

SVA ELEMENTS: HOMINID SPECIFIC RETROTRANSPOSONS

A Dissertation

**Submitted to the Graduate Faculty of
the Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy**

in

The Department of Biological Sciences

By

Hui Wang

B.S., Shanghai Jiao Tong University, P.R.China, 2000

M.S., Shanghai Jiao Tong University, P.R.China, 2003

December 2006

ACKNOWLEDGEMENTS

I would like to thank Dr. Mark A. Batzer, my advisor, for his continuous guidance, constant encouragement, and financial support during my Ph.D study at LSU. It is my great pleasure to learn from and work with such a knowledgeable, considerate, and helpful scholar.

I want to express my sincere gratitude to my committee members, Dr. David Donze, Dr. Joomyeong Kim, Dr. Stephania Cormier and Dr. Karin Peterson. I would like to thank all of them for kindly serving on my committee.

Also I want to thank members in the Batzer's laboratory for their scientific guidance and friendship, especially to Dr. Jinchuan Xing, who offered a lot of help during my Ph. D study.

Last but not least, I would like to thank my family for their spiritual support. This dissertation would not appear without their consistent encouragement, love, and patience.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	ii
LIST OF TABLES.....	iv
LIST OF FIGURES.....	v
ABSTRACT.....	vi
CHAPTER ONE: INTRODUCTION.....	1
CHAPTER TWO: SVA ELEMENTS: A HOMINID SPECIFIC RETROPOSON FAMILY... ..	15
CHAPTER THREE: EMERGENCE OF PRIMATE GENES BY RETROTRANSPOSON-MEDIATED SEQUENCE TRANSDUCTION.....	51
CHAPTER FOUR: SUMMARY.....	73
APPENDIX A: SUPPLEMENTAL DATA.....	78
APPENDIX B: LETTERS OF PERMISSION.....	100
VITA.....	103

LIST OF TABLES

Table 2.1 Repeat content and G+C content in the SVA flanking regions	21
Table 2.2 Age estimates of SVA subfamilies	24
Table 3.1 Affiliation of 3' transduced sequences and SVA subfamilies	54
Table 3.2 PAML analysis	63
Table 3.3 Putative positive selection sites on <i>AMAC</i> genes	64

LIST OF FIGURES

Fig. 1.1 Full-length L1 structure.....	3
Fig. 1.2 <i>Alu</i> element structure.....	7
Fig. 1.3 Full-length SVA structure.....	8
Fig. 2.1 Structure analysis of SVA element. (A) Structure of a full-length SVA element with the target site duplications. (B) Size distribution of the SVA VNTR regions	18
Fig. 2.2 Human genomic distribution of SVA elements.....	19
Fig. 2.3 SVA subfamily composition in the human genome.....	23
Fig. 2.4 The relationship between VNTR lengths and age.....	25
Fig. 2.5 Correlation between SVA subfamily distribution and chromosomal G+C content	27
Fig. 2.6 SVA subfamily and genomic G+C content.....	28
Fig. 2.7 Correlation between SVA subfamilies and chromosomal gene density.....	29
Fig. 2.8 L1 and <i>Alu</i> densities flanking SVA elements and G+C content.....	30
Fig. 2.9 Median-joining network of the SVA subfamilies.....	32
Fig. 2.10 SVA copy number in the primate lineage.....	33
Fig. 2.11 Amplification dynamics of the SVA family of retroposons in primates.....	42
Fig. 3.1 Identification of SVA 3' transduction events and their source elements.....	54
Fig. 3.2 Length distribution of 3' transduction events.....	55
Fig. 3.3 SVA 3' transduction events.....	57
Fig. 3.4 SVA transduction mediated gene duplication. (A) Schematic diagram of the H17_76 transduction group in the human genome. (B) Schematic diagrams for putative evolutionary scenarios of the SVA transduction mediated gene duplications. (C) The phylogenetic relationships among various species used in d_N/d_S analysis.....	59
Fig. 3.5 Expression analysis of <i>AMAC</i> gene duplicates in humans. (A) Agarose gel chromatograph of RT-PCR products derived from human testis (T) and placental (P) RNA templates. (B) Relative expression levels of four human <i>AMAC</i> gene duplicates in human testis and placenta.....	65

ABSTRACT

SVA is a composite repetitive element named after its main components, SINE, VNTR and Alu. There are ~3000 SVA elements in the human genome. A genomic distribution analysis indicates that SVA elements are enriched in G+C-rich regions but have no preferences for inter- or intra-genic regions. A phylogenetic analysis of these elements resulted in the recovery of six subfamilies that were named SVA_A to SVA_F. The composition, age and genomic distribution of the different subfamilies have been examined. Subfamily age estimates indicate that the expansion of four SVA subfamilies (SVA_A, SVA_B, SVA_C and SVA_D) began before the divergence of human, chimpanzee and gorilla, while subfamilies SVA_E and SVA_F are restricted to the human lineage. Furthermore, I examined the amplification dynamics of SVA elements throughout the primate order and traced their origin back to the beginnings of hominid primate evolution, approximately 18 to 25 million years ago, which makes SVA elements the youngest family of retrotransposons in the primate order. Gene duplication is one of the most important mechanisms for creating new genes and generating genomic novelty. Retrotransposon-mediated sequence transduction (*i.e.* the process by which a retrotransposon carries flanking sequence during its own mobilization) has been proposed as a gene duplication mechanism. SVA elements are capable transducing 3' flanking sequence during retrotransposition. I examined all the full-length SVA elements in the human genome to assess the frequency and impact of SVA-mediated 3' sequence transduction. The results showed that ~53 kb of genomic sequence has been duplicated by 143 different SVA-mediated transduction events. In particular, I identified one group of SVA elements that has duplicated the entire *AMAC* gene three times in the human genome via SVA-mediated transduction events, which happened before the divergence of humans and African great apes. In addition to the original *AMAC* gene, the three transduced

AMAC copies contain intact open reading frames (ORFs) in the human genome and at least two are actively transcribed in different human tissues. Thus, duplication of entire genes and creation of new gene families via retrotransposon-mediated sequence transduction represent an important mechanism by which mobile elements impact their host genomes.

CHAPTER ONE:
INTRODUCTION

Mobile elements are interspersed DNA sequences that were first discovered by Barbara McClintock in her study of the controlling element at the dissociation locus in the maize genome (McClintock 1956; McClintock 1984). They constitute substantial portions of almost all eukaryotic genomes sequenced to date. Transposons and retrotransposons are two major types of mobile elements utilizing the nucleic acid intermediate for their transposition process. DNA transposons possess inverted terminal repeats and encode a transposase protein that they use to self-excise from the genome (Mizuuchi 1992; Smit and Riggs 1996). There are no known active DNA transposons in the human genome, although active DNA transposons are present in the genomes of bacteria, plants and flies (Lander et al.2001). Retrotransposons replicate themselves via an RNA intermediate and generate a new copy in the genome (Feng et al. 1996; Luan et al. 1993; Moran et al. 1996). Due to their “copy and paste” behavior, retrotransposons can accumulate fast in the genome and have had a large impact on genomic composition and architecture (Deininger and Batzer 2002).

Retrotransposable elements can be classified as either autonomous or non-autonomous. Those elements that encode the machinery necessary for their own mobility are considered autonomous. Autonomous retrotransposons are further divided into two subclasses of elements depending on whether or not they are flanked by long terminal repeats (LTRs). LTR-containing retrotransposons resemble retroviruses in structure, except for the absence of a functional envelope (*env*) gene, the protein product of which is used to transport elements between cells (Ono et al. 1987). In humans, the most abundant members of this class are the human endogenous retroviruses (HERVs) which comprise ~1-2% of the genome (Lower et al. 1996).

The non-LTR retrotransposons utilize a promoter sequence located within the 5' end of the coding sequence and make polyadenylated RNAs during their retrotransposition. These

RNAs differ from traditional mRNAs in that they are generally bicistronic RNAs that code for an RNA-binding protein (ORF1) and an ORF2 protein with endonuclease and reverse transcriptase domains (Kazazian 2000). The major non-LTR retrotransposon in humans is the long interspersed element (LINE).

Three different families of LINEs exist in the human genome (i.e., LINE1 (L1), LINE2 (L2) and LINE3 (L3)) (Lander et al. 2001). L1s have a long evolutionary history dating back to the beginnings of eukaryotic existence. About 520,000 L1s are dispersed throughout the human genome to date, accounting for 17% of the genome by mass (Smit 1999). Full-length L1s contain about 6kb nucleotides and have a poorly characterized internal promoter at the 5' end; two open reading frames (ORF1 and ORF2); a functional AATAAA polyadenylation (polyA) signal and are usually flanked by short duplications of genomic DNA, called target site duplications (TSDs) (Moran et al. 1999) (Figure1.1).



Figure 1.1: Full-length L1 structure. Full-length L1 is ~6kb in length. TSDs are shown as arrows. 5'UTR, ORF1, ORF2, 3'UTR, polyadenylation signal and polyA tail are indicated.

L1 ORF1 encodes an approximately 40-kDa protein with nucleic acid binding activity. The exact size of the ORF1 protein varies among species and sometimes even within species (Hohjoh and Singer 1997). L1 ORF2 encodes an approximately 150-kDa protein with

endonuclease (EN), reverse transcriptase (RT), and zinc knuckle (a zinc finger-like motif) domains (Kazazian and Moran 1998). The L1 EN domain cleaves one strand of double-stranded DNA at a large number of genomic sites characterized by the loose consensus sequence, AA/TTTT (Feng et al. 1996). The L1 RT domain is thought to function in the nucleus, use genomic DNA as primer, and carry out reverse transcription through the relatively simple process of target primed reverse transcription (TPRT) (Luan et al. 1993). In humans, the two L1 ORFs are in frame and are separated by a 63-bp non-coding spacer region that contains stop codons in all three reading frames. Unlike humans, mice L1s have two non-overlapping ORFs in different reading frames, while rat L1s have two overlapping ORFs (Ostertag and Kazazian 2001a).

The active L1 consensus element also contains a functional AATAAA polyadenylation (polyA) signal, which is required for RNA polymerase II (Pol II) termination, and proper cleavage and polyadenylation. The polyadenylation of L1 after the AATAAA signal suggests that Pol II transcribes L1 (Hirose and Manley 1998; McCracken et al. 1997; Osheim et al. 1999). However, there are two unusual features of the L1 polyadenylation signal: (i) the AATAAA polyadenylation signal of L1 elements is immediately followed by the presumed polyA tail; (ii) L1s usually lack important sequences downstream of the polyadenylation site. So L1 elements frequently bypass their own polyA signal and use a downstream signal (Moran et al. 1999). This process results in the retrotransposition of additional genomic sequence located 3' of the L1 element and is called L1-mediated transduction (Boeke and Pickeral 1999).

L1s are reverse transcribed and integrated into the genome by a process termed target-primed reverse transcription (TPRT). TPRT was first demonstrated for the R2 element, a site-specific, non-LTR retrotransposon found in arthropods (Luan et al. 1993). R2 retrotransposons

only have one ORF with both Type II restriction endonuclease and reverse transcriptase activities. *In vitro* experiments have demonstrated that the endonuclease domain of the R2 element cleaves the non-coding strand of its target site, a sequence in the 28S rRNA gene. A free 3' hydroxyl terminal at the DNA nick serves as a primer and the R2 RNA is used as a template for the reverse transcription reaction. Reverse transcription of the RNA is followed by cleavage of the coding strand and integration. Perfect duplications of small stretches of DNA (usually 7-20 bp) surrounding the original target site, which flank the newly inserted element, are produced by TPRT (Luan et al. 1993).

Several lines of evidence support the hypothesis that L1s also use TPRT to reverse transcribe and integrate into the genome. First, *in vitro* experiments showed that the full-length L1 ORF2 protein produces limited TPRT reactions (Cost et al. 2002). Second, most L1s in the genome are flanked by perfect 7-20 bp TSD, which is a typical consequence of the TPRT reaction. Third, the predicted cleavage site is often located in T-rich region, which is complementary to the polyA tail at the 3' end of an L1 element, suggesting that they could indeed be used as a primer for reverse transcription of the L1 RNA.

The human genome contains 3000-4000 full length L1s. Among these, only about 80-100 full length L1s are retrotranspositionally competent (Brouha et al. 2003; Lander et al. 2001). The retrotransposition mechanism of L1s shows *cis*-preference: the protein products of a particular L1 usually bind to the RNA from the same L1, although low levels of *trans*-complementation cannot be ruled out (Ostertag and Kazazian 2001a). The vast majority of L1 elements are truncated at their 5' ends. This might be due to an inability of the L1 reverse transcriptase to copy the entire L1 RNA before disassociating from the RNA. About 25% of recently inserted L1s also contain an inversion within their inserted portions, size-ranging from

several hundred to fifteen hundred nucleotides in length (Ostertag and Kazazian 2001b). A proposed model called “twin priming” is used to explain this phenomenon. In this model, the second DNA strand is sometime nicked before the reverse transcription is completed, the additional 3' hydroxyl is then used as the primer to invade the L1 RNA internally. Then the L1 RNA template is primed by two different primers at two separate locations and the resolution of the RNA/cDNA structure causes the typical L1 inversion with a 5' truncation (Ostertag and Kazazian 2001b).

There are several classes of non-autonomous elements that appear to rely on the L1-generated enzymatic machinery for their retrotransposition. The most abundant of these are the short interspersed elements (SINEs). SINEs are small elements, usually 90-300 bp in length, and are transcribed by RNA polymerase III. These elements are either ancestrally derived from *tRNA* genes or the *7SL RNA* gene (Daniels and Deininger 1985; Ullu and Tschudi 1984). SINEs use internal RNA polymerase III promoters for transcription, which allows these elements to carry their promoters to new genomic locations. However, these promoters are also subsequently dependent on flanking sequences to stimulate expression levels *in vivo* (Chesnokov and Schmid 1996; Roy et al. 2000). SINEs do not contain any protein-coding sequence, and they are integrated into the host genome by borrowing a LINE-encoded protein.

Alu elements are the only known active SINEs within the human genome (Batzer and Deininger 2002). These successful elements have freeloading wildly on the back of their partner L1, to produce over 1.2 million copies per haploid genome during the 65 million years of evolutionary time (Lander et al. 2001). The full-length *Alu* element is about 300 bp long, composed of two monomers connected by a 19-bp A-rich linker (Figure 1.2). Like other non-LTR retrotransposons, *Alu* elements are flanked by short TSDs that are remnants of the

retrotransposition process. Due to their evolutionary origin as pseudogenes of the *7SL RNA* gene, *Alu* elements contain a split RNA polymerase III promoter that is crucial for their amplification.



Figure 1.2: *Alu* element structure. *Alu* element is ~300 bp in length. Internal RNA polymerase III promoter is indicated as A and B boxes. TSDs are shown as arrows. Middle polyA linker and polyA tail are indicated.

Although a great number of *Alu* elements exist in the human genome, only a small subset of *Alu* elements is thought to be retrotranspositionally competent “source genes” (Batzer and Deininger 2002; Cordaux et al. 2004). While promoter integrity and the length and homogeneity of the polyA tail have been suggested as principal factors in determining the retrotranspositional capability of *Alu* elements, the criteria required for successful retrotransposition are still not fully resolved (Cordaux et al. 2004; Roy-Engel et al. 2002). It is also possible that some post-transcriptional selections of *Alu* transcripts are involved in the retrotranspositional activity (Sinnott et al. 1992). Over the course of primate evolution, mutations within *Alu* source genes have created nearly 30 subfamilies, generating a hierarchy of elements that have amplified over defined periods of time (Price et al. 2004).

Unlike L1s and *Alu* elements, which have been extensively studied (Batzer and Deininger 2002; Deininger and Batzer 1999; Kazazian 2000; Ostertag and Kazazian 2001a), SVA elements represent a new kind of non-autonomous retrotransposons, and only a few studies have been conducted to date to elucidate their properties (Ono et al. 1987; Ostertag et al. 2003). The SVA element was originally named SINE-R, with the R indicating its retroviral origin (Ono et al. 1987). In 1994, Shen *et al.* identified a new composite retroposon, which consisted of the SINE-R element together with a stretch of sequence sharing similarity with *Alu* sequences, when they studied the structure of the RP gene (Shen et al. 1994). Thus, they named the element SVA after its main components, SINE, VNTR and Alu (Shen et al. 1994).

The full-length SVA element can be divided into five components (Figure 1.3): (1), a $(CCCTCT)_n$ hexamer simple repeat region located at the 5' end; (2), an *Alu* homologous region, usually composed of two anti-sense *Alu* fragments and additional sequence of unknown origin; (3), a variable number of tandem repeat (VNTR) region, composed of a variable number of copies of a 35-50 bp repeated sequence; (4), a ~490 bp long SINE region, which is derived from



Figure 1.3: Full-length SVA structure. Full-length SVA element is ~2kb in length. Various Regions of the SVA element are denoted. TSDs are shown as arrows.

the 3' end of the *env* gene and the 3' LTR of the endogenous retrovirus HERV-K10 (Ono et al. 1987); and (5) a polyA tail immediately following a putative polyadenylation signal (AATAAA).

SVA elements are thought to have been mobilized by L1s (Ostertag et al. 2003). Similar to L1s, SVA elements also lack downstream motifs that are important for efficient transcription termination. Therefore, when they are transcribed, the RNA transcription machinery sometimes skips the element's own weak polyadenylation signal and terminates transcription using a downstream polyadenylation site located in the 3' flanking genomic sequence. The transcript containing the retrotransposon along with the extra genomic sequence is subsequently integrated back into the genome through retrotransposition, a process termed transduction.

SVA elements share the sequence similarity with *Alu* elements, but lack the internal RNA polymerase III promoter of the *Alu* family. SVA elements may be transcribed by RNA polymerase II based on the following reasons: first, SVA elements are too long to be transcribed by RNA polymerase III. Second, SVA element has a putative polyadenylation site. According to the sequence structure of both types (with and without transduced sequence), polyA tail is an added sequence to the element during the retrotransposition event. In the first type with 3' transduced sequence, the polyA tail at the end of the transduced sequence was not in the genomic sequence but was post-transcriptionally added. In the second type with 3' truncation, an alternative polyadenylation site in the SVA element sequence was used, so the transcripts of these elements were truncated and the polyA tail was added after the alternative site. The third evidence supporting the RNA polymerase II transcription of SVA elements is that an extra G nucleotide is present at the beginning of about one third of the SVA elements. This extra G may represent the 5' capping modification of the SVA RNA sequence by RNA polymerase II. Similar addition of G caps has been observed in L1s, and L1 reverse transcriptase has been known to

reverse transcribe the 5' cap structure of the mRNA (Hirzmann et al. 1993; Lavie et al. 2004; Volloch et al. 1995).

SVA elements remain active in the human genome, as demonstrated by their involvement in the creation of various diseases. To date, five diseases, caused by either *de novo* insertion or recombination between pre-existing elements, have been reported that are associated with SVA elements (Callinan and Batzer 2006; Kobayashi et al. 1998; Legoix et al. 2000; Ostertag et al. 2003; Rohrer et al. 1999; Wilund et al. 2002). So SVA element is the third known category of retrotransposons currently expanding in the human lineage, along with L1 and *Alu* elements (Batzer and Deininger 2002; Ostertag and Kazazian 2001a).

This dissertation is a detailed study of the SVA family. In chapter two, I primarily analyzed the SVA elements on a genome-wide scale. First, I estimated the copy number in the human genome and provided the genomic distribution pattern of the SVA elements. In this study, I not only examined the subfamily structures of SVA elements within the human genome, but also analyzed the relationship between SVA elements and their flanking genomic sequences. Second, I traced the origin of the entire SVA family back to the beginning of hominid primate radiation and determined the copy numbers of SVA elements in different non-human primate genomes.

In chapter three, I investigated a less well-characterized mechanism that can potentially duplicate genes, namely the transduction of flanking genomic sequence associated with the retrotransposition of mobile elements. I examined the extent and properties of SVA-mediated transduction events to evaluate their evolutionary impact on the human genome. And I identified the first case of the entire gene duplication caused by mobile element-mediated transduction in the human genome. The results demonstrate that retrotransposon-mediated sequence transduction

is not only a mechanism for exon shuffling, but also serves as a novel mechanism for gene duplication and the creation of new gene families.

References

- Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370-379.
- Boeke, J.D. and O.K. Pickeral. 1999. Retroshuffling the genomic deck. *Nature* **398**: 108-109, 111.
- Brouha, B., J. Schustak, R.M. Badge, S. Lutz-Prigge, A.H. Farley, J.V. Moran, and H.H. Kazazian, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100**: 5280-5285.
- Callinan, P.A. and M.A. Batzer. 2006. Transposable elements and human disease In *Genome Dynamics* (ed. J.-N. Volff), pp. 104-115. S Karger AG, Basel (Switzerland).
- Chesnokov, I. and C.W. Schmid. 1996. Flanking sequences of an Alu source stimulate transcription in vitro by interacting with sequence-specific transcription factors. *J Mol Evol* **42**: 30-36.
- Cordaux, R., D.J. Hedges, and M.A. Batzer. 2004. Retrotransposition of Alu elements: how many sources? *Trends Genet* **20**: 464-467.
- Cost, G.J., Q. Feng, A. Jacquier, and J.D. Boeke. 2002. Human L1 element target-primed reverse transcription in vitro. *Embo J* **21**: 5899-5910.
- Daniels, G.R. and P.L. Deininger. 1985. Repeat sequence families derived from mammalian tRNA genes. *Nature* **317**: 819-822.
- Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol Genet Metab* **67**: 183-193.
- Deininger, P.L. and M.A. Batzer. 2002. Mammalian retroelements. *Genome Res* **12**: 1455-1465.
- Feng, Q., J.V. Moran, H.H. Kazazian, Jr., and J.D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905-916.
- Hirose, Y. and J.L. Manley. 1998. RNA polymerase II is an essential mRNA polyadenylation factor. *Nature* **395**: 93-96.
- Hirzmann, J., D. Luo, J. Hahnen, and G. Hobom. 1993. Determination of messenger RNA 5'-ends by reverse transcription of the cap structure. *Nucleic Acids Res* **21**: 3597-3598.

- Hohjoh, H. and M.F. Singer. 1997. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *Embo J* **16**: 6034-6043.
- Kazazian, H.H., Jr. 2000. Genetics. L1 retrotransposons shape the mammalian genome. *Science* **289**: 1152-1153.
- Kazazian, H.H., Jr. and J.V. Moran. 1998. The impact of L1 retrotransposons on the human genome. *Nat Genet* **19**: 19-24.
- Kobayashi, K., Y. Nakahori, M. Miyake, K. Matsumura, E. Kondo-Iida, Y. Nomura, M. Segawa, M. Yoshioka, K. Saito, M. Osawa et al. 1998. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* **394**: 388-392.
- Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
- Lavie, L., E. Maldener, B. Brouha, E.U. Meese, and J. Mayer. 2004. The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res* **14**: 2253-2260.
- Legoix, P., H.D. Sarkissian, L. Cazes, S. Giraud, F. Sor, G.A. Rouleau, G. Lenoir, G. Thomas, and J. Zucman-Rossi. 2000. Molecular characterization of germline NF2 gene rearrangements. *Genomics* **65**: 62-66.
- Lower, R., J. Lower, and R. Kurth. 1996. The viruses in all of us: characteristics and biological significance of human endogenous retrovirus sequences. *Proc Natl Acad Sci U S A* **93**: 5177-5184.
- Luan, D.D., M.H. Korman, J.L. Jakubczak, and T.H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**: 595-605.
- McClintock, B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21**: 197-216.
- McClintock, B. 1984. The significance of responses of the genome to challenge. *Science* **226**: 792-801.
- McCracken, S., N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S.D. Patterson, M. Wickens, and D.L. Bentley. 1997. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature* **385**: 357-361.
- Mizuuchi, K. 1992. Transpositional recombination: mechanistic insights from studies of mu and other elements. *Annu Rev Biochem* **61**: 1011-1051.

- Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530-1534.
- Moran, J.V., S.E. Holmes, T.P. Naas, R.J. DeBerardinis, J.D. Boeke, and H.H. Kazazian, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**: 917-927.
- Ono, M., M. Kawakami, and T. Takezawa. 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res* **15**: 8725-8737.
- Osheim, Y.N., N.J. Proudfoot, and A.L. Beyer. 1999. EM visualization of transcription by RNA polymerase II: downstream termination requires a poly(A) signal but not transcript cleavage. *Mol Cell* **3**: 379-387.
- Ostertag, E.M., J.L. Goodier, Y. Zhang, and H.H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73**: 1444-1451.
- Ostertag, E.M. and H.H. Kazazian, Jr. 2001a. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35**: 501-538.
- Ostertag, E.M. and H.H. Kazazian, Jr. 2001b. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* **11**: 2059-2065.
- Price, A.L., E. Eskin, and P.A. Pevzner. 2004. Whole-genome analysis of Alu repeat elements reveals complex evolutionary history. *Genome Res* **14**: 2245-2252.
- Rohrer, J., Y. Minegishi, D. Richter, J. Eguiguren, and M.E. Conley. 1999. Unusual mutations in Btk: an insertion, a duplication, an inversion, and four large deletions. *Clin Immunol* **90**: 28-37.
- Roy-Engel, A.M., A.H. Salem, O.O. Oyeniran, L. Deininger, D.J. Hedges, G.E. Kilroy, M.A. Batzer, and P.L. Deininger. 2002. Active Alu element "A-tails": size does matter. *Genome Res* **12**: 1333-1344.
- Roy, A.M., N.C. West, A. Rao, P. Adhikari, C. Aleman, A.P. Barnes, and P.L. Deininger. 2000. Upstream flanking sequences and transcription of SINES. *J Mol Biol* **302**: 17-25.
- Shen, L., L.C. Wu, S. Sanlioglu, R. Chen, A.R. Mendoza, A.W. Dangel, M.C. Carroll, W.B. Zipf, and C.Y. Yu. 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J Biol Chem* **269**: 8466-8476.
- Sinnett, D., C. Richer, J.M. Deragon, and D. Labuda. 1992. Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J Mol Biol* **226**: 689-706.

- Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9**: 657-663.
- Smit, A.F. and A.D. Riggs. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* **93**: 1443-1448.
- Ullu, E. and C. Tschudi. 1984. Alu sequences are processed 7SL RNA genes. *Nature* **312**: 171-172.
- Volloch, V.Z., B. Schweitzer, and S. Rits. 1995. Transcription of the 5'-terminal cap nucleotide by RNA-dependent DNA polymerase: possible involvement in retroviral reverse transcription. *DNA Cell Biol* **14**: 991-996.
- Wilund, K.R., M. Yi, F. Campagna, M. Arca, G. Zuliani, R. Fellin, Y.K. Ho, J.V. Garcia, H.H. Hobbs, and J.C. Cohen. 2002. Molecular mechanisms of autosomal recessive hypercholesterolemia. *Hum Mol Genet* **11**: 3019-3030.

CHAPTER TWO:
SVA ELEMENTS: A HOMINID SPECIFIC RETROPOSON
FAMILY*

*Reprinted by permission of Journal of Molecular Biology

Introduction

Transposons and transposon-like repetitive elements collectively occupy about 44% of the human genome. *Alu* and L1 (long interspersed elements-1) elements account for ~30% of the genome sequence and are the most abundant transposable elements in humans (Lander et al. 2001), while human endogenous retroviruses (HERVs) represent another ~1% of the human genome. In addition to the major retrotransposon families, there are smaller families of transposons such as SVA, which are receiving increased attention lately due to their recent retrotransposition activity in the human genome (Bennett et al. 2004; Ostertag et al. 2003).

The SVA element was originally named SINE-R, with the R indicating its retroviral origin (Ono et al. 1987). In 1994, Shen *et al.* identified a new composite retroposon when they studied the structure of the RP gene (Shen et al. 1994). This new retroposon consisted of the SINE-R element together with a stretch of sequence that shares sequence similarity with *Alu* sequences. Thus, it was named “SVA” after its main components, SINE, VNTR and *Alu* (Shen et al. 1994).

SVA elements contain the hallmarks of retrotransposons, in that they are flanked by target site duplications (TSDs), terminate in a poly(A) tail and they are occasionally truncated and inverted during their integration into the genome (Ostertag et al. 2003). In addition, they can transduce 3' sequences during their movement. Therefore, it has been proposed that SVA elements are non-autonomous retrotransposons that are mobilized by L1 encoded proteins *in trans* (Ostertag et al. 2003).

SVA elements remain active in the human genome, as demonstrated by their involvement in the creation of various diseases. To date, at least four diseases have been reported related to SVA insertions (Callinan and Batzer 2006; Kobayashi et al. 1998; Legoix et al. 2000; Ostertag et

al. 2003; Rohrer et al. 1999; Wilund et al. 2002). This makes SVA the third known category of retrotransposons currently expanding in the human lineage, along with L1 and *Alu* elements (Batzer and Deininger 2002; Ostertag and Kazazian 2001).

To assess the distribution and impact of SVA elements in the human genome, we examined all the SVA elements in the human genome reference sequence (Lander et al. 2001). Six SVA subfamilies were identified and characterized. For the two human-specific subfamilies, the associated insertion presence/absence human genomic diversity was analyzed. Furthermore, we traced the origin of the entire SVA family back to the beginning of hominid primate radiation and determined the copy number of SVA elements in different non-human primate genomes. The overall distribution of SVA elements showed a significant correlation with genomic G+C content and gene density.

Results

Copy Number and Genomic Distribution

Copy Number of the SVA Elements

In total, 2762 SVA elements were identified in the human genome draft sequence (hg17, May 2004 freeze). Together, they account for 4.2 Mb of the genome, with an average density of one element per 1.03 Mb. Among them, 1752 elements are full-length, composing 63% of the group. The copy number of SVA elements in the chimpanzee genome (panTro1, Nov 2003 freeze) was also determined. A total of 2637 elements were obtained, with 42% being full-length. After examining the truncated SVA elements in the chimpanzee genome draft sequence, we found that a large proportion of the SVA elements were truncated by stretches of Ns, corresponding to unsequenced/unassembled regions of the genome. Therefore, the lower proportion of full-length elements in the chimpanzee genome may be due to the lower quality of

the chimpanzee draft sequence compared to its finished human counterpart. Detailed wet bench estimates of the copy number and phylogenetic distribution of SVA elements throughout the primate order are outlined below.

