a 14-kb deletion that included the entire HLA-A locus in three unrelated families, leading to leukemia in one individual from each of the families. Analysis of the deletion site identified an SVA insertion, hereafter referred to as SVAHLA-A. Using the SVAHLA-A DNA sequence, we located an SVA insertion on chromosome 3p21.31 as the likely progenitor of the Alu-like region (black line) likely occurred resulting in the original SVAf1, (3) an AluSp, and then (4) a 3' transduction directly in front of a truncated AluS (153 nt) and the 3' duplication following a polyA tail downstream from a non-RepeatMasker annotated sequence. The entire insertion between the AluS and the terminal 3' polyA tail nearest the 3' TSD consisted of (1) the AluS, followed by (2) a SVAf1, (3) an AluSp, and then (4) a 3' transduction of 82 nt 3' to the AluSp. When using BLAT (Kent 2002) to identify the source locus for the CH3 3' transduction, 13 hits in addition to the CH3 query sequence were obtained (Supplemental Table 2). The source locus was identified on CH10 (Fig. 2C) due to the absence of a polyA tail 3' of the transduced sequence present on CH3. Interestingly, the SVA on CH10 was flanked by a 5' AluS (320 nt) and a 3' AluSp (299 nt). Overall, 13 SVAF1 insertions contained the 3' transduction from chromosome 10 (Supplemental Table 2), and all 13 had the AluSp and were variably truncated with three containing the AluS, four containing some portion of the MAST2 sequence and no AluS, five truncated in the VNTR, and one truncated in the SVA polyA tail. Furthermore, one of the SVAs, which

![Figure 1. SVA and alternative splicing at the MAST2 locus. (A) The canonical SVA is displayed, consisting of some number of CCCTCT hexameric repeats followed by the Alu-like region, a variable number of tandem repeats, and then the SINE-R region, followed by a polya signal with the entire SVA flanked by target-site duplications (black arrowheads). Two black arrows over the Alu-like region indicate sequence homology in the SVA to two ancestral Alu elements. (B) Shown is an ancestral CH1 MAST2 locus, sometime after the human–chimp split, containing a full-length SVA within intron 1 (top). An alternative splicing event into the Alu-like region (black line) likely occurred resulting in the original SVAf1, (3) an AluSp, and then (4) a 3' transduction directly in front of a truncated AluS (153 nt) and the 3' duplication following a polyA tail downstream from a non-RepeatMasker annotated sequence. The entire insertion between the AluS and the terminal 3' polyA tail nearest the 3' TSD consisted of (1) the AluS, followed by (2) a SVAf1, (3) an AluSp, and then (4) a 3' transduction of 82 nt 3' to the AluSp. When using BLAT (Kent 2002) to identify the source locus for the CH3 3' transduction, 13 hits in addition to the CH3 query sequence were obtained (Supplemental Table 2). The source locus was identified on CH10 (Fig. 2C) due to the absence of a polyA tail 3' of the transduced sequence present on CH3.

1984  Genome Research  www.genome.org
CH10 was the source locus for, had a 160-nt 3′ transduction. This element represents a transcript from the CH10 locus that bypassed the polyA signal at which the other 12 elements terminated. The sequence directly after the AluSp, the original source for these transductions, contains two canonical polyA signals, AATAAA (Colgan and Manley 1997), which are 15–20 nt upstream of the polyA tails of the SVAs derived from the CH10 locus (Fig. 2C).

To distinguish whether the SVA on CH10 inserted alone, or with the AluSc and/or AluSp, we searched for TSDs of each element and examined the chimpanzee reference sequence (Chimpanzee Sequencing and Analysis Consortium 2005). Only the AluSp, hereafter referred to as 3′ Alu′, was present in the chimpanzee reference sequence, suggesting that it was the first insertion on CH10 (Fig. 3A) and that the SVA insertion occurred, with or without the AluSc, hereafter referred to as 5′ Alu, sometime since our last common ancestor with chimp. Furthermore, the 3′ Alu on CH10 is at least 25 million years (Myr) old because it is present in the Rhesus macaque genome sequence (Rhesus Macaque Genome Sequencing and Analysis Consortium 2007). The 5′ Alu on CH10 could be traced back to a locus on CH9, due to 185 nt present directly upstream of the 5′ Alu on CH10 (Fig. 3E), which represents a 5′ transduction from the CH9 AluSc source locus (Fig. 3C). On CH10 there is a 13-nt TSD flanking the 5′ Alu containing the 5′ transduction and the SVA, suggesting that the 5′ Alu and SVA retrotransposed as one unit (Fig. 3C). However, at the CH9 locus there is no SVA downstream from the AluSc in the human reference sequence.

