2005

Alu retrotransposition-mediated genomic variation within the primate order

Pauline Ann Callinan
Louisiana State University and Agricultural and Mechanical College

ALU RETROTRANSPOSITION-MEDIATED GENOMIC VARIATION WITHIN THE PRIMATE ORDER

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
Pauline Ann Callinan
B.Sc., Loughborough University, 2000
May 2005

ACKNOWLEDGEMENTS

I would like to express my gratitude to C.B. Davis, my best friend and partner, for his unfailing love and encouragement throughout my doctoral studies. I thank my close friends and family back at home in England, who graciously accepted my career decision to relocate to the USA, yet still continue to reward with me with unconditional love and support.

I thank my mentor, Prof. Mark Batzer, for his support over the last four years, financially, scientifically and as a friend. I extend my gratitude to my committee, Dr. David Donze, Dr. Craig Hart, Dr. Marcia Newcomer, and finally Dr. David Baker, for their advice on the preparation of this dissertation. Lastly, I would like to thank members of the Batzer laboratory for their scientific guidance and friendship.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Retrotransposons are an active family of mobile elements within primate genomes and the Short INterspersed Element (SINE) *Alu* is the most abundant member. These non-autonomous elements are responsible for introducing genomic diversity on an intra- and inter-species level that is useful in studies of forensic identity, population genetics, and evolutionary biology. In a computational survey of the human sex chromosomes, 344 recently integrated *Alu* elements were detected and subjected to empirical testing by polymerase chain reaction to determine presence/absence polymorphism. Sixteen elements were found to be polymorphic on the X chromosome, and only one polymorphic element on the Y chromosome (previously termed YAP, Y chromosome *Alu* Polymorphism), across four geographically diverse populations. In line with previous research using other types of genetic markers, these results indicate a low *Alu*-associated diversity level on the human sex chromosomes, presumably due to reduced recombination rates and lower effective population sizes on the sex chromosomes.

Alu elements often contribute to genomic instability via insertional and recombinational mutagenesis. Recently, a novel mechanism of retrotransposon-associated genomic instability was discovered, termed retrotransposition-mediated deletion. A computational search within the draft human and chimpanzee genomes found evidence of 33 retrotransposition-mediated deletion events that have eliminated approximately 9,000 nucleotides of genomic DNA. During the course of primate evolution, *Alu* retrotransposition may have contributed to over 3000 deletion events, eliminating approximately 900,000 bp of DNA in the process. Potential mechanisms for the creation of *Alu* retrotransposition-mediated deletions include L1 endonuclease-dependent retrotransposition, L1 endonuclease-independent retrotransposition, internal priming on DNA breaks, and promiscuous target primed reverse transcription (pTPRT).

Approximately 0.27% of all human disease mutations are attributable to the activity of Long INterspersed Element (LINE) L1, *Alu* and SVA (SINE-R/VNTR/*Alu*) retrotransposons within our genomes. Although researchers in the field of human genetics have discovered many mutational mechanisms for retrotransposable elements, including retrotranspositional insertion, recombination, retrotransposition-mediated and gene conversion-mediated deletion, in addition to 3' transduction, their individual contribution to genetic variation within humans is still being resolved.

CHAPTER 1:

INTRODUCTION

Mobile Genetic Elements

Mobile genetic elements are interspersed DNA sequences that were first discovered by Barbara McClintock in the 1950s as being responsible for the variegating color patterns in maize kernels (McClintock 1956). Since then, whole genome sequence analysis has shown that mobile genetic elements occupy a vast array of genomes, from bacteria to human (Lander *et al.* 2001). Mobile genetic elements are linear DNA fragments that transfer within the genome from one location to another (McClintock 1956). In order to move, mobile elements may excise themselves entirely from the genome and transfer to another location, or produce duplicates that integrate elsewhere, in a copy and paste fashion. Mobile elements are classed as either DNA transposons or retrotransposons.

DNA transposons possess inverted terminal repeats and encode a transposase protein that they use to self-excise from the genome (Mizuuchi 1992, Smit *et al.* 1996). Although DNA transposons are active in the genomes of bacteria, plants and flies, no known active DNA transposons are present in the human genome (Lander *et al.* 2001). Mobile elements that move via an RNA intermediate are termed retrotransposons. Two general types of retrotransposable element exist depending on whether they have or lack long terminal direct repeats. In mammals, Long Terminal Repeat (LTR) retrotransposons are similar to retroviruses, except for the absence of a functional envelope (*env*) gene, a gene used to transport elements between cells (Ono *et al.* 1987). LTR retrotransposable elements are particularly common in plant and fly genomes, although it is unlikely any functional LTR retrotransposons reside in humans (Ono *et al.* 1987). Non-LTR retrotransposons are represented by Long INterspersed Elements (LINEs), Short INterspersed Elements (SINEs) and SVA (SINE-R/VNTR/*Alu*) elements. All three types of non-LTR retrotransposable elements produce RNA transcripts driven from an internal promoter, and

contain short direct surrounding repeats despite differences in length. LINEs are additionally differentiated from SINEs and SVA elements based on their ability to autonomously mobilize.

The LINE represents the only active non-LTR autonomous element in primate genomes, although their origin extends back to the beginning of eukaryotic evolution (Eickbush 1992). LINEs are approximately 6 kb in length although the majority are 5' truncated, and contain two open reading frames (ORFs) encoding endonuclease and reverse transcriptase (EN/RT) proteins (Feng *et al.* 1996, Jurka 1997). In addition, LINEs possess an RNA polymerase II promoter within their 5' untranslated region (UTR), and terminate in a poly-dA tail coded for by a termination and poly-adenylation signal within its 3' UTR (Moran *et al.* 1996). LINEs mobilize via a mechanism termed TPRT, or Target Primed Reverse Transcription (Luan *et al.* 1993). First discovered of R2 elements in arthropods, TPRT is a mechanism whereby LINE mRNA binds to a free 3'-OH created by staggered LINE EN nicking at 5'-TT/AAAA sites in the genome (Jurka 1997, Luan *et al.* 1993). LINE RT then reverse transcribes LINE mRNA into a cDNA copy, which is integrated at the target site (Luan *et al.* 1995). Short direct repeats flank the newly inserted LINE and are remnants of the integration process.

SINEs are small stretches of DNA (~ 80-300 bases in length) that are ubiquitously dispersed throughout the genomes of eukaryotes. SINEs are thought to be derived from either tRNA or 7 SL RNA genes and contain an RNA polymerase III promoter that drives their duplication. SINEs are non-autonomous non-LTR retrotransposons and purportedly obtain the factors they require to duplicate from their partner LINEs. Recent studies have shown that SINEs are unable to mobilize without LINE-derived proteins (Dewannieux *et al.* 2003) and because they possess similar 3' ends to their partner LINEs, they appear to share the same method of mobilization (Boeke 1997, Kajikawa *et al.* 2002).

SVA elements are the least well-documented retrotransposon residing within the human genome and are young by evolutionary standards, with a time of origin estimated at 15 million years (Ostertag *et al.* 2003). SVA elements are approximately 1500 bp long, and consist of a hexameric repeat region (presumably acting as an internal promoter), an anti-sense *Alu,* a variable number of tandem repeats (VNTR), in addition to a SINE-R element (Ostertag *et al.* 2003). SVA elements, like LINEs and SINEs, possess small direct repeats as a hallmark of their integration process and a poly-dA tail at their 3' end (Ostertag *et al.* 2003). Based on the general sequence similarities between SVA and other non-LTR retrotransposable elements, it is presumed that active SVA elements borrow the necessary retrotransposition proteins from LINEs in order to proliferate (Ostertag *et al.* 2003).

Active Mobile Elements within the Human Genome

The majority of LINEs within mammalian genomes are derived from the L1 family, an element that has successfully amplified to over 500,000 copies and encompasses approximately 20% of the human genome (Lander *et al.* 2001). Retrotranspositionally competent L1 elements, otherwise known as source genes, have created a hierarchy of subfamilies over time that are distinguished by diagnostic mutations. Some L1 source genes have produced copies so recently that the progeny are specific to the human lineage, and some of these new L1 elements are polymorphic across geographically diverse human populations (Myers *et al.* 2002, Salem *et al.* 2003). Approximately 1500 L1 elements are specific to the human genome, 500 of which belong to the youngest, most active subfamily in humans, L1 Ta (transcribed subset a) (Myers *et al.* 2002).

L1 elements are responsible for generating immense genetic change within the human genome. By virtue of L1s active EN/RT proteins, they are indirectly and directly responsible for

all retrotransposable element insertions. On sixteen separate occasions L1 elements have induced genetic diseases by inserting into sensitive regions of the genome (Ostertag *et al.* 2001). But overall, 0.27% (118 retrotransposition-mediated disease events/44,000 human disease mutations) of human genetic diseases are attributed to the activity and presence of L1-driven retrotransposons in our genome (Callinan *et al.* Submitted). In addition, L1 may also be responsible for genomic instability by retrotransposition-mediated deletion, a mechanism first discovered by (Gilbert *et al.* 2002, Symer *et al.* 2002) using *in vitro* retrotransposition assays. Moreover, actively mobilizing L1 Ta elements can shuffle genomic DNA through a process termed 3' transduction, where sequence directly adjacent to the element is transcribed by RNA polymerase read-through to create new chimeric exons or influence the expression of nearby genes (Goodier *et al.* 2000, Moran *et al.* 1999).

Alu elements are the only active SINEs within the human genome. These successful elements have freeloaded wildly on the back of their partner L1, to produce over 1.2 million copies per haploid genome (Lander *et al.* 2001). *Alu* elements were originally derived from the 7 SL RNA gene (Ullu et al. 1984) and arose approximately 65 million years ago, overlapping with the evolution and radiation of the primate order. The *Alu* element is 300 bp long, composed of two arms between which, a middle A-rich tract resides. Due to their evolutionary beginnings as a pseudogene of the 7 SL RNA gene, *Alu* elements contain a split RNA polymerase III promoter that is required for their amplification. However, it has been shown that for efficient transcription, *Alu* elements require additional 5' flanking sequences approximately 37 bases upstream of the transcription initiation site (Ullu et al. 1985). *Alu* elements, like other non-LTR retrotransposons, are flanked by short direct repeats that are a remnant of the retroposition process.

Similar to LINEs, only a small subset of *Alu* elements are thought to be retrotranspositionally competent source genes (Batzer *et al.* 2002, Cordaux *et al.* 2004). Although the criteria required for this function are still not fully resolved, promoter integrity and the length and homogeneity of the poly-dA tail have been suggested as principal factors in determining the retrotranspositional capability of *Alu* elements (Cordaux *et al.* 2004, Roy-Engel *et al.* 2002). It is also possible that some post-transcriptional selection of *Alu* transcripts may be involved in retrotranspositional activity (Sinnett 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.

Human's posses between 6000-9000 lineage-specific elements that are characterized by low sequence divergence, as well as high polymorphism levels (*Alu* insertion presence/absence) across geographically diverse populations. Currently amplifying *Alu* elements within humans derive from the Y (Young) subfamily and include Ya5, Ya5a2, Ya8, Yb8, Yb9, Yc1, and Yc2, in addition to many other small subfamilies. The retrotransposition activity of Y-lineage elements have been documented by their ability to cause human genetic diseases through insertion, such as neurofibromatosis (Wallace et al. 1991), breast cancer (Miki et al. 1996), and severe acute hemophilia A (Ganguly et al. 2003). *Alu* elements also create genetic instability through *Alu*-*Alu* recombination, which is fueled by their vast number and sequence similarity. Recent experimental evidence suggests that elements oriented in a head to head fashion with other closely (<20 bp distance) spaced *Alu*s may contribute to unstable regions in genomic DNA (Lobachev et al. 2000). In fact, computational analysis of the human genome sequence detects a paucity of *Alu* elements in this closely-spaced inverted orientation (Lobachev et al. 2000).

SVA elements are unique mobile elements that are composed from multiple types of repetitive DNA. These elements have amplified to a copy number ranging between 1750 and 3500, and are characterized by low sequence divergence, which suggests that their rarity can be attributed to their young age, rather than their retrotranspositional inactivity. SVA retrotransposons are likely restricted to the genomes of higher primates, however, few details are known about their insertion distribution within primates, including *Homo sapiens*. SVA elements were first discovered following a disease-causing insertion within the α-spectrin gene (SPTA1) (Ostertag *et al.* 2003). Two other cases of disease-causing SVA insertions have also been reported (Ostertag *et al.* 2003). As few data exist on SVA elements, studies are currently investigating their subfamily composition and amplification dynamics within primate genomes.

L1 and *Alu*: Studying Genetic Variation in Primate Genomes

LINE and SINE elements are ideal markers for studies of genetic variation as they are identical by descent (IBD) markers unlike other types of genetic systems that are merely identical by state (IBS). LINE and SINE elements are also unidirectional characters. This means an organism can only gain an element at a locus, and once gained, it is highly unlikely that the element will be cleanly lost. In addition, the chances of two *Alu* elements inserting at the same site in the genomes of related organisms is virtually nil. With thousands of loci already tested, LINE and SINE elements have clearly demonstrated these homoplasy-free characteristics for the study of primate variation and evolution (Callinan *et al.* 2003, Carroll *et al.* 2001, Carter *et al.* 2004, Hedges *et al.* 2004, Ho *et al.* 2005, Myers *et al.* 2002, Salem *et al.* 2003, Vincent *et al.* 2003). As a result, evolutionary studies can determine the true relationships between closely related species with minimal ambiguity (Shedlock *et al.* 2000). The few elements that appear to be homoplasic insertions within primate genomes have turned out to be gene conversion events

(Kass et al. 1995) or insertions at nearby but distinct locations (Roy-Engel et al. 2002). LINE and SINE elements continue to be useful for intra-species studies as they can be used to detect genetic differences between populations, sub-populations and even individuals (Batzer *et al.* 2002). Coupled to easy DNA amplification and gel separation of LINE and SINE elements, the unambiguous insertion presence/absence genotype can reduce turn around times within forensic laboratories, in addition to simplifying data for research scientists.