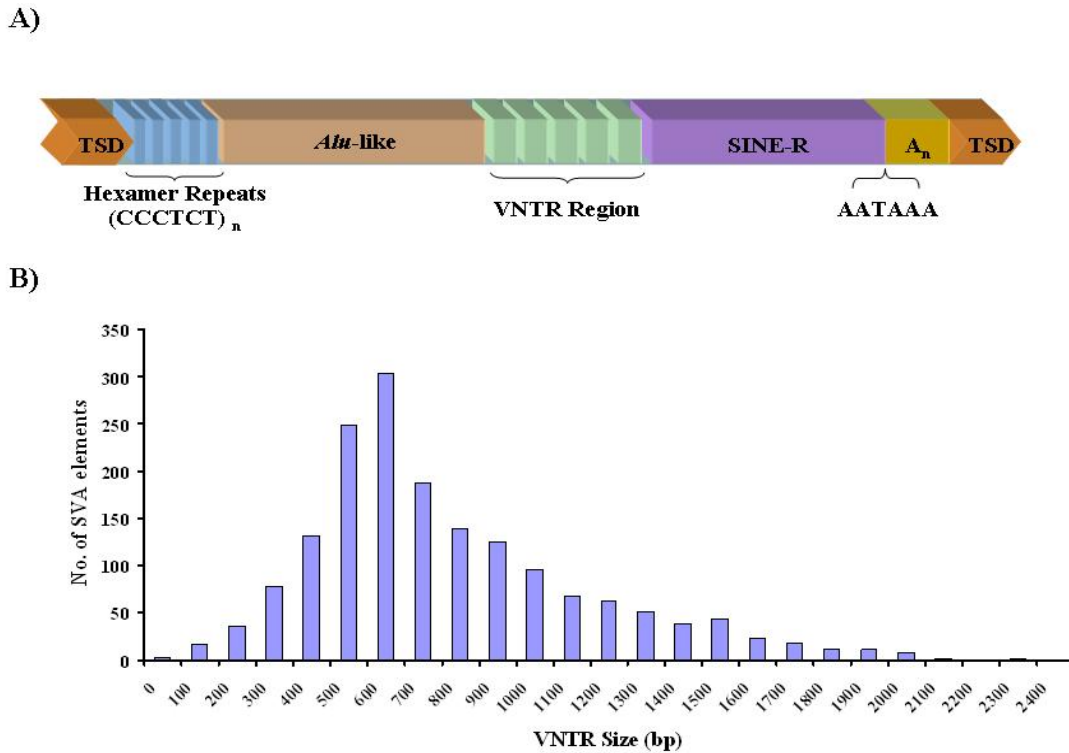


Figure 2.1: Structure analysis of SVA element. (A) Structure of a full-length SVA element with the target site duplications. Various Regions of the SVA element are color-coded and denoted. **(B) Size distribution of SVA VNTR regions.** The VNTR regions of SVA elements are shown in 100 bp intervals or bins.

The Structure of the SVA Element

The full-length SVA element can be divided into five components (Figure 2.1(a)): (1), a $(CCCTCT)_n$ hexamer simple repeat region which is located at the 5' end; (2), an *Alu* homologous region, usually composed of two antisense *Alu* fragments and an additional sequence of unknown origin; (3), a variable number of tandem repeat (VNTR) region, composed of a variable number of copies of a 35-50 bp sequence; (4), a short interspersed element (SINE) region about 490 bp

long, which is derived from the 3' end of the *env* gene and the 3' long terminal repeat (LTR) of the endogenous retrovirus HERV-K10 (Ono et al. 1987); and (5) a poly (A) tail after a putative polyadenylation signal (AATAAA). We extracted the VNTR region from all of the full-length SVA elements and analyzed their length variation (Figure 2.1(b)). The length of the SVA VNTR region varies from 48 to 2306 bp, with an average value of 819 bp. Two-thirds of the SVA elements have VNTR lengths in the range of 400-900 bp.

Genomic Distribution

To examine the genomic distribution of SVA elements, we first analyzed their distribution at the chromosomal level by comparing the observed number of the elements with the expected numbers, assuming an infinite sites (random) insertion model (Figure 2.2). In this

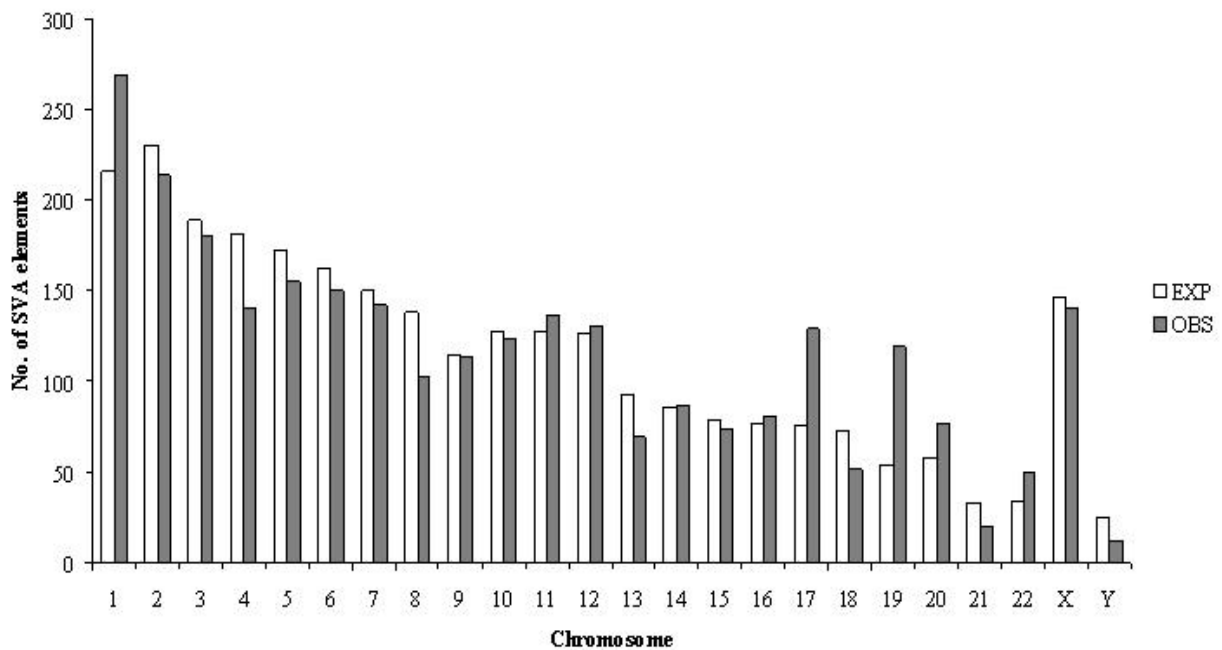


Figure 2.2: Human genomic distribution of SVA elements. The observed and expected numbers of SVA elements from each chromosome are shown. The expected numbers on each chromosome were obtained by multiplying the chromosomal length with average density of these elements in the genome.

model, the number of insertions on each chromosome was solely proportional to the size of the chromosome. A chi-square analysis revealed that the two distributions were significantly different ($\chi^2 = 78.29$, $df = 23$, $p < 0.001$), thus leading us to reject the simple random insertion model. In particular, the density of SVA elements was found to be much higher than expected on chromosomes 1, 17, 19 and 22. On the other hand, chromosomes 4, 5, 13, 18, 21 and Y had far fewer elements than expected. A similar non-random chromosomal distribution has previously been observed for *Alu* elements (Carter et al. 2004; Grover et al. 2004).

Since the properties of the genomic sequence vary greatly among different human chromosomes (Lander et al. 2001), we further analyzed the distribution of SVA elements in relation to other genomic properties, such as G+C content, gene content and repeat content of the flanking nucleotide sequence. At the whole-genome level, we observed a positive correlation between SVA density and both G+C content ($r = 0.59$, $p = 0.002$) and gene density ($r = 0.53$, $p = 0.007$). In terms of distribution, the SVA elements resemble *Alu* elements, which are also preferentially found in high G+C/gene -rich regions of primate genomes, although the extent of correlation is much higher in the case of *Alu* elements (Lander et al. 2001). In terms of vicinity to genes, we observed that 1025 SVA elements reside in intronic regions of the genes and 1737 elements occur in intergenic regions. Statistically, there is no significant difference ($\chi^2 = 11.48$, $df = 23$, $p = 0.98$) between the number of SVA elements in these two regions. This result suggests that there is no preference for SVA insertion in genes or intergenic regions.

To examine the SVA distribution in relation to the flanking repeat content, we extracted nucleotide sequence intervals of 1, 2, 5, 10, 25 and 50 kb from both 5' and 3' flanking regions of the SVA elements and analyzed the repeat content using RepeatMasker (Table 2.1). We found that the repeat content in these different-sized flanking sequences were not significantly different

from each other or the overall repeat content in the genome ($p > 0.2$ in each case). Then, we studied individually the two most abundant repeat families, *Alu* and L1 in these flanking sequences. When compared with their expected distribution frequency throughout the genome, *Alu* elements were found to be overly represented in these regions, whereas L1 elements were underrepresented (in both cases, $p < 0.00001$). A closer analysis of the SVA flanking regions revealed that the G+C content in the flanking regions is much higher than the genomic average ($p < 0.00001$). Given high *Alu* density in G+C-rich, and high L1 density in A+T-rich regions of the genome (Lander et al. 2001), these results are not surprising and indicate that SVA elements are not preferentially distributed in repeat rich regions of the genome.

Table 2.1: Repeat content and G+C content in the SVA flanking regions.

	Total Repeat (%)	Alu (%)	L1 (%)	G+C (%)
1 kb	43.50	12.40	12.25	42.05
2 kb	43.97	12.79	12.62	41.91
5 kb	44.18	12.79	12.76	42.85
10 kb	44.08	12.82	12.74	42.84
25 kb	43.91	13.02	12.67	43.00
50 kb	45.12	13.00	12.52	43.06
Whole Genome	44.83	10.60	16.89	40.91

Subfamily Analysis

Subfamily Structure and Composition

If SVA elements expanded in a process similar to *Alu* and L1 elements, a hierarchical subfamily structure should be formed. To identify possible subfamily structure, multiple alignments of the SVA elements were constructed. Due to the highly variable VNTR region and considerable number of 5' truncated elements, only the LTR derived region (referred to as the S part in the following text) was used for the initial analysis. Examination of the alignments resulted in the recovery of at least six different groups among the elements based on their diagnostic substitutions. Consensus sequences for these groups of SVA elements were generated in BioEdit v7.0 (Hall 1999) using a “majority rules” approach. Some of the CpG sites were reconstructed manually due to their high mutation rate (Jurka et al. 2002; Xing et al. 2004). These subfamilies were named SVA_A to SVA_F. Using a similar approach, one chimpanzee-specific subfamily was identified and named SVA_PtA. The lineage specific distribution of this SVA subfamily was verified by BLAT (blast-like alignment tool) (Kent 2002) comparison to the human genome.

To validate the subfamily definition, multiple alignments were also constructed using the *Alu* related region (referred to as the A part in the following text) together with the S part and same groups were identified. Thus, we constructed the human SVA subfamily consensus including both A and S regions of the element. The multiple alignments of the human SVA subfamily consensus along with the corresponding HERV-K10 sequences are available in supplemental data (Figure A.1).

Next, the relative proportions of each SVA subfamily were calculated (Figure 2.3). Among all the subfamilies, SVA_D represents the largest subfamily and accounts for over 40%

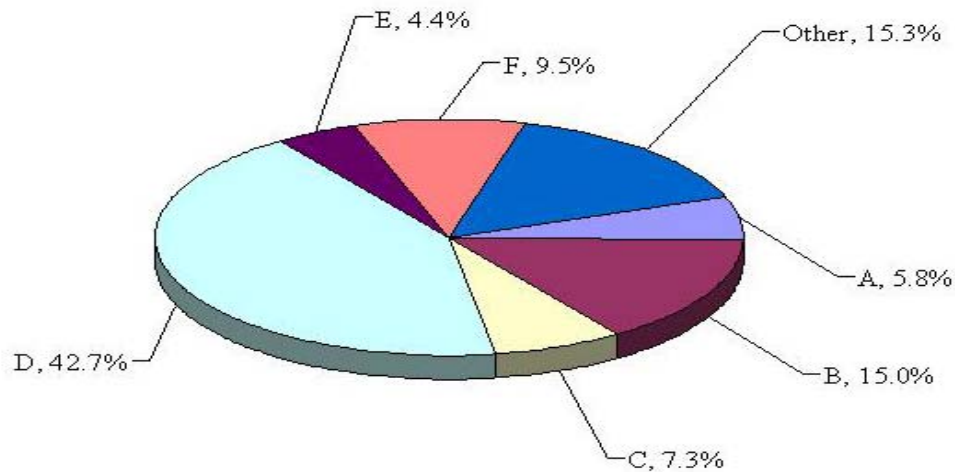


Figure 2.3: SVA subfamily composition in the human genome. The breakdown of SVA elements into various subfamilies is shown as a percentage of all the SVA elements in the human genome.

of the family. The second largest subfamily, SVA_B, accounts for about 15% of the family. The remaining subfamilies (A, C, E and F) have relatively small numbers of elements (<300). Since we recognized the SVA subfamilies based on the intact S part of the elements, we could not group the elements into subfamilies when they lacked a full-length S part. These truncated elements were collectively designated “others” and accounted for 15% of the family.

SVA Subfamily Age Estimates

The ages of different SVA subfamilies were estimated using a method similar to that used for *Alu* subfamilies described previously (Xing et al. 2004). Briefly, the S parts of the SVA elements in each subfamily were aligned with the subfamily consensus. Next, substitutions in this region were divided into CpG and non-CpG substitutions and substitution density was calculated using a Perl script. Separate neutral substitution rates of 0.0015/site/million year (myr) and 0.0090/site/myr were used for non-CpG and CpG substitution, respectively (Xing et al. 2004). The substitution density and age estimate of each subfamily are shown in Table 2.2.

Given the approximate divergence time of hominid primates (Glazko and Nei 2003; Goodman et al. 1998), the age estimates indicate that subfamily SVA_A (13.56 Myrs) may have expanded contemporary to the divergence of the orangutan and the great apes (human, chimpanzee and gorilla) (12-15 million years ago (Mya)). The expansion of subfamilies SVA_B (11.56 Myrs), SVA_C (10.88 Myrs) and SVA_D (9.55 Myrs) may have predated the human, chimpanzee and gorilla divergence (~7 Mya). The relatively young age of subfamilies SVA_E (3.46 Myrs) and SVA_F (3.18 Myrs) suggested these two subfamilies may have expanded after the human and chimpanzee divergence (~4-6 Mya). Indeed, when the human and chimpanzee sequences were compared using BLAT, the members of subfamilies SVA_E and SVA_F were absent at chimpanzee orthologous loci, confirming their human-specific distribution.

Table 2.2: Age estimates of SVA subfamilies.

SVA Subfamily	CpG Density (%)	Non-CpG Density (%)	CpG Age (Myrs)	non-CpG Age (Myrs)	Age average (Myrs)
SVA_A	15.13	1.55	16.81	10.30	13.56
SVA_B	9.53	1.88	10.59	12.53	11.56
SVA_C	9.73	1.64	10.81	10.94	10.88
SVA_D	8.50	1.45	9.45	9.64	9.55
SVA_E	2.21	0.67	2.46	4.47	3.46
SVA_F	2.46	0.54	2.73	3.63	3.18

Having obtained age estimates for the SVA subfamilies, we examined the properties of SVA VNTR repeat regions as a function of subfamily and age. ANOVA results indicate that there are significant ($p \ll 0.001$) differences among subfamilies for VNTR lengths. With respect to age, if the oldest SVA subfamily (SVA_A) is excluded from the analysis, we find a significant negative correlation ($r^2 = 0.96$) of VNTR lengths with time (Figure 2.4). It is unclear why this subfamily deviates from the overall pattern observed.

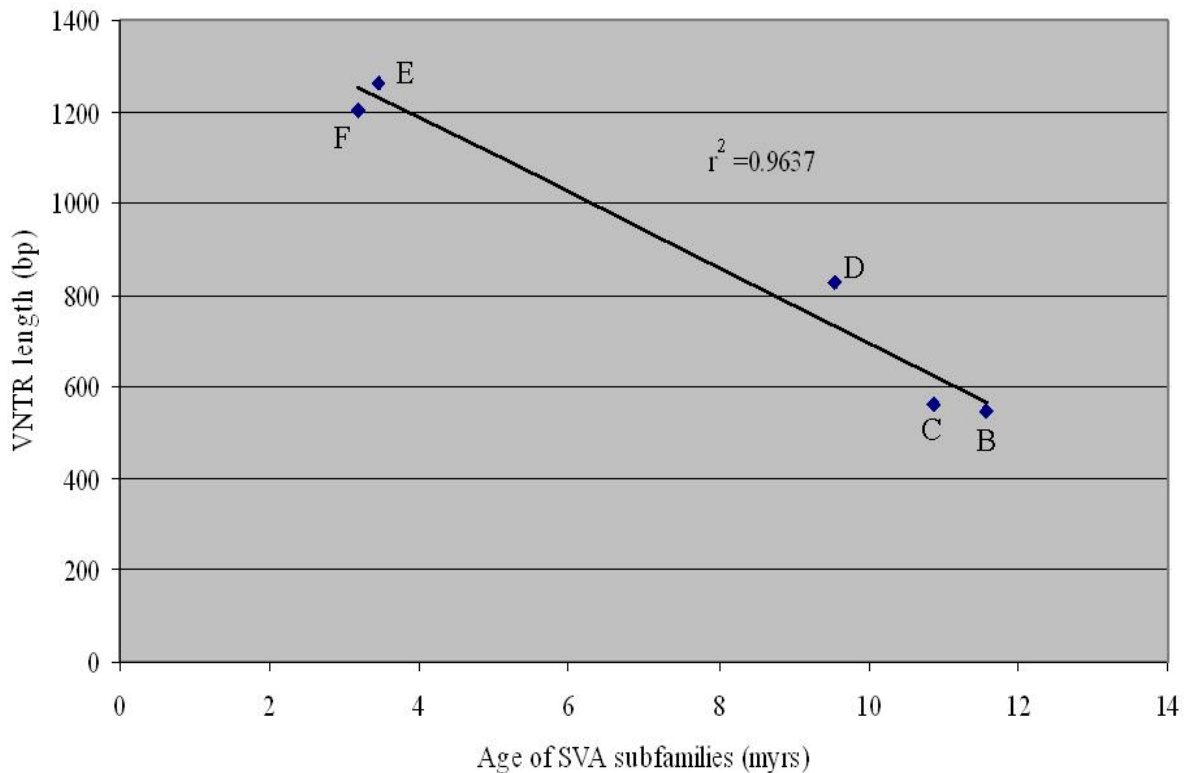


Figure 2.4: The relationship between VNTR lengths and age. Each dot represents one SVA subfamily. Correlation coefficient (r) is shown.

G+C Content Distribution, Gene Density and Repeat Density in the Flanking Regions of the Subfamilies

A correlation analysis between SVA content and G+C content of all human chromosomes was performed to examine the distributions of SVA subfamilies (Figure 2.5). Different levels of correlation were observed, with the highest values for the youngest subfamilies SVA_E ($r = 0.75$, $p = 0.000022$) and SVA_F ($r = 0.57$, $p = 0.0035$), mild correlation for SVA_D ($r = 0.45$, $p = 0.027$), and no significant correlation for SVA_A, SVA_B and SVA_C. The general trend suggests the enrichment of the youngest SVA subfamilies on the high G+C content chromosomes. This distribution is in contrast to *Alu* and L1 elements, whose younger subfamilies are preferentially found in A+T-rich regions (Medstrand et al. 2002).

To examine whether the distribution of SVA subfamilies observed at the chromosomal level also exists in different regions within a chromosome, the human genome was separated into 12 bins based upon their G+C content (Medstrand et al. 2002) and the SVA density in each of the bins was calculated. We pooled the subfamilies into three groups according to their ages: the oldest subfamilies (A, B and C), the intermediate subfamily (D) and young subfamilies (E and F). Their distributions in relation to the genomic G+C content were plotted (Figure 2.6). As shown in the figure, SVA elements reside mainly in medium to medium-high G+C-rich regions of the genome, with maximum density in the bins that correspond to 42-50% G+C content. However, there was a shift in SVA density towards higher G+C bins with a decrease in evolutionary age. The older subfamilies (A, B and C) were quite rich in the 40-42% G+C bin, where the members of young subfamilies (E and F) were found less frequently. The opposite was observed in 48-52% G+C content regions of the genome which were extremely rich in young SVA elements. The SVA_D subfamily exhibited an intermediate pattern for density in these G+C bins.

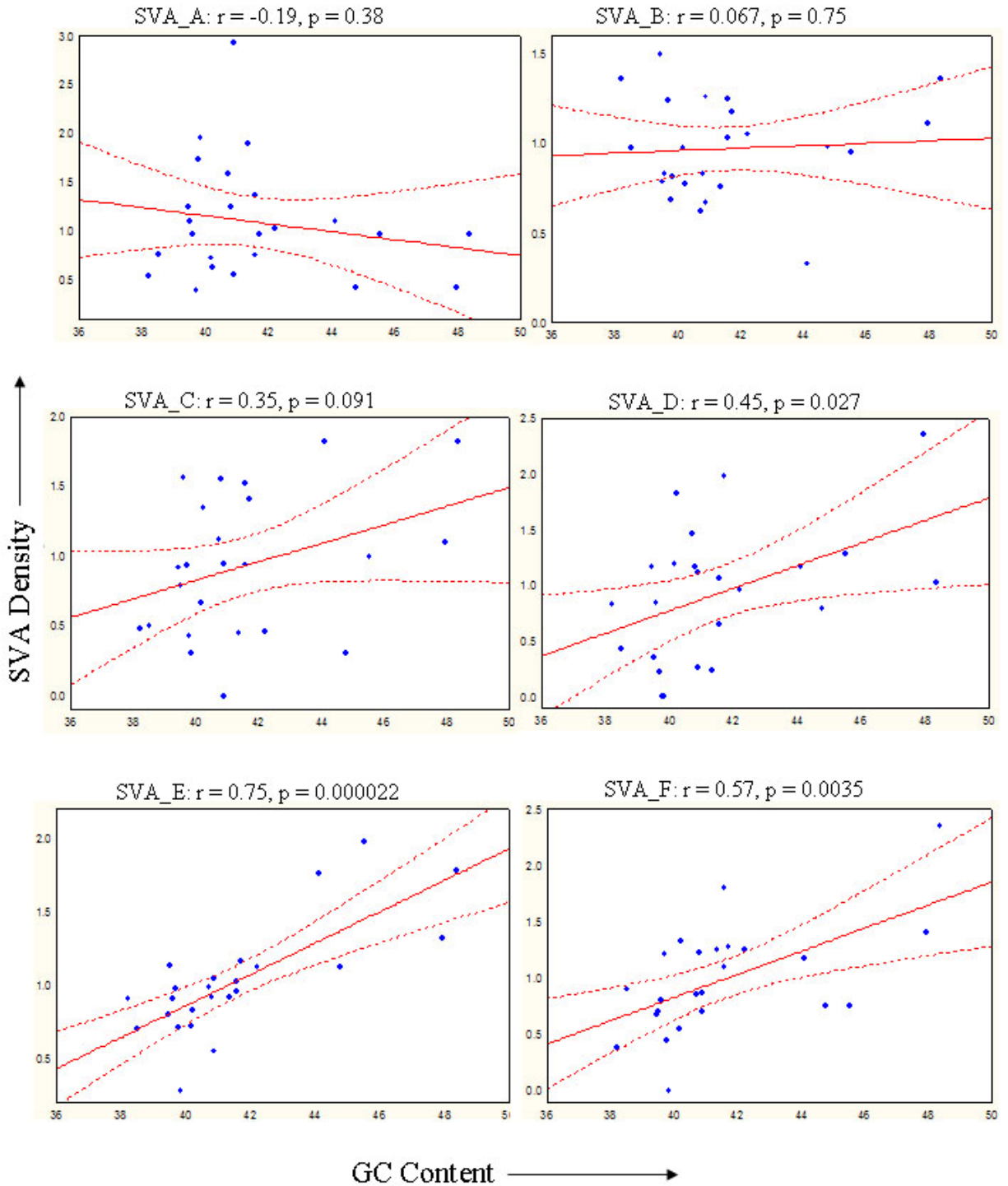


Figure 2.5: Correlation between SVA subfamily distribution and chromosomal G+C content. The linear regression is denoted with a continuous line and the 95% confidence intervals are denoted by the broken lines. Correlation coefficients (r) and p values are shown.

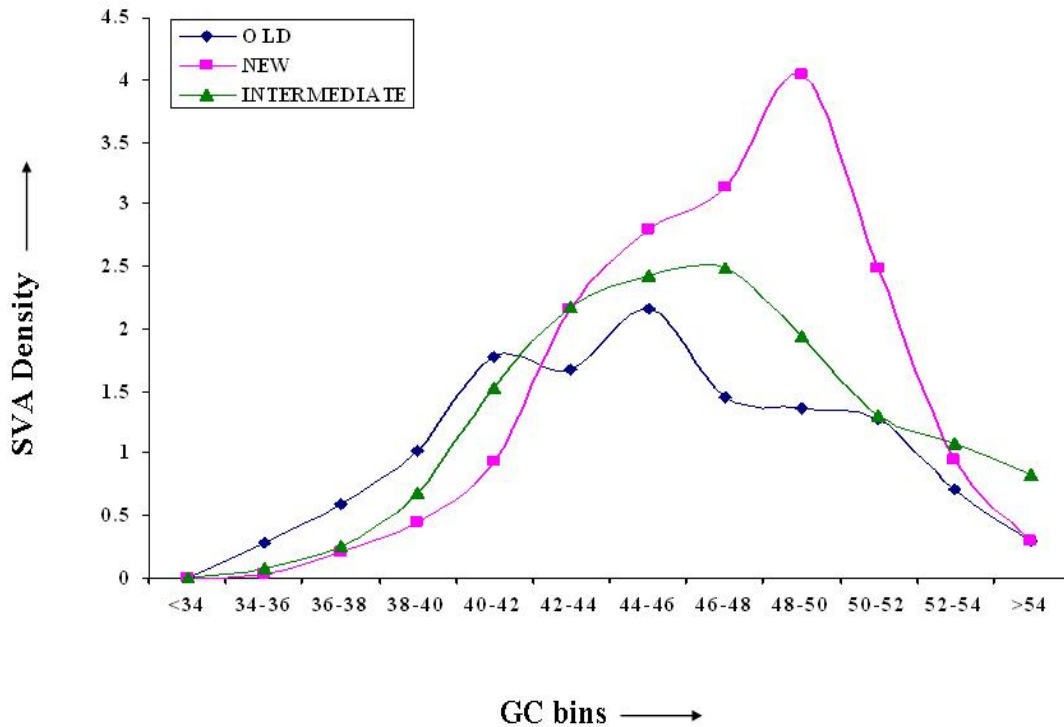


Figure 2.6: SVA subfamily and genomic G+C content. The bin separations are identical with those previously described by Lander *et al.* SVA subfamilies are pooled into old (SVA_A, SVA_B, SVA_C), intermediate (SVA_D) and young (SVA_E and SVA_F) subfamilies.

The correlation analysis between SVA subfamilies and chromosomal gene content were examined and similar results were obtained as in the case of G+C content. Strong correlations were obtained from young subfamilies SVA_E ($r = 0.77$, $p < 0.0001$) and SVA_F ($r = 0.64$, $p < 0.001$); moderate correlation existed for subfamily SVA_D ($r = 0.54$, $p = 0.007$) and SVA_C ($r = 0.47$, $p = 0.02$). The correlations were very low and insignificant for the oldest members of this repeat family (SVA_A and SVA_B) (Figure 2.7). This result is not surprising, given that gene densities and G+C content are highly correlated in human genome (Lander et al. 2001).