Identification of multiple SVA TSSs

SVA is a nonautonomous retrotransposon and was previously thought to rely on an internal promoter to initiate its transcription, similar to LINE-1 (Swergold 1990) and Alu (Di Segni et al. 1981; Duncan et al. 1981; Fritsch et al. 1981). Previous attempts in our laboratory to locate the SVA promoter have led to ambiguous results (MC Seleme and HH Kazazian, unpubl.). We set out to identify the SVA TSS for insight into how SVA mRNA is transcribed.

We used 5′ RACE to identify novel SVA 5′ ends from total RNA extracted from cell lines (see Methods) and chimpanzee testes. Currently, both the requirements for SVA transcription and the repertoire of expressed SVAs are unknown. We identified a total of 56 unique SVA-associated TSSs after sequencing and analysis of human and chimp SVA 5′ RACE products (Table 1). We grouped the TSSs into three classes: (1) internal SVA TSSs (Supplemental Table 3); (2) 5′ TSSs, defined as any position upstream of SVA annotated sequence (Supplemental Table 4); and (3) examples in which part of the sequence aligned within the SVA and part aligned upstream with a large gap in between, representing transcripts where 5′ sequences are spliced into the SVA part of the transcript (Table 2). The 26 class I TSSs are scattered throughout the SVA but tend to cluster toward the 5′ end of the element (Fig. 4). The 14 class II TSSs start 76–440 bp upstream of SVAs in the human or chimp genome (Fig. 4; Supplemental Table 2).

Class III SVA-associated TSSs represent splicing into the SVA

We identified 17 class III TSSs, 16 of which are unique (13 human and four chimp) where the 5′ ends mapped upstream and represented splicing into SVA at 10 different 3′SS (Table 2). SVA splicing events are listed in Table 2, with the 3′SS position and 3′SS sequence annotated relative to SVARep. Twelve of 17 class III TSSs involved exons from known genes/ESTs. One gene, AFF1 on CH4 had two SVA alternative splicing events identified by 5′ RACE, suggesting the same SVA may contain multiple functional 3′SS.

Many SVA splicing events are present in human ESTs

To identify further examples of SVA splicing, we performed a computational survey to identify splicing events. We focused on splicing events involving the 228 SVA elements present in intronic regions oriented in the same direction as the surrounding gene. EST databases (Boguski et al. 1993) were mined for uniquely aligned sequences with evidence of intrinsic SVA expression. Spliced ESTs were selected where one or more blocks aligned within the intronic SVA sequence, and the exon–SVA junctions were examined. The SVA sequences upstream and downstream from the EST junctions were compared to the SS consensus sequence, and those containing the canonical “(T/C)AG” trinucleotide (Wang and Cooper 2007) at the junction preceded by a reasonable polypyrimidine tract were kept. We defined an event as a unique exon 5′SS and a unique SVA 3′SS. If multiple overlapping ESTs having the same 5′SS and SVA 3′SS existed, we called it one event. ESTs with the same 5′SS and two different SVA 3′SS were called two events.