Over the last few years, the human and chimpanzee genome sequence has enabled researchers to identify lineage-specific SINE and LINE elements, test for their insertion presence/absence and answer questions on primate phylogenetics. The contribution of *Alu* and L1 elements to genetic instability within these genomes is also of considerable interest given their proposed roles in disease causation and genome evolution. This dissertation demonstrates the use of both *Homo sapiens* and *Pan troglodytes* genomic sequences in order to address questions regarding mobile element-derived genomic diversity and genomic instability caused by the presence of mobile elements within primate genomes.

Chapter 2 primarily attempts to analyze *Alu* element distribution and diversity on the human sex chromosomes, and secondarily, to determine whether *Alu* elements generate a similar picture of genetic variation on the sex chromosome compared to other types of genetic markers. Literature suggests that mobile elements should accumulate on the sex chromosomes as a direct result of genetic drift, encouraged by a small effective population size and less frequent recombination on the dimorphic human sex chromosomes (Boissinot et al. 2001). However, (Boissinot *et al.* 2001) determined that this phenomenon was only seen for full length LINE elements (6 kb), with no corresponding bias found for *Alu* elements or truncated LINES (<500bp) (Boissinot et al. 2001). Our analysis indicated that, as a whole, young *Alu* elements do not show

any appreciable insertion bias on the human sex chromosomes, in agreement with (Boissinot *et al.* 2001). Previous studies have noted that diversity on the sex chromosomes is reduced in comparison to the autosomes (Begun *et al.* 2000, Nachman 1997, Yu *et al.* 2001). This phenomenon is expected due to a reduced population size and the partial non-recombining nature of the sex chromosomes (Nachman 1997; Begun et al. 2000). Our results agree with current literature. Only 16 polymorphic *Alu* elements were found on the X chromosome and one on Y, resulting in polymorphism frequencies below half that expected for elements of the same age and subfamily. Thus, it was concluded that *Alu* elements are able to capture similar levels of genomic diversity as other DNA markers. The polymorphic elements found in this study will provide useful sex-linked markers in studies of human population genetics and evolution.

Chapter 3 answers the question: What contribution has *Alu* retrotransposition-mediated deletion made to genomic instability and evolution within primate genomes? *In vitro* studies in 2002 revealed that L1 elements were able to induce target site deletions spanning from 1 bp to 70,000 bp upon integration into the genome (Gilbert *et al.* 2002, Symer *et al.* 2002). Given that both L1 and *Alu* have been shown to share mobilization proteins and 3' sequence characteristics (Boeke 1997, Dewannieux *et al.* 2003, Jurka 1997), it was intuitive that *Alu* elements could also induce deletion upon genomic integration. Using computational approaches supported by wet bench experimentation, it was determined that, *in vivo*, *Alu* retrotransposition-mediated deletion is responsible for 33 deletion events in human and chimpanzee, combined. These events have led to the elimination of approximately 9,000 nucleotides of genomic DNA over the last 5 million years since the human-chimp radiation. Moreover, the data suggest that during the course of primate evolution, *Alu* retrotransposition may have contributed to over 3000 deletion events, deleting approximately 900 kb of DNA in the process. Several potential mechanisms

were identified for the creation of *Alu* retrotransposition-mediated deletions. These include L1 endonuclease-dependent retrotransposition, L1 endonuclease-independent retrotransposition, internal priming on fortuitous DNA breaks, and promiscuous target primed reverse transcription (pTPRT).

As an extension of the study into genetic instability in chapter 3, chapter 4 provides an overview of current literature concerning retrotransposable elements and disease within humans. In this review paper, L1, *Alu* and SVA are estimated to contribute to 0.27% of all currently known human disease mutations. A number of different mechanisms of genome alteration by retrotransposable elements are discussed and include insertional mutagenesis and recombination, in addition to retrotransposition-mediated, gene conversion-mediated deletion and 3' transduction. It was concluded that although researchers in the field of human genetics have discovered many mutational mechanisms of retrotransposable elements, their contribution to genetic variation within humans is still being fully defined.

References

Batzer, M. A. and Deininger, P. L. (2002). "Alu repeats and human genomic diversity." Nat Rev Genet 3 (5): 370-9.

Begun, D. J. and Whitley, P. (2000). "Reduced X-linked nucleotide polymorphism in Drosophila simulans." Proc Natl Acad Sci U S A 97 (11): 5960-5.

Boeke, J. D. (1997). "LINEs and Alus--the polyA connection." Nat Genet 16 (1): 6-7.

Boissinot, S., *et al.* (2001). "Selection against deleterious LINE-1-containing loci in the human lineage." Mol Biol Evol 18 (6): 926-35.

Callinan, P. A. and Batzer, M.A. (Submitted). "Retrotransposable elements and disease." Genome Dynamics

Callinan, P. A., *et al.* (2003). "Comprehensive analysis of Alu-associated diversity on the human sex chromosomes." Gene 317 (1-2): 103-10.

Carroll, M. L., *et al.* (2001). "Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity." J Mol Biol 311 (1): 17-40.

Carter, A. B., *et al.* (2004). "Genome wide analysis of the human Yb lineage." Human Genomics 1 167-168.

Cordaux, R., *et al.* (2004). "Retrotransposition of Alu elements: how many sources?" Trends Genet 20 (10): 464-7.

Dewannieux, M., *et al.* (2003). "LINE-mediated retrotransposition of marked Alu sequences." Nat Genet 35 (1): 41-8.

Eickbush, T. H. (1992). "Transposing without ends: the non-LTR retrotransposable elements." New Biol 4 (5): 430-40.

Feng, Q., *et al.* (1996). "Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition." Cell 87 (5): 905-16.

Ganguly, A., *et al.* (2003). "Exon skipping caused by an intronic insertion of a young Alu Yb9 element leads to severe hemophilia A." Hum Genet 113 (4): 348-52.

Gilbert, N., *et al.* (2002). "Genomic deletions created upon LINE-1 retrotransposition." Cell 110 (3): 315-25.

Goodier, J. L., *et al.* (2000). "Transduction of 3'-flanking sequences is common in L1 retrotransposition." Hum Mol Genet 9 (4): 653-7.

Hedges, D. J., *et al.* (2004). "Differential alu mobilization and polymorphism among the human and chimpanzee lineages." Genome Res 14 (6): 1068-75.

Ho, H. J., *et al.* (2005). "Straightening out the LINEs: LINE-1 orthologous loci." Genomics 85 (2): 201-7.

Jurka, J. (1997). "Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons." Proc Natl Acad Sci U S A 94 (5): 1872-7.

Kajikawa, M. and Okada, N. (2002). "LINEs mobilize SINEs in the eel through a shared 3' sequence." Cell 111 (3): 433-44.

Kass, D. H., *et al.* (1995). "Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution." Mol Cell Biol 15 (1): 19-25.

Lander, E. S., *et al.* (2001). "Initial sequencing and analysis of the human genome." Nature 409 (6822): 860-921.

Lobachev, K. S., *et al.* (2000). "Inverted Alu repeats unstable in yeast are excluded from the human genome." Embo J 19 (14): 3822-30.

Luan, D. D. and Eickbush, T. H. (1995). "RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element." Mol Cell Biol 15 (7): 3882-91.

Luan, D. D., *et al.* (1993). "Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition." Cell 72 (4): 595-605.

McClintock, B. (1956). "Intranuclear systems controlling gene action and mutation." Brookhaven Symp Biol (8): 58-74.

Miki, Y., *et al.* (1996). "Mutation analysis in the BRCA2 gene in primary breast cancers." Nat Genet 13 (2): 245-7.

Mizuuchi, K. (1992). "Transpositional recombination: mechanistic insights from studies of mu and other elements." Annu Rev Biochem 61 1011-51.

Moran, J. V., *et al.* (1999). "Exon shuffling by L1 retrotransposition." Science 283 (5407): 1530-4.

Moran, J. V., *et al.* (1996). "High frequency retrotransposition in cultured mammalian cells." Cell 87 (5): 917-27.

Myers, J. S., *et al.* (2002). "A comprehensive analysis of recently integrated human Ta L1 elements." Am J Hum Genet 71 (2): 312-26.

Nachman, M. W. (1997). "Patterns of DNA variability at X-linked loci in Mus domesticus." Genetics 147 (3): 1303-16.

Ono, M., *et al.* (1987). "A novel human nonviral retroposon derived from an endogenous retrovirus." Nucleic Acids Res 15 (21): 8725-37.

Ostertag, E. M., *et al.* (2003). "SVA elements are nonautonomous retrotransposons that cause disease in humans." Am J Hum Genet 73 (6): 1444-51.

Ostertag, E. M. and Kazazian, H. H., Jr. (2001). "Biology of mammalian L1 retrotransposons." Annu Rev Genet 35 501-38.

Roy-Engel, A. M., *et al.* (2002). "Non-traditional Alu evolution and primate genomic diversity." J Mol Biol 316 (5): 1033-40.

Salem, A. H., *et al.* (2003). "LINE-1 preTa elements in the human genome." J Mol Biol 326 (4): 1127-46.

Shedlock, A. M. and Okada, N. (2000). "SINE insertions: powerful tools for molecular systematics." Bioessays 22 (2): 148-60.

Sinnett, D., *et al.* (1992). "Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences." J Mol Biol 226 (3): 689-706.

Smit, A. F. and Riggs, A. D. (1996). "Tiggers and DNA transposon fossils in the human genome." Proc Natl Acad Sci U S A 93 (4): 1443-8.

Symer, D. E., *et al.* (2002). "Human l1 retrotransposition is associated with genetic instability in vivo." Cell 110 (3): 327-38.

Ullu, E. and Tschudi, C. (1984). "Alu sequences are processed 7SL RNA genes." Nature 312 (5990): 171-2.

Ullu, E. and Weiner, A. M. (1985). "Upstream sequences modulate the internal promoter of the human 7SL RNA gene." Nature 318 (6044): 371-4.

Vincent, B. J., *et al.* (2003). "Following the LINEs: an analysis of primate genomic variation at human-specific LINE-1 insertion sites." Mol Biol Evol 20 (8): 1338-48.

Wallace, M. R., *et al.* (1991). "A de novo Alu insertion results in neurofibromatosis type 1." Nature 353 (6347): 864-6.

Yu, N., *et al.* (2001). "Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1." Mol Biol Evol 18 (2): 214-22.

CHAPTER 2:

COMPREHENSIVE ANALYSIS OF *ALU*-ASSOCIATED DIVERSITY ON THE HUMAN SEX CHROMOSOMES

Reprinted by Permission of *Gene*

Introduction

Recently Integrated *Alu* Insertions in the Human Genome

Alu elements are a class of repetitive mobile sequences that are dispersed ubiquitously throughout the genomes of primates (Batzer *et al.* 2002, Deininger *et al.* 1993, Schmid 1996). As Short INterspersed Elements (SINEs), *Alu* repeats are the largest family of mobile genetic elements within the human genome, having reached a copy number of over one million during the last 65 Myr (million years) (Batzer *et al.* 2002). *Alu* elements have achieved this copy number by duplicating via an RNA intermediate that is reverse transcribed by target primed reverse transcription and integrated into the genome (Kazazian *et al.* 1998, Luan *et al.* 1993). While unable to retropose autonomously, *Alu* elements are thought to appropriate the necessary mobilization machinery from the LINE (Long INterspersed Element) retrotransposon family (Boeke 1997, Sinnett *et al.* 1992), which encodes a protein possessing endonuclease and reverse transcriptase activity (Feng *et al.* 1996, Jurka 1997).

Phylogenetic studies of *Alu* elements suggest that only a small number of *Alu* elements, deemed master or source genes, are retropositionally competent (Deininger *et al.* 1992). Over time, the eventual accumulation of new mutations within master or source genes created a hierarchy of *Alu* subfamilies (Batzer *et al.* 2002, Deininger *et al.* 1992). Diagnostic mutation sites can be used to classify each individual element according to subfamily and to stratify *Alu* subfamily members based upon age from the oldest (designated J) to intermediate (S) and youngest (Y) (Batzer *et al.* 1996). Some young *Alu* subfamilies have amplified so recently that they are virtually absent from the genomes of non-human primates (Batzer *et al.* 2002). As a result, individual humans can be polymorphic for the presence of *Alu* elements at particular loci. Because the likelihood of two *Alu* elements independently inserting into the same location of the

genome is extremely small, and as there are no known biological mechanisms for the specific excision of *Alu* elements from the genome, *Alu* insertions can be considered identical by descent or homoplasy-free characters for the study of human population genetics (Batzer *et al.* 2002, Roy-Engel *et al.* 2002). SINE insertion polymorphisms are generally thought to be homoplasy-free characters for phylogenetic studies (Batzer *et al.* 2002, Shedlock *et al.* 2000) and have been utilized to resolve the relationships of artiodactyls and whales (Nikaido *et al.* 2001, Nikaido *et al.* 1999).

Repetitive Elements and Genetic Variation on the Sex Chromosomes

The aim of the present study is to annotate young *Alu* insertions on the human sex chromosomes in order to assess *Alu*-associated diversity and identify new *Alu* insertion polymorphisms. Several previous studies have focused on the evolutionary dynamics of repetitive elements on the sex chromosomes. Increased accumulation of repetitive elements on the X and Y has been detected in humans and other taxa (Boissinot *et al.* 2001, Charlesworth *et al.* 1994, Erlandsson *et al.* 2000, Smit 1999, Wichman *et al.* 1992). The differential accumulation of mobile elements is thought to result from reduced recombination and lower effective population sizes of the sex chromosomes leading to increased fixation of slightly deleterious insertions. However, Boissinot et al. (2001) found sex chromosome enrichment for full-length and greater-than 500 bp L1 elements, while demonstrating no associated enrichment in SINEs. Their results suggest that, unlike the longer-length L1 mobile elements, *Alu* insertions may not be deleterious enough on average to exhibit a sex chromosome distribution bias.

While no previous research specifically addresses repetitive element generated insertion polymorphisms on the sex chromosomes, studies using other classes of genetic markers have shown reduced genetic variation on the X and Y chromosomes of humans and other organisms

(Begun *et al.* 2000, Nachman 1997, Yu *et al.* 2001). This reduction of observed polymorphism has largely been attributed to reduced recombination and lower effective population sizes of these chromosomes (Begun *et al.* 2000, Nachman 1997). The current study affords the opportunity to assess human sex chromosome variability with a novel class of genetic markers.