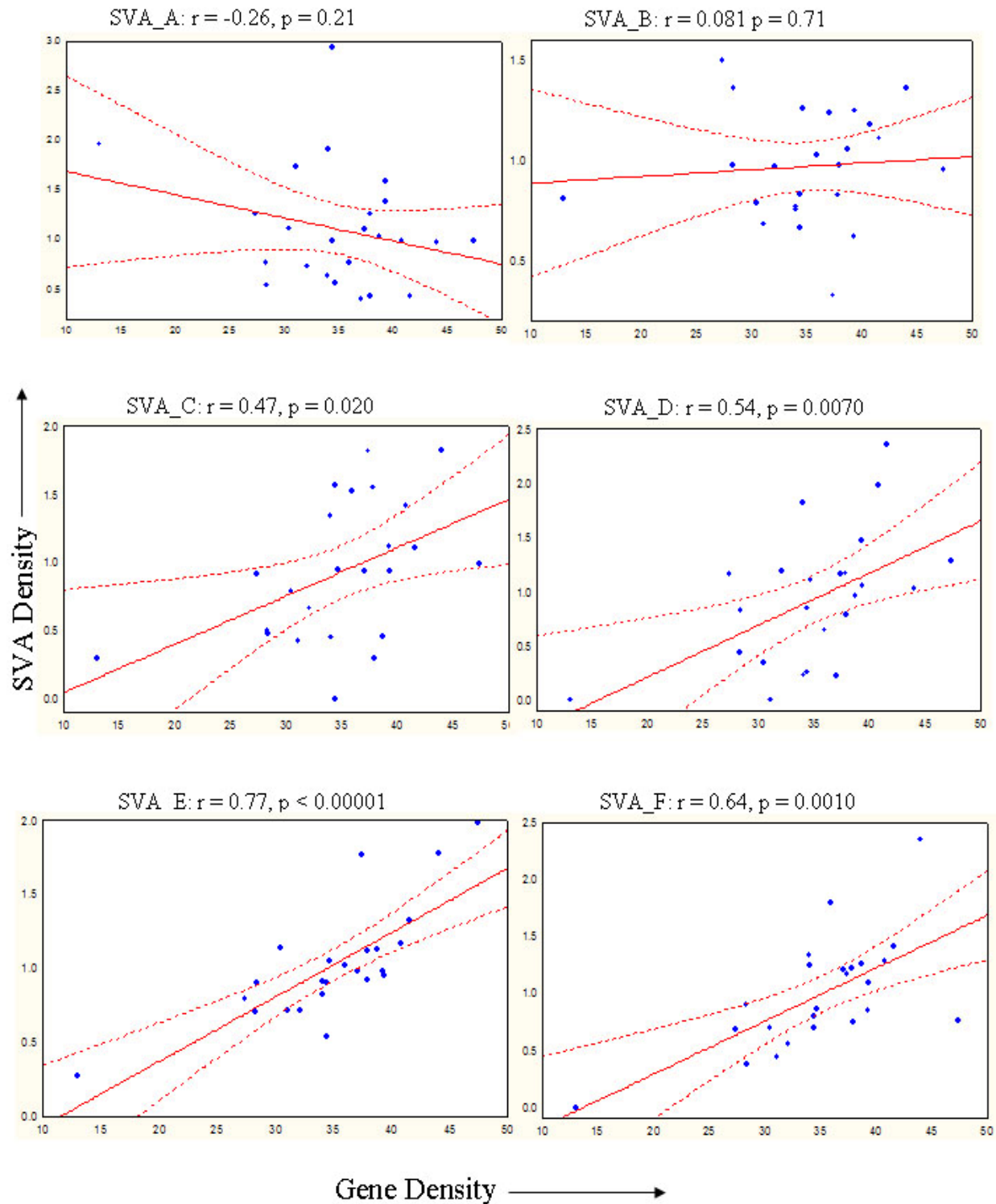


Figure 2.7: Correlation between SVA subfamilies and chromosomal gene density. The linear regression is denoted with a continuous line and the 95% confidence intervals are denoted by the broken lines. Correlation coefficients (r) and p values are shown.

We next analyzed the repeat distribution in the flanking regions of SVA elements belonging to each subfamily (Figure 2.8). The variations in *Alu* and L1 contents across different subfamilies correspond with the G+C content of the flanking regions. In fact, we observed that *Alu* and L1 element density in these regions is significantly correlated with the G+C and A+T content, respectively ($p > 0.001$). These data suggest that SVA elements are not preferentially distributed in either *Al*-rich or L1-poor regions of the genome.

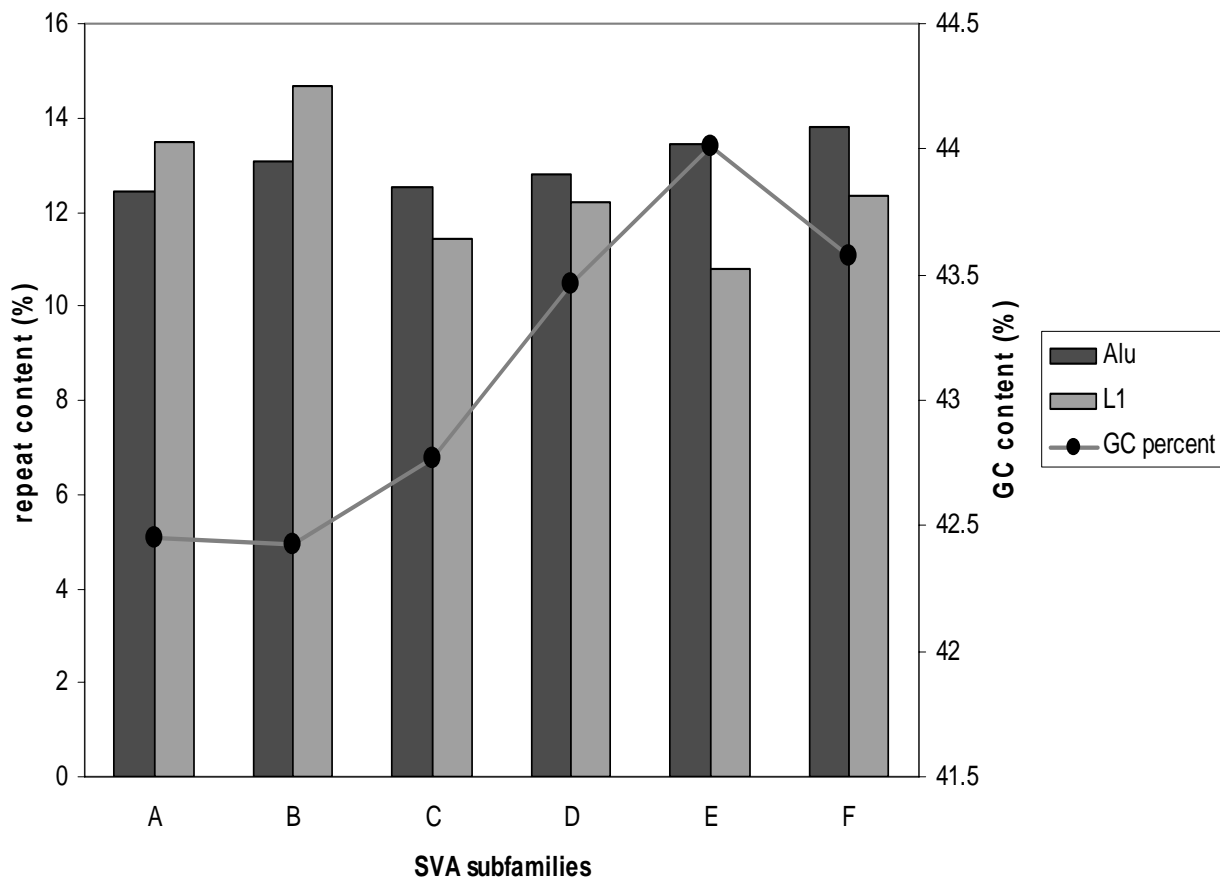


Figure 2.8: L1 and *Alu* densities flanking SVA elements and G+C content. The percentage of *Alu* and L1 elements flanking SVA elements from different subfamilies are shown by the dark bars. The flanking unique sequence G+C content is shown by the line among SVA subfamilies.

Human Genomic Diversity of Two Human-specific Subfamilies

To assess the human genetic diversity related to the SVA elements, elements from SVA_E and SVA_F subfamilies were screened for insertion polymorphism on a diverse human population panel (see Materials and Methods). In total, 48 members of subfamily SVA_E and 58 members of SVA_F were surveyed. For SVA_E, 37.5% (18/48) of the elements showed insertion presence/absence polymorphism on our panel with a 27.6% (16/58) polymorphism rate for SVA_F. The insertion polymorphism rates of these two subfamilies are comparable to the human-specific *Alu* and L1 subfamilies with similar ages (Carter et al. 2004; Myers et al. 2002; Otieno et al. 2004; Salem et al. 2003), further demonstrating the contemporary retrotransposition activity of the SVA family of retroposons. The detailed insertion allele frequencies, heterozygosities and genotypes for the SVA insertion polymorphisms are shown in supplemental data Table A.1.

Evolution of the SVA Elements

Evolution of SVA Subfamilies

To examine the relationship among the subfamilies, a median-joining network (Bandelt et al. 1999; Cordaux et al. 2004) was constructed using the S part of the subfamily consensus and the corresponding region of HERV-K10 (Figure 2.9). The network analysis indicates that the older SVA subfamilies evolved in a single lineage: the SVA_A consensus has the highest sequence similarity to the HERV-K10 counterpart, differing by 27 substitutions. SVA_B differs from SVA_A by 9 substitutions and a 16 bp deletion is present at the 5' end of the SVA_B consensus as compared to HERV-K10 sequence and the SVA_A consensus. This deletion is present in all other subfamily consensus as well. SVA_C is derived from SVA_B and differs from SVA_B by 10 substitutions while SVA_D differs from SVA_C by 15 substitutions. Unlike

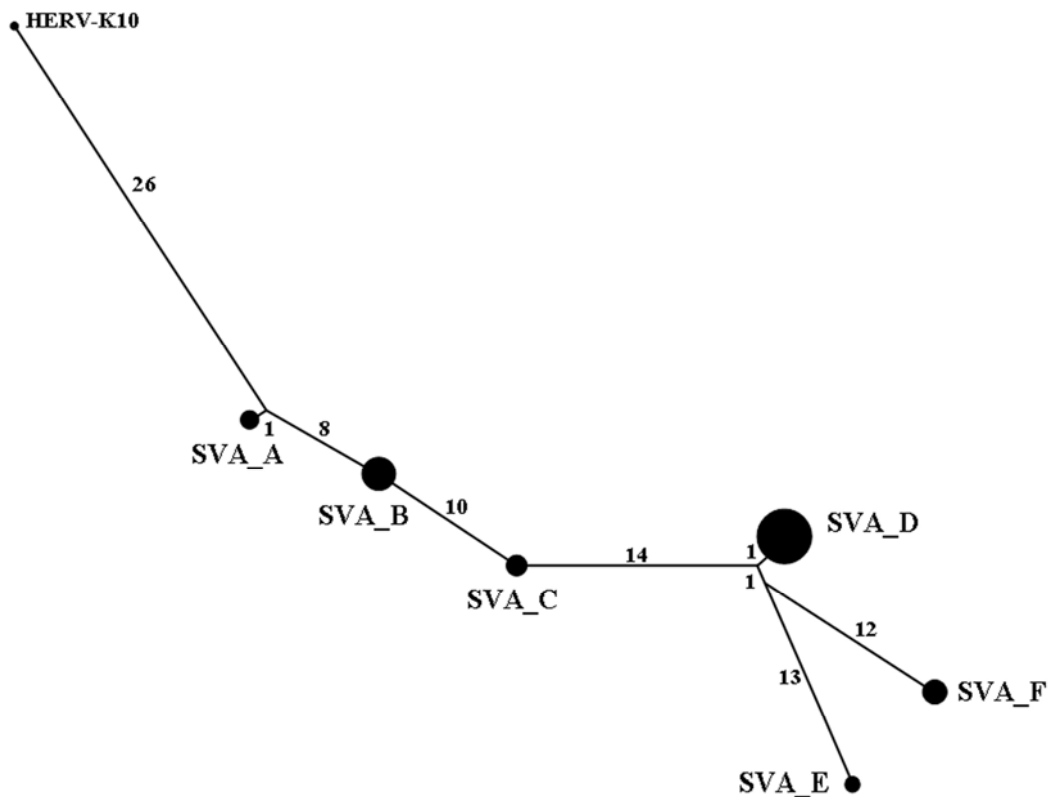


Figure 2.9: Median-joining network of the SVA subfamilies. The network of the SVA subfamily was reconstructed using the S part of the consensus sequences from each subfamily along with the corresponding region of HERV-K10. The lines denote substitution steps, with the number of substitutions shown on the top of the lines. The size of each of the circles corresponds to the relative size of the subfamily in the human genome.

the older subfamilies, the human-specific SVA_E and SVA_F derived independently from a reconstructed ancestral sequence, which is only one substitution different from SVA_D differ and they differ from SVA_D by 15 and 14 substitutions, respectively.

Copy Number of SVA Elements in Different Non-Human Primate Genomes

To further investigate the evolutionary history of the SVA elements, the copy number of SVA elements in different primate genomes was estimated by quantitative PCR (QPCR) using a pair of intra-SVA primers. QPCR results from different primates were normalized based on the

SVA copy number in the human genome (2762) (Figure 2.10). The result indicated that both common chimpanzee (*Pan troglodytes*) (2769) and pygmy chimpanzee (*Pan paniscus*) (2646) had similar number of elements compared to human, while gorilla (*Gorilla gorilla*) (2334) appeared to have about 400 fewer SVA elements. Nevertheless, given the standard deviations in our copy number estimates (90, 163 and 221 in common chimpanzee, pygmy chimpanzee and gorilla, respectively), the number of SVA elements in human, chimpanzees and gorilla may not be appreciably different. By contrast, orangutan (*Pongo pygmaeus*), which diverged from gorilla about 12-15 myrs ago (Glazko and Nei 2003; Goodman et al. 1998), has fewer than 1000 SVA elements. Siamang (*Hylobates syndactylus*) has about 40 elements and no SVA elements were detected in any of the two Old World monkeys (green monkey (*Chlorocebus aethiops*), rhesus macaque (*Macaca mulatta*)) we examined.

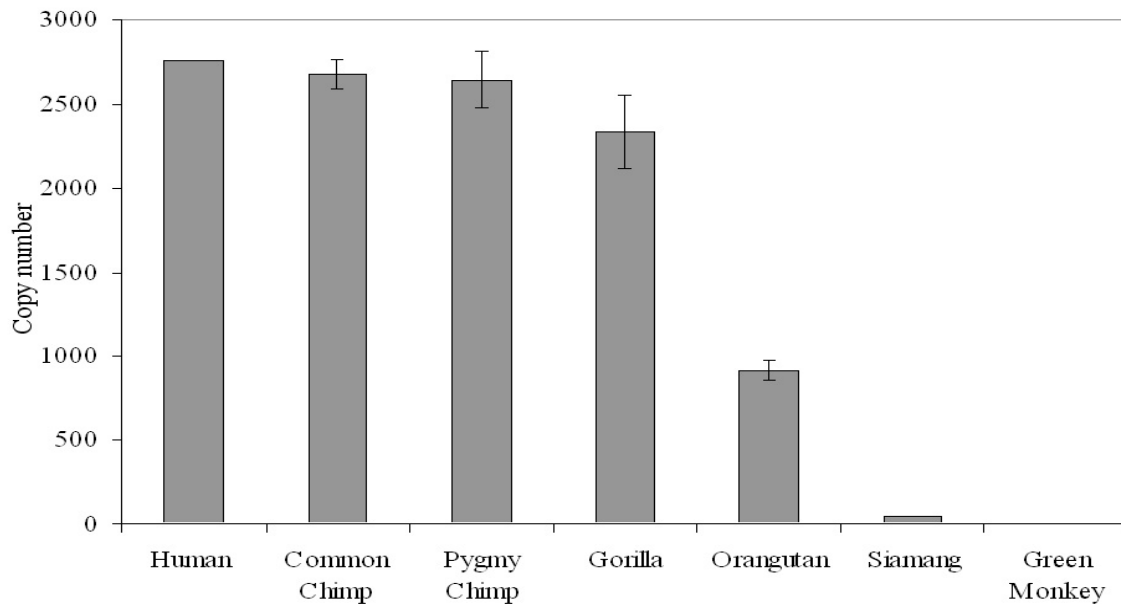


Figure 2.10: SVA copy number in the primate lineage. The signal intensity derived from the number of known SVA elements in human genome (2762) was used as standard for the estimations from quantitative PCR experiments. These experiments utilized QPCR of the S part of each element to estimate the copy number as outlined in the Materials and Methods. The standard deviation of three replicates each estimation is shown in the figure as a bar.

To verify the QPCR results, the available Old World monkey genomic sequences in the NCBI database were searched for SVA elements using BLAST (basic local alignment search tool) (Altschul et al. 1990). In addition, the rhesus macaque draft assembly, Mmul_0.1 (UCSC version rheMac1) was searched using BLAT. In agreement with the QPCR results, no authentic SVA elements were identified from either of the database searches. It is noteworthy that the individual components of the SVA elements (*Alu* region, VNTR region and LTR-derived region) are present in the Old World monkey genomes, although the origin or the “composition” of the full-length SVA element appears to have happened after the divergence of hominid and Old World primates.

Furthermore, a PCR display assay was performed as described in previous study (Han et al. 2005) on the two Old World monkeys (rhesus macaque and green monkey) and four hominid primates (pygmy chimpanzee, common chimpanzee, gorilla and gibbon (*Hylobates lar*)), by using two primers specific for the *Alu* part and SINE-R part of the SVA element (the SVA_*Alu* primer: 5'-ATTGAGCACTGAGTGAACGAGA-3'; and the SVA_SINE-R primer: 5'-AGTACCCAGGGACACAAACT-3'). Although SVA elements were readily isolated from all of the hominid primates, none was recovered from either of the two Old World primates.

Discussion

Copy Number and Subfamily Composition of SVA Elements

The total number of SVA elements reported here was based on a whole-genome analysis; therefore it is more likely to represent the actual number of SVA elements in the human genome compared to previous studies (Ono et al. 1987; Ostertag et al. 2003). It should be noted here that due to the polymorphic nature of some of the young SVA elements, a proportion of the total number of elements will not be retrieved from any single reference genome that has been

sequenced (Hedges et al. 2004). Therefore the copy number reported here represents the lower limit of the number of SVA elements in the human genome. For the common chimpanzee genome, a similar number of elements were identified. However, a large proportion of the SVA elements were truncated or flanked by Ns, preventing further investigation of the SVA family in the chimpanzee genome. With improvement in the assembly of the common chimpanzee genome, a more thorough examination of the SVA elements in chimpanzees can be achieved in the future.

Multiple alignment of the S part of the SVA elements revealed at least six subfamilies based on their diagnostic mutations. Aside from the SVA_B and SVA_D subfamilies, the rest of the subfamilies have relatively small numbers of elements (<300). The smaller size of a particular subfamily could be the result of a relatively short amplification period or the lower retrotransposition activity of the source gene(s). The complex nature of the amplification of retrotransposons has recently been underscored by the “stealth model” of *Alu* amplification (Han et al. 2005). It remains a possibility that SVA subfamilies with currently low copy numbers are simply currently in a state of retrotranspositional quiescence.

Truncated SVA Elements

SVA elements were initially named SINE-R due to their close relationship to the endogenous retrovirus HERV-K10 (Ono et al. 1987). The question remained whether the S part of the element first originated from HERV-K10 and expanded alone similar to LINE-mediated integration of HERV-W (Pavlicek et al. 2002) and the recruitment of the A and VNTR parts happened later, or if the SVA element started the expansion as a whole. To address this issue, we examined all of the truncated SVA elements. By comparing the sequence of truncated elements with the subfamily consensus, we found truncated elements distributed among all the subfamilies

we identified. This suggests that SVA elements expanded as one composite unit. Furthermore, we examined the alignments of the two oldest subfamilies, SVA_A and SVA_B. The results showed that the majority (>70%) of the elements were full-length, containing the A part of the elements. These two lines of evidence suggest that the SVA elements were fully assembled/composed very early during their expansion as a repeated DNA sequence family. One very interesting possibility is that the acquisition of the antisense *Alu* fragments and the hexamer simple repeats changed the properties of the element (e.g. the hexamer simple repeats provided a promoter for the element) and initiated the expansion of the SVA family. Additional evidence, including cell culture based assays need to be gathered to test this hypothesis.

The Mobilization of SVA Elements

Sequence examination of SVA elements showed that, unlike *Alu* elements, SVA elements do not have an RNA polymerase III (pol III) internal promoter. In addition, the full-length transcript (~2kb) is too long for normal RNA pol III transcription. As such, it seems likely that the SVA element is transcribed by RNA polymerase II (pol II), similar to the L1 elements. Indeed, several lines of evidence support this concept: first, SVA has a putative polyadenylation site and the poly(A) tail is added to the element during the retrotransposition. Two types of integrations clearly showed the addition of poly (A) tails: (1), in elements with 3' transduced sequence, the poly (A) tail at the end of the transduced sequence was not in the genomic sequence but was post-transcriptionally added; (2), in 3' truncated elements, an alternative polyadenylation site in the SVA element sequence was used so the transcripts of these elements were truncated and the poly(A) tail was added after the alternative site. Second, an extra G was present at the beginning of about one-third of the SVA elements. This extra G may represent the 5' capping modification of the SVA RNA sequence by RNA pol II. Similar addition of G

residues has been observed for L1 elements (Lavie et al. 2004) and reverse transcriptase has been known to reverse transcribe the 5' cap structure (Hirzmann et al. 1993; Volloch et al. 1995).

If SVA elements are indeed transcribed by RNA pol II, do they have their own promoters or do they rely on the promoter activity of their flanking regions? We believe SVA elements may rely on both approaches. At least 10% of SVA elements have 3' transductions and can be traced to different origination points. The possibility that most of these several hundred elements have fortuitously integrated right after a viable promoter is remote. We believe that these active elements bear a functional promoter region and are capable of transcription by themselves. By contrast, there is at least one 5' transduced SVA element which suggests that SVA may be able to use a 5'-flanking promoter.

Genomic Distribution of SVA Elements

Initial analysis suggested that the genomic distribution of SVA elements was more similar to *Alu* elements than to L1s. However, when we studied each SVA subfamily individually, apparent differences were observed between the distributions of SVA and *Alu* elements. The distribution of the youngest SVA elements showed an apparent enrichment in the G+C-rich regions in the genome, with a peak at 48-50% G+C content. When the age of the SVA elements increased, the enrichment of the elements became less apparent and the peak of distribution shifted towards the more A+T-rich regions. By contrast, young *Alu* subfamilies are found in more A+T-rich regions while the enrichment shifts to more G+C-rich regions over time (Lander et al. 2001). Thus, although the overall distributions of *Alu* and SVA elements are both in G+C-rich region, the shifting of their G+C distribution patterns are opposite.

Several scenarios could explain the data noted above. For instance, if SVA elements prefer to insert in A+T-rich regions, the distribution of the elements may be due to positive

selection acting on the young elements in the G+C-rich regions. The positive selection hypothesis requires selection to be acting on the majority of the SVA elements and will result in the fixation of the elements that are under selection in a short period of time. However, the ~30% polymorphism observed for the two human-specific subfamilies (<4 Myrs old) is consistent with neutral expectations. Therefore, positive selection is unlikely to be the causative factor for the observed scenario. Another possibility is that elements were preferentially removed from the A+T-rich regions. The removal of *Alu* elements from A+T-rich regions due to unequal homologous recombination is thought to contribute to the shifting of the *Alu* distribution pattern over time (Hackenberg et al. 2005; Pavlicek et al. 2001). But given that there is one SVA every 1.03 Mb on average (compared to one *Alu* every 3 kb), the probability of a large scale removal of the elements via this type of recombination seems remote. Nevertheless, due to the relatively low copy number of SVA elements, less recombination-related problems may be generated as compared to L1 elements. This may give the SVA elements a better chance to be fixed in G+C-rich regions. The possibility remains that this fixation advantage of SVA elements in G+C-rich region combining with other unknown mechanisms shifted their distribution over time.

Another possibility is that SVA elements preferentially inserted in the G+C-rich regions. Under this hypothesis, more SVA elements would have to be removed from G+C-rich regions over time to generate the observed shift in the distribution. However, this seems counter-intuitive for at least two reasons. Firstly, SVA elements are quite possibly using the retrotransposition machinery as L1 and *Alu* elements, (Ostertag et al. 2003) and both young *Alu* and L1 elements are found more common in the A+T-rich regions of the genome (Lander et al. 2001). Secondly, the G+C-rich regions are known to have higher gene densities; therefore, the insertion of retroelements in these regions will have a larger chance of influencing gene expression and

causing genetic defects. Nevertheless, we have no evidence to rule out the possibility of the changing insertion preference at this moment. In addition to these scenarios, other mechanisms such as compositional matching (Gu et al. 2000) may also have played a role in shaping the SVA distribution.

Our observations in this study bring up an interesting question: if *Alu*, L1 and SVA elements are indeed all retrotransposed using the L1 enzymatic machinery, why are there such dramatic differences among their distributions? Multiple scenarios have been proposed for the shifting of the *Alu* distribution to G+C-rich regions, yet the topic remains intensively debated (Brookfield 2001; Cordaux et al. 2004; Gu et al. 2000; Hackenberg et al. 2005; Jurka et al. 2004; Lander et al. 2001; Ovchinnikov et al. 2001; Pavlicek et al. 2001; Rynditch et al. 1998). The observed distribution of the SVA elements undoubtedly added another interesting dimension to the overall retroelement genomic distribution puzzle.

The Impact of SVA Elements

Similar to other mobile elements, the insertion of SVA elements in the genome may have a profound impact on the genomic architecture and stability (Batzer and Deininger 2002; Deininger et al. 2003; Ostertag and Kazazian 2001). One of the properties of the SVA elements is their high G+C content. A typical full-length element has about 60% G+C content while the G+C content of the VNTR region may even exceed 70%. This makes each SVA element a potential mobile CpG island and the insertion of the element may influence the surrounding genomic environment. SVA elements are also enriched in potential functional units, including SP1 binding sites (GGCGG) in the VNTR region and hormone responsive elements (HRE) in the S part, to name a few. Since many of these units are harbored in the repetitive regions (both of the hexamer repeat region and the VNTR region), multiple copies can be found in a single

element. If inserted near a gene, SVA elements may influence the gene expression pattern. An examination of SVA insertions adjacent to the 5' end of annotated genes indicated that approximately 200 elements integrated within 5000 bp upstream of annotated genes in the UCSC May 2004 (hg17) human genome assembly. Each of these elements has the potential to alter the transcription of the nearby genes. However, detailed studies of gene expression will be required to accurately determine the impact of these elements on gene expression.

What is certain, however, is that SVA can have a negative impact on the genome. To date, four cases of different diseases have been reported related to the SVA insertions (Kobayashi et al. 1998; Ostertag et al. 2003; Rohrer et al. 1999; Wilund et al. 2002). The existence of diseases caused by SVA insertions, along with the presence of partial SVA transcripts in dbEST (data not shown), provide strong evidence for the ongoing amplification of SVA elements in the human genome. However, as a result of their relatively lower copy number (<3000), SVA elements are likely to have a much lower mutagenic recombination-based impact on the genome as compared to the more abundant *Alu* and L1s. In fact, only one case of disease caused by SVA-mediated recombination has previously been reported (Legoix et al. 2000).

Similar to L1s, SVA elements sometimes bypass their own polyadenylation site and use a downstream site, a phenomenon known as “transduction” (Moran et al. 1999; Ostertag and Kazazian 2001). In our study, we found ~10% of the SVA elements transduce sequence at their 3' end. The transduced sequence ranged from several base-pairs to several thousand base-pairs and some of them contained coding sequences. The ability of SVA elements to utilize 3' transduction to shuffle genetic material represents another impact that SVA elements have on the genome. In some cases, this so-called “exon shuffling” may generate new proteins with novel functions (Moran et al. 1999).

Evolution of SVA Elements

The evolutionary history of SVA retroposons was examined using human SVA elements. Since large amounts of genomic sequence data are not yet available for all primate species, we employed a QPCR-based approach to estimate the copy number of SVA elements in other primate genomes. QPCR has been proven useful in estimating DNA copy number (Alonso et al. 2004; Ginzinger et al. 2000; Walker et al. 2003), because there is a quantitative relationship between the amount of DNA target sequence and the amount of PCR product generated at any given PCR cycle prior to saturation. The copy number estimates in different primates and the age estimates of SVA subfamilies generated congruent scenarios for the evolution of the SVA elements (Figure 2.11): the SVA elements originated before the divergence of hominid primates but after the divergence of hominid and Old World primates. The evolutionary history of SVA elements is characterized by two major expansion periods: subfamilies SVA_A and SVA_B were expanded before the divergence of great apes (orangutan, gorilla, chimpanzee and human). Another major expansion of the SVA family occurred after the divergence of orangutan and the rest of the great apes. This major expansion is characterized by the amplification of subfamily SVA_D, which caused the number of SVA elements in human/chimp/gorilla to be higher than other species. The SVA_C subfamily may have also expanded during this period. After the divergence of gorilla from human and chimpanzee, SVA family showed lineage-specific activity, sprouting two human specific subfamilies. The independent expansion of multiple SVA subfamilies in the human genome indicates the existence of multiple SVA source genes, similar to *Alu* elements in the human genome (Cordaux et al. 2004).

As mentioned earlier, the evolutionary history of SVA elements was examined using the human lineage as a starting point. Even though the QPCR primers were designed in a well-conserved

region of the human SVA consensus, there is still insufficient data about the actual structure of SVA elements in other non-human primate genomes. The possibility remains that species-specific SVA insertions or even subfamilies exist in other hominid primates and can not be detected using QPCR based assays; thus, we can not rule out that our QPCR results may underestimate the SVA copy numbers in more divergent species. Additional studies are needed to focus on the SVA elements in the non-human primate genomes to augment the work presented here.

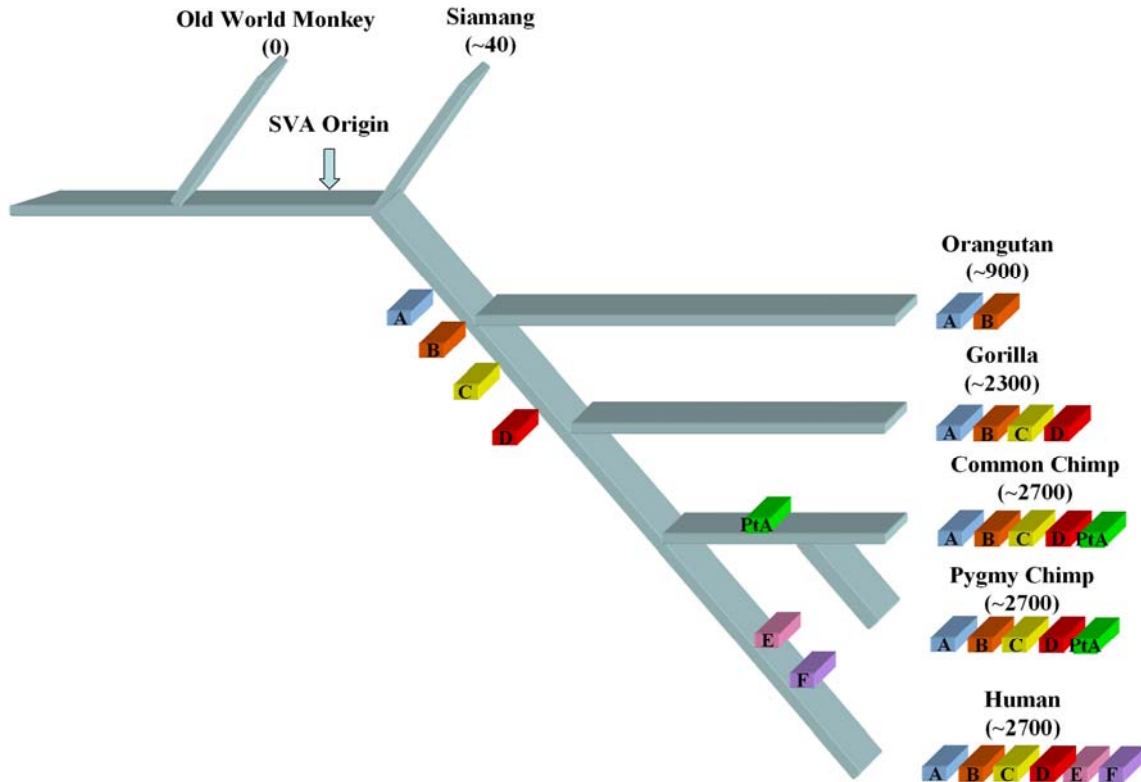


Figure 2.11: Amplification dynamics of the SVA family of retroposons in primates. Putative evolutionary history of the SVA lineage in primate genomes. A schematic of the hominid primate radiation is shown. The estimated copy number of SVA elements in various primate genomes is shown next to their names. The estimated time of origin and period of expansion for each subfamily are shown.