In total, 16 events, involving 14 genes, at eight different SVA 3′SS were detected, supporting the notion that splicing into SVAs

**Figure 2. Identification of the SVA master locus on CH10.** (A) The SVA insertion on CH6. The 40-bp deletion in the MAST2 sequence (△) allowed the identification of the (B) CH6 progenitor element to be identified on CH3. The SVA insertion on CH3 contains both 5′ and 3′ transductions. At the 5′ end, the SVA contains a truncated AluSc while at the 3′ end, an AluSp along with additional sequence, indicated by X, followed by a polyA tail, with the entire insertion flanked by target-site duplications. The 3′ transduction (X, red line) on CH3 allowed the identification of (C) the master element on CH10 along with 12 additional elements derived from the CH10 locus. The SVA on CH10 contains starting at it’s 5′ end: 185 bp transduction derived from CH9, a full-length AluSc, the MAST2-SVA, an AluSp, polyA tail, and then a unique sequence, which was 3′ transduced, and which allowed the identification of CH10 as the source locus. A target-site duplication flanks the insertion, which inserted on CH10. Downstream polyA signals are displayed over the X and Y transduction sequences.
occurs with some frequency across the genome (Table 2). We found one gene, C2CD3, which had ESTs aligning to three different 3’ SS locations at the C2CD3 locus, AG138, AG319, and AG386, further indicating that multiple functional 3’ SS exist within SVA.

**Gene-trapping occurs in primates**

The upstream sequence of all class II TSSs were aligned to either the human or chimp reference sequence using BLAT (Kent 2002) to determine whether the sequence was present elsewhere in the human or chimp reference sequence. These SVA insertions differ in that the 5’ TSS, unique sequence blocks (blue boxes), labeled X and Y, are present in the daughter inserts derived from this locus. The polyA signals utilized in the 3’ transductions are indicated by AATAAA relative to the unique sequence blocks.

![Diagram of SVA insertion and transduction](image)

**Table 1. SVA-associated transcriptional start sites**

<table>
<thead>
<tr>
<th>Class</th>
<th>Human</th>
<th>Chimp</th>
<th>Total unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>I: Internal</td>
<td>16</td>
<td>10</td>
<td>26</td>
</tr>
<tr>
<td>II: Upstream</td>
<td>13</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>III: Splicing</td>
<td>13</td>
<td>4</td>
<td>16</td>
</tr>
</tbody>
</table>

*CH17 SVA splicing event was identified in both species.
original SVA<sub>H<sub>1<sub> event occurred some time after the orangutan diverged from the human–gorilla last common ancestor, aging the insertion between 8 and 15 Myr.

**SVA splicing is not rare**

To study the potential mutagenic potential of SVA splicing, we cloned two SVAs from the human genome, SVA<sub>22369</sub> and SVA<subnant</sub>_1<sub>369</sub>, containing multiple functional 3' SS. We cloned the SVAs into the intron of a splicing minigene construct named pPKC-EGFP (Fig. 6A), hereafter pPKC-SVA. 293T cells were transfected with pPKC-SVA, and total RNA was harvested after 1 d. To characterize SVA splicing and identify functional 3' SS, we performed RT-PCR with a forward primer in the first exon, PKC, and then used three reverse primers (6A) in independent reactions to answer three questions: (1) are SVAs exonized (PKC For + 1R primers); (2) what 3' SS exist in SVA (PKC For + 2R); and (3) can we detect SVA gene-trapping (PKC For + 3R).

PCR products were analyzed on a 2% agarose gel (Fig. 6B); bands were cloned and sequenced. Bands from the lane labeled 1R for pPKC-SVA<sub>22369</sub> corresponded to the normal splicing, PKC exon to EGFP exon, and also to SVA exonization events. Five SVA exonization events utilizing three different 3' SS and three different 5' SS with SVA exons ranging from 159–359 nt (Fig. 6B; Table 3) were identified. Three out of the five SVA exonization events introduce premature stop codons (PTCs) located in the exonized SVA sequence.

**Figure 4.** SVAs contain many transcriptional start sites (TSSs). 5' RACE was performed on total RNA extracted from 293T, HeLa, and nTera cell lines, along with chimpanzee tissues. A nested PCR was performed followed by Sanger sequencing of Topo clones. The 5’ RACE adaptor was identified, and the first nucleotide after it was annotated as the TSS. Human (red triangles) and chimpanzee (blue triangles) TSSs identified by 5’ RACE are shown. The CCCTCT hexamer (Hex), a Location of TPTE gene in HG18. The CCCTCT hexamer (Hex), a Location of MAST2 gene in HG18.