Materials and Methods

Cell Lines and DNA Samples

The DNA samples used in this study were isolated from the cell lines as follows: human (*Homo sapiens*), HeLa (ATCC CCL-2); chimpanzee (*Pan troglodytes*) (NG06939); lowland gorilla (*Gorilla gorilla*) (NG05251). All non-human primate cell lines were obtained from the Coriell Institute for Medical Research, Camden, NJ. Human DNA samples from the African-American, Asian, European and Egyptians were described previously (Carroll *et al.* 2001). Indian DNA samples of defined sex were described previously (Bamshad *et al.* 2001). The South American human DNA samples were part of human diversity panels (HD 17 and 18) purchased from the Coriell Institute for Medical Research, Camden, NJ.

Identification of *Alu* Elements

Alu elements from the recently integrated *Alu* subfamilies Ya5, Ya5a2, Ya8, Yb8, Yb9, Yc1, Yd3, and Yd6 were identified from the August 2001 release of the UC Santa Cruz draft sequence (http://genome.ucsc.edu/). *Alu* subfamily members were located by two complementary methods. A local installation of RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) was used to screen sequences on chromosomes X and Y for the positions of recently integrated *Alu* elements. Exceptions to this were the Yc1 and Yc2 subfamilies, which were not identified by the software at the time of the study. In addition, subfamily specific oligonucleotides (Table 2.1) were

utilized in a local installation of the National Center for Biotechnology Information basic local alignment search tool (BLAST) software (Altschul *et al.* 1990) to identify exact complements within the draft human genomic sequence as previously described. Results from these analyses were pooled and cross-checked to remove duplicate elements. *Alu* elements were then extracted from their locations within the chromosome and aligned with MEGALIGN (DNASTAR V 3.1.7) for subfamily verification and further analysis. Lists of all the *Alu* elements identified in the database searches and full alignments of the recovered *Alu* elements are available under the publications section of our website (http://batzerlab.lsu.edu).

Table 2.1 *Alu* Subfamily-Specific Oligonucleotides [a]

Ya5/Ya5a2	5'-CCATCCCGGCTAAAAC-3'
Ya8	5'-ACTAAAACTACAAAAAATAG-3'
Yb8/Yb9	5'-ACTGCAGTCCGCAGTCCGGCC-3'
Yc1/Yc2	5'-GGGCGTGGTAGCGGGCGCCTG-3'
Yd3/Yd6 [b]	5'-CGAGACCACGGTGAAACCCCGTC-3'

a. Subfamilies Ya5/Ya5a2, Yb8/Yb9, Yd3/Yd6, and Yc1/Yc2 were screened using the same oligonucleotide and subsequently differentiated using multiple alignments and/or RepeatMasker.

b. The Yd3/Yd6 oligonucleotide listed will match all members of the Yd lineage. Yd3 and Yd6 members are subsequently identified by multiple alignments.

Primer Design and Amplification

Oligonucleotide primers for the polymerase chain reaction (PCR) amplification of each *Alu* element were designed using the Primer3 program (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Sequences flanking the *Alu* insertions were first masked with RepeatMasker to remove all repetitive elements. Primer3 was then utilized to design PCR primers within the remaining flanking unique DNA sequences. PCR amplification was

accomplished in 25 µl reactions using either 60 ng of template DNA (human populations) or 15 ng (non-human primates), 0.2 nM of each oligonucleotide primer, 200 µM deoxynucleotide-triphosphates, 1.5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.4) and Taq® DNA polymerase (1 unit). Each sample was subjected to the same amplification cycle as follows: initial denaturation of 150 seconds at 94ºC, 32 cycles of one minute of denaturation at 94ºC, one minute at the specific annealing temperature (shown in appendix B), one minute of extension at 72ºC, followed by a final extension at 72ºC for 10 minutes. For analysis, 20 µl of the PCR products were fractionated on a 2% agarose gel that contained 0.25µg/ml of ethidium bromide. PCR products were visualized using ultra-violet (UV) fluorescence. Twenty individuals from four populations (African-American, Asian, European and either Egyptian or South American) were screened to test each locus for insertion polymorphism. Additional male DNA samples from the following populations; French (8 individuals); Indian (15); African-American (15) were used to confirm polymorphism on the Y chromosome.

Results

Subfamily Copy Number and Distribution

Following a computational search of the human draft sequence, using both diagnostic oligonucleotide queries of the database and RepeatMasker screening, 344 *Alu* repeat elements from eight young *Alu* subfamilies (*Alu* Ya5; *Alu* Ya8; *Alu* Ya5a2; *Alu* Yb8; *Alu* Yb9; *Alu* Yc1; *Alu* Yd3; and *Alu* Yd6) were identified. Of these, 264 recently integrated *Alu* subfamily members were found on human chromosome X, while chromosome Y contained 80. The expected distributions of young *Alu* subfamilies on the sex chromosomes were calculated based on the size of each *Alu* subfamily, and the proportion of the human draft sequence represented by the respective chromosome (chromosome sizes and sequenced base pair totals taken from the

August 2001 freeze UC Santa Cruz summary statistics) as reported previously for human chromosome 19 (Arcot *et al.* 1998). The results of the database screening and expected numbers are given in Table 2.2. While several subfamilies were represented at or near expected levels, some deviated substantially. In particular, the number of *Alu* Ya5 elements was double that expected on the Y chromosome, but nearly equal to that expected on the X. The number of Yb8 subfamily members was consistent with expected numbers on both sex chromosomes. The Yc1 subfamily had approximately twice the expected number of elements on both the X and Y chromosomes. However, the excess of *Alu* Yc1 elements probably reflects the erroneous detection of Y subfamily elements that have had a fortuitous single base pair mutation to the Yc1 consensus sequence (Roy-Engel *et al.* 2001).

Table 2.2 Expected and Observed Distribution of Recently Integrated *Alu* Elements on the X and Y Chromosomes

***Alu* subfamily**	**Genomic copies[a]**	**Expected on X[b]**	**Found on X**	**Expected on Y[b]**	**Found on Y**
Ya5	2640	130.15	119	20.59	45
Ya8	60	2.96	0	0.47	2
Ya5a2	35	1.73	1	0.27	1
Yb8	1852	91.30	91	14.45	19
Yc1	381	18.78	37	2.97	10
Yb9	79	3.89	7	0.62	1
Yd3	198	9.76	7	1.54	0
Yd6	97	4.78	2	0.76	2

[a] Copy numbers based on previous estimated size of the subfamilies (Batzer *et al.* 2002, Xing *et al.* 2003).

[b] Expected number estimated based on the subfamily size and amount of X or Y chromosome sequence in the database, as outlined in the text.

Age of *Alu* Insertions on the Sex Chromosomes

The average age of the recently integrated *Alu* insertions on the X and Y chromosomes were estimated and compared to previous subfamily age estimates to determine if the

amplification dynamics of recently integrated *Alu* elements on the sex chromosomes is comparable to that of the rest of the nuclear genome. In order to estimate the average age for each *Alu* subfamily, the number of substitutions at CpG and non-CpG sites was determined. The mutation density for each of these mutation classes is different as a result of the methylation and subsequent spontaneous deamination of 5-methyl-cytosine bases (Bird 1980) and is approximately 10-fold higher in CpG than in non-CpG base positions within *Alu* elements (Batzer *et al.* 1990, Labuda *et al.* 1989). The average age for each *Alu* subfamily is then estimated by using the mutation density and a neutral rate of evolution of 0.15% per million years for non-CpG sequences (Miyamoto *et al.* 1987) and 1.5% per million years for CpG sequences as described previously. All deletions, insertions, simple sequence repeat expansions, and truncations were eliminated from the age calculations. All of the *Alu* elements that were identified in the draft sequence and were less than 100 bp in length were eliminated from the analysis. The estimated ages of Ya5, Yb8, and Yc1 are in line with the age estimates which were reported previously (Carroll *et al.* 2001, Roy-Engel *et al.* 2001, Xing *et al.* 2003) of 2.1-4.2 Myr and are summarized in Table 2.3. Subfamilies with less than five representatives on the sex chromosomes were excluded as there was not enough sequence for accurate estimates to be made. It is important to note that the mutation rate for X and Y chromosome DNA sequences is different (Huang *et al.* 1997), and these differences may influence these age estimates. However, this difference should be minimal.

An evolutionary analysis of the time of origin of the *Alu* elements located on the human sex chromosomes was determined within the primate lineage by PCR amplification of the individual loci using chimpanzee and gorilla DNA as templates. From the 225 recently integrated *Alu* elements analyzed in this study, three X chromosome loci (Yc1DP26, Yc1DP8

and Ya5DP38) and three Y chromosome loci (Yc1AD168, Yc1AD242, Yc1AD244) contained insertions within the chimpanzee and/or gorilla genomes, confirming that the overwhelming majority of the sex-chromosome specific *Alu* elements inserted in the human genome after the human and African-ape divergence, which is thought to have occurred within the last 4-6 million years. It is interesting to note that most of the putative recently integrated *Alu* elements that were also found in non-human primate genomes were members of the *Alu*Yc1 family. This is not surprising since a single base mutation differentiates this subfamily from the *Alu*Y subfamily as mentioned above (Roy-Engel *et al.* 2001).

Table 2.3 Estimated Ages of Sex-Chromosome Specific *Alu* Subfamilies

***Alu* subfamily**	**Ya5**		**Yb8**		**Yc1**		**Yd3**	
Chromosome	X	Y	X	Y	X	Y	X	Y
Number of loci analyzed	119	36	88	17	32	10	7	0
CpG mutation density (%)	2.53	1.97	3.60	1.74	2.5	2.65	12.1	N/A
Non-CpG mutation density (%)	0.78	0.49	0.53	0.47	0.28	0.24	1.39	N/A
Estimated age from CpG mutations (Myr)	**1.73**	**1.35**	**2.47**	**1.19**	**1.72**	**1.81**	**6.60**	N/A
Estimated age from non-CpG mutations (Myr)	**4.92**	**3.24**	**3.54**	**3.16**	**1.86**	**1.62**	**8.03**	N/A
Variance (between age estimates) (Myr)	5.09	1.77	5.79	1.94	0.01	0.02	1.37	N/A

Human Genomic Diversity

Individual *Alu* elements were screened for polymorphism by amplification of a panel of diverse human DNA samples, which included 20 African-Americans, 20 Europeans, 20 Asians, and either 20 Egyptians or South Americans. A total of eighty individuals were screened,

comprising approximately 120 X chromosomes and 40 Y chromosomes (Table 2.4). One hundred twenty one sex chromosome-specific *Alu* elements were not amplified by PCR, 109 of which were positioned within repeat-saturated regions of the genome, making the design of unique primers impossible. The remaining 12 elements either generated paralogous PCR products, or failed to amplify for unknown reasons that may include mutations within the primer binding sites, small deletions or even larger recombination events between adjacent sequences such as mobile elements.

The number of elements on the X chromosome, which exhibited polymorphism within the human genomes that were surveyed, consisted of nine Ya5's, five Yb8's, one Ya5a2, and one Yd3 element. All young subfamily members analyzed on the Y chromosome were found to be monomorphic, with the exception of one previously identified Yb8 *Alu* insertion, termed YAP (Y A*lu* polymorphism) (Hammer 1994), which is an intermediate frequency *Alu* insertion polymorphism. The remaining *Alu* insertion polymorphisms were classified as high, low or intermediate frequency as previously described and summarized in Table 2.4. Unbiased heterozygosity values for each of the polymorphisms were determined by allele counting. The heterozygosity data suggest that *Alu* insertion polymorphisms on the X chromosome will be useful as genetic markers for human population genetics. A schematic diagram showing the location of all *Alu* insertion polymorphisms located on the human X and Y chromosomes is shown in Figure 2.1.

The levels of *Alu* insertion polymorphism on the X and Y chromosomes were compared to previous data on the detection of autosomal *Alu* insertion polymorphisms. The data in (Carroll *et al.* 2001) was adapted to exclude all elements on the sex chromosomes in order to make comparisons against autosomal loci only. Chromosome X showed 14.06% (9/64) polymorphism

for the Ya5 subfamily, 100% (1/1) for Ya5a2, 20% (1/5) for the Yd3 subfamily and 8.77% (5/57) for the Yb8 subfamily. Compared to previously reported levels of *Alu* insertion polymorphism throughout the genome of 25% (Ya5), 80% (Ya5a2), 20% (Yb8), and 25% (Yc1) (Batzer *et al.* 2002), our data indicate that there is a slight reduction in *Alu* insertion polymorphism on the human sex chromosomes.

Discussion

Distribution of *Alu* Elements

The expected chromosomal distribution of recently integrated *Alu* elements was calculated based on the estimated subfamily size and the relative percentage of the draft sequence constituted by each chromosome. The distribution bias in the observed numbers of *Alu* elements appears to be subfamily-specific and is in good agreement with a recently published analysis of mobile elements on the sex chromosome (Jurka *et al.* 2002). For example, the Ya5 subfamily has approximately twice the number of *Alu* elements expected on the Y chromosome but nearly equal the number expected on the X chromosome. In contrast, the distribution of Yb8 subfamily members was consistent with estimated expectations on both chromosomes. Population genetics theory predicts that smaller effective populations should result in more frequent fixation of slightly deleterious insertions. Similarly, the virtual lack of recombination on the Y and reduced recombination on the X increases the extent of background selection and selective sweeps, further lowering the effective population size. Previous studies have reported a higher percentage of repetitive elements on the Y chromosome relative to autosomes and the X chromosome (Erlandsson *et al.* 2000). Boissinot and coworkers (Boissinot *et al.* 2001) previously reported an over-representation of full length and >500bp LINE elements, but no enrichment of SINEs on the sex chromosomes. In addition, the mobilization of *Alu* repeats has

recently been suggested to be male germline specific (Jurka *et al.* 2002), suggesting yet another mechanism for the differential accumulation of *Alu* repeats within the human genome. Therefore, we conclude the distribution of different classes of mobile elements on the sex chromosomes in different species is the result of a number of complex processes such as mobilization mechanism and integration site preferences that are mobile element specific.