Conclusion

The SVA family of retroposons represents the third known category of mobile elements whose *de novo* mobilization results in human genetic disorders. Until now, little was known about the distribution and properties of these elements in the human genome as compared to *Alu* and L1 elements. In this study, we identified and characterized all of the SVA elements from the draft sequence of the human genome. In addition, the human genomic diversity associated with polymorphic SVA elements as well as the phylogenetic distribution of SVA elements was examined. By adding the contribution of the SVA family to our current knowledge of mobile elements, this study provides a more comprehensive picture and will further enhance our understanding of the impact of mobile elements on primate genomes.

Materials and Methods

Genome Analysis

The RepeatMasker annotations of the human (hg17, May 2004 freeze) and chimpanzee genomes (panTro1, Nov. 2003 freeze) were obtained from the UCSC Genome Bioinformatics Site (<http://genome.ucsc.edu/>). The locations of SVA elements were then extracted and inspected manually. Due to the highly variable VNTR region in SVA elements, some of the annotations did not correctly recognize the composition of the elements. In those cases, the annotations were edited manually.

Next, the SVA elements, along with 2000 bp flanking regions on both sides, were extracted from the human genomic sequence using a Perl script. The LTR derived region of each element was then extracted manually and aligned using Clustal_X (Thompson et al. 1997). For the age estimates, elements in each subfamily were aligned and subjected to further manual adjustment by removing insertions and poly(A) tails. For enhancing the quality of our data, any

element that contained a deletion larger than 50 bp or could not be confidently aligned was also excluded from the alignment.

To study the SVA chromosomal distribution, human chromosome sizes and gap sizes were obtained from summary tables from the UCSC website. The information about genes, G+C content and nucleotide sequence of each chromosome was also downloaded from the same resource. For subfamily G+C analysis, the fraction of elements in each G+C bin were divided by the fraction of genome in that bin and the resulting ratio was used as measure of SVA density (Lander et al. 2001; Medstrand et al. 2002). Custom-made Perl scripts were used to calculate SVA density in G+C bins as well as in genes and intergenic regions, extract flanking sequences of various sizes, extract VNTR regions from full-length SVA elements and calculate their sizes. Repeat identification in the SVA flanking regions was annotated using a local installation of the RepeatMasker program.

Statistical Analysis

STATISTICA (version 6.1) was used in the statistical analysis. The chi-squared test was used to analyze chromosomal distribution of SVA elements and to compare their inter- and intra-genic densities. Correlation analysis was performed to examine the relationship between SVA elements and genes/G+C content at chromosomal level, as well as *Alu* and L1 densities in the SVA flanking regions with the G+C content. Repeat densities in the flanking regions were compared with each other using a pair-wise Student's *t*-test.

Oligonucleotide Primer Design and PCR Analysis

Because the PCR amplicon of a typical SVA locus is usually larger than 2 kb, two separate PCRs were performed in an assay designed for L1 elements as previously described (Myers et al. 2002; Sheen et al. 2000). For the filled site PCR, an SVA-specific internal primer

(located in the element) and a 3' flanking unique primer were used to genotype the presence of the filled alleles of SVA insertions. For the empty site PCR, two flanking unique primers were used to genotype the empty alleles. The SVA presence/absence polymorphism can be determined by combining these two results. In this assay, 20 individuals from each of four geographically diverse human populations (European, African American, Asian and South American) were surveyed for the presence and absence of SVA elements. DNA samples from each of these populations were available from previous studies or were purchased from the Coriell Institute for Medical Research (Camden, New Jersey). The primers and annealing temperatures for each locus are shown in supplemental data Table A.2.

Other DNA samples used in this study including human genomic DNA (HeLa cell line ATCC CCL-2) and nonhuman primate species: DNA samples of *Pan troglodytes* (common chimpanzee), *Pan paniscus* (pygmy chimpanzee), *Gorilla gorilla* (western lowland gorilla), *Pongo pygmaeus* (orangutan) and *Macaca mulatta* (rhesus monkey) are available as a primate phylogenetic panel PRP00001 from Coriell. DNA samples of *Hylobates syndactylus* (siamang) were also purchased from Coriell (PR00721). DNA samples of *Chlorocebus aethiops* (green monkey) were isolated from cell lines ATCC CCL70.

Quantitative PCR

SYBR[®] Green PCR core reagents kits were purchased from Applied Biosystems (Foster City, CA). Quantitative PCR was carried out in 50µl total volume reactions containing 1.25 units AmpliTaq Gold[™] DNA polymerase, 5µl 10X SYBR[®] Green PCR buffer, 3mM MgCl₂, 1mM dNTP, 0.5mM forward (SVA_SF 5'-ACAAACACTGCGG AAGGCC-3') and reverse (SVA_SR 5'-AGGTCTCTGGTTTTCTAGGCA-3') primers and 1.0 µl of DNA sample. Four serial dilutions of template DNA (10ng, 1ng, 100pg, 10pg) were used for each primate species.

Amplification reactions were performed in an ABI Prism 7000 Real Time PCR System (Applied Biosystems) following the manufacturer's instructions with conditions set as follows: 95°C, 12 min; 40 cycles of 95°C, 15 sec and 68°C, 1min. A standard curve was constructed by using serial dilutions of known human genomic DNA samples ranging from 10ng to 10pg. The human DNA samples were included as standard with each batch of quantitation reactions. The data from three identical reactions were exported from the ABI Prism 7000 System SDS Software (Applied Biosystems) into a Microsoft Excel spreadsheet and the copy number of the elements in each species was calculated based on the standard curve in each reaction. The mean values and standard deviations were calculated based on the three replications.

References

- Alonso, A., P. Martin, C. Albarran, P. Garcia, O. Garcia, L.F. de Simon, J. Garcia-Hirschfeld, M. Sancho, C. de La Rua, and J. Fernandez-Piqueras. 2004. Real-time PCR designs to estimate nuclear and mitochondrial DNA copy number in forensic and ancient DNA studies. *Forensic Sci Int* **139**: 141-149.
- Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.
- Bandelt, H.J., P. Forster, and A. Rohl. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37-48.
- Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3**: 370-379.
- Bennett, E.A., L.E. Coleman, C. Tsui, W.S. Pittard, and S.E. Devine. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168**: 933-951.
- Brookfield, J.F. 2001. Selection on Alu sequences? *Curr Biol* **11**: R900-901.
- Callinan, P.A. and M.A. Batzer. 2006. Transposable elements and human disease In *Genome Dynamics* (ed. J.-N. Volff), pp. 104-115. S Karger AG, Basel (Switzerland).
- Carter, A.B., A.H. Salem, D.J. Hedges, C.N. Keegan, B. Kimball, J.A. Walker, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2004. Genome-wide analysis of the human Alu Yb-lineage. *Hum Genomics* **1**: 167-178.

- Cordaux, R., D.J. Hedges, and M.A. Batzer. 2004. Retrotransposition of Alu elements: how many sources? *Trends Genet* **20**: 464-467.
- Deininger, P.L., J.V. Moran, M.A. Batzer, and H.H. Kazazian, Jr. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13**: 651-658.
- Ginzinger, D.G., T.E. Godfrey, J. Nigro, D.H. Moore, 2nd, S. Suzuki, M.G. Pallavicini, J.W. Gray, and R.H. Jensen. 2000. Measurement of DNA copy number at microsatellite loci using quantitative PCR analysis. *Cancer Res* **60**: 5405-5409.
- Glazko, G.V. and M. Nei. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol* **20**: 424-434.
- Goodman, M., C.A. Porter, J. Czelusniak, S.L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C.P. Groves. 1998. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* **9**: 585-598.
- Grover, D., M. Mukerji, P. Bhatnagar, K. Kannan, and S.K. Brahmachari. 2004. Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition. *Bioinformatics* **20**: 813-817.
- Gu, Z., H. Wang, A. Nekrutenko, and W.H. Li. 2000. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* **259**: 81-88.
- Hackenberg, M., P. Bernaola-Galvan, P. Carpena, and J.L. Oliver. 2005. The biased distribution of Alus in human isochores might be driven by recombination. *J Mol Evol* **60**: 365-377.
- Hall, T.A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*: 95-98.
- Han, K., J.C. Xing, H. Wang, D.J. Hedges, R.K. Garber, R. Cordaux, and M.A. Batzer. 2005. Under the genomic radar: the stealth model of Alu amplification. *Genome Res* **15**: 655-664.
- Hedges, D.J., P.A. Callinan, R. Cordaux, J.C. Xing, E. Barnes, and M.A. Batzer. 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* **14**: 1068-1075.
- Hirzmann, J., D. Luo, J. Hahnen, and G. Hobom. 1993. Determination of messenger RNA 5'-ends by reverse transcription of the cap structure. *Nucleic Acids Res* **21**: 3597-3598.
- Jurka, J., O. Kohany, A. Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* **101**: 1268-1272.

- Jurka, J., M. Krnjajic, V.V. Kapitonov, J.E. Stenger, and O. Kokhanyy. 2002. Active Alu elements are passed primarily through paternal germlines. *Theor Popul Biol* **61**: 519-530.
- Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.
- Kobayashi, K., Y. Nakahori, M. Miyake, K. Matsumura, E. Kondo-Iida, Y. Nomura, M. Segawa, M. Yoshioka, K. Saito, M. Osawa et al. 1998. An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* **394**: 388-392.
- Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
- Lavie, L., E. Maldener, B. Brouha, E.U. Meese, and J. Mayer. 2004. The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res* **14**: 2253-2260.
- Legoix, P., H.D. Sarkissian, L. Cazes, S. Giraud, F. Sor, G.A. Rouleau, G. Lenoir, G. Thomas, and J. Zucman-Rossi. 2000. Molecular characterization of germline NF2 gene rearrangements. *Genomics* **65**: 62-66.
- Medstrand, P., L.N. van de Lagemaat, and D.L. Mager. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12**: 1483-1495.
- Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530-1534.
- Myers, J.S., B.J. Vincent, H. Udall, W.S. Watkins, T.A. Morrish, G.E. Kilroy, G.D. Swergold, J. Henke, L. Henke, J.V. Moran et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71**: 312-326.
- Ono, M., M. Kawakami, and T. Takezawa. 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res* **15**: 8725-8737.
- Ostertag, E.M., J.L. Goodier, Y. Zhang, and H.H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73**: 1444-1451.
- Ostertag, E.M. and H.H. Kazazian, Jr. 2001. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35**: 501-538.
- Otieno, A.C., A.B. Carter, D.J. Hedges, J.A. Walker, D.A. Ray, R.K. Garber, B.A. Anders, N. Stoilova, M.E. Laborde, J.D. Fowlkes et al. 2004. Analysis of the Human Alu Ya-lineage. *J Mol Biol* **342**: 109-118.

- Ovchinnikov, I., A.B. Troxel, and G.D. Swergold. 2001. Genomic Characterization of Recent Human LINE-1 Insertions: Evidence Supporting Random Insertion. *Genome Res* **11**: 2050-2058.
- Pavlicek, A., K. Jabbari, J. Paces, V. Paces, J.V. Hejnar, and G. Bernardi. 2001. Similar integration but different stability of Alus and LINES in the human genome. *Gene* **276**: 39-45.
- Pavlicek, A., J. Paces, D. Elleder, and J. Hejnar. 2002. Processed pseudogenes of human endogenous retroviruses generated by LINES: their integration, stability, and distribution. *Genome Res* **12**: 391-399.
- Rohrer, J., Y. Minegishi, D. Richter, J. Eguiguren, and M.E. Conley. 1999. Unusual mutations in Btk: an insertion, a duplication, an inversion, and four large deletions. *Clin Immunol* **90**: 28-37.
- Rynditch, A.V., S. Zoubak, L. Tsyba, N. Tryapitsina-Guley, and G. Bernardi. 1998. The regional integration of retroviral sequences into the mosaic genomes of mammals. *Gene* **222**: 1-16.
- Salem, A.H., J.S. Myers, A.C. Otieno, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2003. LINE-1 pre-Ta elements in the human genome. *J Mol Biol* **326**: 1127-1146.
- Sheen, F.M., S.T. Sherry, G.M. Risch, M. Robichaux, I. Nasidze, M. Stoneking, M.A. Batzer, and G.D. Swergold. 2000. Reading between the LINES: human genomic variation induced by LINE-1 retrotransposition. *Genome Res* **10**: 1496-1508.
- Shen, L., L.C. Wu, S. Sanlioglu, R. Chen, A.R. Mendoza, A.W. Dangel, M.C. Carroll, W.B. Zipf, and C.Y. Yu. 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J Biol Chem* **269**: 8466-8476.
- Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876-4882.
- Volloch, V.Z., B. Schweitzer, and S. Rits. 1995. Transcription of the 5'-terminal cap nucleotide by RNA-dependent DNA polymerase: possible involvement in retroviral reverse transcription. *DNA Cell Biol* **14**: 991-996.
- Walker, J.A., G.E. Kilroy, J. Xing, J. Shewale, S.K. Sinha, and M.A. Batzer. 2003. Human DNA quantitation using Alu element-based polymerase chain reaction. *Anal Biochem* **315**: 122-128.

- Wilund, K.R., M. Yi, F. Campagna, M. Arca, G. Zuliani, R. Fellin, Y.K. Ho, J.V. Garcia, H.H. Hobbs, and J.C. Cohen. 2002. Molecular mechanisms of autosomal recessive hypercholesterolemia. *Hum Mol Genet* **11**: 3019-3030.
- Xing, J.C., D.J. Hedges, K. Han, H. Wang, R. Cordaux, and M.A. Batzer. 2004. Alu element mutation spectra: molecular clocks and the effect of DNA methylation. *J Mol Biol* **344**: 675-682.

CHAPTER THREE:

**EMERGENCE OF PRIMATE GENES BY
RETROTRANSPOSON-MEDIATED SEQUENCE
TRANSDUCTION***

*Reprinted by permission of Proceedings of the National Academy of Sciences, U.S.A.

Introduction

The emergence of new genes and biological functions is crucial to the evolution of species (Long et al. 2003). Several mechanisms for creating new genes are known, the best characterized pathway being through duplication of preexisting genes (Long et al. 2003; Ohno 1970). Several types of duplications leading to genetic innovation have been investigated, including segmental duplication (Johnson et al. 2001) and gene retrotransposition (Marques et al. 2005; Nisole et al. 2004; Sayah et al. 2004; Vinckenbosch et al. 2006). Here, we investigate a less well characterized mechanism that can potentially duplicate genes, namely the transduction of flanking genomic sequence associated with the retrotransposition of mobile elements.

Retrotransposons usually do not carry downstream motifs that are important for efficient transcription termination. Therefore, when they are transcribed, the RNA transcription machinery sometimes skips the element's own weak polyadenylation signal and terminates transcription using a downstream polyadenylation site located in the 3' flanking genomic sequence. The transcript containing the retrotransposon along with the extra genomic sequence is subsequently integrated back into the genome through retrotransposition, a process termed 3' transduction (Holmes et al. 1994; Moran et al. 1996). In principle, this could lead to the duplication of coding sequences located in the transduced flanking genomic sequence. Indeed, the exon shuffling and genetic diversity of the L1 mediated 3' transduction has been demonstrated in cell culture assays (Moran et al. 1999; Moran et al. 1996). However, among all the studies investigating L1-mediated exon shuffling (Ejima and Yang 2003; Goodier et al. 2000; Lander et al. 2001; Pickeral et al. 2000; Rozmahel et al. 1997; Szak et al. 2003), only two putative examples of exon transduction have been reported in the human genome (Ejima and Yang 2003; Rozmahel et al. 1997).

The SVA family of retrotransposons originated less than 25 million years ago (mya) and has increased to about 3,000 copies in the human genome (Ostertag et al. 2003; Wang et al. 2005). Similar to L1 elements, SVA elements are thought to be transcribed by RNA polymerase II and have the ability to transduce downstream sequence (Ostertag et al. 2003). About 10% of human SVA elements appear to have been involved in sequence transduction events (Wang et al. 2005). Here, we examined the extent and properties of SVA-mediated transduction events to evaluate their evolutionary impact on the human genome. Our results demonstrate that retrotransposon-mediated sequence transduction is not only a mechanism for exon shuffling, but also serves as a novel mechanism for gene duplication and the creation of new gene families.

Results and Discussion

Genomic Analysis for SVA 3' Transductions

To investigate the extent of transduction events associated with SVA elements, we examined all 1752 full-length SVA elements in the human genome reference sequence. In total, 143 SVA elements with putative transduced sequences were identified according to our validation criteria (Figure 3.1, see materials and methods for details). Most of these loci (123/143) also displayed typical AATAAA or ATTAAA polyadenylation signals located 5-52 nucleotides upstream of the start of the poly(A) tail, further supporting that these loci represent authentic SVA-mediated transduction events. The size of the transduced sequences ranged from 35 to 1853 bp, with an average of 340 bp (Figure 3.2). Overall, 52,740 bp of genomic sequence was duplicated by these SVA mediated transductions. To determine if the transduction events are specific characteristics of particular SVA subfamilies, SVA elements with transduced sequence were aligned and compared to all known SVA subfamily consensus sequences (Wang et al. 2005). We found transduction events involving all previously identified SVA families,

suggesting that 3' transduction is a common phenomenon among SVA members (Table 3.1). Given that full-length SVA elements comprise 63% of the family, we extrapolate that SVA elements may have transduced a total of ~84 kb of genomic sequence during their expansion.

Table 3.1: Affiliation of 3' transduced sequences and SVA subfamilies.

Subfamily	A	B	C	D	E	F	Total
Number of transduction events	4	29	20	72	11	7	143
Number of full length elements	104	323	151	931	95	148	1752
Rate of transduction	3.8%	9.0%	13.2%	7.7%	11.6%	4.7%	8.2%

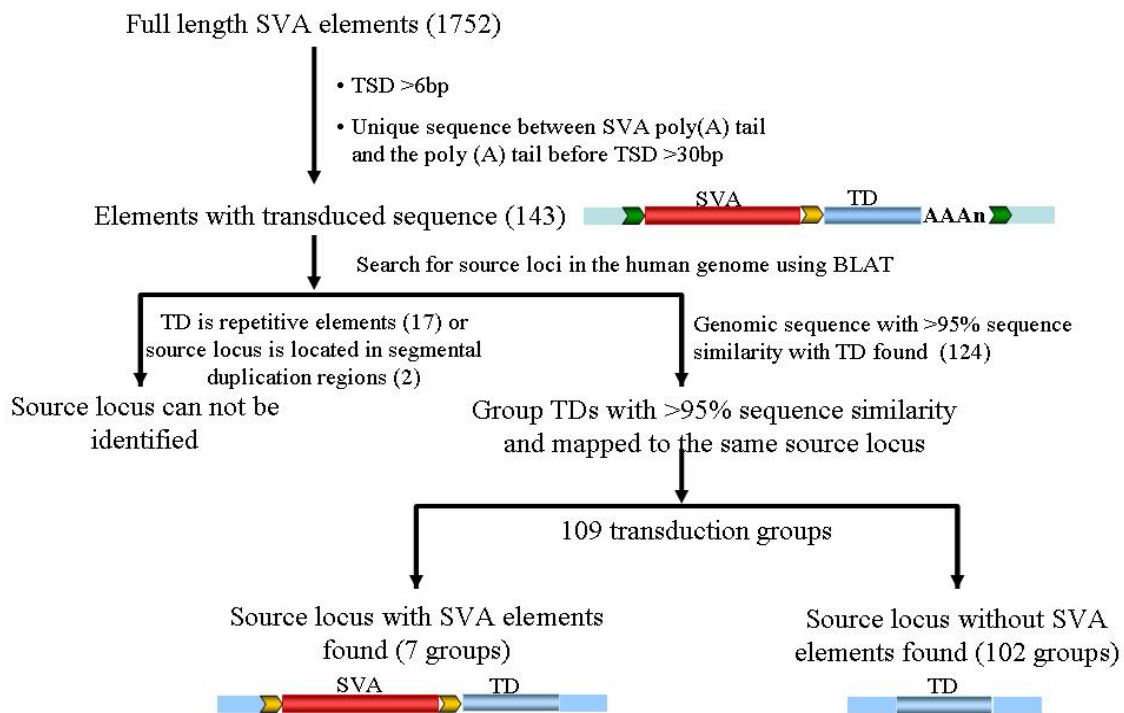


Figure 3.1: Identification of SVA 3' transduction events and their source elements. schematic diagrams for the identification process. Flanking sequences of the source locus are shown as blue boxes; TSDs are shown as yellow and green arrows. SVA elements are depicted as red bars, and the transduced sequences are shown as blue bars and labeled as “TD”. SVA element poly (A) tails are shown as “(AAA)n”. The numbers in parentheses correspond to total number of SVA elements/groups identified in each step.

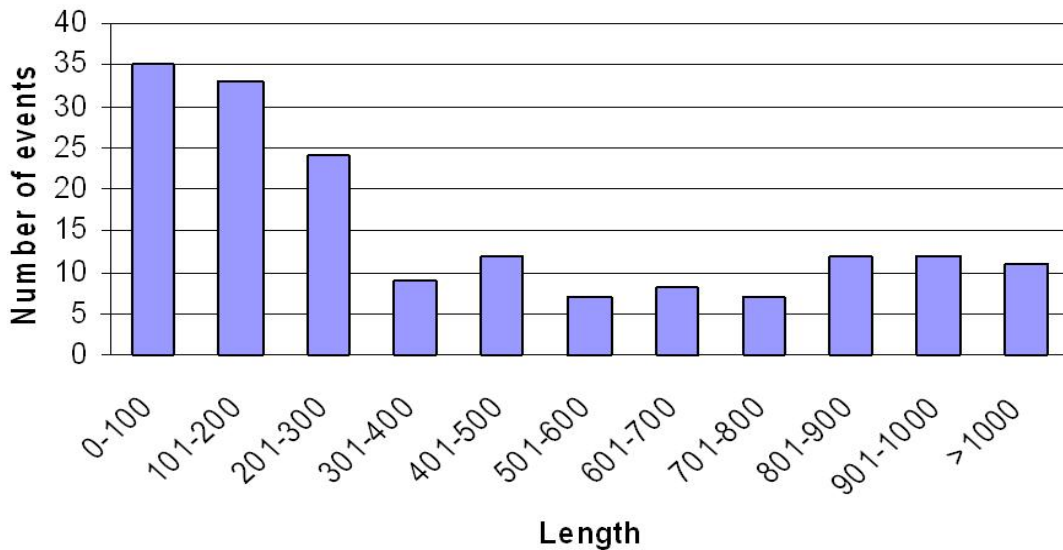


Figure 3.2: Length distribution of 3' transduction events. Number of human SVA mediated 3' transduction events in each 100 bp size interval.

The rate of SVA-mediated transduction events (8.2%) is similar to the L1 transduction rates reported in previous studies (Goodier et al. 2000; Pickeral et al. 2000; Szak et al. 2003). Our results are likely to be an underestimate because the method we used to validate candidate transductions relied on the detection of perfect target site duplications (TSDs). This requirement may miss transduction events as a result of substitutions in their TSDs. Detailed information concerning all the SVA transduction events reported in this paper can be found in supplemental data Table A.3.

Identification of Transduction Source Loci

To identify the source loci of the transduced sequences, we searched the human genome using the BLAST-like alignment tool (BLAT) (<http://genome.ucsc.edu/cgi-bin/hgBLAT>). With the exception of 19 loci in which the transduced sequence was totally composed of repetitive

sequences or source loci were located in segmental duplications (Figure 3.1) and thus could not be mapped precisely to any location in the human genome, we were able to identify the source loci for each of the other 124 transductions. While 98 transduced loci could be uniquely linked to their source locus and each of them was treated as a unique group, the remaining 26 transduced loci showed >95% sequence identity with at least one other transduced locus and were mapped to the same source locus. Hence, the 26 transduced loci were assigned to 11 transduction groups each of which contained two or more transduced loci.

Seven out of the total 109 source loci contained an SVA element (Figure 3.1). These source SVA elements could be unambiguously identified because only the SVA elements were surrounded by TSDs. By comparison, in the transduced loci, the TSDs included both the SVA element and the transduced flanking genomic sequence. The source SVA element for the transduction group H3_186 located on human chromosome 2p11.2 represented one of the most active elements we identified, given that it generated at least nine transduced copies (Figure 3.3). Sequence comparisons showed that the source locus and all nine transduced loci were absent from the chimpanzee genome. This result suggests that the source locus inserted in the human genome after the human-chimpanzee divergence and generated all the loci with transductions within the last ~6 million years. Among the nine transduced loci, five were typical SVA transduction loci (*i.e.* having both the SVA element and the transduced sequence) while four loci contained only the transduced genomic sequences. Although there is no SVA element upstream of these sequences, a poly (A) tail can be found downstream of the transduced sequence and the TSDs surrounding the transduced sequence are identifiable. Presumably, these loci were generated via SVA-mediated transduction events associated with 5' truncation during the integration process or through incomplete reverse transcription during retrotransposition. SVA represents one of the

most active retrotransposon families in humans (CSAC 2005) and little is known about their retrotransposition mechanism. Because the source loci with SVA elements may be capable of retrotransposition, detailed analysis of the seven transduction source loci we identified in this study may shed new light on the underlying mechanism of SVA retrotransposition.

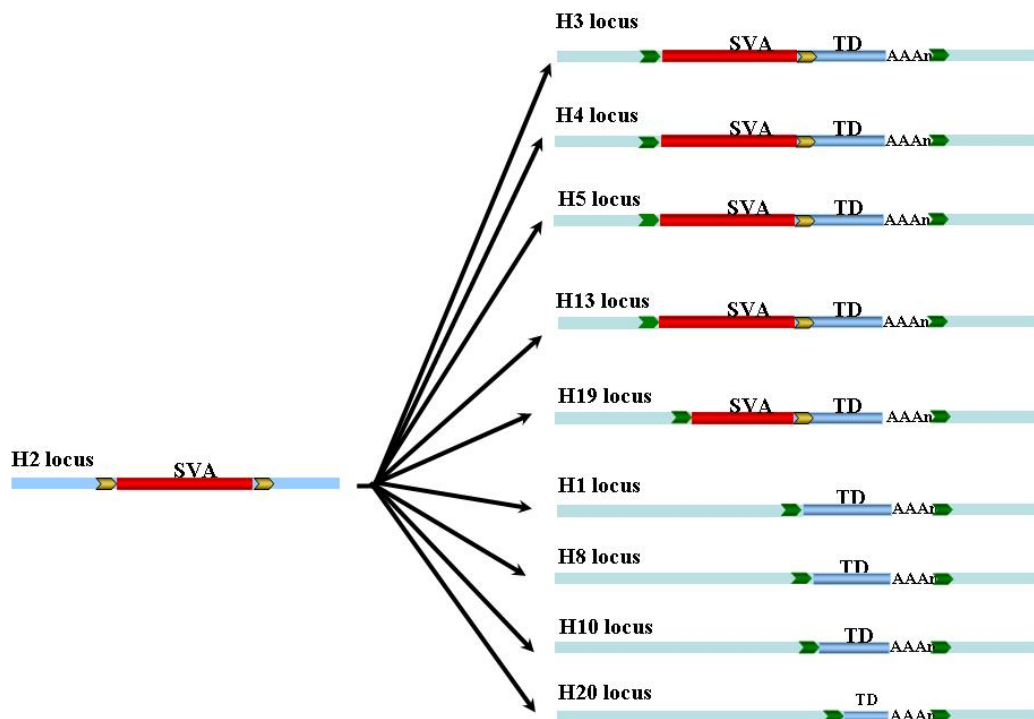


Figure 3.3 : SVA 3' transduction events. One group of SVA 3' transduction events (H3_186). Flanking sequences of the original locus are shown as blue boxes while the flanking sequences of the transduced loci are shown as light blue boxes. TSDs are shown as yellow and green arrows. SVA elements are depicted as red bars and the transduced sequences are shown as blue bars and labeled as “TD”. SVA element poly (A) tails are shown as “(AAA)n”.

For the other 102 source loci, only sequences corresponding to transduced sequences were present. Each of the loci was devoid of the SVA elements, TSDs and poly (A) tails. The simplest explanation for this genomic configuration is that the retrotransposition events carrying transduced sequence occurred while the source SVA element was polymorphic for insertion

presence/absence in the population. Subsequently, the original locus with the source SVA element may have been lost in the population, or may still be polymorphic in the human population but absent from the human genome reference sequence. To test source loci for insertion polymorphism, we genotyped 30 source loci in 80 diverse human genomes using polymerase chain reaction (PCR) assays as described previously (Wang et al. 2005). No source loci containing an SVA element were present in any of the diverse human genomes that were surveyed for the 30 SVA source elements. These results suggest that most of the SVA elements that generated transduction events were subsequently lost from the human genome. This observation is not surprising since similar results have been observed in L1-mediated transduction studies (Boissinot et al. 2001; Szak et al. 2003). In fact, the majority of new mobile element insertions are expected to be lost from the population due to genetic drift under neutral evolution. A second potential reason for the disappearance of these source elements may be moderate negative selection as a result of their transcription and retrotransposition capacity (Boissinot et al. 2001).