<table>
<thead>
<tr>
<th>Splice acceptor</th>
<th>Sequence</th>
<th>Gene/EST</th>
<th>Species</th>
<th>UCSC Genome Browser</th>
<th>SVA subfamily</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>52</td>
<td>CTCTCTTCCACAG</td>
<td>n/a</td>
<td>Hsa, C2CD3</td>
<td>chr17: 7800466–78005622</td>
<td>A</td>
<td>S' RACE</td>
</tr>
<tr>
<td>52</td>
<td>CTCTTTTCCACAG</td>
<td>FAM122C</td>
<td>Hsa</td>
<td>chrX: 133800359–133802069</td>
<td>D</td>
<td>EST analysis</td>
</tr>
<tr>
<td>70</td>
<td>CTCCTCATGTCCAG</td>
<td>CF145918</td>
<td>Hsa</td>
<td>chr2: 27435097–27436626</td>
<td>F</td>
<td>S' RACE</td>
</tr>
<tr>
<td>52</td>
<td>CTCCTCATGTCCAG</td>
<td>n/a</td>
<td>Hsa</td>
<td>chr3: 9556016–9557408</td>
<td>D</td>
<td>S' RACE</td>
</tr>
<tr>
<td>99</td>
<td>ATGCCGGAGCCAG</td>
<td>SgK494</td>
<td>Hsa</td>
<td>chr17: 2395462–23956451</td>
<td>D</td>
<td>S' RACE</td>
</tr>
<tr>
<td>138</td>
<td>TGTCTTCTTCCAG</td>
<td>AFFI</td>
<td>Hsa</td>
<td>chr4: 68140008–68141613</td>
<td>D</td>
<td>S' RACE</td>
</tr>
<tr>
<td>138</td>
<td>TGTCTTCTTCCAG</td>
<td>D427153</td>
<td>Hsa</td>
<td>chr7: 10461381–10462912</td>
<td>B</td>
<td>EST analysis</td>
</tr>
<tr>
<td>138</td>
<td>TGTCTTCTTCCAG</td>
<td>SCYE1</td>
<td>Hsa</td>
<td>chr14: 24997705–25000164</td>
<td>F</td>
<td>EST analysis</td>
</tr>
<tr>
<td>138</td>
<td>TGTCTTCTTCCAG</td>
<td>C2CD3</td>
<td>Hsa</td>
<td>chr: 73517261–73518822</td>
<td>D</td>
<td>EST analysis</td>
</tr>
<tr>
<td>138</td>
<td>TGTCTTCTTCCAG</td>
<td>MAPKAP1</td>
<td>Hsa</td>
<td>chr: 127260318–12726193</td>
<td>D</td>
<td>EST analysis</td>
</tr>
<tr>
<td>138</td>
<td>TGTCTTCTTCCAG</td>
<td>ESCO1</td>
<td>Hsa</td>
<td>chr: 17:426562–17427548</td>
<td>D</td>
<td>EST analysis</td>
</tr>
</tbody>
</table>

*Location of TPTE gene in HG18.
**Location of MAST2 gene in HG18.
PKC-EGFP was 0.12:1.

that if an SVA loses the CCCTCT hexamer and most of the duplication. (A) CH3 SVA containing four processed TPTE exons that utilized AG 336. (B) CH15 SVA containing five processed exons that utilized AG 386. (C) The structure of SVA insertions on CH13, CH18, and CH21 containing the six processed exons from the RHOT1 gene on CH17 with identical target-site duplications.

Figure 5. SVA gene-trapping has occurred in other hominids. (A,B) Two different SVA splicing events followed by retrotransposition derived from the TPTE locus on chromosome 21 flanked by a target-site duplication. (A) CH3 SVA containing four processed TPTE exons that utilized AG 336. (B) CH15 SVA containing five processed exons that utilized AG 386. (C) The structure of SVA insertions on CH13, CH18, and CH21 containing the six processed exons from the RHOT1 gene on CH17 with identical target-site duplications.

SVA splicing events using pPKC-SVA and verified by sequencing are listed in Table 3. It is noteworthy that a 3' SS site was identified in the SINE-R domain of SVA (Fig. 6B, lane 3, lower band). These PCR results suggest that SVA splicing is not rare and that both SVA exonization and SVA gene-trapping can occur in the same SVA.