Age of *Alu* Subfamily Members

The ages of recently integrated *Alu* elements on the sex chromosomes was estimated based upon CpG and non-CpG mutation densities and are in good agreement with those reported previously (Carroll *et al.* 2001, Roy-Engel *et al.* 2001). It is possible that the higher mutation rate in the male germline (Huang *et al.* 1997) would result in increased divergence and therefore higher estimated ages for *Alu* subfamily members on the Y chromosome. This effect, however, may be more detectable in older *Alu* subfamilies that have had more time to acquire mutations than in the recently integrated *Alu* subfamilies and certainly should not act selectively upon a single family of elements. This is in good agreement with a previous computational analysis of Y chromosome-specific mobile elements which demonstrated that the older *Alu* J and *Alu* S subfamilies showed significantly higher divergence on the Y chromosome, while the younger *Alu* Y subfamily divergence did not exhibit a significant difference (Erlandsson *et al.* 2000). Similarly, due to the increased male mutation rate, X-linked loci should theoretically exhibit a lower mutation rate than their autosomal counterparts since only one out of three X chromosomes is transmitted through the male germline each generation. However, this effect is likely minimal and is not reflected in the ages of the young *Alu* elements.

Table 2.4 X Chromosome *Alu* Insertion Polymorphism, Genotypes and Heterozygosity

	African American							Asian							European							Egyptian							
	Genotypes					*fAlu*	Het[a]	Genotypes					*fAlu*	Het[a]	Genotypes					*fAlu*	Het[a]	Genotypes					*fAlu*	Het[a]	Avg Het[b]
	Female			Male				Female			Male				Female			Male				Female			Male				
Name	+/+	+/-	-/-	+	-			+/+	+/-	-/-	+	-			+/+	+/-	-/-	+	-			+/+	+/-	-/-	+	-			
A. *Intermediate frequency*																													
Ya5a2DP1	2	0	4	3	7	0.32	0.47	3	0	3	10	1	0.37	0.45	0	1	4	1	12	0.09	0.18	6	1	1	8	0	0.09	0.18	0.32
Yb8DP2	5	2	0	9	3	0.81	0.34	0	3	8	1	8	0.13	0.23	0	3	9	1	7	0.13	0.23	2	4	6	2	6	0.31	0.43	0.31
Yd3JX437	1	2	4	5	0	0.33	0.48	3	6	2	6	0	0.58	0.50	0	2	10	0	8	0.07	0.08	0	5	8	1	6	0.18	0.29	0.34
B. *High frequency*																													
Ya5DP57	3	0	4	1	10	0.28	0.41	5	2	0	11	2	0.85	0.27	3	2	0	13	2	0.84	0.31	8	1	0	9	0	0.96	0.06	0.26
Ya5DP62	5	2	0	7	5	0.73	0.43	7	0	0	12	1	0.96	0.08	4	0	0	8	5	0.76	0.36	5	4	0	6	2	0.77	0.38	0.31
Ya5DP77	2	3	2	4	9	0.41	0.52	2	4	0	11	3	0.73	0.43	5	0	0	15	0	1.00	0	5	2	0	9	1	0.88	0.23	0.30
Ya5NBC98	5	2	0	8	5	0.74	0.42	7	0	0	12	1	0.96	0.08	5	1	0	6	6	0.71	0.45	5	4	0	5	1	0.79	0.33	0.32
Ya5NCB491	3	0	4	6	3	0.52	0.53	6	0	1	10	0	0.92	0.14	5	0	0	12	0	1.00	0	10	0	0	7	0	1.00	0	0.17
Yb8DP49	6	1	0	9	3	0.78	0.38	8	3	0	9	0	0.90	0.13	8	4	0	7	1	0.85	0.26	10	2	1	7	0	0.94	0.08	0.21
Yb8NBC102	7	1	0	10	3	0.86	0.27	7	0	0	13	0	1.00	0	5	0	0	15	9	0.74	0.34	10	0	0	10	0	1.00	0	0.15
Yb8NBC578	3	4	0	8	5	0.67	0.48	6	0	0	11	2	0.92	0.16	5	0	0	15	0	1.00	0	10	0	0	6	1	0.96	0.14	0.19
Ya5NDP13	7	0	0	12	1	0.96	0.08	7	0	0	13	0	1.00	0	5	0	0	15	0	1.00	0	9	0	0	10	0	1.00	0	0.02
Yb8NBC634	4	2	1	9	0	0.93	0.26	7	0	0	7	0	1.00	0	7	0	0	5	0	1.00	0	7	0	0	10	0	1.00	0	0.07
C. *Low frequency*																													
Ya5DP3	0	2	4	3	10	0.20	0.35	0	4	3	6	7	0.37	0.50	0	1	4	1	12	0.09	0.18	0	0	8	2	4	0.09	0.30	0.33
Ya5DP4	0	1	6	3	10	0.15	0.28	0	0	6	0	13	0	0	0	0	5	1	11	0.05	0.09	0	2	7	0	6	0.08	0.11	0.12
Ya5NBC37	2	3	2	4	9	0.41	0.52	2	2	3	5	8	0.41	0.52	0	3	1	3	13	0.25	0.46	0	3	6	0	7	0.12	0.16	0.42

The level of insertion polymorphism was determined as: Low frequency - the absence of the element from all individuals tested, except one or two homozygous or heterozygous individuals. Intermediate frequency - the *Alu* element is variable as to its presence or absence in at least one population. High frequency – the element is present in all individuals in all populations tested, except for one or heterozygous individuals.

a. This is the unbiased heterozygosity, which takes into account sex differences within the calculation.

b. Average heterozygosity is the average of the population heterozygosity across all four populations.

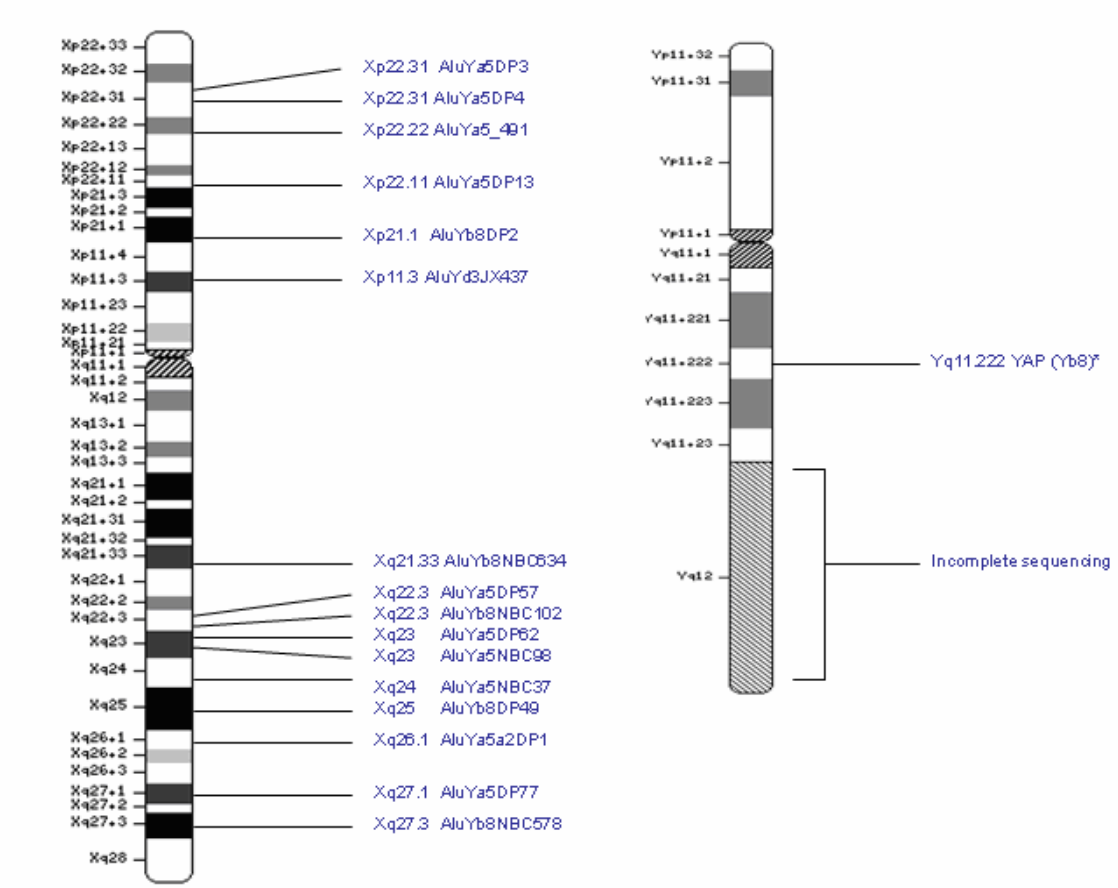

Figure 2.1 Idiogram of Human Sex Chromosome-Specific *Alu* Insertion Polymorphisms The physical location of each *Alu* insertion polymorphism was determined using the sequence map from each chromosome as a framework to localize the elements. The sequence from the q12 portion of the human Y chromosome has not yet been completed and therefore the *Alu* elements within this portion of the Y chromosome have not been analyzed. All of the *Alu* insertion polymorphisms from the recently integrated subfamilies of elements are shown in the figure. The * denotes the previously reported YAP *Alu* element (Hammer 1994).

Population Dynamics

The recently integrated *Alu* subfamily members on the X and Y chromosomes exhibited reduced polymorphism as compared to their autosomal counterparts. Age estimates and data from orthologous inserts in non-human primates indicate that this reduction in polymorphism is not the result of increased age of *Alu* insertions found on the sex chromosomes. Rather, the results are consistent with neutral theory, given that lower effective population size should result

in more rapid fixation of elements, lowering overall polymorphism levels on the sex chromosomes. Reduced recombination on the X and Y chromosomes may exacerbate this effect by increasing the extent of background selection and selective sweeps which further remove polymorphism (Charlesworth *et al.* 1994, Lander *et al.* 2001). The current findings are in agreement with several previously published studies in humans and other organisms that have found reduced polymorphism on the sex chromosomes (Hammer 1994, Jorde *et al.* 2000, Lander *et al.* 2001, Yu *et al.* 2001).

Aside from the previously identified YAP *Alu* element, all of the *Alu* loci located in the non-recombining portion of the Y chromosome were monomorphic for the presence of the *Alu* repeat in diverse populations. This suggests that the *Alu*-associated variation currently on the human Y chromosome is very low, probably existing as low frequency insertions which were not detected in this study, as the young *Alu* elements were ascertained from a single genome. Thus, our data points to an evolutionarily recent event that dramatically reduced *Alu*-associated Y chromosome diversity or to an effective population size for the human Y chromosome that has not been large enough to harbor appreciable *Alu* polymorphism.

References

Altschul, S. F., *et al.* (1990). "Basic local alignment search tool." J Mol Biol 215 (3): 403-410.

Arcot, S. S., *et al.* (1998). "High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils." J Mol Biol 281 (5): 843-856.

Bamshad, M., *et al.* (2001). "Genetic evidence on the origins of Indian caste populations." Genome Res 11 (6): 994-1004.

Batzer, M. A. and Deininger, P. L. (2002). "Alu repeats and human genomic diversity." Nat Rev Genet 3 (5): 370-9.

Batzer, M. A., *et al.* (1996). "Standardized nomenclature for Alu repeats." J Mol Evol 42 (1): 3-6.

Batzer, M. A., *et al.* (1990). "Structure and variability of recently inserted Alu family members." Nucleic Acids Res 18 (23): 6793-6798.

Begun, D. J. and Whitley, P. (2000). "Reduced X-linked nucleotide polymorphism in Drosophila simulans." Proc Natl Acad Sci U S A 97 (11): 5960-5965.

Bird, A. P. (1980). "DNA methylation and the frequency of CpG in animal DNA." Nucleic Acids Res 8 (7): 1499-1504.

Boeke, J. D. (1997). "LINEs and Alus--the polyA connection." Nature Genetics 16 6-7.

Boissinot, S., *et al.* (2001). "Selection against deleterious LINE-1-containing loci in the human lineage." Mol Biol Evol 18 (6): 926-935.

Carroll, M. L., *et al.* (2001). "Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity." J Mol Biol 311 (1): 17-40.

Charlesworth, B., *et al.* (1994). "The evolutionary dynamics of repetitive DNA in eukaryotes." Nature 371 (6494): 215-220.

Deininger, P. L. and Batzer, M. A. (1993). "Evolution of retroposons." Evolutionary Biology 27 157-196.

Deininger, P. L., *et al.* (1992). "Master genes in mammalian repetitive DNA amplification." Trends Genet 8 (9): 307-311.

Erlandsson, R., *et al.* (2000). "Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans." Mol Biol Evol 17 (5): 804-812.

Feng, Q., *et al.* (1996). "Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition." Cell 87 (5): 905-916.

Hammer, M. F. (1994). "A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies." Mol Biol Evol 11 (5): 749-761.

Huang, W., *et al.* (1997). "Sex differences in mutation rate in higher primates estimated from AMG intron sequences." J Mol Evol 44 (4): 463-465.

Jorde, L. B., *et al.* (2000). "The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data." Am J Hum Genet 66 (3): 979-88.

Jurka, J. (1997). "Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons." Proc Natl Acad Sci U S A 94 (5): 1872-1877.

Jurka, J., *et al.* (2002). "Active Alu elements are passed primarily through paternal germlines." Theor Popul Biol 61 (4): 519-30.

Kazazian, H. H., Jr. and Moran, J. V. (1998). "The impact of L1 retrotransposons on the human genome." Nat Genet 19 (1): 19-24.

Labuda, D. and Striker, G. (1989). "Sequence conservation in Alu evolution." Nucleic Acids Res 17 (7): 2477-2491.

Lander, E. S., *et al.* (2001). "Initial sequencing and analysis of the human genome." Nature 409 (6822): 860-921.

Luan, D. D., *et al.* (1993). "Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition." Cell 72 (4): 595-605.

Miyamoto, M. M., *et al.* (1987). "Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region." Science 238 (4825): 369-373.