SVA Transduction Mediated Gene Duplication

One particularly interesting example of SVA-mediated 3' transduction is the H17_76 group (Figure 3.4A), in which four related loci were identified in the human genome. The source locus (chr17_A) contained only the transduced sequence without the SVA element and the other three loci (chr17_B, chr18 and chr8) contained an SVA element along with the transduced sequence resulting from retrotransposition. This group of transduced sequences is among the longest of all the elements we recovered (1682 bp, 1245 bp and 1257 bp for loci chr17_B, chr18 and chr8 respectively). Subsequent analysis of the transduced sequences resulted in the identification of the gene *AMAC1* (acyl-malonyl condensing enzyme 1, Entrez Gene ID: 146861)

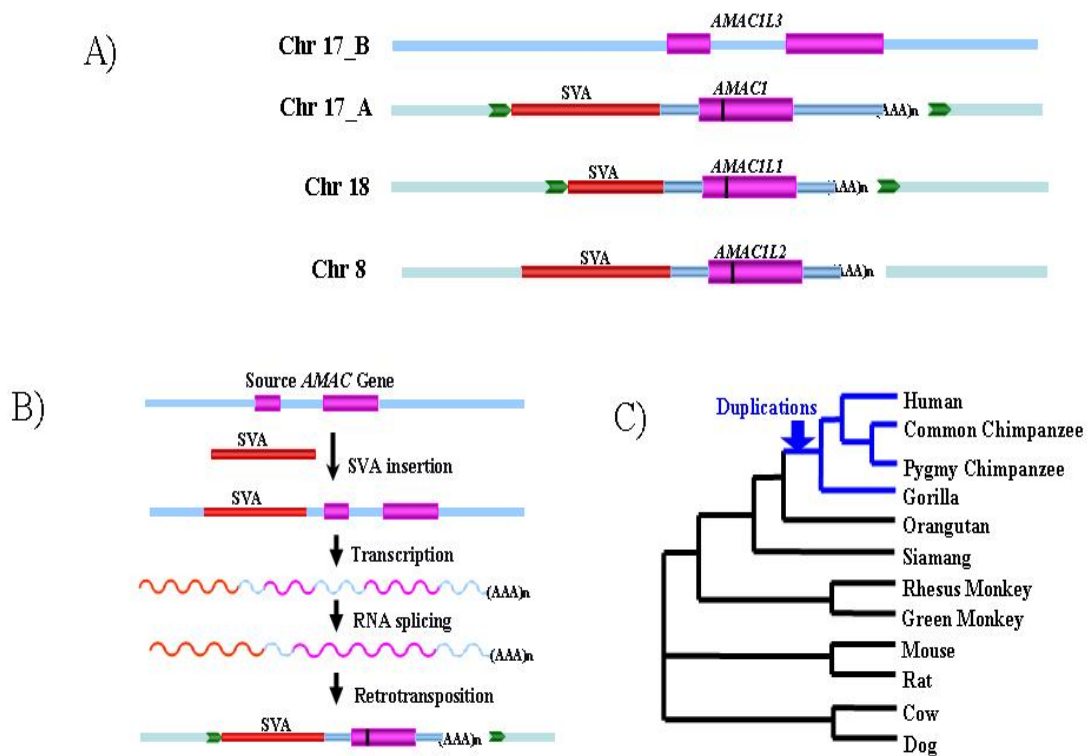


Figure 3.4: SVA transduction mediated gene duplication. (A) Schematic diagram of the H17_76 transduction group in the human genome. Flanking sequences of the original locus are shown as blue boxes while the flanking sequences of the transduced loci are shown as light blue boxes. TSDs are shown as yellow and green arrows. SVA elements are depicted as red bars and the transduced sequences are shown as blue bars and coding regions are shown as purple bars. SVA element poly (A) tails are shown as "(AAA)n". **(B) Schematic diagrams for putative evolutionary scenarios of the SVA transduction mediated gene duplications.** About 7 to 14 mya ago, one active SVA element inserted upstream of the original AMAC gene locus. Then, transcription of this active SVA element transduced the full length AMAC gene sequence. During the retrotransposition process, the intron of the gene was removed by RNA processing machinery. Finally, the SVA element along with the intronless AMAC gene sequence retrotransposed into new genomic locations. The original retrotransposition competent SVA element upstream of the source locus was eventually lost in the population. The predicted RNA transcripts are shown as curved lines. **(C) The phylogenetic relationships among various species used in d_N/d_S analysis.**

at the chr17_B locus, *AMACIL1* (AMAC1-like 1, Entrez Gene ID: 492318) at the chr18 locus, *AMACIL2* (AMAC1-like 2, Entrez Gene ID: 83650) at the chr8 locus and *AMACIL3* (AMAC1-like 3, Entrez Gene ID: 404029) at the chr17_A locus. Interestingly, the source locus (chr17_A) contains 467 bp of extra sequence in the middle of the transduced sequence as compared to the other copies. Examination of the gene structure showed that *AMACIL3* at the inferred source locus had two exons separated by an intron (the extra DNA sequence) while the three SVA transduced loci contained intronless versions of *AMACIL3* (Figure 3.4A). These results suggest that the intron was spliced out during the retrotransposition process, providing further evidence for the underlying mechanism that created the three duplicated copies (Figure 3.4B). Although the exact function of human *AMAC* genes has not been determined, studies in bacteria showed that *AMAC* is an enzyme involved in fatty acid synthesis, in which it condenses a two-carbon unit from malonyl-(acyl carrier protein) to fatty acyl-(acyl carrier protein), adding two carbons to the fatty acid chain with the release of carbon dioxide (Toomey and Wakil 1966).

Evolutionary Analyses of *AMAC* Genes

To determine the evolutionary history of *AMAC* transduction events, we first used BLAT to search for orthologous loci corresponding to all four human *AMAC* genes in four available mammalian genome sequences (mouse, rat, cow, dog). The search resulted in the identification of a single *AMAC* locus in all four species examined, and it was orthologous to the human *AMACIL3*. In particular, the mouse orthologous region is annotated as the gene *AMAC1* (Entrez GeneID: 56293). The murine *AMAC1* gene contains two exons similar to human *AMACIL3* gene. Only the human *AMACIL3* locus was co-linear with the mouse *AMAC1* genomic sequence, while the three SVA transduced loci were co-linear with the mouse *AMAC1* mRNA sequence.

Next, we investigated the origin of the multiple *AMAC* gene copies within the primate lineage by analyzing seven non-human primate species (*i.e.* pygmy chimpanzee, common chimpanzee, gorilla, orangutan, siamang, African green monkey and rhesus monkey). PCR and DNA sequence analyses showed that the intron-containing locus *AMACIL3* is present in all species examined. By contrast, the *AMAC1* and *AMACIL2* loci were present only in human, pygmy chimpanzee, common chimpanzee and gorilla genomes. Due to the presence of repetitive elements in the flanking regions, the *AMACIL1* locus could not be successfully amplified using PCR and was excluded from subsequent analyses. Taken together, our results suggest that the transduction events happened after the divergence of African apes from orangutans but before the divergence of humans, chimpanzees and gorillas, approximately 7 to 14 mya based on the estimated divergence time of primates (Goodman et al. 1998).

To determine the functional status of the *AMAC* genes, we first examined all available *AMAC* copies for intact ORFs. Examination of the DNA sequences showed that the *AMAC1* loci in both common chimpanzee and gorilla contained a premature stop codon in their ORF regions, located at amino acid position 29 and 32, respectively. Therefore, these two copies have lost their coding capacity and have become processed pseudogenes. The ORFs of all the other *AMAC* copies, including all four human copies, remained intact.

Next, we examined the selective constraints on all *AMAC* copies using the maximum likelihood-based program PAML (Yang 1997). PAML estimates the non-synonymous (d_N) and synonymous (d_S) substitution rate ratios ($\omega=d_N/d_S$) as measures of selective pressure, where the value of ω indicates the type of selection (<1 , purifying; $=1$, neutral; >1 , positive). Likelihood ratio tests can then be used to compare the different models of evolution. First, a maximum-likelihood tree was constructed using all available *AMAC* ORF sequences (see methods), and ω

values were estimated according to a model of a single ω among branches of the tree (Model 0) and a model where ω is allowed to vary among branches (Model 1). The results showed that Model 1 fit the data significantly better than Model 0 ($p < 0.0001$, Table 3.1), suggesting that different selective pressures are acting on the various *AMAC* copies within and between species. We then separated the branches of the tree into two groups, corresponding to the sequences predating the gene duplication events (Figure 3.4C, black branches) and sequences postdating duplication events (Figure 3.4C, bold blue branches) and estimated ω values for the two groups separately. The ω in the branches predating the duplications was estimated to be 0.13 (significantly different from $\omega=1$, $p < 0.0001$), suggesting that the *AMAC* gene is under strong purifying selection in the species lacking duplicated copies. By contrast, the ω was estimated to be 1.24 (not significantly different from $\omega=1$, $p=0.25$) in the species that possess multiple copies of *AMAC*. In addition, the comparison between Model 0 and Model 1 for branches after the duplications showed that ω values among branches are not significantly different ($p=0.69$). Together, these results indicate that the *AMAC* gene was under purifying selection in all the species prior to the duplication events and that, after the gene duplication events, all *AMAC* copies experienced a relaxation of selective constraints.

These observations are in good agreement with the predictions of classical gene duplication theory (Ohno 1970; Prince and Pickett 2002), which suggests that the functional redundancy of newly duplicated genes will initially result in free evolution of all gene copies. The long-term evolutionary fate of the new gene copies includes loss of function (non-functionalization), evolution of a new function (neo-functionalization) or maintenance of the duplicated copies for the original function (sub-functionalization). The stop codon present in different positions in the chimpanzee and gorilla *AMAC1* copies shows the non-functionalization

Table 3.2: PAML analysis.

	ω	lnL	np	p	Significance
All AMAC copies					
Model 0 (Single ω)	0.273	-3915	36		
Model 1 (Free ω)	Varies	-3812.7	69		
LRT Model 0 vs. Model 1				0.000	**
AMACs Before Duplication					
Model 0 (Single ω)	0.13	-2866.7	15		
Model 0 ($\omega=1$)	1	-3013	14		
Model 1 (Free ω)	Varies	-2831.4	27		
LRT Model 0 (Single ω) vs. Model 0 ($\omega=1$)				0.000	**
LRT Model 0 (Single ω) vs. Model 1 (free ω)				0.000	**
Duplicated AMAC copies					
Model 0 (Single ω)	1.245	-2165.1	21		
Model 0 ($\omega=1$)	1	-2165.7	20		
Model 1 (Free ω)	Varies	-2157.8	39		
LRT Model 0 (Single ω) vs. Model 0 ($\omega=1$)				0.251	No
LRT Model 0 (Single ω) vs. Model 1 (free ω)				0.694	No
Site Specific Models for duplicated copies					
Model 0 (NS site Model1_Neutral)	p: 0.107 0.893 w: 1.000 1.000	-2165.7	22		
Model 0 (NS site Model2_PosSel)	p: 0.824 0.000 0.176 w: 0.533 1.000 5.442	-2154.9	24		
LRT NS site M1_Neutral vs. M2_PosSel				0.000	**

of these particular gene copies. Furthermore, we also analyzed the *AMAC* gene sequences for possible sites or domains that are (or were) under positive selection and developed new functions (neo-functionalization) (Tables 3.2 and 3.3). However, the paucity of functional studies on the human *AMAC* genes prevented detailed validation and comparisons of the potential functional role of candidate sites.

Table 3.3: Putative positive selection sites on *AMAC* genes.

Site	AA	NEB	BEB
28	C	0.99*	0.95*
88	R	0.99**	0.96*
269	G	0.98*	0.94
309	T	0.98*	0.9
324	R	0.99*	0.95

Putative positive selection sites were identified using PAML under Naive Empirical Bayes (NEB) analysis or Bayes Empirical Bayes (BEB) analysis.
*: P>95%; **: P>99%.

AMAC Expression Studies

To further investigate the functional status of human *AMAC* gene duplicates, we examined the expression pattern of the four human *AMAC* gene copies. RT-PCR analysis was performed using poly(A)-selected RNA from human testis and placenta and oligonucleotide primers designed to match conserved regions in all four gene copies (supplemental data Figure A.2). We first amplified the target sequences from human genomic DNA (HeLa), cloned the PCR products and sequenced ~100 clones. Based on nucleotide substitutions specific for each *AMAC* gene duplicate, we determined the origin of each clone. Our results showed that all four gene copies were recovered in a comparable manner, showing that our approach amplifies the four human *AMAC* copies with similar efficiency (Figure 3.5B). Using the same primer set, RT-PCR generated a product with the expected size in both tissues (Figure 3.5A). We cloned the RT-PCR products and sequenced ~100 clones derived from each tissue (Figure 3.5B). Sequence analysis showed that *AMAC1*, *AMAC1L2* and *AMAC1L3* were expressed in testis while only *AMAC1L2* and *AMAC1L3* were expressed in placenta. In both tissues, the SVA-transduced *AMAC1L2*

transduced *AMAC* gene duplicates are currently expressed in humans and that they may have differential tissue expression patterns.

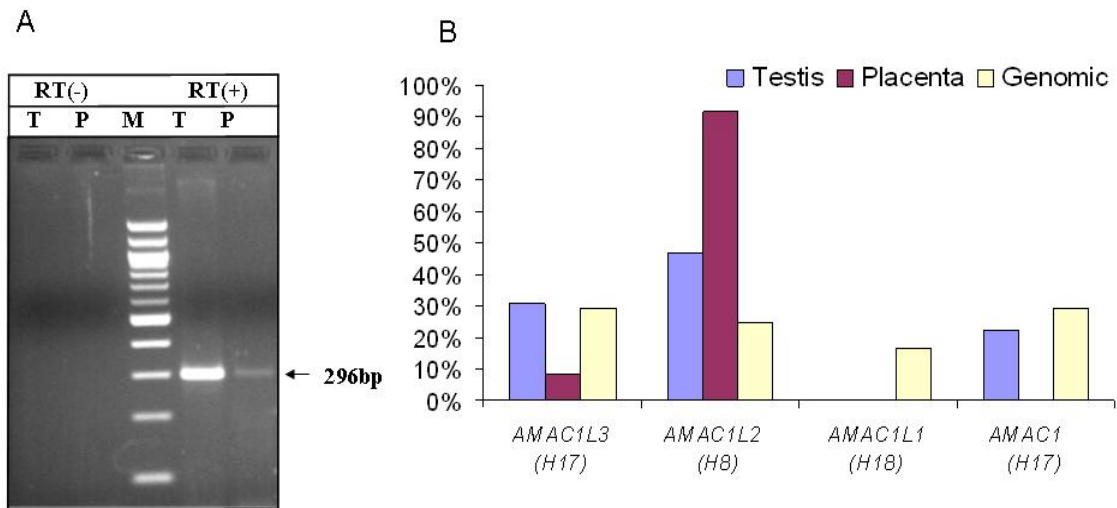


Figure 3.5. Expression analysis of *AMAC* gene duplicates in humans. (A) Agarose gel chromatograph of RT-PCR products derived from human testis (T) and placental (P) RNA templates. Negative controls with no reverse transcriptase (RT -) are on the left, a 100 bp marker (M) is in the middle and the sizes of the correct fragments are indicated. (B) Relative expression levels of four human *AMAC* gene duplicates in human testis and placenta. Human genomic DNA (HeLa) amplification is the control for uniform amplification of all gene duplicates.

To further investigate the expression pattern of *AMAC* gene copies in human tissues, we searched the NCBI EST database for *AMAC* related transcripts. Consistent with the RT-PCR results, two of the SVA-transduced *AMAC* copies (*AMAC1* and *AMAC1L2*) were recovered from the EST database and their full length cDNAs have already been sequenced (AK097473 and

AJ291677 for *AMAC1* and *AMACIL2* respectively). Both mRNA sequences start downstream from the SVA elements, suggesting that their promoters have been duplicated along with the gene copies themselves. By contrast, the two transcripts contained unique 3' UTR sequences specific to their new genomic locations. Thus, it appears that these two gene copies have acquired new downstream polyadenylation signals subsequent to their integration in the genome.

Concluding Remarks

Our results represent the first example in the primate lineage of gene duplications derived from SVA-mediated 3' transduction. One factor that may increase the potency of SVA elements with regard to the generation of new genes is that, in addition to duplicating genes, SVA elements may also provide promoters for newly integrated gene duplicates. SVA elements contain a long terminal repeat (LTR) derived region that is used as promoter in endogenous retroviruses, and several studies have shown that LTRs in the genome can function as promoters for downstream genes (Dunn et al. 2003; Dunn et al. 2005). Therefore, the SVA LTR-derived tail region might function as an alternative promoter for genes involved in 3' transduction events whenever the original gene-specific promoter was not transduced.

Retrotransposon-mediated 3' transduction represents a novel mechanism for entire gene duplication that can lead to the rapid generation of new gene families. Although the gene duplications by retrotransposon-mediated transduction reported here created intronless gene copies similar to duplications resulting from gene retrotransposition, there is one major difference between these two duplication mechanisms. Gene retrotransposition generally does not carry the promoter and regulatory region of the retrotransposed gene to its new location due to the process by which the gene is reverse transcribed into cDNA. Thus, the newly transposed gene must acquire new regulatory sequences to be functional. By contrast, retrotransposon-

mediated 3'-transduction events not only duplicate whole genes, but can also duplicate promoter regions (as demonstrated by the *AMAC* gene duplicates). Consequently, duplications resulting from 3'-transduction retain their functional potential after inserting into their new genomic locations, allowing immediate release of the gene copies (the original as well as the duplicates) from selective constraints.

Mobile elements are already known for creating genomic novelty in a variety of ways. For example, L1 elements provide the molecular machinery necessary for gene retrotransposition (Esnault et al. 2000; Wei et al. 2001) and new fusion genes can be generated during the process (Nisole et al. 2004; Sayah et al. 2004). DNA transposons can also mediate exon shuffling and gene duplication, as demonstrated by mutator-like elements and helitron-like elements in plants (Jiang et al. 2004; Morgante et al. 2005). Further, mobile elements themselves can serve as raw material for the generation of new functions by their incorporation into existing genes (Cordaux et al. 2006; Krull et al. 2005; Nekrutenko and Li 2001). By serving as a mechanism for gene duplication and generating new gene families, retrotransposon-mediated sequence transduction represents an important mechanism by which mobile elements impact their host genomes.

Materials and Methods

Genome Analysis

The RepeatMasker annotations of the human genome reference sequence (hg17, May 2004) were obtained from the UCSC Genome Bioinformatics Site (<http://genome.ucsc.edu>). The full-length SVA elements with 2000 bp flanking sequence on each side were exacted using a perl script and the TSDs were identified manually. SVA elements were considered as candidates for containing 3' transduction sequence if they had (1) unambiguous TSDs (>6bp) and (2) more than 30 bp sequence between the end of the SVA sequence and the poly(A) tail. The extra sequence

was then used for a sequence similarity BLAT search against the human genomic database. For genomic loci exhibiting >95% identity to the putative transduced sequence, 5 kb extra sequence was extracted on both ends of the locus and a locally installed RepeatMasker program was used to determine their repeat content. In some cases, segmental duplications may result in a similar pattern as the 3' transduction. To remove such false positive hits, we further compared the sequence directly flanking the newly identified locus with the flanking sequence of the query locus. If the flanking sequences from both loci showed >90% sequence similarity for more than 2 kb in length, the newly identified locus was excluded from the further analysis.

PCR/RT-PCR and DNA Sequence Analysis

The human population panel used in the polymorphism analysis were described previously (Wang et al. 2005). Other DNA samples used in this study included human genomic DNA (HeLa cell line ATCC CCL-2) and the following nonhuman primate species available as a primate phylogenetic panel PRP00001 from Coriell: DNA samples of *Pan troglodytes* (common chimpanzee), *P. paniscus* (bonobo or pygmy chimpanzee), *Gorilla gorilla* (western lowland gorilla), *Pongo pygmaeus* (orangutan) and *Macaca mulatta* (rhesus monkey). DNA samples of *Hylobates syndactylus* (siamang) was also purchased from Coriell (PR00721). DNA samples of *Chlorocebus aethiops* (green monkey) were isolated from a cell line ATCC CCL70.

The primers and annealing temperatures for each locus are shown in Table A.4. For the *AMAC* gene analysis, *AMAC* gene related loci were amplified using PCR with different primates as templates using different primer sets and annealing temperatures (Table A.5).

For RT-PCR, 1 µg of poly (A) selected RNA (Ambion) from human testis and placenta was used to perform the reverse transcription (RT) reaction using the Reverse Transcription System kit (Promega) with conserved primer 4(-) (5'-CAGATAGGAAGGCCACTGTTG-3')

according to manufacturer's protocol. After completion the volume of the RT reaction was brought to a final volume of 100 μ l with nuclease free water. 10 μ l of the RT reaction was used for PCR with conserved primer 1(+) (5'-ATTGCCCTGCTACTTAAACTGC-3')/conserved primer 1(-) (5'-TGTAGTGTCCAGAGTCCAGGTC-3') for 32 cycles at annealing temperature of 60°C. PCR products were fractionated on a 1.5% low melting agarose gel and extracted with QIAquick Gel Extraction kit (QIAGEN).

For sequencing analysis, individual RT-PCR/PCR products were cloned using the TOPO-TA cloning kit (Invitrogen) and sequenced using chain termination sequencing on an ABI 3100 Genetic Analyzer. All sequences were deposited in GenBank under accession numbers DQ482900-DQ482914.

Evolutionary Analysis

All available *AMAC* gene coding region homologous sequences were aligned using ClustalX (Thompson et al. 1997), followed by manual adjustments. Next, a maximum-likelihood tree were constructed under HKY85+G model using PAUP* (Swofford 2003) The resulting tree (supplemental data Figure A.3), was used for the d_N/d_S ratio analysis.

References

- Boissinot, S., A. Entezam, and A.V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* **18**: 926-935.
- Cordaux, R., S. Udit, M.A. Batzer, and C. Feschotte. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A* **103**: 8101-8106.
- Chimpanzee Sequencing and Analysis C. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69-87.
- Dunn, C.A., P. Medstrand, and D.L. Mager. 2003. An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc Natl Acad Sci U S A* **100**: 12841-12846.

- Dunn, C.A., L.N. van de Lagemaat, G.J. Baillie, and D.L. Mager. 2005. Endogenous retrovirus long terminal repeats as ready-to-use mobile promoters: the case of primate beta3GAL-T5. *Gene* **364**: 2-12.
- Ejima, Y. and L. Yang. 2003. Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. *Hum Mol Genet* **12**: 1321-1328.
- Esnault, C., J. Maestre, and T. Heidmann. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **24**: 363-367.
- Goodier, J.L., E.M. Ostertag, and H.H. Kazazian, Jr. 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* **9**: 653-657.
- Goodman, M., C.A. Porter, J. Czelusniak, S.L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C.P. Groves. 1998. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* **9**: 585-598.
- Holmes, S.E., B.A. Dombroski, C.M. Krebs, C.D. Boehm, and H.H. Kazazian, Jr. 1994. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat Genet* **7**: 143-148.
- Jiang, N., Z. Bao, X. Zhang, S.R. Eddy, and S.R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**: 569-573.
- Johnson, M.E., L. Viggiano, J.A. Bailey, M. Abdul-Rauf, G. Goodwin, M. Rocchi, and E.E. Eichler. 2001. Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**: 514-519.
- Krull, M., J. Brosius, and J. Schmitz. 2005. Alu-SINE exonization: en route to protein-coding function. *Mol Biol Evol* **22**: 1702-1711.
- Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.
- Long, M., E. Betran, K. Thornton, and W. Wang. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865-875.
- Marques, A.C., I. Dupanloup, N. Vinckenbosch, A. Reymond, and H. Kaessmann. 2005. Emergence of Young Human Genes after a Burst of Retroposition in Primates. *PLoS Biol* **3**: e357.

- Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530-1534.
- Moran, J.V., S.E. Holmes, T.P. Naas, R.J. DeBerardinis, J.D. Boeke, and H.H. Kazazian, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**: 917-927.
- Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo, and A. Rafalski. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* **37**: 997-1002.
- Nekrutenko, A. and W.H. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17**: 619-621.
- Nisole, S., C. Lynch, J.P. Stoye, and M.W. Yap. 2004. A Trim5-cyclophilin A fusion protein found in owl monkey kidney cells can restrict HIV-1. *Proc Natl Acad Sci U S A* **101**: 13324-13328.
- Ohno, S. 1970. *Evolution by gene duplication*. Springer-Verlag, Heidelberg.
- Ostertag, E.M., J.L. Goodier, Y. Zhang, and H.H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73**: 1444-1451.
- Pickeral, O.K., W. Makalowski, M.S. Boguski, and J.D. Boeke. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* **10**: 411-415.
- Prince, V.E. and F.B. Pickett. 2002. Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet* **3**: 827-837.
- Rozmahel, R., H.H. Heng, A.M. Duncan, X.M. Shi, J.M. Rommens, and L.C. Tsui. 1997. Amplification of CFTR exon 9 sequences to multiple locations in the human genome. *Genomics* **45**: 554-561.
- Sayah, D.M., E. Sokolskaja, L. Berthoux, and J. Luban. 2004. Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**: 569-573.
- Swofford, D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.0b10. Sinauer Associates, Inc., Sunderland, Massachusetts.
- Szak, S.T., O.K. Pickeral, D. Landsman, and J.D. Boeke. 2003. Identifying related L1 retrotransposons by analyzing 3' transduced sequences. *Genome Biol* **4**: R30.

- Thompson, J.D., T.J. Gibson, F. Plewniak, F. Jeanmougin, and D.G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876-4882.
- Toomey, R.E. and S.J. Wakil. 1966. Studies on the mechanism of fatty acid synthesis. XVI. Preparation and general properties of acyl-malonyl acyl carrier protein-condensing enzyme from *Escherichia coli*. *J Biol Chem* **241**: 1159-1165.
- Vinckenbosch, N., I. Dupanloup, and H. Kaessmann. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A* **103**: 3220-3225.
- Wang, H., J. Xing, D. Grover, D.J. Hedges, K. Han, J.A. Walker, and M.A. Batzer. 2005. SVA Elements: A Hominid-specific Retroposon Family. *J Mol Biol* **354**: 994-1007.
- Wei, W., N. Gilbert, S.L. Ooi, J.F. Lawler, E.M. Ostertag, H.H. Kazazian, J.D. Boeke, and J.V. Moran. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**: 1429-1439.
- Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555-556.

CHAPTER FOUR:
SUMMARY

Mobile elements contribute greatly to the fluidity of the genomes by introducing genetic variation. The SVA family has been relatively less well documented with respect to other retrotransposons residing within the human genome. In this study, I provided a comprehensive overview of the SVA elements found in the human genome for the first time.

In chapter two, I first determined the SVA copy number in the human genome sequence. To examine the genomic distribution of SVA elements, I analyzed their distribution at the chromosomal level by comparing the observed number of the elements with the expected numbers based on size of the chromosomes. A chi-square analysis revealed that the two distributions are significantly different ($\chi^2 = 78.29$, $df = 23$, $p < 0.001$), which indicated that SVA elements are non-randomly distributed in the human genome.

Since the properties of the genomic sequence vary greatly among different human chromosomes, we further analyzed the distribution of SVA elements in relation to other genomic properties, such as G+C, gene and repeat contents of the SVA-flanking genomic regions. Genomic distribution analysis indicated that the SVA elements are enriched in G+C-rich regions but have no preferences for inter- or intra-genic regions.

A phylogenetic analysis of the elements resulted in the recovery of six subfamilies, which were named SVA_A to SVA_F. Subfamily age estimates based upon nucleotide divergence indicate that subfamily SVA_A (13.56 Myrs) may have expanded contemporary to the divergence of the orangutan and the great apes (human, chimpanzee and gorilla) (12-15 million years ago (Mya)). The expansion of subfamilies SVA_B (11.56 Myrs), SVA_C (10.88 Myrs) and SVA_D (9.55 Myrs) may have predated the human, chimpanzee and gorilla divergence (~7 Mya). The relatively young age of subfamilies SVA_E (3.46 Myrs) and SVA_F (3.18 Myrs) suggested these two subfamilies may have expanded after the human and chimpanzee divergence

(~4-6 mya). In congruence with the age estimates, a survey of human genomic diversity associated with SVA_E and SVA_F subfamily members showed insertion polymorphism frequencies of 37.5% and 27.6%, respectively.

Next, we use a median-joining network to examine the relationship among the subfamilies. Results from this analysis indicate that the older SVA subfamilies evolved in a single lineage: the SVA_A consensus has the highest sequence similarity to the counterpart of the ancestral HERV-K10 consensus. SVA_B differs from SVA_A by 9 substitutions and a 16 bp deletion is present at the 5' end of the SVA_B consensus as compared to HERV-K10 sequence and the SVA_A consensus. This deletion is present in all other subfamily consensus as well. SVA_C is derived from SVA_B, with SVA_D subsequently having evolved from SVA_C. Unlike the older subfamilies, the human-specific SVA_E and SVA_F were derived independently from a re-constructed ancestral sequence which is only one substitution different from SVA_D. In order to investigate the evolutionary history of the SVA elements, the copy number of SVA elements in different primate genomes was estimated by quantitative PCR (QPCR) using a pair of intra-SVA primers. Both the age estimates and the QPCR results showed that the full-length SVA element may have appeared in the Hominid lineage after the divergence of Hominids and Old World Monkeys.

Both during and after integration into the genome, mobile elements frequently serve as substrates for creating genomic variation. In chapter three, I analyzed a specific mechanism (i.e. sequence transduction), by which SVA elements can impact their host genome.

To examine the SVA-related transduction events, 1752 full-length SVA elements in the reference sequence of human genome were studied. In total, 143 SVA elements containing >6 bp TSD were identified with putative transduced sequences. Overall, 52,740 bp of genomic

sequence was duplicated by these SVA mediated transductions. The sizes of the transduced sequences ranged from 35 to 1853 bp in length, with an average of 340 bp. I also found that 3' transduction is a common phenomenon among SVA members since all previously identified SVA families were involved in transduction.