Semi-quantitative RT-PCR followed by Southern blotting using amplicons from a PKC for and 1R PCR (Fig. 6A,C) was carried out for pPKC-EGFP, pPKC-SVA<sub>CH1A</sub>, and pPKC-SVA<sub>CH1R</sub> to estimate SVA exonization. The intensities for normal splicing varied across the samples, so the Southern blot was exposed overnight in order to ensure no bands were present in the vector-only lane (data not shown). The ratio of the higher molecular weight bands indicative of SVA exonization relative to PKC-EGFP splicing within that lane was determined using a phosphorimager (Fig. 6C, bottom panel). The ratio of total SVA<sub>CH1A</sub> exonization relative to PKC-EGFP splicing was 0.19:1, while total SVA<sub>CH1R</sub> exonization to PKC-EGFP was 0.12:1.

Discussion
These data are the first to provide insight into how SVA retrotransposons are expressed and how they might impact gene expression. Our data suggest that SVAs are expressed in a variety of ways in humans and chimps. Recently, a study identified many TSSs in LINE-1s and SINEs in human and mouse tissues and cell lines. (Faulkner et al. 2009). Whether or not internal TSSs identified here represent retrotransposition-competent SVA transcripts is unknown. Many SVAs have transduced sequence 5' of their location in the genome to other locations (Damert et al. 2009); this is consistent with our observation of upstream TSSs. What is unclear is whether upstream TSSs represent solely upstream promoters driving SVA expression or whether something inherent to SVA directs transcriptional initiation upstream.

The SVA<sub>A</sub> subfamily, SVA<sub>TPTE</sub> and SVA<sub>RHOT1</sub> together indicate that if an SVA loses the CCCTCT hexamer and most of the Alu-like region due to alternative splicing into it, the remaining SVA sequence is able to retrotranspose. Furthermore, the lack of most of the Alu-like region suggests that the model suggested by Mills et al. (2007), adapted from Boeke's model (Boeke 1997), where the SVA Alu-like region hybridizes to Alu RNAs at the ribosome in order to compete for the LINE-1 ORF2 reverse transcriptase, may not be case. However, it is possible that SVA RNA may be located at the ribosome where competition for the LINE-1 ORF2 takes place, but that it is not hybridizing to Alu RNA.

The lack of SVA<sub>CH1</sub> with a complete MAST2 5' UTR and first exon suggests that a full-length MAST2 5' UTR first exon is not required for transcription or retrotransposition. Exactly, how the MAST2 sequence contributed to the expansion of SVA<sub>CH1</sub> in the absence of the CCCTCT hexamer and the Alu-like region still needs to be determined. One possibility is that the MAST2 sequence, in combination with certain SVA sequence variants or in a specific genomic context, enhances transcription or retrotransposition relative to a canonical SVA. It is worthwhile to note that TPTE is a testis-specific gene (Chen et al. 1999) and that we found the CH3 SVA<sub>TPTE</sub> by 5' RACE, and the TSS was in exon 1 of the transduced TPTE sequence.

We have identified 11 3' SS throughout SVA, in addition to multiple 3' SS in the VNTR, including examples from all subfamilies except E, and we have shown that exonization can occur. Whether or not older SVAs are still retrotransposing in humans is currently unknown; however, the older SVAs are still able to be transcribed and may influence transcription if residing within an intron in the same orientation as the gene.

Ss within retrotransposons are not uncommon. Alus are primate-specific retrotransposons that are known to exonized (Sorek et al. 2002; Lev-Maor et al. 2003). In addition to Alu, internal splicing has been observed in the human L1 (Belancio et al. 2006) and the zebrafish LINE (Tamura et al. 2007). Although, it appears that if an SVA undergoes a splicing event, it can still carry out subsequent rounds of retrotransposition, as indicated by SVA<sub>CH1</sub>.

SVA splicing followed by retrotransposition may be rare based on only four examples identified in the human genome, three of which are present in the chimpanzee. Additional splicing followed by retrotransposition events may have occurred, but the results are undetectable due to truncation upon insertion or low allele frequency. On the contrary, splicing into the SVA is not rare, as indicated by our semi-quantitative PCR data, which suggest that SVA exonization events may occur 12%–19% of the time relative to normal splicing in our minigene.