Nachman, M. W. (1997). "Patterns of DNA variability at X-linked loci in Mus domesticus." Genetics 147 (3): 1303-1316.

Nikaido, M., *et al.* (2001). "Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins." Proc Natl Acad Sci U S A 98 (13): 7384-7389.

Nikaido, M., *et al.* (1999). "Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales." Proc Natl Acad Sci U S A 96 (18): 10261-10266.

Roy-Engel, A. M., *et al.* (2002). "Non-traditional Alu evolution and primate genomic diversity." J Mol Biol 316 (5): 1033-1040.

Roy-Engel, A. M., *et al.* (2001). "Alu insertion polymorphisms for the study of human genomic diversity." Genetics 159 (1): 279-290.

Schmid, C. W. (1996). "Alu: structure, origin, evolution, significance and function of one-tenth of human DNA." Prog Nucleic Acid Res Mol Biol 53 283-319.

Shedlock, A. M. and Okada, N. (2000). "SINE insertions: powerful tools for molecular systematics." Bioessays 22 (2): 148-160.

Sinnett, D., *et al.* (1992). "Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences." J Mol Biol 226 (3): 689-706.

Smit, A. F. (1999). "Interspersed repeats and other mementos of transposable elements in mammalian genomes." Curr Opin Genet Dev 9 (6): 657-63.

Wichman, H. A., *et al.* (1992). "Transposable elements and the evolution of genome organization in mammals." Genetica 86 (1-3): 287-293.

Xing, J., *et al.* (2003). "Comprehensive analysis of two Alu Yd subfamilies." J Mol Evol 57 Suppl 1 S76-89.

Yu, N., *et al.* (2001). "Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1." Mol Biol Evol 18 (2): 214-22.

CHAPTER 3:

ALU RETROTRANSPOSITION-MEDIATED DELETION

Reprinted by Permission of *Journal of Molecular Biology*

Introduction

Alu repeats are the most prolific SINEs (Short INterspersed Elements) in primate genomes, accumulating approximately 1.2 million members over the last 65 million years of evolution (Lander *et al.* 2001). True to their moniker as "genomic parasites", *Alu* elements rely on the cellular machinery of other mobile elements, such as LINEs (Long INterspersed Elements), for their successful transmission through the germline (Boeke 1997, Dewannieux *et al.* 2003, Sinnett *et al.* 1992). Not all *Alu* elements are capable of using the borrowed commodities however. Some hypotheses suggest that only a few *Alu* source genes are retrotranspositionally competent (Deininger *et al.* 1999, Deininger *et al.* 1992). Over time, the source *Alu* elements accumulate sequence mutations, and this has resulted in an array of *Alu* subfamilies distinguished by diagnostic mutations (Batzer *et al.* 2002, Deininger *et al.* 1992). Although the peak of *Alu* amplification occurred some 40-60 million years ago, lineage-, population- and individual-specific insertion events in modern primate genomes are indicated in recent studies (Callinan *et al.* 2003, Carroll *et al.* 2001, Carter *et al.* 2004, Hedges *et al.* 2004, Otieno *et al.* 2004, Xing *et al.* 2003).

Alu elements are a unique source of genomic instability among primates. As a direct result of their abundance and sequence identity, they promote genetic recombination events that are responsible for large-scale deletions, duplications and translocations (Bailey *et al.* 2003, Chen *et al.* 1989, Iafrate *et al.* 2004, McNeil 2004, Sebat *et al.* 2004). Some *Alu*-mediated recombination events that have occurred within and nearby coding regions are instigators of disease. Currently, *Alu*-*Alu* recombination events have been linked to approximately 50 human diseases including hypercholesterolemia, α-thalassemia and BRCA1 related breast cancer, see review (Deininger *et al.* 1999). The disruptive consequences of newly integrated *Alu* insertions

within genic regions of the human genome have also been documented in several studies. *Alu* elements may disrupt splicing by integrating within introns, alter patterns of gene expression by inserting within promoter regions or regions upstream of genes, or even silence gene function by inserting within the gene itself (Deininger *et al.* 1999). Mutagenesis via *Alu* insertion accounts for approximately 0.1% of all human diseases and is responsible for cases of familial cancer, metabolic disease and blood disorders (Deininger *et al.* 1999).

Recently, novel consequences of *Alu*-induced genomic instability have come to light. An example by (Hayakawa *et al.* 2001) documents the deletion of an exon caused by gene conversion of an older *Alu*Sx element to a younger *Alu*Y element, specifically within the human lineage. The consequential loss of the CMP-*N*-acetylneuraminic acid hydroxylase gene produces a biochemical difference between humans and non-human primates. Although other gene conversion-associated deletions are documented in the literature (Carter *et al.* 2004, Salem *et al.* 2003), this mechanism has yet to be explored on a large scale.

Alu Retrotransposition-mediated Deletion (ARD), the focus of our paper, is another novel type of genetic instability mediated by *Alu* elements. The initial evidence for this mechanism derived from studies by (Gilbert *et al.* 2002) and (Symer *et al.* 2002), who independently determined that 10% of L1 integrations within cultured human cells resulted in target site deletions spanning from 1 bp to 70,000 bp. L1 insertions associated with the deletion of target DNA had characteristics not typical of usual L1 integrants. In addition to the lack of target site duplications (TSDs), deletion-inducing L1 elements integrated at non-canonical L1 EN (endonuclease) nick sites and sometimes lacked poly-A tails (Gilbert *et al.* 2002, Morrish *et al.* 2002, Symer *et al.* 2002).

Because *Alu* repeats and LINEs share the mobilization machinery needed to retrotranspose (Boeke 1997, Dewannieux *et al.* 2003, Sinnett *et al.* 1992), it was presumed that *Alu* elements also possessed the same ability to introduce genomic instability through retrotransposition-mediated deletion (Gilbert *et al.* 2002, Symer *et al.* 2002). Even though ARD has not been investigated *in vitro*, some examples from natural genomes are present in the current literature (Carter *et al.* 2004, Salem *et al.* 2003). In the first case, documented by (Salem *et al.* 2003), the insertion of an *Alu*Yg6 into human chromosome 3 was accompanied by a deletion of approximately 300 bp of DNA. The second event involved the insertion of a young Yb7 subfamily member, again associated with a deletion of 300 nucleotides (Carter *et al.* 2004). Given that *Alu* elements have reached copy numbers in excess of one million per haploid genome, it is likely that significant genomic alteration resulting from ARD will be found within the primate order. Despite the intriguing preliminary evidence for this unusual mechanism of genomic instability, no comprehensive studies have attempted to quantify the rate of *Alu* retrotransposition-mediated deletion within primate genomes.

In this study, we employ a sensitive computational screening approach to compare the draft genomic sequences of *Homo sapiens* and *Pan troglodytes* in order to assess the occurrence of deletions associated with *Alu* retrotransposition. Our findings, further supported by wet bench verification methods, indicate that *Alu* retrotransposition may have generated over 3000 deletion events during the course of primate evolution, removing nearly a megabase of DNA in the process.

Materials and Methods

DNA Samples

DNA cell lines used in this study were obtained from the following sources: DNA samples from the African-American, European, and Asian populations were isolated as described in previous studies (Callinan *et al.* 2003, Carroll *et al.* 2001, Hedges *et al.* 2004, Otieno *et al.* 2004, Roy *et al.* 1999, Roy-Engel *et al.* 2001). DNA for the South American population group (HD 17 and 18) and for a lowland gorilla (*Gorilla gorilla gorilla* AG05253A) was purchased from the Coriell Institute for Medical Research. Green monkey (*Chlorocebus aethiops* ATCC CCL70) and orangutan (*Pongo pygmaeus* ATCC CR6301) DNA was obtained from the American Type Culture Collection. A chimpanzee panel comprising 12 unrelated chimpanzees of unknown subspecies membership was obtained from the Southwest Foundation for Biomedical Research.

Computational Analysis

The human July 2003 freeze and the *Pan troglodytes* November 2003 freeze from the University of California Santa Cruz (http://genome.ucsc.edu) were analyzed in this study. To identify ARD events, 100 bases of 5' and 3' *Alu* flanking sequence in human were extracted and joined together into 200 bp fragments. These 200 bp fragments were used as query against the common chimpanzee genomic sequence using the Parcel BlastMachine at the Genome Core Facility at Columbia University. Due to random sequence match at the ends, we often see that the matches for the 5' flanking region extend past the first 100 bp and the matches for the 3' flanking region start before the 101 bp position. Therefore, the end-point for the 5' flanking sequence and the start-point for the 3' flanking sequence have to be re-adjusted in order to obtain the correct start- and end-points in the target sequence. Following this, the sequences in the

target chimpanzee genome between the 5' and 3' flanking sequences were extracted and used to compare with the corresponding human *Alu* sequences using the bl2seq program. To identify an ARD event, the corresponding criteria were met: 1) bl2seq did not produce a match between the query and the target sequence; or, 2) bl2seq produced one or several hits (from deleted unrelated *Alu* fragments) but the aligned region(s) were at least 5 bases away from at least one end of the target sequence. Then the computational comparison was reversed, comparing the chimpanzee genome against the human target sequence. Manual verification was performed using the Blast Like Alignment Search Tool (BLAT) and Basic Local Alignment Search Tool (BLAST) software (Altschul *et al.* 1990, Kent 2002). This eliminated instances of deletion due to *Alu* gene conversion deletion, which appear as a replacement of an *Alu* in one lineage over another, accompanied by deleted sequence in the derived state. All of the manually verified ARD candidates were subjected to experimental verification using the polymerase chain reaction analyses of the loci.

To determine whether deleted sequences in the human or chimpanzee genome contained coding or regulatory regions, the experimentally verified deleted sequence data retrieved from the computational comparison above was queried against BLAT (Kent 2002) and TRANSFAC software (www.gene-regulation.com).

Polymerase Chain Reaction Analysis

To authenticate the ARD events, oligonucleotide primers were designed within the 400-1000 nucleotide long flanks surrounding each locus of interest using the primer design software Primer3 (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Primer sequences, annealing temperatures, PCR product sizes and chromosomal locations are located in the publications

section of our website (http://batzerlab.lsu.edu) and in Appendix C. Each locus was amplified from the genomes of 80 humans (20 from each of 4 geographically diverse populations), 12 chimpanzees, 1 Western Lowland gorilla, 1 orangutan and 1 green monkey.

PCR analysis was performed in 25 μL reactions using between 10-30 ng DNA, 200 nM of each oligonucleotide primer, 200 μM of dNTPs in 50 mM KCl, 1.5 mM $MgCl_2$, 10 mM Tris-HCl (pH 8.4) and 2.5 units *Taq* DNA polymerase. Each sample reaction was subjected to an initial denaturation step of 94°C for 120 seconds, followed by 32 amplification cycles of 30 seconds at 94°C, 30 seconds at the specific annealing temperature and 60 seconds at 72°C, followed by one round of extension at 72°C for 5 minutes. The PCR products were separated on a 2% agarose gel and stained with ethidium bromide. Following separation, DNA fragments were visualized with UV fluorescence to assess the status of each locus.

DNA Sequence Analysis

To verify the existence of the retrotransposition-mediated deletions, individual PCR products were either sequenced using chain termination sequencing methodology (Sanger *et al.* 1977) with ABI Big Dye v.3.1 (ABI Biosystems) after gel extraction and cloning with the TOPO-TA cloning vector (Invitrogen), or directly from PCR products purified by the Wizard gel and PCR clean up system as directed by the manufacturer (Promega). All sequenced PCR products were analyzed on an Applied Biosystems 3100 automated DNA sequencer. DNA sequence data were analyzed using the Seqman program in the DNAstar suite and aligned with BioEdit. The sequences of the orthologous non-human primates loci analyzed in this study have been assigned accession numbers (AY881293-AY881325, AY900585-AY900619).

Results

Alu Retrotransposition-Mediated Deletions

To detect lineage-specific ARD events, data from the National Center for Biotechnology (NCBI) draft sequence of the human genome were compared to the draft genomic sequence of the common chimpanzee *Pan troglodytes* (for program details, see methodology). The program was designed to detect lineage-specific *Alu* elements in one genome that are associated with extra (non-homologous) genomic sequences in the other, (see alignment, Figure 3.1). To eliminate the presence of *Alu* gene-conversion mediated deletions in our dataset, manual verification of the sequence was performed (see materials and methods). The remaining putative *Alu* retrotransposition-mediated deletion events were verified as authentic deletions rather than independent insertions through polymerase chain reaction (PCR) amplification of the locus in outgroup taxa (gorilla, orangutan and green monkey), Figure 3.2.

In total, 19 young *Alu* insertion events specific to the human lineage were associated with deleted target site DNA; in the chimpanzee genome, 14 such events were recovered (Table 3.1). Among the human data, we recovered the two ARD events detected in prior studies (Carter *et al.* 2004, Salem *et al.* 2003), thereby validating our computational methods. One of the human-specific ARD events, HuARD9, could not be experimentally verified due to a lack of unique flanking sequence, but it was included in the total *Alu* insertion number due to its structural authenticity. Our data indicate that humans possess 1.36 times as many detectable ARD events than do chimpanzees. Adjusting this number to account for polymorphisms missed by sampling a single sequenced genome, as described in (Hedges *et al.* 2004), we conclude that ARD levels in the human genome are approximately 1.1 times greater than in the chimpanzee (Table 3.1).

Figure 3.1 Alignment of an *Alu* Retrotransposition-Mediated Genomic Deletion
Sequence alignment of HuARD10 (a 5'-truncated *Alu* element) in human and chimpanzee. Letters in black capital indicate shared flanking unique sequence. The human-specific *Alu* insertion is featured in red; the extra portion in chimpanzee (representing that sequence deleted in human) is shown in blue.