One particularly interesting example of SVA-mediated 3' transduction is the H17_76 group. Detailed analysis of their transduced sequence at this locus resulted in the identification of the *AMAC1* gene and *AMAC1-like* genes, which code for an enzyme that may be involved in the fatty acid synthesis process. In total, there were four related loci in the human genome. The source locus (chr17_A) contained only the transduced sequence without the SVA element and the other three loci (chr17_B, chr18 and chr8) contained an SVA element along with the transduced sequence. Only the source locus has two exons separated by an intron, while the three SVA transduced loci contained intronless versions of *AMACIL3*. All these results suggest that the intron was spliced out during the retrotransposition process, which provides further evidence for a genomic mechanism that created the three duplicated copies.

Evolutionary analysis of the *AMAC1* gene showed that the source locus existed in all examined primates and other mammals whose genomic sequence are available. On the other hand, the transduction events seem to have happened after the divergence of African apes from orangutans but before the divergence of humans, chimpanzees and gorillas, approximately 7 to 14 Mya based on the estimated divergence time of primates. I further examined the selective constraints on all *AMAC* copies using the maximum likelihood-based program PAML. The results are in good agreement with the predictions of classical gene duplication theory, which suggests that the functional redundancy of newly duplicated genes will initially result in free evolution of all gene copies. Lastly, I examined the expression pattern of the four human *AMAC*

gene copies by RT-PCR. The results showed that at least two SVA-transduced duplicates of the *AMAC* gene are currently expressed in humans and that they may have differential tissue expression patterns.

In summary, by analyzing the structure of SVA subfamilies and the amplification mechanism during primate evolution, this study provides a good understanding of the biology of SVA elements within the primate order. This study also includes the first example of gene duplication derived from SVA-mediated 3' transduction. This offers strong evidence that mobile elements can serve as raw material for the generation of new genetic functions by being incorporated into existing genes.

The possibility that most of these hundred elements have fortuitously integrated right after a viable promoter is remote. So in the future study, I think the most important and interesting thing is to figure out the promoter region of the SVA element.

APPENDIX A:
SUPPLEMENTAL DATA

```

HERVK
SVA_A TCTCCCTCTGTT --- GCGAGGCTGGACTGACTGCCGTGATCTGGCTCGCTGCAACCTCCCTGCCTGGGCTCCCGTGAATTCCTGCTCGCCCTGCCAGTGGC
SVA_B TCTCCCTCTGATGCCAGCCGAGGCTGGACTGTACTGCCCCATCTCGGCTCACTGCAACCTCCCTGCCT --- GATTCCTCTGCTCAGCCTGCCAGTGGC
SVA_C TCTCCCTCTGATGCCAGCCGAGGCTGGACTGTACTGCCCCATCTCGGCTCACTGCAACCTCCCTGCCT --- GATTCCTCTGCTCAGCCTGCCAGTGGC
SVA_D TCTCCCTCTGATGCCAGCCGAGGCTGGACTGTACTGCCCCATCTCGGCTCACTGCAACCTCCCTGCCT --- GATTCCTCTGCTCAGCCTGCCAGTGGC
SVA_E TCTCCCTCTGATGCCAGCCGAGGCTGGACTGTACTGCCCCATCTCGGCTCACTGCAACCTCCCTGCCT --- GATTCCTCTGCTCAGCCTGCCAGTGGC
SVA_F TCTCCCTCTCATGCCAGCCGAGGCTGGACTGTACTGCCATCTCGGCTCACTGCAACCTCCCTGCCT --- GATTCCTCTGCTCAGCCTGCCAGTGGC

HERVK
SVA_A TGGGATTGCAAGCCGCGCCGCCACGCTGACTGGTTTTTTGATTTTTT --- GGTGGAGACGGGGTTTCGCCGTGTTGCCGGGCTGGTCTCCAGCTCCTGACCTCGAGTGA
SVA_B TGGGATTGCAAGCCGCGCGCCGCCACGCTGACTGGTTTTTTGATTTTTT --- GGTGGAGACGGGGTTTCGCCGTGTTGCCGGGCTGGTCTCCAGCTCCTGACCCGAGTGA
SVA_C TGGGATTGCAAGCCGCGCGCCGCCACGCTGACTGGTTTTTTGATTTTTT --- GGTGGAGACGGGGTTTCGCCGTGTTGCCGGGCTGGTCTCCAGCTCCTAACCAGTGA
SVA_D TGGGATTGCAAGCCGCGCGCCGCCACGCTGACTGGTTTTTTGATTTTTT --- GGTGGAGACGGGGTTTCGCCGTGTTGCCGGGCTGGTCTCCAGCTCCTAACCAGTGA
SVA_E TGGGATTGCAAGCCGCGCGCCGCCACGCTGACTGGTTTTTTGATTTTTT --- GGTGGAGACGGGGTTTCGCCGTGTTGCCGGGCTGGTCTCCAGCTCCTAACCAGTGA
SVA_F TGGGATTGCAAGCCGCGCGCCGCCACGCTGACTGGTTTTT --- GGTGGAGACGGGGTTTCGCCGTGTTGCCGGGCTGGTCTCCAGCTCCTAACCAGTGA

HERVK
SVA_A TCTGCCCGCTCGGCTCCCGAGGTGCTGGGATTGCAGACGGAGTCTCGTCACTCAATGCTCAATGTTGCCAGGCTGGAGTGCAGTGGCGTGAATCTCGGCTCGCTACA
SVA_B TCTGCCAGCTCGGCTCCCGAGGTGCGGGATTGCAGACGGAGTCTCGTCACTCAATGCTCAATGTTGCCAGGCTGGAGTGCAGTGGCGTGAATCTCGGCTCGCTACA
SVA_C TCTGCCAGCTCGGCTCCCGAGGTGCGGGATTGCAGACGGAGTCTCGTCACTCAATGCTCAATGTTGCCAGGCTGGAGTGCAGTGGCGTGAATCTCGGCTCGCTACA
SVA_D TCCGCCAGCTCGGCTCCCGAGGTGCGGGATTGCAGACGGAGTCTCGTCACTCAATGCTCAATGTTGCCAGGCTGGAGTGCAGTGGCGTGAATCTCGGCTCGCTACA
SVA_E TCCGCCAGCTCGGCTCCCGAGGTGCGGGATTGCAGACGGAGTCTCGTCACTCAATGCTCAATGTTGCCAGGCTGGAGTGCAGTGGCGTGAATCTCGGCTCGCTACA
SVA_F TCCGCCAACCTCGGCTCCCGAGGTGCGGGATTGCAGACGGAGTCTCGTCACTCAATGTTGCCAGGCTGGAGTGCAGTGGCGTGAATCTCGGCTCGCTACA

HERVK
SVA_A ACGCTCCACCTCCAGCCGCTGCTTGGCTCCCAAAGTCTAAACAGCTCCGAAGAGACAGCAACCTCGAGAACGGGCCATGATGACGATGGCGTTTTGTGCGAAAGA
SVA_B ACGCTCCACCTCCAGCCGCTGCTTGGCTCCCAAAGTCTAAACAGCTCCGAAGAGACAGCAACCTCGAGAACGGGCCATGATGACGATGGCGTTTTGTGCGAAAGA
SVA_C ACGCTCCACCTCCAGCCGCTGCTTGGCTCCCAAAGTCTAAACAGCTCCGAAGAGACAGCAACCTCGAGAACGGGCCATGATGACGATGGCGTTTTGTGCGAAAGA
SVA_D ACGCTCCACCTCCAGCCGCTGCTTGGCTCCCAAAGTCTAAACAGCTCCGAAGAGACAGCAACCTCGAGAACGGGCCATGATGACGATGGCGTTTTGTGCGAAAGA
SVA_E ACC --- ACCCTCCAGCCGCTGCTTGGCTCCCAAAGTCTAAACAGCTCCGAAGAGACAGCAACCTCGAGAACGGGCCATGATGACGATGGCGTTTTGTGCGAAAGA
SVA_F ACCTACACCTCCAGCCGCTGCTTGGCTCCCAAAGTCTAAACAGCT --- CATTGAGAACGGGCCAGGATGACAAATGGCGCTTTGTGGAATAGA

HERVK
SVA_A AAAGGGGAAATGTGGGAAAGCAAGAGAGATCAAAATTGACTGTCTGTGTAGAAAAGTAGACATAGGAGAC --- TCCATTTTGTATGTCTAAGAAAAATTC
SVA_B AAAGGGGAAATGTGGGAAAGCAAGAGAGATCAAGATTGACTGTGTGTAGAAAAGTAGACATAGGAGAC --- TCCATTTTGTCTGTACTAAGAAAAATTC
SVA_C AAAGGGGAAATGTGGGAAAGCAAGAGAGATCAAGATTGACTGTGTGTAGAAAAGTAGACATAGGAGAC --- TCCATTTTGTCTGTACTAAGAAAAATTC
SVA_D AAAGGGGAAATGTGGGAAAGCAAGAGAGATCAAGATTGACTGTGTGTAGAAAAGTAGACATAGGAGAC --- TCCATTTTGTCTGTACTAAGAAAAATTC
SVA_E AAAGGGGAAAGGTGGGAAAGCAAGAGAGATCAAGATTGACTGTGTGTAGAAAAGTAGACATAGGAGAC --- TCCATTTTGTCTGTACTAAGAAAAATTC
SVA_F AAAGGGGAAAGGTGGGAAAGCAAGAGATCAAGATTGACTGTGTGTAGAAAAGTAGACATAGGAGAC --- TCCATTTTGTCTGTACTAAGAAAAATTC

HERVK
SVA_A TCTGCTTGAGATCTGTTAATCTATGACCTTACCCCAACCCCGTCTCTCTGAACATGTGCTGTCAACTCAGGGTTAAATGGATTAAAGGCGGTGCAAGATGTGC
SVA_B TCTGCTTTGGGATGCTGTTAATCTATGACCTTACCCCAACCCCGTCTCTCTGAACATGTGCTGTCAACTCAGGGTTAAATGGATTAAAGGCGGTGCAAGATGTGC
SVA_C TCTGCTTTGGGATGCTGTTAATCTATGACCTTACCCCAACCCCGTCTCTCTGAACATGTGCTGTCAACTCAGGGTTAAATGGATTAAAGGCGGTGCAAGATGTGC
SVA_D TCTGCTTTGGGATGCTGTTAATCTATGACCTTACCCCAACCCCGTCTCTCTGAACATGTGCTGTCAACTCAGGGTTAAATGGATTAAAGGCGGTGCAAGATGTGC
SVA_E TCTGCTTTGGGATGCTGTTAATCTATGACCTTACCCCAACCCCGTCTCTCTGAACATGTGCTGTCAACTCAGGGTTAAATGGATTAAAGGCGGTGCAAGATGTGC
SVA_F TCTGCTTTGGGATGCTGTTAATCTATGACCTTACCCCAACCCCGTCTCTCTGAACATGTGCTGTCAACTCAGGGTTAAATGGATTAAAGGCGGTGCAAGATGTGC

HERVK
SVA_A TTTGTTAAACAGATGCTTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCCTCCCTAATCTCAAGTACCCAGGGACACAAACCTGCGGAAGCCGCGAGGGTCCCTGCCC
SVA_B TTTGTTAAACAGATGCTTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCCTCCCTAATCTCAAGTACCCAGGGACACAAACCTGCGGAAGCCGCGAGGGTCCCTGCCC
SVA_C TTTGTTAAACAGATGCTTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCCTCCCTAATCTCAAGTACCCAGGGACACAAACCTGCGGAAGCCGCGAGGGTCCCTGCCC
SVA_D TTTGTTAAACAGATGCTTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCCTCCCTAATCTCAAGTACCCAGGGACACAAACCTGCGGAAGCCGCGAGGGTCCCTGCCC
SVA_E TTTGTTAAACAGATGCTTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCCTCCCTAATCTCAAGTACCCAGGGACACAAACCTGCGGAAGCCGCGAGGGTCCCTGCCC
SVA_F TTTGTTAAACAGATGCTTTGAAGGCAGCATGCTCGTTAAGAGTCATCACCCTCCCTAATCTCAAGTACCCAGGGACACAAACCTGCGGAAGCCGCGAGGGTCCCTGCCC
0

HERVK
SVA_A TAGGAAAGCCAGAGACCTTTGTTACGTTGTTGCTGCTGACCTCTCCCACAATGTTGTTGACCCCTGACACATCCCCCTTTGAGAAACCCCAAGATGATCAA
SVA_B TAGGAAAGCCAGAGACCTTTGTTACGTTGTTGCTGCTGACCTCTCCCAATATTATCTATGACCCCTGCCACATCCCCCTCTCCGAGAAACCCCAAGATGATCAA
SVA_C TAGGAAAGCCAGAGACCTTTGTTACGTTGTTGCTGCTGACCTCTCCCAATATTGTTCTATGACCCCTGCCAAATCCCCCTCTCCGAGAAACCCCAAGATGATCAA
SVA_D TAGGAAAGCCAGAGACCTTTGTTACGTTGTTGCTGCTGACCTCTCCCAATATTGTTCTATGACCCCTGCCAAATCCCCCTCTCCGAGAAACCCCAAGATGATCAA
SVA_E TAGGAAAGCCAGAGACCTTTGTTACGTTGTTGCTGCTGACCTCTCCCAATATTGTTCTATGACCCCTGCCAAATCCCCCTCTCCGAGAAACCCCAAGATGATCAA
SVA_F TAGGAAAGCCAGAGACCTTTGTTACGTTGTTGCTGCTGACCTCTCCCAATATTGTTCTATGACCCCTGCCAAATCCCCCTCTCCGAGAAACCCCAAGATGATCAA

HERVK
SVA_A TAAATACTAA
SVA_B TAAATACTAA
SVA_C TAAAAAAAAA
SVA_D TAAAAAAAAA
SVA_E TAAAAAAAAA
SVA_F TAAAAAAAAA

```

Figure A.1: Multiple alignments of human SVA subfamily consensuses along with the corresponding HERV-K10 sequence

```

      10      20      30      40      50      60      70      80
AMAC1  |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L1 .....C.....T.....A.....A...
AMAC1L2 .....C.....T.....
AMAC1L3 .....C..G...C.....

      90     100     110     120     130     140     150     160
AMAC1  |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L1 GTACCAACGCTGCCAGCCCTCTGATGCCACCAGTGGCCTGCTGGTGGCCCTGCTGGGTGGGGCCTGCCTGCTGGCTTCG
AMAC1L1 .C...G.....A.....
AMAC1L2 .C...G.....G.....A.....
AMAC1L3 .C...GT.....A.....

     170     180     190     200     210     220     230     240
AMAC1  |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L1 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L2 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L3 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
          G.....T.....
          G.....T.....
          A.....T.....

     250     260     270     280     290     300     310     320
AMAC1  |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L1 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L2 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L3 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
          CTCCCTATTGCCCTGCTACTTAACTGCGTGGCGACCCCTTCTGGGAACTCCTGACATCCGAAGCCGGGCCTTCTTCTG
          C.....G...CA..G.....
          C.....G..T...G.....
          C.....G.....A...A

     330     340     350     360     370     380     390     400
AMAC1  |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L1 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L2 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L3 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
          TGCCCTGCTCAACATCCTCAGCATTGGATGTGCCTACAGTGCGGTTCAGGTGGTGCCCGCTGGCAACGCTGCCACTGTTC
          G.....A.....T.....A.....T.....
          G.....A.....
          G.....

     410     420     430     440     450     460     470     480
AMAC1  |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L1 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L2 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1L3 |.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
          GCAAAGGTTCTTCCACCGTCTGCTCCGCCGTCTCACTCTCTGCCTTGAGAGCCAGGGTCTCAGTGGCTACGACTGGTGT
          CA.....A.....C.....T.....
          A.....T.....C.....G.....G.....
          C.....

```

```

          490      500      510      520      530      540      550      560
    .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1    GGACTGTTGGGCTGCATCCTAGGACTAATCATCATTGTGGGACCTTGGACTCTGGACACTACAGGAGGGGACCACGGGTGT
AMAC1L1    .....A.....A.....
AMAC1L2    .....A.....C.....A.....
AMAC1L3    .....A.....T.....

          570      580      590      600      610      620      630      640
    .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1    CTACACCGCCCTGGGCTATGTGGAGGCTTTCCTGGGAGGCCTGGCGCTGTCCCTGAGGCTTCTGGTCTATCGTTCTCTGC
AMAC1L1    .....G.....C.....A.....G.....C.....
AMAC1L2    .....A.....C.....G.....
AMAC1L3    .....G.C.....G.....A.....G.....

          650      660      670      680      690      700      710      720
    .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1    ACTTTCCCCCTGCCTCCCAACAGTGGCCTTCTATCTGGCTTGGTGGGGCTGCTGGGCTCTGTGCCAGGCCTCTTTGTG
AMAC1L1    .....T.....
AMAC1L2    .....T.....G.....
AMAC1L3    .....T.....

          730      740      750      760      770      780      790      800
    .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1    CTGCAGGCCCCCGTGTGGCCAGTGACCTCCTGAGTTGGAGTTGTGTGGGGCAGTGGGGATCCTCGCCTTGGTCTCCTT
AMAC1L1    .....T.....A.....
AMAC1L2    .....A.....A.....
AMAC1L3    .....C.....T..C.....

          810      820      830      840      850      860      870      880
    .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1    CACATGTGTGGGCTATGCGGTACCAAGGCCACCCTGCCCTGGTGTGCGCTGTCCTACATTCCGAGGTGGTTGTGGCCC
AMAC1L1    .....T.....G.....GA.....
AMAC1L2    .....G.....
AMAC1L3    .....A.....G.....G.....

          890      900      910      920      930      940      950      960
    .....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|
AMAC1    TTATACTGCAGTATTATATGCTCCATGAGACTGTGGCACCTTCTGACATCGTGGCGGCAGGGGTTGTGCTGGGCAGCATT
AMAC1L1    .....T.....A..G.....
AMAC1L2    .....T.....A..G.....
AMAC1L3    .....G.....

```

```

          970      980      990      1000      1010
.....|.....|.....|.....|.....|.....|.....|.....|.....|.....|..
AMAC1      GCCATCATTACAGCCCAGAACCTCAGCTGTGAGAGGACAGGGAGGGTGGAGGAGTGA
AMAC1L1      .....G.....T.....A.....
AMAC1L2      .....G.....A.....
AMAC1L3      .....C.....TG.....GA.....A.....

```

Figure A.2: Multiple sequence alignment of human *AMAC* coding regions. Sequence of the coding region of human *AMAC1* is shown at the top and the sequence of other *AMAC* copies are shown below. Dots represent the same nucleotides as the *AMAC1* sequence. The conserved primer pair used in the *AMAC* expression studies is shown in boldface and underlined.

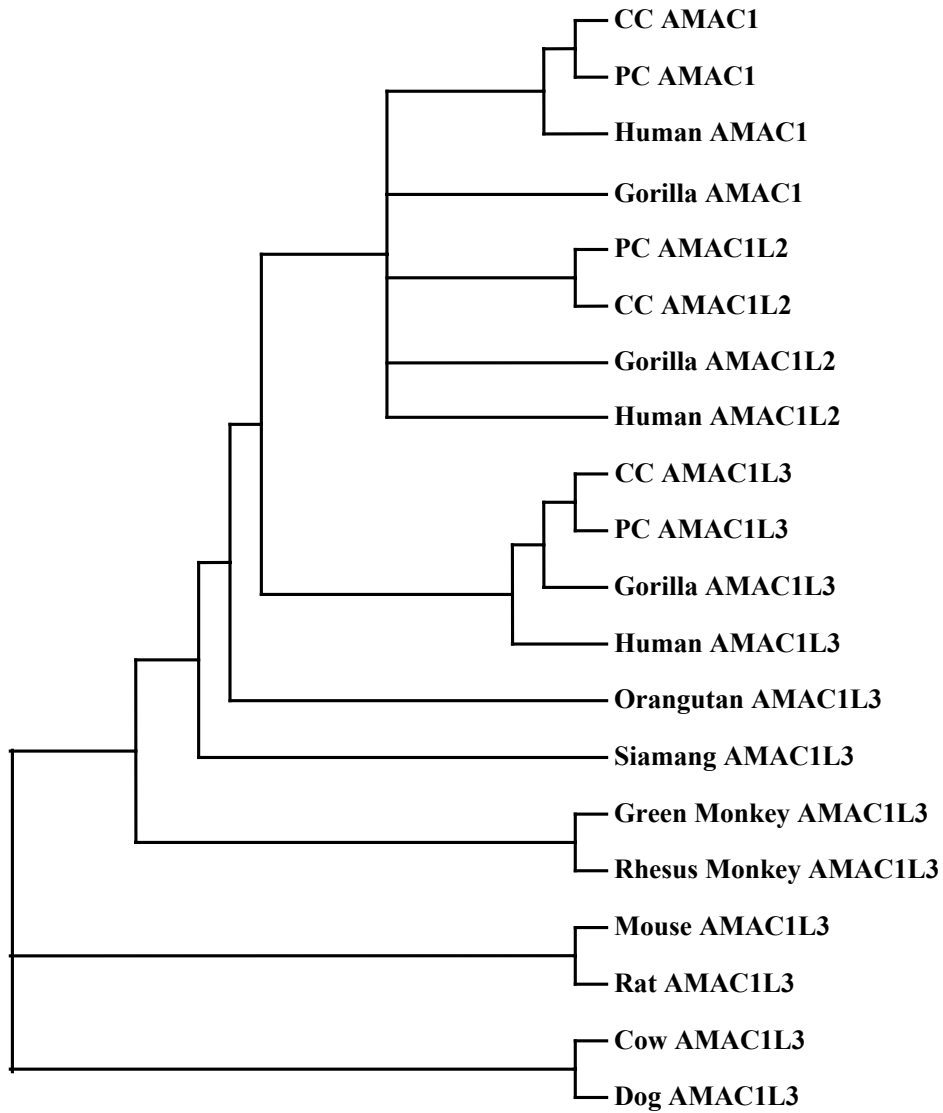


Figure A.3: Phylogeny of the AMAC coding regions used in PAML analysis. Human AMAC gene nomenclature was used in all species. CC: common chimpanzee; PC: pygmy chimpanzee.

Table A.1: Insertion allele frequencies, heterozygosities and genotypes of the SVA insertion polymorphisms.

Elements	African American					Asian					European					South American					Avg Het
	Genotypes					Genotypes					Genotypes					Genotypes					
	+/+	+/-	-/-	<i>f</i>	Het	+/+	+/-	-/-	<i>f</i>	Het	+/+	+/-	-/-	<i>f</i>	Het	+/+	+/-	-/-	<i>f</i>	Het	
H1_F_95	19	1	0	0.98	0.05	20	0	0	1	0	20	0	0	1	0	20	0	0	1	0	0.01
H1_F_127	18	2	0	0.95	0.1	20	0	0	1	0	20	0	0	1	0	20	0	0	1	0	0.03
H1_F_163	4	2	14	0.25	0.38	3	6	8	0.35	0.47	9	9	2	0.68	0.45	7	10	3	0.6	0.49	0.33
H2_F_9	11	9	0	0.78	0.36	2	18	0	0.55	0.51	12	8	0	0.8	0.33	14	6	0	0.85	0.26	0.3
H2_F_56	1	8	11	0.25	0.38	6	12	2	0.6	0.49	3	5	12	0.28	0.41	2	4	14	0.2	0.33	0.32
H2_F_76	6	10	4	0.55	0.51	10	6	4	0.65	0.47	11	8	1	0.75	0.38	13	6	1	0.8	0.33	0.34
H2_F_129	3	5	9	0.32	0.45	4	12	4	0.5	0.51	3	4	12	0.26	0.4	2	6	11	0.26	0.4	0.34
H5_F_9	20	0	0	1	0	19	1	0	0.98	0.05	19	1	0	0.98	0.05	18	2	0	0.95	0.1	0.03
H5_F_171	0	17	3	0.43	0.5	6	9	5	0.53	0.51	6	11	3	0.58	0.5	7	8	5	0.55	0.5	0.38
H6_F_74	16	0	0	1	0	20	0	0	1	0	19	0	0	1	0	18	0	0	1	0	0
H6_F_172	2	9	9	0.33	0.45	0	1	19	0.03	0.05	1	8	11	0.25	0.38	0	2	18	0.05	0.1	0.22
H7_F_22	3	11	6	0.43	0.5	15	5	0	0.88	0.22	12	8	0	0.8	0.33	5	10	5	0.5	0.51	0.26
H7_F_35	10	1	2	0.81	0.32	13	2	0	0.93	0.13	3	6	7	0.38	0.48	9	3	5	0.62	0.49	0.23
H7_F_61	0	10	10	0.25	0.38	0	6	14	0.15	0.26	2	10	8	0.35	0.47	4	9	7	0.43	0.5	0.28
H7_F_148	3	11	6	0.43	0.5	0	0	20	0	0	4	7	9	0.38	0.48	1	8	11	0.25	0.38	0.25
H8_F_89	20	0	0	1	0	20	0	0	1	0	19	0	0	1	0	20	0	0	1	0	0
H1_E_43	0	4	16	0.1	0.18	0	0	20	0	0	0	0	20	0	0	0	1	19	0.03	0.05	0.05
H1_E_89	5	15	0	0.63	0.48	10	10	0	0.75	0.38	3	17	0	0.58	0.5	11	9	0	0.78	0.36	0.34
H1_E_271	2	3	15	0.18	0.3	3	13	4	0.48	0.51	1	12	7	0.35	0.47	1	11	8	0.33	0.45	0.32
H2_E_239	0	2	17	0.05	0.1	0	0	20	0	0	2	2	16	0.15	0.26	0	0	19	0	0	0.09
H2_E_245	0	3	17	0.08	0.14	6	11	3	0.58	0.5	1	5	6	0.29	0.43	2	6	7	0.33	0.46	0.27
H2_E_256	0	2	18	0.05	0.1	0	2	18	0.05	0.1	1	8	11	0.25	0.38	0	6	14	0.15	0.26	0.15
H5_E_27	20	0	0	1	0	20	0	0	1	0	20	0	0	1	0	19	1	0	0.98	0.05	0
H5_E_128	0	17	3	0.43	0.5	0	9	11	0.23	0.36	0	12	6	0.33	0.46	3	4	13	0.25	0.38	0.33
H6_E_1	2	13	5	0.43	0.5	0	1	19	0.03	0.05	1	7	12	0.23	0.36	0	5	15	0.13	0.22	0.23
H6_E_146	8	11	1	0.68	0.45	16	4	0	0.9	0.18	9	10	1	0.7	0.43	15	3	2	0.83	0.3	0.27
H7_E_5	1	14	5	0.4	0.49	2	7	9	0.31	0.44	1	6	9	0.25	0.39	2	9	8	0.34	0.46	0.33
H7_E_27	20	0	0	1	0	19	1	0	0.98	0.05	20	0	0	1	0	20	0	0	1	0	0.01

(Table A.1 cont.)

H7_E_82	3	13	3	0.5	0.51	0	12	8	0.3	0.43	0	20	0	0.5	0.51	0	15	5	0.38	0.48	0.36
H7_E_171	0	20	0	0.5	0.51	4	16	0	0.6	0.49	4	16	0	0.6	0.49	4	16	0	0.6	0.49	0.37
H8_E_35	19	1	0	0.98	0.05	20	0	0	1	0	20	0	0	1	0	20	0	0	1	0	0.01
H8_E_45	20	0	0	1	0	19	1	0	0.98	0.05	20	0	0	1	0	20	0	0	1	0	0.01
H14_E_66	11	7	2	0.93	0.41	6	11	3	0.58	0.5	0	4	16	0.1	0.18	2	9	9	0.33	0.45	0.27
H20_E_91	18	2	0	0.95	0.1	20	0	0	1	0	19	1	0	0.98	0.05	19	1	0	0.98	0.05	0.04

Table A.2: PCR primers, annealing temperature, amplicon sizes of human specific SVA elements.

Name	Chr	5' Primer Sequence (5'-3')	3' Primer Sequence (5'-3')	Annealing	Product Sizes	
				Temp	Empty	Filled
H1_E_8IN	1	TCTGCCTAGGAAAACCAGAGAC	GGCTGACTTAGAGGACAAGTGC	60	-	395
H1_E_8OUT	1	ATTACTGTGAGCAGGACCAACC	GGCTGACTTAGAGGACAAGTGC	60	345	3086
H1_E_27IN	1	TGTGTCCACTCAGGGTAAATG	TGTCACCTTCCCTGATACCTCT	60	-	407
H1_E_27OUT	1	GCCAACACTGCTTTTTCATCTT	TGTCACCTTCCCTGATACCTCT	60	236	2786
H1_E_43IN	1	AGTACCCAGGGACACAAACT	TTCATCGTTTCTGGAATTCTT	60	-	408
H1_E_43OUT	1	AGCCATAGCCAATGATCAAGTT	TTCATCGTTTCTGGAATTCTT	60	301	3121
H1_E_50IN	1	TGTGTCCACTCAGGGTAAATG	CAAGAAGGGCAGAGCCTATCTA	60	-	386
H1_E_50OUT	1	GAAATGTCCTGCAGTTGTTGA	CAAGAAGGGCAGAGCCTATCTA	60	396	2566
H1_E_89IN	1	AACCCTGTGCTCTCTGAAACAT	AGCTCATGAGGCACCATTATT	60	-	439
H1_E_89OUT	1	GGAGCTGCTAACGCTAATGAAC	AGCTCATGAGGCACCATTATT	60	276	1187
H1_E_101IN	1	ACCTTCCCTCCACTATTGTCCT	TGCATAAGAATCACCCAGAAAC	60	-	469
H1_E_101OUT	1	TTTGTCTAAAGGCTCTGAGACA	TGCATAAGAATCACCCAGAAAC	60	828	3448
H1_E_115IN	1	GGATCCTGTTGATCTGTGACCT	CAACATTTACATTTGTGGATCA	60	-	396
H1_E_115OUT	1	TCCATTATTTTCGGTTTGTCC	CAACATTTACATTTGTGGATCA	60	303	2812
H1_E_124IN	1	TGTGTCCACTCAGGGTAAATG	GAAATGCCAAATTCCTCTGTA	60	-	639
H1_E_124OUT	1	TCCAAGACAATCTTCTAAACCTCA	GAAATGCCAAATTCCTCTGTA	60	752	2959
H1_E_189IN	1	TGTGTCCACTCAGGGTAAATG	CTCTGCCATCTTTGGTGTATCA	60	-	415

(Table A.2 cont.)