Currently, the ratio of SVA gene-trapping to SVA exonization has not been determined. Here we provide a low-end estimate for SVA splicing by assessing SVA exonization using a splicing minigene. Our SVA exonization estimate may be an underestimate because SVAs contain more than 10 nonsense codons in each reading frame on the sense strand and exonization of these sequences may induce nonsense-mediated decay if the exonized SVA sequence is more than 50–55 nt upstream of the 3' most
exon–exon junction (Nagy and Maquat 1998). Be that as it may, an SVA splicing event, exonization or trapping (Fig. 7), will likely lead to a dead-end to the protein-coding capacity of the mRNA because either event has the capability to produce truncated proteins.

Most SSs identified using the splicing minigene were not identified by 5′ RACE, such as the 3′ SSs in the VNTR and SINE-R. This is likely due to the nested PCR approach utilized in 5′ RACE. However, downstream SVA 3′ SS may be selected for in the splicing minigene due to the small size of the intron. Each SVA was cloned into pPKC-EGFP with less than 100 bp of flanking DNA to ensure splicing was inherent to the SVA and not due to intronic splicing enhancers. If SVA is cloned in a larger fragment, one may see 3′ SS selection shift toward the 5′ end of SVA residing in the Alu-like domain.

If SVAs impact gene expression by being alternatively spliced, then one would expect to observe either a depletion of SVAs in genes or on the coding strand of genes. An underrepresentation was observed for SVAs on the coding-strand in the human genome; 1060/2772 SVAs are in RefSeq genes, with 228/1060 on the coding strand (introns or exons) and 832/1060 on the antisense strand. This underrepresentation of intronic SVA insertions on the coding strand is highly significant \( (P < 2.2 \times 10^{-16}) \) under a null hypothesis of random orientation. Likewise, a similar significant underrepresentation is observed in the chimpanzee genome, with 228 (partial overlap with the human 228) of 1024 intronic SVAs oriented on the sense strand with respect to the surrounding gene. Nevertheless, this SVA strand bias may be due to a factor other than selection, such as SVA insertional preference.

Altogether, these data show that SVAs are alternatively spliced in cell culture, in tissue, and in vivo. We speculate that SVAs may influence local gene expression by providing alternative SSs and might even account for some of the variation in gene expression observed within and across hominids. As more primate genomes are sequenced along with more studies on SVA, the impact of this retrotransposon will become clear. Thus, although SVAs effect on genome evolution may be less than that of L1 and Alu because of their smaller numbers, SVA has had recent effects that are likely growing with their continued expansion as indicated by the SVAF1 subfamily and the CH10 subgroup. In another 50 Myr, the SVA effect on genome evolution may be much greater than that of L1 and Alu.

**Figure 6.** SVA gene-trapping and exonization are not rare. (A) Two SVAs were cloned into PKC-EGFP (Newman et al. 2006) to test the mutagenic potential of SVA splicing. Primers used for RT-PCR are marked. (B) RT-PCR was performed on total RNA extracted from 293T cells transiently transfected with pPKC-EGFP containing one of two different SVAs cloned into the intron. SVA exonization events (left panel) are annotated with the first and last nucleotide of the SVA exon, all of which occur within the Alu-like and VNTR domains. A representative agarose gel displaying SVA alternative splicing events is shown (right panel) (see Table 3). (*) Indicates bands verified by DNA sequencing to be SVA splicing events. (C) Semi-quantitative PCR to determine the frequency of SVA exonization. Ten cycles of PCR on cDNA from individual pPKC-EGFP, pPKC-SVA\(_{C2CD3}\) and pPKC-SVA\(_{MTFR1}\) transfections were carried out using PKC For and 1R. PCR products were resolved on a 2% agarose gel, followed by overnight transfer to a membrane, and subsequent probing using a DNA probe targeting the PKC exon (top panel). (*) Indicates bands quantified by a phosphorimager. Total SVA exonization was normalized to PKC-EGFP splicing within each respective lane and graphed (bottom panel).