Levels of *Alu* Retrotransposition-Mediated Deletion Polymorphism

To assess the level of polymorphism in *Homo sapiens* for ARD events, we used PCR to amplify loci from 20 unrelated individuals from each of four geographically diverse populations (80 total individuals). Eleven percent (2/18) (one locus, HuARD9, could not be amplified) of the tested loci were polymorphic; this value translates to a polymorphism rate of 19% after an adjustment for single genome sampling (Table 3.1). The polymorphism level obtained appears

to be lower than what is typical for recently integrated *Alu* elements. Fourteen of the 18 events were insertions of elements from either the *Alu* Yb or the Ya5 lineages, which have insertion polymorphism rates of 20-25% across diverse human populations (Callinan *et al.* 2003, Carroll *et al.* 2001, Hedges *et al.* 2004, Otieno *et al.* 2004, Roy-Engel *et al.* 2001). This is a conservative estimate of polymorphism for these subfamilies considering the figures are unadjusted for single genome sampling. We believe that the reduced polymorphism in our dataset is a result of the relatively small sample sizes as compared to the previous analyses of thousands of young *Alu* insertions.

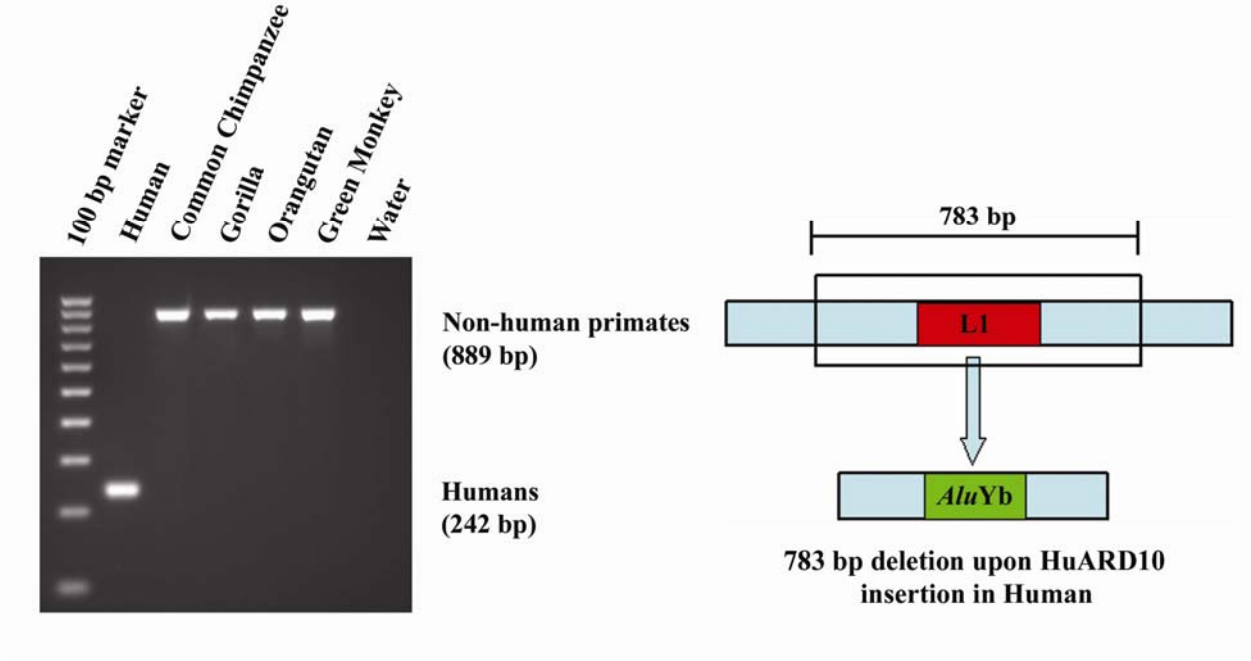

Figure 3.2 Chromatograph and Schematic of an *Alu* Retrotransposition-Mediated Genomic Deletion
Agarose gel chromatograph of a phylogenetic PCR analysis with an adjacent schematic diagram depicting the insertion of the HuARD10 element and the deletion of 783 bases of DNA including a LINE element.

Table 3.1 Retrotransposition-Mediated Deletion Frequency and Polymorphism Levels within the Human and Chimpanzee Lineages

	Human	Chimp	Human to Chimp Ratio
Observed Deletion Events Total	19	14	1.36
P.C.R. Tested	18	14	----
Fixed Present	16	9	----
Polymorphic Loci	2	5	----
Polymorphic Fraction	0.11	0.36	0.31
Adjusted Polymorphic Loci	4	10	----
Adjusted Polymorphic Fraction	0.19	0.53	0.36
Adjusted Deletion Events Total	21	19	1.11

Using a DNA panel of twelve unrelated chimpanzee individuals, all 14 chimpanzee loci were successfully amplified by polymerase chain reaction. We determined the *Alu* insertion polymorphism to be 36% (5 polymorphic loci; Table 3.1), similar to the polymorphism level of 37% recently reported by (Hedges *et al.* 2004) (who used the same DNA panel). After adjusting the value for sampling from a single sequenced genome, our chimpanzee diversity rose to 53%, again similar to the adjusted 59% polymorphism level reported by (Hedges *et al.* 2004).

However, we found that two highly variable chimpanzee DNA donors accounted for four of the five polymorphic loci represented in the dataset. Another study from our laboratory has also found these two chimpanzee genomes to be highly polymorphic (Han In press). Although information on sub-species membership for these chimpanzees is unavailable, recent nucleotide diversity data suggest that central African chimpanzees possess between 1.5 and 2.5 times more

variability than do other chimpanzee subspecies (Fischer *et al.* 2004, Yu *et al.* 2003). Without these two individuals, our chimpanzee insertion polymorphism levels would have appeared considerably lower. Therefore, care should be taken when assessing polymorphism using small datasets and DNA of unknown subspecies membership. Further research to identify the four putative sub-species of chimpanzee through genetic testing will help improve primate genomic diversity sampling strategies.

From our PCR screening of 160 human chromosomes (80 human individuals) and 24 chimpanzee chromosomes (12 chimpanzee individuals), we did not detect evidence of individual variation in the presence/absence of extra sequence alongside the newly inserted *Alu* elements.

Nucleotides Lost through *Alu* Retrotransposition-Mediated Deletion

The number of nucleotides deleted per retrotransposition event varied considerably within and between species. The number of nucleotides eliminated from the human genome totaled 8,550 bp, with a range of 1,546 bases between the largest and the smallest deletion (Table 3.2). Deletions associated with *Alu* retrotransposition occurring in chimpanzee totaled 466 bp (range = 204 bp), considerably fewer bases than in human even considering the smaller quantity of chimp-specific insertion events.

Table 3.2 Genomic Alteration through *Alu* Retrotransposition-Mediated Deletion

	Human	**Chimpanzee**
Total bp Deleted	8550	466
Mean (bp)	450	33
Range (bp)	1546	204

Target Site Duplications

Target site duplications were absent from the ARD loci detected in human and chimpanzee genomes, consistent with previous examples of L1 retrotransposition-mediated genomic deletions. Potential TSDs were present in only one ARD event, HuARD15. However,

the sequences were not a perfect match. Given that HuARD15 is a young *Alu* element, (0.6% diverged from consensus), there has been insufficient time for originally perfect TSDs to mutate to the current sequences, suggesting that this element did not possess TSDs from the integration process. Therefore, we conclude that hallmarks identified from retrotransposition-mediated deletion events using a cell culture system to study L1 retrotransposition (Gilbert *et al.* 2002) closely mirror the characteristics of element retrotransposition associated with deletion *in vivo*.

Cleavage Site Preferences

In our data set, only eight out of the 33 ARD events (HuARD7, HuARD15, HuARD19, ChARD3, ChARD6, ChARD7, ChARD9 and ChARD12) possessed an integration site sequence similar to that preferred by L1 endonuclease, the endonuclease purportedly used by *Alu* elements during mobilization (Boeke 1997, Dewannieux *et al.* 2003) (Table 3.3). The remaining 25 events exhibited noncanonical integration sites that may indicate L1 EN-independent integration, as postulated in previous studies (Gilbert *et al.* 2002, Morrish *et al.* 2002, Symer *et al.* 2002) (Table 3.3). However, these non-canonical integration sites may also be characteristic of L1 EN-dependent nicking, followed by promiscuous target primed reverse transcription (pTPRT, see later section).

Genomic Location

Alu insertions associated with genomic deletion localized to 12 of the 24 human chromosomes, and to 11 of the 25 chromosomes in chimpanzee (Figures 3.3 and 3.4, respectively). In both cases, the *Alu* elements appear to be scattered widely among the chromosomes. Deletions within gene-rich (typically GC-rich) regions would most likely be detrimental to the survival of an organism. Therefore, we would expect *Alu* retrotransposition-mediated deletions to be located in more AT-rich regions of the genome. To investigate this

hypothesis, 10,000 nucleotides directly surrounding each element were analyzed for GC content using sequence analysis software (DNAstar v.5). The young deletion-associated *Alu* inserts in the human genome were more common in regions with lower GC content (~38% GC; genome-wide average = 42% GC), similar to chimpanzee-specific *Alu* element insertions (36.4% GC; genome-wide average 40% GC). Thus, our dataset indicates that deletions in the human and chimpanzee genomes are more tolerated in regions with higher AT content, rather than in regions of high GC content. Approximately 75% of the genomic deletions detected in our study occurred within the introns of genes, rather than between genes. In one instance, a 1002 bp deletion at the HuARD6 locus induced the functional loss of a retroviral transforming gene, *c-rel,* within the human lineage. Research indicates that *c-rel* may have important roles in regulating cell proliferation and differentiation (Bishop 1982).

Table 3.3 *Alu* Element Integration Sites

Human Retrotransposition-Mediated Deletion	Locus Name	Target Integration Site[b]	Chimp Retrotransposition-Mediated Deletion	Locus Name	Target Integration Site[b]
	1	5'-aaat/a		1	5'-aagt/a
	2	5'-gaat/a		2	5'-aacc/a
	3	5'-tttt/t		3	5'-tttt/a[a]
	4	5'-tttc/t		4	5'-acac/c
	5	5'-ttga/t		5	5'-ttat/t
	6	5'-ttct/g		6	5'-ttct/a[a]
	7	5'-tttc/a[a]		7	5'- tctt/a[a]
	8	5'-gccc/t		8	5'-tttt/g
	9	5'-gtct/t		9	5'-ttct/a[a]
	10	5'- atgc/t		10	5'-gttt/g
	11	5'-ttgt/t		11	5'-ttcc/a
	12	5'-tgta/t		12	5'-ttct/a[a]
	13	5'-aaat/t		13	5'-gaat/a
	14	5'- ttca/t		14	5'-tact/a
	15	5'-tctt/a[a]			
	16	5'-tttt/t			
	17	5'-cttc/t			
	18	5'-tatc/t			
	19	5'-tttc/a[a]			

[a] Indicates typical L1 EN nick sites.
[b] Target integrations sites are presented on the anti-sense strand in the 5'-3' direction

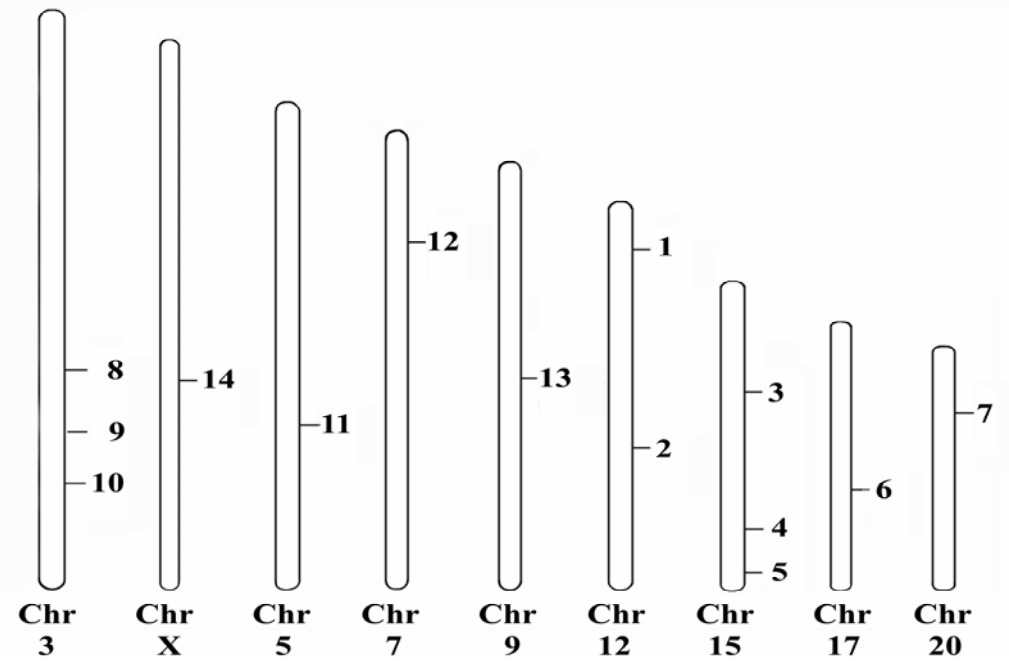

Figure 3.3 *Alu* Retrotransposition-Mediated Deletions within the Chimpanzee Genome
A partial schematic of the chimpanzee genome including those chromosomes occupied by *Alu* retrotransposition-mediated deletions. The labels indicate the locus number.