H1_E_189OUT	1	CATGAGAGAAAAGCTGATGCAAG	CTCTGCCATCTTTGGTGTATCA	60	189	2659
H1_E_271IN	1	TGAGAGTCATCACCCTCCCTA	GGGCAAGTAAGCCAGTATTGAA	60	-	425
H1_E_271OUT	1	GTTCAATTCCTGGAAGCACAA	GGGCAAGTAAGCCAGTATTGAA	60	709	2115
H1_E_302IN	1	TCTGAAACATGTGTTGTGTCCA	GGCAGAACTAACAAAGCAGACA	60	-	778
H1_E_302OUT	1	CCGTATGATATTCCATCGTGTG	GGCAGAACTAACAAAGCAGACA	60	805	2374
H2_E_11IN	2	TGTGTCCACTCAGGGTTAAATG	ATGAGTCTTTTGCAGGCATTCT	60	-	349
H2_E_10OUT	2	GGTTGTGTGACTTGGCTATCA	ATGAGTCTTTTGCAGGCATTCT	60	360	1837
H2_E_13IN	2	TGTGTCCACTCAGGGTTAAATG	TTAAATGGCTTaCAGCgAACTG	60	-	500
H2_E_13OUT	2	ACTTCCCTGAAGTTTGTGGTT	TTAAATGGCTTaCAGCgAACTG	60	299	1269
H2_E_33IN	2	TGTGTCCACTCAGGGTTAAATG	GAGGTATTGGGGTGAATGTTTG	60	-	425
H2_E_33OUT	2	GGTCTCAGCCATTAATTTGAGC	GAGGTATTGGGGTGAATGTTTG	60	426	2760
H2_E_39IN	2	TGTGTCCACTCAGGGTTAAATG	GCAGGTGGCAGTCTTAGTACCT	60	-	455
H2_E_39OUT	2	TCCAAATCAAAGAGGACTGGAT	GCAGGTGGCAGTCTTAGTACCT	60	319	2259
H2_E_140IN	2	TGTGTCCACTCAGGGTTAAATG	AACCCAATGTACATCCAAGACC	60	-	689
H2_E_140OUT	2	TTCCAAATGTCAGGGTAAACAGA	AACCCAATGTACATCCAAGACC	60	673	2509
H2_E_219IN	2	TGTGTCCACTCAGGGTTAAATG	TAGGCCCTGTTGTCAATAATCC	60	-	392
H2_E_219OUT	2	TGTGAAACCATGGTCAGTAAGC	TAGGCCCTGTTGTCAATAATCC	60	183	2423
H2_E_239IN	2	TGTGTCCACTCAGGGTTAAATG	CATGCTTTTTGCCTTAATTCA	60	-	748
H2_E_239OUT	2	ATACAGAAGGTACTGGGGAAGC	CATGCTTTTTGCCTTAATTCA	60	627	3027
H2_E_245IN	2	TGTGTCCACTCAGGGTTAAATG	AGGGAAGGTGTGGAAAAAGATT	60	-	768
H2_E_245OUT	2	AAACCAACAACCTTGCATAGGG	AGGGAAGGTGTGGAAAAAGATT	60	551	3109
H2_E_256IN	2	TGTGTCCACTCAGGGTTAAATG	CTCTTTTGTCTGGCCTGTTAC	60	-	330
H2_E_256OUT	2	AATGTGAAGGTCTCCACTGACC	CTCTTTTGTCTGGCCTGTTAC	60	155	2645
H2_E_261IN	2	TCTGCCTAGGAAAACCAGAGAC	TGGCCTGGAACAAACATACATA	60	-	669
H2_E_261OUT	2	CCAGAAAACAATGGAGCAAAAT	TGGCCTGGAACAAACATACATA	60	616	2123
H5_E_22IN	5	TGTGTCCACTCAGGGTTAAATG	GAAGACAGGGtGGAAGAAAATG	55	-	405
H5_E_22OUT	5	TGCCAGTCTCTTCGTGAAACTA	GAAGACAGGGtGGAAGAAAATG	55	430	1873
H5_E_27IN	5	TCTGCCTAGGAAAACCAGAGAC	ACAGCATGAGGCCAAACTCTCTT	55	-	432

(Table A.2 cont.)

H5_E_27OUT	5	AATTGAGGTTAGATAGACCTGTTCA	GCATGAGGCAAACCTCTCTTTTT	55	180	3075
H5_E_128IN	5	TGTGTCCACTCAGGGTAAATG	GCTGAGAGTCTCCTCCCTGTTA	55	-	412
H5_E_128OUT	5	CCTTGTAGTTTGTGGCATGAA	GCTGAGAGTCTCCTCCCTGTTA	55	470	1795
H6_E_11IN	6	AGTACCCAGGGACACAAACACT	ATTCAAATCCCTCTGCAAAGAA	55	-	553
H6_E_10OUT	6	TCCGGTAAAGCTTCATAACAGTT	ATTCAAATCCCTCTGCAAAGAA	55	640	2572
H6_E_146IN	6	TGCTCTCTGAAACATGTGCTGT	TTTCTCCCTTCCCATTTTTGTA	55	-	419
H6_E_146OUT	6	TTAACAGCACGTAGCACAGTGA	TTTCTCCCTTCCCATTTTTGTA	55	580	3196
H6_E_157IN	6	TGTGTCCACTCAGGGTAAATG	GAAATGGACTGTGGATTGTGAA	55	-	438
H6_E_157OUT	6	CCAACCTAGTAGATGGCTCCTC	AGGTTCCCcTCTGTAAATGCT	55	580	2269
H7_E_5IN	7	AGTACCCAGGGACACAAACACT	TCATTTCCATCAGGtAGGCTTT	55	-	558
H7_E_5OUT	7	AATTTTGAAACTGGGTGGTCAG	TCATTTCCATCAGGtAGGCTTT	55	510	3099
H7_E_27IN	7	AGTACCCAGGGACACAAACACT	GCCTATGCTGATAACCACTCCTC	55	-	420
H7_E_27OUT	7	GTGCACAGAAGGATACATGGAA	GCCTATGCTGATAACCACTCCTC	55	480	2579
H7_E_31IN	7	TCTGCCTAGGAAAACCAGAGAC	CATTGGGAGAGGGGACATTTAT	55	-	490
H7_E_31OUT	7	TTGCCAATTTCTAACCTGACT	CATGGCAATCACCTCCTTTAT	55	180	3211
H7_E_45IN	7	CTGTAGGTTCTGATCGCCCTAT	CCAGAGAGAATTTGGCAAGACT	55	-	508
H7_E_45OUT	7	CATGCCAGCTCAGCTATGTAAG	CCAGAGAGAATTTGGCAAGACT	55	700	2518
H7_E_82IN	7	AAGTACCCAGGGACACAAACAC	ACAGAGGATTGGCGTAATGACT	55	-	783
H7_E_82OUT	7	TTCTGGAATACCATCTTTGGA	ACAGAGGATTGGCGTAATGACT	55	800	3250
H7_E_156IN	7	TGTGTCCACTCAGGGTAAATG	GCCTTCCCATCATATTTGTCAC	55	-	431
H7_E_156OUT	7	CCCTCTGAAGCTCACAGAAATC	GCCTTCCCATCATATTTGTCAC	55	750	2998
H7_E_171IN	7	TGAACACTCAGGGTAAATGGA	TTGATTTCTtGCAGCAATGTTc	55	-	409
H7_E_171OUT	7	GAATAGAAGGGTGCTTCCTCAA	TTGATTTCTtGCAGCAATGTTc	55	500	3112
H8_E_15IN	8	TGTGTCCACTCAGGGTAAATG	AGGTGGGGGAGGTGTATAATAA	55	-	477
H8_E_15OUT	8	TGCTTTTACCATATTTCCACA	AGGTGGGGGAGGTGTATAATAA	55	190	2553
H8_E_17IN	8	AGTACCCAGGGACACAAACACT	AAGGGAAATATTGGGAGAAGGA	55	-	531
H8_E_17OUT	8	TGGCTGCAGGGACTTAATTTAT	AAGGGAAATATTGGGAGAAGGA	55	580	2602
H8_E_35IN	8	TGTGTCCACTCAGGGTAAATG	CAAGAGCGATTGAAGTTTCT	55	-	423

(Table A.2 cont.)

H8_E_35OUT	8	TTGTGTCCTCTTCTGCTAACGA	CAAGAGCGATTGAAGGTTTTCT	55	350	2999
H8_E_45IN	8	TGTGTCCACTCAGGGTTAAATG	CAGTAAAAGAGAGCGAATGCAG	55	-	480
H8_E_45OUT	8	CCACAACCTCTGGCTTAATGTGA	CAGTAAAAGAGAGCGAATGCAG	55	710	2998
H8_E_94IN	8	AGTACCCAGGGACACAAACACT	TGCTTCATCAGTGAGATTGACA	55	-	487
H8_E_94OUT	8	GCTGAAAACATCTCAAGTCACG	CCATTGTCTCTTGGCCTATGTT	55	260	2393
H8_E_103IN	8	GGTGCAAGATGTGCTTTGTTA	CCTCATCACGATCATTATTTAAGC	55	-	435
H8_E_103OUT	8	TACCAAGCTCATGTTCTTTTCA	CCTCATCACGATCATTATTTAAGC	55	270	3272
H8_E_113IN	8	AACCCTGTGCTCTCTGAAACAT	CATAACCCCATACAAGGCTGTT	55	-	438
H8_E_113OUT	8	ACACAAGTCACATGGTCAAACC	CATAACCCCATACAAGGCTGTT	55	200	1000
H14_E_33IN	8	AGTACCCAGGGACACAAACACT	GGGAATGAAAACCACAAcTGAT	60	-	399
H14_E_33OUT	8	ATTTGCAGGTGCTTTTGGTTAT	GGGAATGAAAACCACAAcTGAT	60	308	2868
H14_E_66IN	8	AGTACCCAGGGACACAAACACT	AAAACGTAGGATTGGCCAGAA	60	-	703
H14_E_66OUT	8	CTACCAGTCTTGCCTTCAGCTT	AAAACGTAGGATTGGCCAGAA	60	225	3375
H15_E_24IN	15	TCTGCCTAGGAAAACCAGAGAC	AAAATCAGGTTGTTGCCATCAT	60	-	342
H15_E_24OUT	15	TGCAAAGTGTGCGAAAATAATC	AAAATCAGGTTGTTGCCATCAT	60	808	2471
H15_E_29IN	15	AACCCTGTGCTCTCTGAAACAT	CACAATGATTGGTATGGTCAGG	60	-	619
H15_E_29OUT	15	TGTTGGGTAGGAAGAAATGAAA	TCTCCTTTGTGCTCCTTTAT	60	379	2881
H15_E_48IN	15	AACCCTGTGCTCTCTGAAACAT	GCATCTTTGAAGCTGAATCTCAT	60	-	372
H15_E_48OUT	15	CTTGCAAACACTACAAATGGCTTG	GCATCTTTGAAGCTGAATCTCAT	60	360	2878
H16_E_86IN	16	AGTACCCAGGGACACAAACACT	ATCTGCCCTGAATTTTATGCAC	60	-	561
H16_E_86OUT	16	AGCTGTAGAAAAGGCTTTTGGAA	ATCTGCCCTGAATTTTATGCAC	60	457	1183
H20_E_91IN	20	ACCTTCCTCCACTATTGTCCT	AATGACATAAAGGCTGCCAACTT	60	-	466
H20_E_91OUT	20	GATGAGTGGCAAGTTCATCAAA	AATGACATAAAGGCTGCCAACTT	60	363	1456
H1_F_28IN	1	AACCCTGTGCTCTCTGAAACAT	TCCTTGTTTTGCCTACACATGA	60	-	497
H1_F_28OUT	1	AAACATTAGGCACGCAAAAACACT	TCTCTGGCATTACTGACGTGTT	60	504	2178
H1_F_32IN	1	GCTCGTTAAGAGTCATCACCAA	AGGGTGGATGATTCATGAGTTT	60	-	465
H1_F_32OUT	1	CAGTACAGAAGCCGACCAAAA	AGGGTGGATGATTCATGAGTTT	60	241	3101
H1_F_41IN	1	TGTGCTTTGCTAAACAGATGCT	TGACACTTGCATCAAGAGGAAT	60	-	314

(Table A.2 cont.)

H1_F_41OUT	1	CGGTGTCTGTTCTTGCTAAGTG	TGCTTCCCCTTTGACTTCTATC	60	490	2945
H1_F_58IN	1	TAATCAGGGACACAAACACTGC	CAAGCATATCCCCACCTTAAAC	60	-	470
H1_F_58OUT	1	GAATATGCCTGCATACGGAATC	CAAGCATATCCCCACCTTAAAC	60	501	3471
H1_F_67IN	1	TAATCAGGGACACAAACACTGC	CACACATCCTGTCCATTCAACT	60	-	388
H1_F_67OUT	1	GTTACAATGCCACGACTAGCAG	CACACATCCTGTCCATTCAACT	60	303	2848
H1_F_95IN	1	TGTGTCCACTCAGGGTTAAATG	GaAGGAATGAGGATCACCAAAA	60	-	760
H1_F_95OUT	1	GTGGTGAAAAACAACAAATG	GaAGGAATGAGGATCACCAAAA	60	1345	3791
H1_F_127IN	1	TGTGTCCACTCAGGGTTAAATG	TCTTCAGCAGATTCTCATTTGG	60	-	412
H1_F_127OUT	1	TGCGCAATTCATAGTGTTAGG	TCTTCAGCAGATTCTCATTTGG	60	465	2660
H1_F_159IN	1	TAATCAGGGACACAAACACTGC	TTGATTGGGCTCTCAGTAGTCA	60	-	618
H1_F_159OUT	1	TCTGCTTCACATCCTTTGAATG	TTGATTGGGCTCTCAGTAGTCA	60	716	2247
H1_F_160IN	1	TGTGTCCACTCAGGGTTAAATG	GaAATACCATCTGGAGGGAGAA	60	-	667
H1_F_160OUT	1	ACCACCTTTATGCCCTAGTCAA	GaAATACCATCTGGAGGGAGAA	60	710	2396
H1_F_163IN	1	AAGTACCCAGGGACACAAACAC	TCTAACCTGTATCCTCCGCAGT	60	-	788
H1_F_163OUT	1	TTGTAAGGGATTGGGGTTTATG	TCTAACCTGTATCCTCCGCAGT	60	956	3290
H1_F_171IN	1	TAATCAGGGACACAAACACTGC	GGCGAACAGAAAACCAAAAATA	60	-	495
H1_F_171OUT	1	CAGCAGCTGTTGAGTAAAGTGC	GGCGAACAGAAAACCAAAAATA	60	683	3075
H1_F_186IN	1	TGTGTCCACTCAGGGTTAAATG	CCCTTTATATTGGGCATACAGTG	60	-	369
H1_F_186OUT	1	AACAAACAGGCCTCTCAGTTTC	CCCTTTATATTGGGCATACAGTG	60	473	2532
H1_F_187IN	1	TGTGTCCACTCAGGGTTAAATG	CACTTTTTCACAACGTGTGGTT	60	-	770
H1_F_187OUT	1	TTCAAACAACCTGACACCGAAAG	CACTTTTTCACAACGTGTGGTT	60	712	2410
H1_F_227IN	1	TGTGTCCACTCAGGGTTAAATG	TTCATTTGCCCCACTTCTAAAC	60	-	348
H1_F_227OUT	1	TTTGGGTAGAAAGATGAACCTG	TTCATTTGCCCCACTTCTAAAC	60	139	2789
H1_F_240IN	1	TAATCAGGGACACAAACACTGC	TTAATTTTCCGAGAAACCTCCA	60	-	780
H1_F_240OUT	1	GCAGACTATCCACCTGAAAAGG	TTAATTTTCCGAGAAACCTCCA	60	796	2773
H1_F_264IN	1	TAATCAGGGACACAAACACTGC	TTCTATTTAGCCCTCCCTTTC	60	-	348
H1_F_264OUT	1	TTCCTGATGCATAGTAGCCTCTC	TTCTATTTAGCCCTCCCTTTC	60	530	1283
H1_F_281IN	1	TGTGTCCACTCAGGGTTAAATG	TTTTCTACCTGTCCCACAAAT	60	-	442

(Table A.2 cont.)

H1_F_281OUT	1	TCCATTGATGACAAACTTCCTG	TTTCCTACCTGTCCCACAAAT	60	402	2502
H1_F_306IN	1	TCTGCCTAGGAAAACCAGAGAC	CTGATGGGCAACCTTTTATCTC	60	-	459
H1_F_306OUT	1	TTGTATTTTGGTCTTCCCATC	CTGATGGGCAACCTTTTATCTC	60	640	2275
H1_F_309IN	1	TAATCAGGGACACAAACACTGC	CATTTAAATCACCACGCTGAAG	60	-	485
H1_F_309OUT	1	AGAGTGAAAGGTGACTCCAAGC	CATTTAAATCACCACGCTGAAG	60	478	1055
H2_F_9IN	2	TCTGCCTAGGAAAACCAGAGAC	GCTGTTCAtGCCAATAATGTA	60	-	577
H2_F_9OUT	2	TTTGCTAGACGGCAAATGATG	GCTGTTCAtGCCAATAATGTA	60	585	2544
H2_F_56IN	2	TGTGTCCACTCAGGGTTAAATG	ACCAAACAGGAAaTACCCTGAA	60	-	368
H2_F_56OUT	2	GGGTAGGCTAACTGCTACAACG	ACCAAACAGGAAaTACCCTGAA	60	114	2834
H2_F_62IN	2	AACCCTGTGCTCTCTGAAACAT	AGTCTCCCATCCACTTTGTGTT	60	-	532
H2_F_62OUT	2	CTACAAAGCAGCACCCACATAC	CTCCCATCCACTTTGTGTTAAA	60	358	2938
H2_F_65IN	2	TAATCAGGGACACAAACACTGC	TCAAGAAGTCCCAAAGGAATGT	60	-	423
H2_F_65OUT	2	CATTTCCCTATCCTTCCTACCC	TCAAGAAGTCCCAAAGGAATGT	60	429	2142
H2_F_76IN	2	TGTGTCCACTCAGGGTTAAATG	CCACGTATGCGTTTCTACTCAA	60	-	342
H2_F_76OUT	2	GTCTCCAAGTCTATGGGGAGTG	CCACGTATGCGTTTCTACTCAA	60	150	3152
H2_F_86IN	2	TGGGATCTTGTTGATCTGTGAC	ATGCCTGTTTCCAGGTTTATGT	60	-	438
H2_F_86OUT	2	GGGAATGTCACCCAAAATAAGA	ATGCCTGTTTCCAGGTTTATGT	60	155	1957
H2_F_89IN	2	AGTACCCAGGGACACAAACACT	ATGGAGATTCTGTTCTGGGAGA	60	-	397
H2_F_89OUT	2	CTGTACCCTCTTACCCTTCT	ATGGAGATTCTGTTCTGGGAGA	60	451	2044
H2_F_93IN	2	TAATCAGGGACACAAACACTGC	ACCATCACCCATGGAActCTAA	60	-	303
H2_F_93OUT	2	TCTCATTTGAGAAGAGCATGGA	CCATCACCCATGGAActCTAA	60	147	2477
H2_F_117IN	2	TAATCAGGGACACAAACACTGC	ATAGAATGGCCAGTTTCTGAG	60	-	363
H2_F_117OUT	2	GGTAAGGTCTTAGTGGGTCGTG	ATAGAATGGCCAGTTTCTGAG	60	435	1167
H2_F_129IN	2	TGTGTCCACTCAGGGTTAAATG	TCATGTTTGTGGATGGGATaAA	60	-	854
H2_F_129OUT	2	TTGCTGACTGCTATCAGGGTTA	TCATGTTTGTGGATGGGATaAA	60	687	2865
H2_F_132IN	2	TAATCAGGGACACAAACACTGC	CAGTACTGATCCTGTGCTTGG	60	-	306
H2_F_132OUT	2	CTCCTGGGACTATGTGTTAGGC	CAGTACTGATCCTGTGCTTGG	60	229	2039
H2_F_134IN	2	TAATCAGGGACACAAACACTGC	AATGCTCAATTCAATGGGTTCT	60	-	337

(Table A.2 cont.)

H2_F_134OUT	2	TGTAGACTGTTTCCCCATACCC	AATGCTCAATTCAATGGGTTCT	60	337	2031
H2_F_136IN	2	TGTGTCCACTCAGGGTTAAATG	GAAGCTTTCTTGTAGATTCCTTGG	60	-	386
H2_F_136OUT	2	TCCTCCCTAAAGTTTGAACAG	GAAGCTTTCTTGTAGATTCCTTGG	60	170	2981
H2_F_164IN	2	TGTGTCCACTCAGGGTTAAATG	GCAACCAACAGGTGAAAATGTA	60	-	452
H2_F_164OUT	2	AGGGAGTCACAAAAGGGTGATTA	GCAACCAACAGGTGAAAATGTA	60	386	1821
H2_F_214IN	2	CAAGTAACCAGGGACACAAAACA	TGTCAGACAGTTCTGCCAAGTT	60	-	480
H2_F_214OUT	2	TTCTGGCCTTCTGATTTGAGTT	TGTCAGACAGTTCTGCCAAGTT	60	480	2555
H2_F_237IN	2	TGCTCCTTAAGAGTCATCACCA	GTCCTAAGAAATGGCCATCAAAA	60	-	773
H2_F_237OUT	2	ATATGCCCCCTTTGGATAATTT	GTCCTAAGAAATGGCCATCAAAA	60	564	3184
H5_F_9IN	5	CTCTGTGAGAAACACCCAAGAA	TCATCTGGGAGTCCAAATACCT	55	-	473
H5_F_9OUT	5	TGTCTGCCCTACCGAGTAATTT	GCTCTATTCAGCTGCATTTCT	55	620	2696
H5_F_90IN	5	TAATCAGGGACACAAAACACTGC	GTGGTAGGCAGGAAACACATTT	55	-	406
H5_F_90OUT	5	AGCTGTTTCCATTTGACCATCT	CACATCAAAAGaTAAACCCCAAAA	55	360	2763
H5_F_158IN	5	TAATCAGGGACACAAAACACTGC	GTGCACTAGGAAATGGTTGGTT	55	-	418
H5_F_158OUT	5	TATCACCTCACATGCAAACCTCC	GTGCACTAGGAAATGGTTGGTT	55	320	2870
H5_F_161IN	5	TGACCTTCCCTCCACTATTGTC	CACATGGCCAATGCACTTATAG	55	-	499
H5_F_161OUT	5	AGCATCAATGGGACAACCTAAA	TGGGACTTAGACAAAAAGCTCA	55	790	2681
H5_F_163IN	5	ATTGTGGTTTCGATTTGCATTT	TGATCTCATTCTGACCCTACC	55	-	431
H5_F_163OUT	5	ACTTGGGAGTTTGCTTTTATGC	TGATCTCATTCTGACCCTACC	55	230	3064
H5_F_171IN	5	TAATCAGGGACACAAAACACTGC	GAGGACATTTCTGTGATGACCA	55	-	597
H5_F_171OUT	5	CATGAAAGCATGGCAGAAATAA	TAAC TGTTTTTGGCATCACTG	55	950	3553
H6_F_57IN	6	GGATCCTGTTGATCTGTGACCT	GCAACAATTTTACTTCTTTGCTAGT	55	-	396
H6_F_57OUT	6	AGGTGGTAGCATTGGTGAAATC	GCAACAATTTTACTTCTTTGCTAGT	55	650	3065
H6_F_70IN	6	TAATCAGGGACACAAAACACTGC	TGTGATTGGCATCAGAAGTAGG	55	-	403
H6_F_70OUT	6	ATGCCAGGAAAACACTGTAATTC	TGTGATTGGCATCAGAAGTAGG	55	560	2587
H6_F_74IN	6	AGGAAAACGAGACCTTTGTTC	CAGGGGTACCTTGAGATACTG	55	-	484
H6_F_74OUT	6	AGCTCACATGCTACAGGGAAAT	GGCCGACTGTATTTTATTCTGC	55	430	2295
H6_F_119IN	6	CTCAGGGTTAAACGGATTAAGG	AATCACAAATCGTGTCTCCTCCT	55	-	456

(Table A.2 cont.)

H6_F_119OUT	6	GCCTCTAAACTGATCCAGGAAA	AATCACAATCGTGTCTCCTCCT	55	530	2361
H6_F_124IN	6	TAATCAGGGACACAAACACTGC	AGCCTCCAAGAGCTTTTCATTT	55	-	250
H6_F_124OUT	6	CCATGAACAACACCCATAAGAT	AGCCTCCAAGAGCTTTTCATTT	55	330	2913
H6_F_172IN	6	TAATCAGGGACACAAACACTGC	TTGTGCCAGATATCCACACATT	55	-	442
H6_F_172OUT	6	ATGGCAAACCTGCTCTCCTAT	TTGTGCCAGATATCCACACATT	55	390	1189
H7_F_10IN	7	TGTGTCCACTCAGGGTTAAATG	TGCTGCTaGtAGAAGAGGGTGA	55	-	441
H7_F_10OUT	7	AATGAATCCCCAGAAATCTGAA	TGCTGCTaGtAGAAGAGGGTGA	55	490	2751
H7_F_22IN	7	AACCCTGTGCTCTCTGAAACAT	GCATTGTGCTTCTGTATAGCC	55	-	403
H7_F_22OUT	7	ACTGTCATGGAACCTGTCTTGA	GCATTGTGCTTCTGTATAGCC	55	290	2125
H7_F_35IN	7	GTGTCCACTCAGGGTTAAATGG	GAGCCCATCTGAACGATAAAAAG	55	-	527
H7_F_35OUT	7	TTTAGGCCCAAATATGCAAAA	ATCAATgTCTTCATCTGCTTGG	55	760	3164
H7_F_43IN	7	AACCCTGTGCTCTCTGAAACAT	CAACCATGTAAGAATGGACCTG	55	-	724
H7_F_43OUT	7	AAAGTCGTGTGGACTCCTCATT	CAACCATGTAAGAATGGACCTG	55	570	1372
H7_F_61IN	7	TAATCAGGGACACAAACACTGC	TGAGTCTGTTTCTGAAGTGA	55	-	342
H7_F_61OUT	7	GCTAGTGGGCTTTCTCAACTA	TGAGTCTGTTTCTGAAGTGA	55	380	2700
H7_F_95IN	7	ACCTTCCCTCCACTATTGTCCT	TCCTGTCCATTGACACAGTC	55	-	1050
H7_F_95OUT	7	CCTCCTTGATTGAAGGTTGGA	ATACCAGCCTTCTGTGATGTT	55	1020	3426
H7_F_148IN	7	TAATCAGGGACACAAACACTGC	TCTGGTTGCTTGACAGAGATAA	55	-	771
H7_F_148OUT	7	CATGCTTCAAGAGAACATCAGG	TCTGGTTGCTTGACAGAGATAA	55	630	1401
H8_F_62IN	8	TAATCAGGGACACAAACACTGC	AGTTCCCAAAGAAATTCTGCAA	55	-	418
H8_F_62OUT	8	TGACTGTTACCATTTCTCTGG	AAATTTCTCCTCCACTTgACCA	55	440	2783
H8_F_65IN	8	TAATCAGGGACACAAACACTGC	CCCTCATCCCTAGGTTACCACT	55	-	448
H8_F_65OUT	8	CAATGACAAGTTTTACCCAGGA	GATTCCTTCCACATTTGGCTAC	55	240	2258
H8_F_89IN	8	TGTGTCCACTCAGGGTTAAATG	GGTCAAAGGCAAAGATTACACC	55	-	486
H8_F_89OUT	8	CAACGACAAAATCACCTAACGA	GGTCAAAGGCAAAGATTACACC	55	670	2944
H8_F_116IN	8	GCTCGTTAAGAGTCATCACCAA	CCACTTACTTAGGGTGGCAGAG	55	-	482
H8_F_116OUT	8	GAACTAGAAGATGCATCCCCTC	CACAGCCTACTTCTCAAGCAGA	55	200	2764

Table A.3: SVA-mediated transduction events.