---

**Table 3.** Splice sites identified in SVA splicing mini-gene

<table>
<thead>
<tr>
<th>Event</th>
<th>3′ SS (C2CD3)</th>
<th>3′ SS Sequence</th>
<th>5′ SS (C2CD3)</th>
<th>Exon size</th>
</tr>
</thead>
<tbody>
<tr>
<td>exon</td>
<td>386 (445)</td>
<td>CTCCTAATCCCTCAG</td>
<td>VNTR (786)</td>
<td>339</td>
</tr>
<tr>
<td>exon</td>
<td>468 (411)</td>
<td>TTGCCCTAATCACAG</td>
<td>VNTR (786)</td>
<td>316</td>
</tr>
<tr>
<td>exon</td>
<td>VNTR (625)</td>
<td>GCCATCCCATCTAG</td>
<td>VNTR (986)</td>
<td>359</td>
</tr>
<tr>
<td>exon</td>
<td>VNTR (625)</td>
<td>GCCATCCCATCTAG</td>
<td>VNTR (786)</td>
<td>159</td>
</tr>
<tr>
<td>exon</td>
<td>VNTR (625)</td>
<td>GCCATCCCATCTAG</td>
<td>VNTR (902)</td>
<td>275</td>
</tr>
<tr>
<td>trap</td>
<td>VNTR (625)</td>
<td>GCCATCCCATCTAG</td>
<td>VNTR (982)</td>
<td>159</td>
</tr>
<tr>
<td>trap</td>
<td>VNTR (625)</td>
<td>GCCATCCCATCTAG</td>
<td>VNTR (902)</td>
<td>275</td>
</tr>
<tr>
<td>trap</td>
<td>1372 (1352)</td>
<td>CGGATGCTCAATCAG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N/A</td>
<td>386 (445)</td>
<td>CTCCTAATCCCTCAG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N/A</td>
<td>VNTR (625)</td>
<td>GCCATCCCATCTAG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N/A</td>
<td>VNTR (625)</td>
<td>GCCATCCCATCTAG</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>N/A</td>
<td>VNTR (629)</td>
<td>TCCATCTCAAGGAG</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>
PCR. PCR cycling parameters consisted of variations on touch-
to the Alu-like region to the SINE-R region of SVARep. Primers complementary to the Alu-like region were used for the second round of PCR. PCR cycling parameters consisted of variations on touch-

**Truncated Protein**

![Diagram of Truncated Protein and Nonsense-Mediated Decay](image)

**Figure 7.** SVA alternative splicing outcomes. An intronic truncated SVA is shown (middle). The SVA is truncated because these SVAs are still likely to be spliced. If SVAs are exonized, they will likely generate a truncated protein or subject the mRNA to nonsense-mediated decay due to the inclusion of SVA nonsense codons (top). If SVAs mimic an endogenous gene-trap, that is provide a 3' SS followed by termination at the SVA or downstream polyA signal, this may result in truncated proteins, but more importantly the retrotransposition of exons.

### Methods

#### Sequence analysis

BLAT (Kent 2002) and BLAST (Altschul et al. 1990) were used in mapping sequences to the reference genomes. Censor (Kohany et al. 2006) and RepeatMasker (Smit et al. 1996) were used to identify relative positions in SVA and subfamily classification, respectively.

#### Cell culture

293T and HeLa cells were grown in a humidified, 5% CO2 incubator at 37°C in DMEM (GIBCO) supplemented with 10% fetal bovine serum, 2 mM L-glutamine, and 100 U/mL penicillin, 0.1 mg/mL streptomycin. nTera cells were grown as described above except that the media was supplemented with nonessential amino acids.

#### 5' RACE, cDNA synthesis, and PCR

RNA extraction was performed using the RNaseasy kit (Qiagen) according to the manufacturer's instructions. DNase treatment consisted of using twice the recommended amount of RQ1 RNase-Free DNase (Promega) followed by ethanol precipitation of the RNA. Chimps testis was used for RNA extraction (Department of Veterinary Medicine and Surgery, University of Texas M.D. Anderson Cancer Center, Houston, TX).