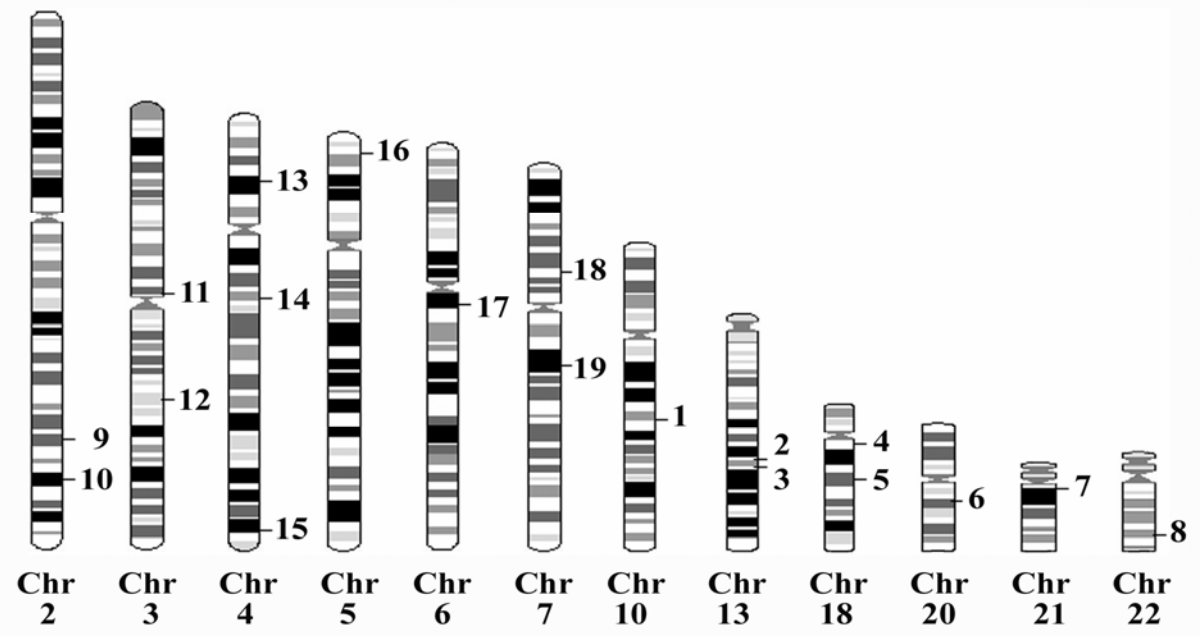

Figure 3.4 *Alu* Retrotransposition-Mediated Deletions within the Human Genome
A partial schematic of the human genome including those chromosomes occupied by *Alu* retrotransposition-mediated deletions. The labels indicate the locus number.

Unusual Loci: Internal Priming

Within our human dataset, we found an example of a tail-less *Alu* repeat element. A member of the *Alu*Ya5 subfamily, the element (HuARD8) lacked approximately 20 bp of its 3'-end as well as the characteristic oligo dA-rich tail. This *Alu* element inserted at a non-canonical integration site and induced a small target site deletion of 21 bp. Plausible explanations for these unusual structural characteristics include internal priming and, alternatively, deletion of the tail via unequal recombination subsequent to the element's insertion. Internal priming appears more plausible than does A-tail recombination, given that the lineage-specific element has resided only briefly in the human genome. This hypothesis is supported by evidence that shows tail-less *Alu* sequences in only four elements (0.1%, one Yb8 and 3 Ya5) out of over 4000 lineage-specific *Alu* elements that have been analyzed in the human genome (Carter *et al.* 2004, Garber *et al.* in press, Otieno *et al.* 2004, Roy-Engel *et al.* 2001). Therefore, to determine if internal priming could account for the tail-less nature of HuARD8, we used the 3'-end of the Ya5 consensus sequence to simulate the missing portion of the *Alu* RNA transcript. Using this approach, we found 11 bases at the 3'-end of the reconstructed HuARD8 RNA transcript to be complementary to the putative primer-binding site located within the first 25 bases downstream of the nick site (Figure 3.4). These data suggest that internal priming occurred during this particular *Alu* integration/deletion event.

Discussion

Our study offers the first genome-wide attempt to quantify the contribution of *Alu* retrotransposition-mediated genomic deletion to the instability of the primate genome. Using computational comparisons supported by wet bench methodologies, we provide evidence from the genomes of human and chimpanzee for 33 independent retrotransposition-mediated deletion

events that have deleted approximately 9,000 bases of DNA during the last 5 million years. These deletions may have been created independently of the *Alu* insertions or as a direct result of the insertion process. However, as we found no non-deleted alleles across 80 human genomes and 24 chimpanzee genomes, we believe that it is highly unlikely that the deletions were created independently of the insertion of the mobile elements. Therefore, we conclude that the deletion of adjacent genomic sequences occurred prior to, or more likely, tightly associated with the insertion of the *Alu* elements. Further, our study indicates that *Alu* elements are able to use non-typical insertion sites in order to proliferate.

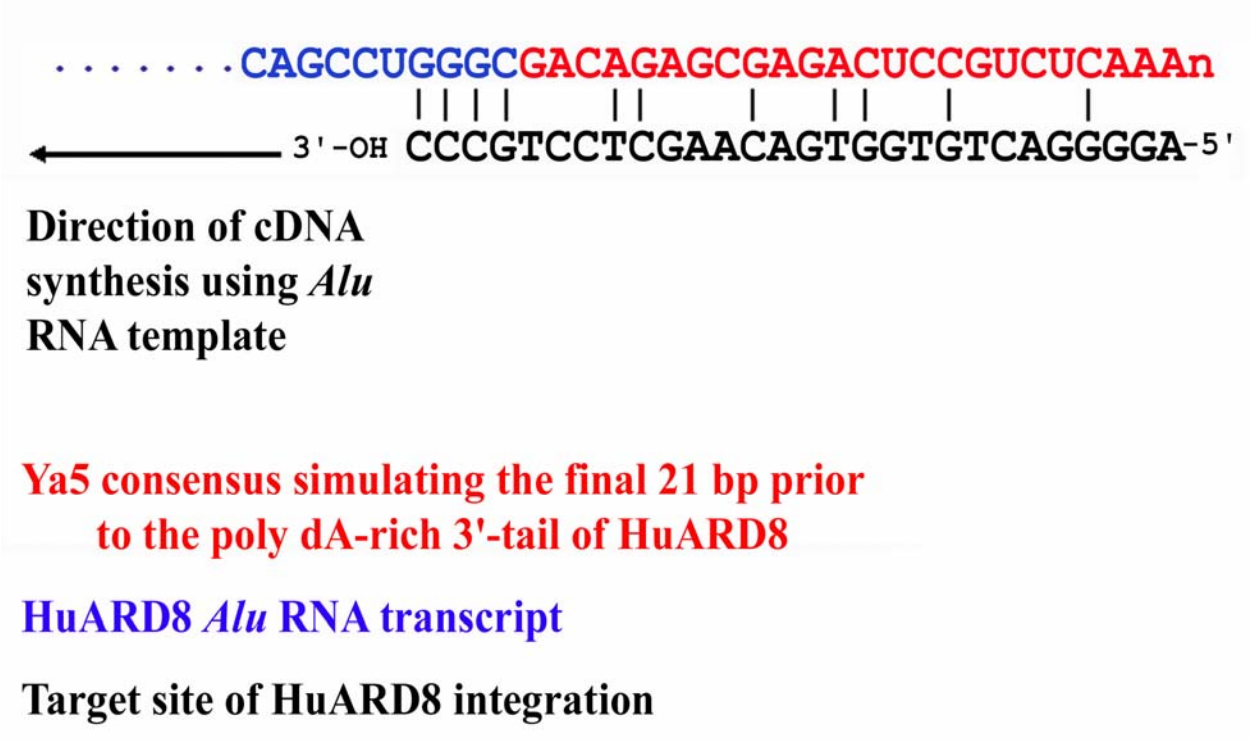

Figure 3.5 Internal Priming of HuARD8
To determine if internal priming could account for the tail-less nature of HuARD8, we used the 3'-end of the Ya5 consensus sequence to simulate the missing portion of the *Alu* RNA transcript. This diagram indicates that 11 bases at the 3'-end of the reconstructed HuARD8 RNA transcript are complementary within the first 25 bases downstream of the nick site.

Insertion Frequency and Polymorphism of *Alu* Retrotransposition-Mediated Deletion Events *in Vivo*

We determined that the human genome has suffered approximately 1.1 times more ARD events than has the chimpanzee. The direction of this adjusted *Alu* insertion ratio agrees with other comparisons of human and chimpanzee sequence data (Hedges *et al.* 2004, Liu *et al.* 2003), although it is somewhat lower than the insertion ratios of 1.8-2.0 detected in those studies. However, our program specifically searched for rare *Alu* retrotransposition-mediated genomic deletion events, so we would not necessarily expect to fully replicate results gathered from larger datasets. It is likely then, that our small data set in human did not fully capture the true level of polymorphism associated with these *Alu* element insertions, which would lead to a lower adjusted *Alu* insertion ratio. This bias would occur because sampling a single genome misses ~ 50% of the polymorphic insertion events that are present in the species as a whole (Hedges *et al.* 2004).

Although our chimpanzee sample size was smaller than that for human, we still obtained chimpanzee *Alu* insertion polymorphism levels consistent with other published studies (Hedges *et al.* 2004, Yu *et al.* 2003). By comparing the chimpanzee polymorphism rate to that of human, we determine chimpanzees to be three times more diverse than humans, in terms of retrotransposition-mediated deletion events. However, this comparison of polymorphism is skewed upwards by the low level of human *Alu* insertion polymorphism captured in our data.

The Rate of Retrotransposition-Mediated Deletions in Primate Genomes

We estimate that 0.28% (14 ARD events/5000 total chimpanzee-specific *Alu* insertion events; 0.38%, if adjusted for single genome sampling) of all *Alu* insertions in chimpanzee are non-typical and involve deletions of genomic material during retrotransposition. The rate of *Alu* retrotransposition-mediated deletion in humans is about 0.21% (19 ARD events/9000 total

human-specific *Alu* insertion events; 0.23%, if adjusted). For each species, the total number of lineage-specific *Alu* elements is based on a previous study (Hedges *et al.* 2004).

The estimated frequencies of retrotransposition-mediated deletion in our data are lower than previously published reports of between 0.8% and 8% (Gilbert *et al.* 2002, Salem *et al.* 2003, Symer *et al.* 2002). However, those studies generated biased estimates of retrotransposition-mediated deletion frequency in native genomes by using retrotransposition assays in cell culture from L1 element integrations (Gilbert *et al.* 2002, Kazazian *et al.* 2002, Symer *et al.* 2002), or by exclusively studying one or two small *Alu* subfamilies (Salem *et al.* 2003). These biases are outlined as follows. First, cell culture assays do not assess the viability of cells suffering the effect of large deletions. Second, the effect of natural selection on the afflicted genome is essentially ignored under experimental conditions, thereby skewing estimates of deletion event frequency in naturally occurring genomes. Third, cells grown in culture may suffer from genomic repair insufficiencies that provide many more opportunities for mobile element integration and genomic deletion. Finally, deletion events drawn from small subfamilies of *Alu* elements rather than from the entire *Alu* family of elements might provide unrepresentative frequency estimations. The genome-wide search in this study provides a relatively unbiased estimate of tolerable *Alu* retrotransposition-mediated deletion in primate genomes.

The Size of Deleted Sequence *in Vivo*

It is intriguing that human deletions are approximately 400 bp larger on average per deletion event than those found in chimpanzee. However, there are no known mechanisms to account for this consistent disparity. In any event, the largest deletions retrieved from the genome sequence comparison accounted for 1556 (human) and 210 (chimpanzee) nucleotides.

These deletions are small in comparison to those detected by L1 retrotransposition assays in HeLa cells in prior studies (Gilbert *et al.* 2002, Symer *et al.* 2002) which found deletions of up to 11,000 bp (and even 70,000 bp, empirically unconfirmed) that were presumably generated upon genomic integration of LINE cDNA transcripts. Whether such massive deletions are tolerable at the organismal level can only be determined by examining existing genomes, and our data suggest that they are not. Further studies to investigate whether human-specific L1 retrotransposition-mediated deletion events *in vivo* are smaller than those found *in vitro* will be informative.

Different Mechanisms of *Alu* Retrotransposition-Mediated Deletion: L1 EN-Dependent Retrotransposition-Mediated Deletion

The *Alu* insertions recovered during our study possess features uncommon to typical *Alu* elements, including the absence of surrounding TSD sequences and unusual target site preference. Experimental retrotransposition assays have documented similar characteristics within deletion-producing L1 element integrations (Gilbert *et al.* 2002, Morrish *et al.* 2002, Symer *et al.* 2002). From these *in vitro* studies, two putative mechanisms were put forward to explain the unique hallmarks of retrotransposition-mediated deletion. The first mechanism presumes that slight variations in L1 EN nicking can account for the absence of TSDs in addition to the insertion site deletions (Gilbert *et al.* 2002, Symer *et al.* 2002). The authors proposed that L1 EN sometimes nicks the second strand a few bp to the left of its initial nick site on the bottom strand, creating a substrate for exonuclease 5'-3' digestion at the target site. L1 EN-dependent nicking is evident in the datasets of (Gilbert *et al.* 2002) and (Symer *et al.* 2002) through L1 integration site preferences for sequences such 3'- A/TTTT. Our data suggest that L1 EN-dependent retrotransposition-mediated deletion, as determined through analysis of integration

site preference, may account for 25% of the combined ARD events in native human and chimpanzee genomes.

L1 EN-Independent Retrotransposition-Mediated Deletion

In contrast to the studies by (Gilbert *et al.* 2002) and (Symer *et al.* 2002), 75% of the ARD events in our data did not integrate at typical AT-rich L1 cleavage sites. This result provides an argument for the existence of an L1 EN-independent integration mechanism for *Alu* elements, similar to that previously suggested for L1 (Morrish *et al.* 2002). In this second model of retrotransposition-mediated deletion, it is likely that reverse transcriptase exploits existing breaks in the genome for TPRT initiation, not depending on L1 EN for the initial nick. The 3' overhangs are presumably created prior to host repair of the lesion, generating the characteristic target site deletion. Thus, it appears that, similar to L1 elements, *Alu* repeats may be able to facilitate the patching of lesions in the genome. Whether EN-free insertion indicates a true function of retroelements or just a fortuitous portal into the genome is unknown. Regardless, confirmation of the L1 EN-independent integration of *Alu* elements requires further investigation using cell culture-based *Alu* retrotransposition assays (Dewannieux *et al.* 2003) within DNA repair-deficient cells (Morrish *et al.* 2002).

Promiscuous TPRT: A New Model for *Alu* Retrotransposition-Mediated Deletion

Here we introduce a new mechanism to explain the unique characteristics associated with *Alu* retrotransposition-mediated deletion events (Fig 3.6). The alternative priming system, promiscuous Target Primed Reverse Transcription (pTPRT), is named after the promiscuous initiation of reverse transcription from sites downstream of genomic breaks. In this model, genomic breaks lead to the unwinding of the double DNA strand, binding of *Alu* RNA transcript at a downstream homologous region and initiation of reverse transcription. Removal of the

unwound DNA strand may be resolved by mechanical force or through enzymatic degradation. This particular mechanism can account for the integration of elements at non-canonical sites without TSDs, in addition to the generation of target site deletions. However, the exact means by which the second strand breaks and the lesion is resolved are still unknown factors in this model.