Locus Name	Locus Position	Length	TargetSiteDuplication	SVA Sub-family	Poly(A) Signal	Source Locus
H1_12	Chr1:11072683-11078215	154	AAGATTAATTCTA	C	AATAAAA	Chr6:107203895-107204048
H1_26	Chr1:13344563-13350269	75	AAGATATTTTAA	D	AATAAAA	Chr9:19122832-19122907
H1_E1	Chr1:23924000-23926310	218	CAAAAAACTC	D	AATAAAA	chr2: 85659919-85660140
H1_71	Chr1:34047471-34053166	888	AAAAGGCAGTGGTC	B	AATAAAA	Chr3:110299525-110300418
H1_133	Chr1:91960730-91966266	185	GAAAATAGTTTC	B	ATTAAA	Chr6:88709515-88709693
H1_167	Chr1:114643178-114648726	290	AAAAATGGCCATGCATT	C	ATTAAA	Chr2: 80683450-80683738
H1_177	Chr1:144037937-144043412	200	AAGAAATTCCTTT	B	AATAAAA	Chr16:20258896-20259001
H1_182	Chr1:146493810-146498800	35	AATAATGA	E	AATAAAA	chr3: 184810276-184810313
H1_220	Chr1:165361782-165367271	126	AATAACCTACAATG	F	AATAAAA	Chr5:64049246- 64049371
H1_225	Chr1:170819088-170824816	1275	AAAAAAGACAATGAC	C	ATTAAA	Chr7:77055695-77056973
H1_238	Chr1:179216698-179222278	63	AGAAGAATGAACTGG	C	AATAAAA	Chr8:41355248-41355312
H1_275	Chr1:209952047-209957618	55	AAAAAACCTAGCC	B	AATAAAA	Chr7:51270016-51270078
H1_292	Chr1:222955922-222961212	171	CTAAATG	B	ATTAAA	Chr15:48628124-48628291
H1_297	Chr1:227026280-227031574	90	AATTACTC	A	ATTAAA	Chr1:152189751-152189840
H1_311	Chr1:243470451-243475964	268	AAAAACCTGAATT	C	AATAAAA	Cannot be located
H2_5	Chr2:3225256-3230915	546	AAGAATGTCCAG	D	AATAAAA	Chr15:57437220-57437771
H2_57	Chr2:42194679-42199868	52	AATACTGTCCTTGGC	D	AATAAAA	chr5:77478208-77478255
H2_68	Chr2:44266043-44272053	60	AAAAAACAAAACCTG	D	AATAAAA	Chr1:24080384-24080452
H2_88	Chr2:62829327-62834890	356	AAGAAAATG	C	AATAAAA	Chr8:71064456-71064812
H2_103	Chr2:74994781-75000744	355	AAAAGAAGACATTTAT	D	AATAAAA	Chr9:13260233-13260501
H2_171	Chr2:161505334-161510901	694	CATTAA	B	AATAAAA	Chr11:116107770-116108489
H2_183	Chr2:171206117-171211975	271	AGAAATGTA	D	AATAAAA	chr7:138616963-138620752
H2_195	Chr2:183451095-183456704	282	AAAGAAAGAAAAAAGAAAGA	D	-	Cannot be located
H2_199	Chr2:187537495-187543478	47	AAAAAGTAGACAAAGG	D	AATAAAA	Cannot be located
H2_204	Chr2:191495474-191501019	342	AAAGAACAAGA	D	AATAAAA	Cannot be located
H3_3	Chr3:5258054-5261221	404	GAAAATGACCATAGTC	B	AATAAAA	Chr1:203445306-203445687
H3_25	Chr3:20562739-20566286	223	AAGAGCATAGGAATA	B	-	Chr3:109158020-109158237
H3_49	Chr3:47309857-47313060	546	AAGAACAGTTAA	C	AATAAAA	Chr17:39303775-39304321
H3_65	Chr3:57328376-57331772	219	AAAAAATTG	A	-	Cannot be located
H3_71	Chr3:60868969-60872340	254	AAAATCTAG	B	AATAAAA	Cannot be located
H3_85	Chr3:82552887-82556759	571	AAGAAAGAAGTATTTTG	D	AATAAAA	Chr2:38522750-38523319
H3_88	Chr3:95082344-95084344	1156	AAGAATCCTTGAATA	D	AATAAAA	Chr15:68900010-68901191
H3_112	Chr3:120280678-120284294	110	AAGAATCACTCAGCTTTTC	B	AATAAAA	Chr9:94622971-94623507
H3_121	Chr3:123818278-123822338	1276	AAGAAAACATATTCAAG	D	ATTAAA	chr11:82522905-82524171
H3_E1	Chr3:130562098-130566782	55	TAAATTTGACTGTCCTGCT	D	AATAAAA	chr6:157524769-157525308

(Table A.3 cont.)

H3_160	Chr3:151857099-151860070	215	AAAAAAAATACCAAACCTG	E	AATAAA	Chr4:39841974-39842190
H3_164	Chr3:165821594-165825014	131	AAATGCTTCTTG	B	ATTAAA	Chr19:35077442-35077593
H3_E2	Chr3:188062000-188070000	1853	GAAAAATATTGCAAATAG	E	-	chr7:64165000-64173000
H3_186	Chr3:191510367-191513912	938	AAAAAGCAACAAAGTA	D	AATAAA	chr2:85625536-85628105
H4_E1	Chr4:3575241-3578880	955	AAGAGTGCCTCTGG	D	AATAAA	chr2:85625536-85628105
H4_23	Chr4:37406486-37409815	111	AAAAATGCAAC	B	AATAAA	Chr7:139337393-139337496
H4_40	Chr4:53244851-53248159	480	AAAAGCCAAAATTG	D	AATAAA	Chr2:38924964-38925428
H4_79	Chr4:99708467-99711905	60	AAAAGTATTTAC	B	AATAAA	Cannot be located
H4_97	Chr4:124190129-124194178	81	AAAAGAAGTGCCA	D	AATAAA	Chr20:21052132-21052212
H4_127	Chr4:160379188-160382794	54	AAATGAAGCCAAAGA	C	AATAAA	Chr3:180495477-180495530
H4_135	Chr4:166517000-166520716	381	AAAATGAACCTCTA	D	AATAAA	Chr15_random:286803-287188
H5_10	Chr5:34348618-34353881	280	TGAAAAAAAAA	F	AATAAA	Chr8:37709894-37710171
H5_12	Chr5:37056065-37061859	288	GAAATTTAAAGATTAAG	D	-	Chr15:32520492-32521791
H5_27	Chr5:43131689-43136768	191	AAGAGATAACAAG	E	AATAAA	Chr6:70514549-70514744
H5_34	Chr5:45933546-45939230	430	AAATACACATGCTT	C	AATAAA	Chr1:208414788-208415222
H5_37	Chr5:53690457-53695993	353	GAAAAATTTT	D	-	Chr5:174548887-174549239
H5_47	Chr5:59740254-59745831	654	AAAAAAAACATAATC	D	AATAAA	Chr17: 59739365-59740046
H5_49	Chr5:61822320-61828351	803	AAGAATATCCTTGGGGATG	D	ATTAAA	Chr13:97480307-97480398
H5_50	Chr5:62392931-62398379	580	AAGAATGCAAGC	D	AATAAA	Chr4:78730367-78730945
H5_84	Chr5:79678438-79684240	190	AGAAGGTGTGGTG	D	AATAAA	Chr13:45244291-45244483
H5_108	Chr5:111770194-111775451	105	AATAATAAACTGAC	B	AATAAA	Chr7:65763902-65765531
H5_113	Chr5:117691892-117697488	1120	AAGAAATAAAACATGC	C	AATAAA	Chr10:23694301-23695466
H5_142	Chr5:149075606-149081514	950	AATAACAGGTGGG	D	AATAAA	chr2:85625536-85628105
H6_4	Chr6:13239485-13245030	1364	GAAAAATCTATTTGG	B	-	Chr15:91533114-91534496
H6_12	Chr6:21902346-21907323	510	AGAAAAATCTACAC	D	ATTAAA	Chr9:19122832-19122907
H6_62	Chr6:36953618-36959120	995	GAAAAAAACACCCTAGA	C	AATAAA	chr5:151836834-151841126
H6_88	Chr6:57219875-57225701	31	AAACAGTAGAAAAAG	A	AATAAA	Cannot be located
H6_106	Chr6:74553753-74559319	233	GAAAATACATCATTT	D	AATAAA	Cannot be located
H6_109	Chr6:74957835-74963309	488	AAATATCTATAGAT	C	AATAAA	Chr1:45671251-45671658
H6_117	Chr6:86541229-86547014	49	AAGAAATACCTACATCA	D	AATAAA	Chr17:57403012-57403741
H6_128	Chr6:106344556-106350393	827	AAGAAATAGGGTTTGG	D	AATAAA	Chr9:36907182-36908013
H6_132	Chr6:108634010-108639720	1263	AAAAATC	D	AATAAA	Chr20:18307880-18309168
H7_8	Chr7:1943077-1948579	95	TAATGGA	B	AATAAA	Chr10:98911789-98911852
H7_31	Chr7:23415795-23421089	108	AAGACTGTCCCTG	E	AATAAA	Chr2:208186339-208186453
H7_40	Chr7:27766522-27772100	711	GTTCATATGG	B	ATTAAA	Chr9:120582327-120583037
H7_48	Chr7:35975193-35980666	487	AGAAATCTTTCAGCTC	D	AATAAA	Chr4:117160949-117161579
H7_131	Chr7:111749636-111755222	487	GAAACTGCACCATT	D	-	Chr8:1209659810-120966424
H8_15	Chr8:28757716-28762851	28	AGAAAAATGTAGACATA	E	ATTAAA	Chr19:20262373-20262400
H8_53	Chr8:57231903-57237542	159	AACAATTTGGG	D	ATTAAA	Chr13:41214642-41214798
H8_56	Chr8:59636341-59641814	105	TAAACTGCACAAAAG	D	ATTAAA	Chr4:113957645-113957746

(Table A.3 cont.)

H8_66	Chr8:66576335-66581826	214	AAAAACTGTAACAGTA	C	AATAAA	Chr4:189147677-189147890
H8_121	Chr8:142180678-142186349	81	AACAGGAAGGGGG	D	AATAAA	Chr2:232560962-232561059
H9_12	Chr9:15429389-15435216	119	AAAGACGTTA	B	ATTAAA	Cannot be located
H9_22	Chr9:29176016-29182063	213	AAAAATTGAAATGTA	F	-	Multiple hits
H10_54	Chr10:51457637-51462948	123	ATAAAAAATGA	B	AATAAA	Chr6:113948949-113949071
H10_E1	Chr10:95719506-95721700	190	AAAAATGGCAAAGT	D	AATAAA	chr13:45244291-45244483
H10_110	Chr10:99263314-99268809	126	AGAAAGCCTTAATA	D	AATAAA	Chr4:79922959-79923097
H10_111	Chr10:99574252-99579972	168	AAGAAATTCCAG	D	AATAAA	Cannot be located
H11_12	Chr11:8961191-8967424	174	AAAAACATCAATACGT	E	AATAAA	Chr7:55948692-55948864
H11_30	Chr11:41286985-41292377	78	AGAAAAATCACTGCATTA	F	-	Cannot be located
H11_59	Chr11:60031723-60037248	52	AAAATGCAGTATG	C	AATAAA	Chr6:63709499-63709555
H11_72	Chr11:63712698-63718315	35	AAAAAAAAA	F	AATAAA	Chr6:33944792-33944830
H11_95	Chr11:73370963-73377067	92	AAAATCTCATCTC	D	ATTAAA	Chr13:97480307-97480398
H11_108	Chr11:77368542-77373765	193	TAAAAGTG	B	AATAAA	Chr13:42089242-42089434
H11_147	Chr11:118913375-118919266	169	AAATCCTCAAG	F	AATAAA	Cannot be located
H12_16	Chr12:12455107-12460407	180	AAGAAAACCAAAAAATC	E	AATAAA	Cannot be located
H12_83	Chr12:55640962-55646367	869	GAAAATGGCCA	D	AATAAA	Chr7:127798873-127798988
H12_93	Chr12:64852465-64857782	255	AAAACATCATTG	C	AATAAA	Cannot be located
H12_97	Chr12:68437039-68442483	210	AAAGGAATGAACACACAGA	C	AATAAA	Chr2:60842597-60842805
H12_107	Chr12:74619053-74624984	218	TAAATTGAAGAAGTG	D	AATAAA	Chr2:85659919-85660140
H12_139	Chr12:109799626-109803710	726	AAAGAAAATTTTTGT	D	AATAAA	Chr6:121778220-121778955
H12_157	Chr12:121276675-121282528	277	AAAGAAATGAATTGG	D	AATAAA	Chr3:101328106-101328389
H12_160	Chr12:122936805-122942646	45	AAAGAAATAGTCAAT	D	-	Chr1:25935624-25935675
H13_29	Chr13:39324557-39330730	910	GAAAGTGTGTAG	D	AATAAA	chr2:85625536-85628105
H13_47	Chr13:49447257-49453060	159	AAATGAGATTCTGGGTTGA	E	-	Chr2:228557924-228558081
H13_64	Chr13:94275916-94281454	137	AAATGTAT	B	ATTAAA	ChrX:64611735-64611874
H14_66	Chr14:63350597-63355414	423	GAAAATTCCT	E	AATAAA	Chr12:117479181-117479602
H14_74	Chr14:69831319-69836783	202	AAGAAGATTA	B	-	Chr13:36588352-36588508
H14_94	Chr14:84283545-84289072	667	AACAGCTCAGA	D	AATAAA	Cannot be located
H15_E1	Chr15:26892141-26896040	47	AGTAAGTGA	D	AATAAA	chr6:157524769-157525308
H15_44	Chr15:54115274-54121928	687	AAGAACTACAAAACCT	D	AATAAA	Chr12:45927804-45928495
H15_67	Chr15:72734404-72739991	58	ATCTTTAGCAG	D	AATAAA	Chr16:48605488-48605545
H15_81	Chr15:82207766-82213760	59	AAAAATATTTTTTCAG	D	AATAAA	Chr19:11600774-11600831
H16_54	Chr16:46993352-46999197	177	GAAAATATGGTATTATTG	D	AATAAA	Chr6:96087742-96087929
H16_57	Chr16:48226052-48231417	57	AGAAGGCCCT	B	AATAAA	Chr6:157695447-157695503
H17_4	Chr17:1510528-1515946	55	TAAAAAAAAAAAAAAAAAAG	D	AATAAA	Chr6:157524769-157525308
H17_E1	Chr17:16152950-16157899	160	AAAGAAAAAACAATA	D	-	chr8:48331385-48331632
H17_76	Chr17:33665047-33670581	1666	AAGTACCTGA	C	-	Chr17:7325026-7328043
H18_4	Chr18:6814686-6820405	191	AACTAAAGAG	D	ATTAAA	Chr2:122302608-122302805
H18_14	Chr18:11947018-11952697	79	AAACAAAACAAA	D	AATAAA	Chr7:20516061-20516138

(Table A.3 cont.)

H18_33	Chr18:28977316-28982431	189	AGAAATCAGCTACCAGA	C	AATAAA	Chr5:75862647-75862842
H18_35	Chr18:43603811-43609310	432	AAAAAAAGCACACA	B	-	Chr3:134725327-134725908
H18_55	Chr18:72481830-72487404	861	AAAAAGCAGAGTCTGGC	D	AATAAA	chr15:32520492-32521791
H19_4	Chr19:6303799-6309513	218	AAAAAACTTTAAAAACA	D	AATAAA	chr2: 85659919-85660140
H19_13	Chr19:11646074-11651703	50	CTGATGCTGA	B	-	Chr15:70755113-70755162
H19_23	Chr19:20107110-20112434	209	AAGAAAAAGCTCAG	D	AATAAA	Chr19:12181220-12181427
H19_54	Chr19:21838664-21843485	439	AAAAACTA	F	AATAAA	Chr11:105262106-105262551
H19_73	Chr19:23505479-23511030	145	AAGAGTACTCAGGTG	C	-	Chr1:39567915-39568024
H19_130	Chr19:58003981-58009847	462	AAAAAAGA	D	ATTAAA	Chr14:65277341-65277807
H20_8	Chr20:4005511-4011746	1687	AAAAATCGTAAATGAG	E	ATTAAA	chr7:64165000-64173000
H20_15	Chr20:10367960-10373476	42	AAGAACCAATGAAATGG	C	AATAAA	Chr5:139966458-139967147
H20_34	Chr20:29418677-29424377	77	AAAAAAAAGAAG	D	-	Chr11:75679877-75679955
H20_35	Chr20:29531868-29538199	193	AAAGTATGTT	D	AATAAA	ChrX:29892360-29892499
H20_49	Chr20:32837153-32843015	827	GCTAAAAAGA	D	AATAAA	chr9:36907182-36908013
H20_50	Chr20:32999960-33006196	79	AAAAATATTTTTT	D	ATTAAA	Cannot be located
H20_94	Chr20:60911415-60916916	59	TCTTGATTTT	D	AATAAA	Chr7:23373870-23373922
H21_17	Chr21:39447115-39452520	177	AAAAGAAAATTTTAG	D	-	Chr8:121214904-121215086
H21_19	Chr21:44095698-44101133	192	GATGGGGT	B	AATAAA	Chr1: 67406283-67406479
H22_27	Chr22:29349549-29355067	266	ATAAATTTTATT	D	ATTAAA	Chr2:220001533-220001800
H22_36	Chr22:33343826-33349519	221	AAGAAATTCCTCTCT	D	ATTAAA	Chr4:40374497-40374687
HX_E1	ChrX:41900446-41900529	60	AAAAAGCAGT	D	AATAAA	Chr1:24080384-24080452
HX_49	ChrX:51820024-51825615	319	GAAATATATTTTGTG	D	AATAAA	Multiple hits
HX_75	ChrX:67505999-67511635	1248	GAAACTACTAGTCTA	B	AATAAA	ChrX:134898910-134900153
HX_87	ChrX:69759789-69766198	444	AAAACCTGTGA	A	AATAAA	Chr13:19286787-19287231
HX_90	ChrX:70643679-70649223	364	AAGAATATACCTA	B	ATTAAA	Chr2:191900881-191901253
HX_100	ChrX:72496594-72502004	343	AAGAAGTCACAGTC	D	-	Chr8:48331385-48331632
HX_153	ChrX:143182789-143188347	1483	AGAAATGCAGTTTAA	D	AATAAA	Chr19:54502025-54503544
HX_160	ChrX:152871790-152877218	58	AAAAAGGCTAG	B	AATAAA	ChrX:40908434-40908501

Table A.4: Loci names, primers, annealing temperatures and amplicon sizes of the source loci without SVA elements.

Locus	Primer Name	Chr.	5' Primer Sequence (5'-3')	3' Primer Sequence (5'-3')	Annealing Temp	Product Size Without SVA
H1_12	IN_H1_12	6	AGTACCCAGGGACACAAACT	CAGCACTCAACTAAAGCCATTG	60	458
	OUT_H1_12	6	ATCTTGCTTCACTTCGTCATCA	CAGCACTCAACTAAAGCCATTG	60	581
H1_26	IN_H1_26	9	CGCAGGGTCTCTGCCTA	TGGTTAACCTGATAATCCACGA	60	492
	OUT_H1_26	9	TCATACATTCTTTTCTTGAGGTC	TGGTTAACCTGATAATCCACGA	60	394

(Table A.4 cont.)

H1_133	IN_H1_133	6	AGTACCCAGGGACACAAAACAC	CCCAGCATTATTTTCCTCAAC	60	595
	OUT_H1_133	6	TTAATGGATCCTAATGGGCAAT	ACTTTCTGCGTCAGCAAAGATT	60	577
H1_167	IN_H1_167	2	AGTACCCAGGGACACAAAACACT	AGCCTCTATGGATCACGGATTA	60	503
	OUT_H1_167	2	GGAGGATTGGATCACCAGATTA	AGCCTCTATGGATCACGGATTA	60	434
H1_220	IN_H1_220	5	AGTGCTCTCTGAAACATGTGCT	TATTTTCCTCATGGAAGCATGG	60	499
	OUT_H1_220	5	GATTTGTGTTGCATCCATTGTT	CCTTTGCTCATTITGGACTCAT	60	598
H1_275	IN_H1_275	7	TCATCACCACCTCCCTAATCTCA	AAGCAGCTTCAACATGGAAAAT	60	480
	OUT_H1_275	7	GAATGTTGCCATTTCTGAAACA	TGAAGAAGTTCTGGTGGACTTG	60	442
H2_68	IN_H2_68	1	AGTACCCAGGGACACAAAACACT	ACTCCATGGTGTGACACTCGTA	60	525
	OUT_H2_68	1	GACTACTTCCCCAAGGGACTC	ACTCCATGGTGTGACACTCGTA	60	447
H2_103	IN_H2_103	9	AGTACCCAGGGACACAAAACACT	TCTGGCACTCAAAAGACTAAACA	60	587
	OUT_H2_103	9	GATCATCAAACACCCCAAGAGT	ATAATCCAGAAAGGACATCTGC	60	597
H5_12	IN_H5_12	15	AGTACCCAGGGACACAAAACACT	ACTGGAATAGGGGAAGTGATGA	60	353
	OUT_H5_12	15	AGAACCGAGTGAAGGTAGCATT	ACTGGAATAGGGGAAGTGATGA	60	547
H5_27	IN_H5_27	6	AGTACCCAGGGACACAAAACACT	TCCCTTGTATGCAAACCTGAAT	60	531
	OUT_H5_27	6	CTGAATATGCCACCTTCAAAAA	TCCCTTGTATGCAAACCTGAAT	60	423
H5_49	IN_H5_49	8	AGTACCCAGGGACACAAAACACT	GGAATTCTTAAAAGCTGCAACAA	60	932
	OUT_H5_49	8	AACATTCTTTTTCCGGGAGGT	GGAATTCTTAAAAGCTGCAACAA	60	837
H5_50	IN_H5_50	4	AGTACCCAGGGACACAAAACACT	TAACATGGGGGAAGGACATAGT	60	898
	OUT_H5_50	4	CTCTGTGCTGGAACCTCTGAATG	TAACATGGGGGAAGGACATAGT	60	848
H5_84	IN_H5_84	13	CGCAGGGTCCTCTGCCTA	TTGGAATATGATACTATGTGGCGAGA	55	392
	OUT_H5_84	13	TGTCCGCACTCTTAGAACAAAA	TGGAATATGATACTATGTGGCGAGA	60	567
H6_109	IN_H6_109	1	ATGCTCTCTGAAACATGTGCTG	CAAGCTAAGGAAGGAGGTTGAA	55	872
	OUT_H6_109	1	TGATGATGTCGGGACTAGAAAAG	TCAAGCTAAGGAAGGAGGTTGA	60	699
H7_31	IN_H7_31	2	AGTACCCAGGGACACAAAACACT	CACGTGTCACCTCTTTGTAGC	60	884
	OUT_H7_31	2	TGAGATTTCTTTTGTGGCTTGTT	GGTTTTATTAACCAATTCTATGAAGG	60	552
H7_48	IN_H7_48	4	AGTACCCAGGGACACAAAACACT	GTGATTCCAGCACGCTTTAGAT	55	789
	OUT_H7_48	4	AGAAAAAGATTAAGGAAGATAATTGC	TTGGAATATCTGACGCTTTAACA	60	974
H7_131	IN_H7_131	8	AGTACCCAGGGACACAAAACACT	TTGCTGATTTAAGGACAACCAA	60	696
	OUT_H7_131	8	ATTCAGTGAAAAGAGGGTCCAA	TTGCTGATTTAAGGACAACCAA	60	664
H10_110	IN_H10_110	4	AGTACCCAGGGACACAAAACACT	GCGGAAAGTGTAAGGAAGTTTG	60	383
	OUT_H10_110	4	TGTTGCAAAAATAGTGCATTTT	GCGGAAAGTGTAAGGAAGTTTG	60	405
H11_12	IN_H11_12	7	AGTACCCAGGGACACAAAACACT	TGCCTGAATTTAECTTTTGCTG	60	436
	OUT_H11_12	7	GAAGTTCATTGGAAGAGCAGGA	ATGGTTACTGTCCCAATTCTGG	60	541
H11_72	IN_H11_72	6	TGTGTCCACTCAGGGTTAAATG	TGTGGATTCACTCACAGAAAGG	60	358

(Table A.4 cont.)

	OUT_H11_72	6	TTCAGGAAAAATCCCACAAAG	TGTGGATTCACACAGAAAGG	60	394
H11_95	IN_H11_95	13	AGTACCCAGGGACACAAACT	GAAAATATGCTCCCAATGAGA	60	678
	OUT_H11_95	13	CTTGACTAAATTATCTGAAAAGGTTT	GAAAATATGCTCCCAATGAGA	60	524
H11_108	IN_H11_108	13	AGTACCCAGGGACACAAACT	TAAATAGCTGGCTCCTCACTCC	60	533
	OUT_H11_108	13	CTGTGTGCCAATCCTGTTAATG	TAAATAGCTGGCTCCTCACTCC	60	469
H13_64	IN_H13_64	X	AGTACCCAGGGACACAAACT	TTCTGTTCTCCATGACCACTTG	60	898
	OUT_H13_64	X	AGGCATTCTCAGTAACCTCAGC	CTTGCATTCTGGACACTTCACT	60	558
H14_66	IN_H14_66	12	GGATCCTGTTGATCTGTGACCT	AATGACTCCCTTGCACTTTTC	60	861
	OUT_H14_66	12	GCTAACCCATAACTGGTCCTTTC	AATGACTCCCTTGCACTTTTC	60	846
H16_54	IN_H16_54	6	GACACAAACTGCGGAAAGG	TCTTGCCAAAAACCTTAGCAT	60	400
	OUT_H16_54	6	TGGAAGAACAGATACGCCATAA	TCTTGCCAAAAACCTTAGCAT	60	515
H18_14	IN_H18_14	7	AGTACCCAGGGACACAAACT	TACAGTCTGCACCGGACAGTAG	60	380
	OUT_H18_14	7	ATTGCCTATGCTGGAAAACAAT	TACAGTCTGCACCGGACAGTAG	60	432
H19_73	IN_H19_73	1	AGTACCCAGGGACACAAACT	ATGTTGGATAGGCAAAGTGCA	60	523
	OUT_H19_73	1	TCCCTGTGGAGTGAGGATAAGT	ATGTTGGATAGGCAAAGTGCA	60	510
H20_94	IN_H20_94	7	AGTACCCAGGGACACAAACT	ATTGAGGGGAAATTAGAAAGC	60	432
	OUT_H20_94	7	TTTGCATAGAAAACAAACCTTT	ATTGAGGGGAAATTAGAAAGC	60	318
H22_36	IN_H22_36	4	AGTACCCAGGGACACAAACT	CCTTCTCTTCATTTCGCACTA	60	484
	OUT_H22_36	4	GACTTTGAAATTCCAGGACACC	CCTTCTCTTCATTTCGCACTA	60	369
HX_30	IN_HX_30	1	AGTACCCAGGGACACAAACT	TCCATGGTGTGACACTCGTAAG	60	504
	OUT_HX_30	1	GTGCAGCCATCACCATTAAGTA	TCCATGGTGTGACACTCGTAAG	60	369

Table A.5: Primer names, annealing temperatures and amplicon sizes for PCR amplification and sequencing of AMAC gene loci.

Name	Chr	5' Primer Sequence (5'-3')	3' Primer Sequence (5'-3')	Annealing		
				Temp	Empty	Filled
H17_AMAC1_SVA	17	GCAATTAAGATGGGTAACCTCC	CGAGATGGCAGCAGTACAGT	55		467
H17_AMAC1_TD	17	TGTTCCGGCAAAGACAGAG	ATGGATGCAAACTCTTTTCATTTT	55		599
H17_AMAC1_BOTH	17	GCAATTAAGATGGGTAACCTCC	ACTTCAATGGATGCAAACTCTT	55	810	4059
H8_AMAC1L2_SVA	8	ATATTGGTGAAGTTTGGATGG	GCAGCAGTACAGTCCAGCTTT	55		1211
H8_AMAC1L2_TD	8	GTGGTTGTGGCCCTTATACTG	TCTCAATATTATAATGTGCTATGGAAG	55		301
H8_AMAC1L2_TD1	8	ACCAGAGACCTTTGTTCACTTGT	TCTCAATATTATAATGTGCTATGGAAG	50		1465
H8_AMAC1L2_TD2	8	GCACTCCAATGAGGTCACAAT	TCTCAATATTATAATGTGCTATGGAAG	50		1252

(Table A.5 cont.)

H8_AMAC1L2_BOTH	8	ATATTGGTGAAGGTTTGGATGG	TCTCAATATTATAATGTGCTATGGAAG	48	1200	4098
H17_AMAC1L3_TD1	17	CAGACAAACTTTGGAGTCATGG	GGTTCAAGTAGGGGTGACTGC	55		1190
H17_AMAC1L3_TD3	17	CTATAGCTCCTGGTTGCTCCAT	TGCAGAGAACGATAGACCAGAA	50		1071
H17_AMAC1L3_TD4	17	GGACCTGGACTCTGGACACTAC	AATGAGAACGGTCGTAGGACTT	55		1055
H17_AMAC1L3_RHESUS1*	17	TCAGAGCTCCTTGCCTTAAAC	AATGGAGCAACCAGGAGCTAT	55		986
H17_AMAC1L3_RHESUS2*	17	TATAGCTCCTGGTTGCTCCATT	AAGCCAAATAGGAAGGCCACT	55		1074
H17_AMAC1L3_RHESUS3*	17	AACAGTGGCCTTCCTATTTGG	ACGGTCGTAGGACTTCCACTC	50		848

* Used for PCR and sequence AMAC1L3 locus in African green monkey and rhesus monkey genomes.

APPENDIX B:
LETTERS OF PERMISSION



ELSEVIER
7 August 2006

Our Ref: CT/jj/Aug06/J001

Hui Wang
Louisiana State University
107 Life Sciences Building, LSU
Baton Rouge 70803
USA

Dear Hui Wang

JOURNAL OF MOLECULAR BIOLOGY, Vol 354, No 4, 2005, pp 994-1007, Wang et al: "SVA Elements: A Hominid Specific"

As per your letter dated 8 August 2006, we hereby grant you permission to reprint the aforementioned material at no charge **in your thesis** subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.
2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:
"Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier".
3. Reproduction of this material is confined to the purpose for which permission is hereby given.
4. This permission is granted for non-exclusive world **English** rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form. Should you have a specific electronic project in mind please reapply for permission.
5. This includes permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

Yours sincerely

Jennifer Jones
Rights Assistant

Your future requests will be handled more quickly if you complete the online form at www.elsevier.com/permissions

PNAS

From: "Lashomb, Chris" <CLashomb@nas.edu>

To: "Hui Wang" hwang7@lsu.edu

Date: Wed, 6 Sep 2006 12:39:01 -0400

Dear Dr. Wang,

Permission is granted for your use of the figure/article as described in your message below. Please cite the full journal references and "Copyright (Copyright year) National Academy of Sciences, U.S.A."

Best regards,

Chris Lashomb for

Diane Sullenberger

Executive Editor

PNAS

VITA

Hui Wang attended Shanghai Jiao Tong University, China, in September of 1996. There she received her Bachelor of Science degree in biotechnology in 2000 and Master of Science degree in molecular biology and biochemistry in 2003. Ms. Wang will graduate with the degree of Doctor of Philosophy in biological sciences from Louisiana State University in December 2006.