5' RACE was performed using the GeneRacer Kit (Invitrogen) with 5 µg DNase-treated RNA as the starting material. First-strand cDNA synthesis was performed using the supplied SuperScript III RT kit with random hexamer primers or Array Script Reverse Transcriptase (Ambion). All steps were carried out according to the manufacturer's instructions. A two-round PCR scheme was utilized in order to enrich for SVA containing transcripts using GoTaq (Promega) or Expand Long (Roche) according to the manufacturer's instructions. Total RNA was isolated 1 d after transfection as described above. Five micrograms of DNase-treated RNA was reverse-transcribed with Array Script (Ambion) using an olio dT primer. Touchdown PCR from 59°C to 51°C over 40 cycles was performed with elongation at 72°C for 2 min using GoTaq Master Mix (Promega) with 1 µL of cDNA as template and primers at a final concentration of 0.2 µM per reaction. Amplicons were analyzed on 2% agarose gels.

#### EST analysis

EST locations corresponding to human genome assembly hg18 were obtained from the UCSC Genome Browser and stored in a local relational database along with SVA and RefSeq gene locations. ESTs with blocks aligning unambiguously within SVA locations present on the same strand as a RefSeq intron were further analyzed using Perl scripts locate EST splice junctions within the SVA. Junctions corresponding to splicing patterns consistent with SVA gene-trapping were compared to the SS consensus sequence to ensure presence of the relevant nucleotides.

#### SVA splicing minigene, transfection, and RT-PCR

pPKC-EGFP has been previously reported (Newman et al. 2006). SVAs were amplified from human genomic DNA and subcloned into Topo (Invitrogen). The SVA was then amplified as a Xhol fragment and cloned into the XhoI site within the intron.

293T cells were seeded into T-75 flasks in order to be 50%–80% confluent upon transfection. Twenty-four hours later, 8 µg of each splicing minigene was transfected using 24 µL of Fugene6 (Roche) according to the manufacturer's instructions. Total RNA was isolated 1 d after transfection as described above. Five micrograms of DNase-treated RNA was reverse-transcribed with Array Script (Ambion) using an olio dT primer. Touchdown PCR from 59°C to 51°C over 40 cycles was performed with elongation at 72°C for 2 min using GoTaq Master Mix (Promega) with 1 µL of cDNA as template and primers at a final concentration of 0.2 µM per reaction. Amplicons were analyzed on 2% agarose gels.

#### Semi-quantitative PCR

One microliter of oligo dT primed cDNA derived from total RNA from splicing minigene transfections was amplified by 10 cycles of PCR (20 sec at 94°C, 30 sec at 57°C, 1 min at 72°C) using GoTaq MasterMix (Promega) with PKC forward and 1R primers in a 25 µL reaction. The entire reaction was resolved on a 2% agarose gel. Overnight alkaline transfer to N+ hybrid membrane (Amersham) was performed followed by overnight hybridization with a 182-bp DNA probe labeled with [α-32P]dCTP targeting the PKC exon at 65°C. Reaction products were imaged using a Storm 840 phosphorimager (GE Healthcare) and quantified with ImageQuant 5.2 (GE Healthcare). The intensity of each band was determined followed by the subtraction of background. SVA exonization band
intensities were summed followed by normalization to PKC-EGFP splicing.

Acknowledgments
This work was supported by grants from the NIH. D.C.H. is funded by NIH training grant T32-M007229-27. We thank Dr. Yoshihide Ishikawa, Japanese Red Cross Central Blood Institute, for HLA-A deletion samples. We also thank Claude Warzecha and Dr.Russ Carstens for the PKC-EGFP construct.

References

Received February 25, 2009; accepted in revised form July 20, 2009.
Exon-trapping mediated by the human retrotransposon SVA


Access the most recent version at doi:10.1101/gr.093153.109

**Supplemental Material**
http://genome.cshlp.org/content/suppl/2009/09/21/gr.093153.109.DC1.html

**References**
This article cites 45 articles, 19 of which can be accessed free at:
http://genome.cshlp.org/content/19/11/1983.full.html#ref-list-1

**Creative Commons License**
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**Email Alerting Service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Copyright © 2009 by Cold Spring Harbor Laboratory Press