Internal Priming of *Alu* Elements

We recovered one example of a 3'-truncated *Alu* repeat element (HuARD8) in the human dataset. Similar sequence hallmarks have been attributed to the mechanism of internal priming and were previously documented within L1 element *in vitro* assays and in the human genome sequence (Morrish *et al.* 2002, Ovchinnikov *et al.* 2001). We determined that internal priming is consistent with the sequence hallmarks of HuARD8; further, regions of homology existed between the site of integration and the 3'-end of the simulated *Alu* Ya5 transcript, making internal priming possible. Although the primer binding site was not 100% complementary to the RNA transcript, empirical evidence suggests that initiation of cDNA synthesis does occur, if less efficiently, with RNA transcripts having low homology to the site of integration (Chambeyron *et al.* 2002, Luan *et al.* 1995). Hence, we believe this study provides the first published analysis of internal priming in the reverse transcription of an *Alu* repeat element.

The mechanism of internal priming is a potential alternative to the classical L1 EN-independent integration presented earlier. *Alu* and L1s do not require L1 EN to nick at AT-rich sites because the RNA transcript can bind internally at the site of genomic breaks, even without 100% homology. Although this mechanism is rarely exploited (less than 0.1% of events), it represents an effective way for *Alu* elements to enter the genome by DNA breaks.

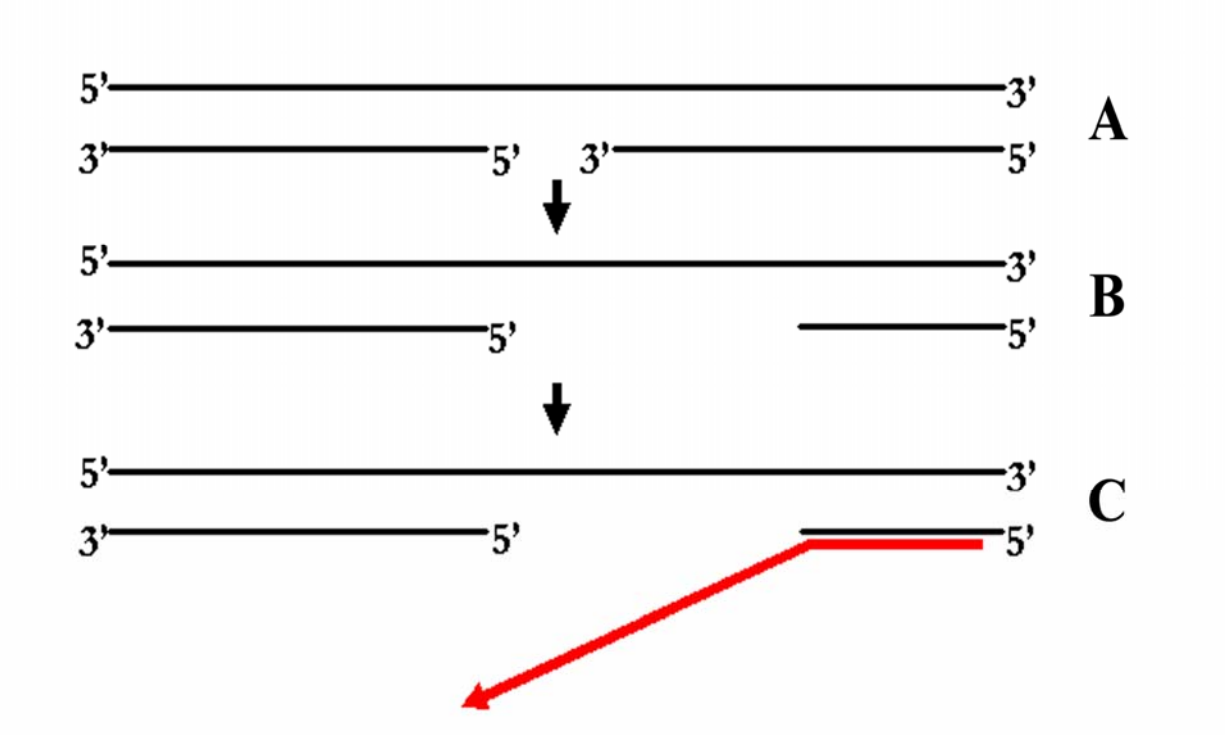

Figure 3.6 Model of Genomic Deletion Mediated by Promiscuous TPRT
In this model, genomic breaks lead to the unwinding of the double DNA strand (A). Removal of the unwound DNA strand may be resolved by mechanical force, or through enzymatic degradation (B). Following this, TPRT is initiated from binding sites downstream of the initial break (C). This particular mechanism can account for the integration of elements at non-canonical sites without TSDs, in addition to the generation of target site deletions. However, the exact means by which the second strand breaks and the lesions are resolved are still unknown factors in this model.

Contribution of *Alu* Retrotransposition-Mediated Deletion to Primate Genomic Instability

We have provided the first genome-wide study to quantify the contribution of *Alu* retrotransposition-mediated deletion to the instability of the human and chimpanzee genomes, with an estimate of approximately 0.21-0.28% (0.23 – 0.38%, if adjusted for single genome sampling) of all *Alu* element integrations over the last 5 million years being responsible for target site genomic deletions. If we assume the occurrence of retrotransposition-mediated deletion has been constant throughout the evolution of all primate orders, approximately 2,520 to 3,360 (2760-4560, if adjusted) of all *Alu* insertion events (1.2 million) have eliminated around 687,926

to 917,280 bp (753,480-1,244,880, if adjusted) of DNA from primate genomes (based on the observed ARD rate data and a human-chimp average of 273 bp per deletion event). Even conservative amplification rates of one *Alu* insertion every 250 births (Deininger *et al.* 1993) suggest that retrotransposition-mediated deletion could induce significant future changes to the overall architecture of primate genomes.

Although only one *Alu* retrotransposition-mediated deletion event (*c-rel*) *a*ppears to have caused a coding difference between humans and chimpanzees over the last 5 million years of evolution, the potential contribution of ARD to primate genomic instability as a whole is undeniable. The true extent of collateral effects caused by *Alu* mobilization will require sequencing the genomes of representative members throughout the primate order.

References

Altschul, S. F., *et al.* (1990). "Basic local alignment search tool." J Mol Biol 215 (3): 403-10.

Bailey, J. A., *et al.* (2003). "An Alu transposition model for the origin and expansion of human segmental duplications." Am J Hum Genet 73 (4): 823-34.

Batzer, M. A. and Deininger, P. L. (2002). "Alu repeats and human genomic diversity." Nat. Rev. Genet. 3 (5): 370-9.

Boeke, J. D. (1997). "LINEs and Alus--the polyA connection." Nat. Genet. 16 (1): 6-7.

Callinan, P. A., *et al.* (2003). "Comprehensive analysis of Alu-associated diversity on the human sex chromosomes." Gene 317 (1-2): 103-10.

Carroll, M. L., *et al.* (2001). "Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity." J. Mol. Biol. 311 (1): 17-40.

Carter, A. B., *et al.* (2004). "Genome wide analysis of the human Yb lineage." Human Genomics 1 167-168.

Chambeyron, S., *et al.* (2002). "Tandem UAA repeats at the 3'-end of the transcript are essential for the precise initiation of reverse transcription of the I factor in Drosophila melanogaster." J Biol Chem 277 (20): 17877-82.

Chen, S. J., *et al.* (1989). "Structural alterations of the BCR and ABL genes in Ph1 positive acute leukemias with rearrangements in the BCR gene first intron: further evidence implicating Alu sequences in the chromosome translocation." Nucleic Acids Res 17 (19): 7631-42.

Deininger, P. L. and Batzer, M. A. (1999). "Alu repeats and human disease." Mol Genet Metab 67 (3): 183-93.

Deininger, P. L. and Batzer, M. A. (1993). "Evolution of retroposons." Evolutionary Biology 27 157-196.

Deininger, P. L., *et al.* (1992). "Master genes in mammalian repetitive DNA amplification." Trends Genet. 8 (9): 307-11.

Dewannieux, M., *et al.* (2003). "LINE-mediated retrotransposition of marked Alu sequences." Nat Genet 35 (1): 41-8.

Fischer, A., *et al.* (2004). "Evidence for a complex demographic history of chimpanzees." Mol Biol Evol 21 (5): 799-808.

Garber, R. K., *et al.* (in press). "The Alu Yc1 subfamily: sorting the wheat from the chaff." Cytogenetics and Genome Research

Gilbert, N., *et al.* (2002). "Genomic deletions created upon LINE-1 retrotransposition." Cell 110 (3): 315-25.

Han, K., Xing, J., *et al.* (In press). "Extended retrotranspositonal quiescence supports a "Back Seat Driver" model of Alu evolution." Genome Research

Hayakawa, T., *et al.* (2001). "Alu-mediated inactivation of the human CMP- N-acetylneuraminic acid hydroxylase gene." Proc. Natl. Acad. Sci. U S A 98 (20): 11399-404.

Hedges, D. J., *et al.* (2004). "Differential alu mobilization and polymorphism among the human and chimpanzee lineages." Genome Res 14 (6): 1068-75.

Iafrate, A. J., *et al.* (2004). "Detection of large-scale variation in the human genome." Nat Genet 36 (9): 949-51.

Kazazian, H. H., Jr. and Goodier, J. L. (2002). "LINE drive. retrotransposition and genome instability." Cell 110 (3): 277-80.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res 12 (4): 656-64.

Lander, E. S., *et al.* (2001). "Initial sequencing and analysis of the human genome." Nature 409 (6822): 860-921.

Liu, G., *et al.* (2003). "Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome." Genome Res 13 (3): 358-68.

Luan, D. D. and Eickbush, T. H. (1995). "RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element." Mol Cell Biol 15 (7): 3882-91.

McNeil, N. (2004). "AluElements: Repetitive DNA as Facilitators of Chromosomal Rearrangement." J Assoc Genet Technol 30 (2): 41-47.

Morrish, T. A., *et al.* (2002). "DNA repair mediated by endonuclease-independent LINE-1 retrotransposition." Nat Genet 31 (2): 159-65.

Otieno, A. C., *et al.* (2004). "Analysis of the human Alu Ya-lineage." J Mol Biol 342 (1): 109-18.

Ovchinnikov, I., *et al.* (2001). "Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion." Genome Res 11 (12): 2050-8.

Roy, A. M., *et al.* (1999). "Recently integrated human Alu repeats: finding needles in the haystack." Genetica 107 (1-3): 149-61.

Roy-Engel, A. M., *et al.* (2001). "Alu insertion polymorphisms for the study of human genomic diversity." Genetics 159 (1): 279-90.

Salem, A. H., *et al.* (2003). "Recently integrated Alu elements and human genomic diversity." Mol Biol Evol 20 (8): 1349-61.

Sanger, F., *et al.* (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A 74 (12): 5463-7.

Sebat, J., *et al.* (2004). "Large-scale copy number polymorphism in the human genome." Science 305 (5683): 525-8.

Sinnett, D., *et al.* (1992). "Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences." J. Mol. Biol. 226 (3): 689-706.

Symer, D. E., *et al.* (2002). "Human l1 retrotransposition is associated with genetic instability in vivo." Cell 110 (3): 327-38.

Xing, J., *et al.* (2003). "Comprehensive analysis of two Alu Yd subfamilies." J Mol Evol 57 Suppl 1 S76-89.

Yu, N., *et al.* (2003). "Low nucleotide diversity in chimpanzees and bonobos." Genetics 164 (4): 1511-8.

CHAPTER 4:

RETROTRANSPOSABLE ELEMENTS AND DISEASE

Transposable Elements in the Human Genome

Almost the entire human genome is ubiquitously littered with the skeletons of mobile elements, which all told, account for a staggering 45% of the sequence content (Lander *et al.* 2001). Mobile elements successfully accumulated in genomes during eukaryotic evolution and are grouped into one of two different classes: DNA transposons or retrotransposons.

DNA transposons constitute 3% of the human genome (Lander *et al.* 2001) and although they are represented by inactive fossils in humans, DNA transposons remain active in the genomes of plants, flies and bacteria (Kaminker *et al.* 2002, Kleckner 1981, Wessler 2001). Retrotransposons, on the other hand, are currently actively mobilizing within the human genome and comprise approximately 40% of the DNA sequence (Lander *et al.* 2001). Due to the current propagation of retrotransposons in humans, they will be the focus of this review.

Retrotransposons, by definition, mobilize via an RNA intermediate that is subsequently reverse transcribed into a cDNA copy using a mechanism termed Target Primed Reverse Transcription (TPRT) (Batzer *et al.* 2002). This copy and paste mechanism of mobilization results in the spread of retrotransposons to new genomic locations. Retrotransposable elements are categorized based on their ability to mobilize. Long INterspersed Elements (LINEs) are autonomous retrotransposons that encode the enzymatic machinery required for their propagation (Ostertag *et al.* 2001). Short INterspersed Elements (SINEs), such as *Alu*, and SVA (SINE/VNTR/*Alu*) elements, are non-autonomous and thus require the enzymatic machinery of LINE elements for retrotransposition (Boeke 1997, Ostertag *et al.* 2003).

Over the last quarter century, many ideas concerning the function of mobile elements have been put forth. Orgel and Crick were proponents of the idea that mobile elements served no function and resided as parasitic entities within the genome, without contributing to the

evolutionary well-being of the organism (Orgel *et al.* 1980). Others have hypothesized that mobile elements function as origins of replication (Jelinek *et al.* 1980), chromosomal band-aids (Morrish *et al.* 2002) and mediators of translational activation (Chu *et al.* 1998).

Despite disagreement over the function of mobile elements, they constitute an interesting source of human genomic variation and occasionally, disease. Here we present an overview of the contribution of mobile elements, in particular, retrotransposable elements, to genetic disease in *Homo sapiens.*

Autonomous Retrotransposons and Disease

Long INterspersed Elements (LINEs)

Computational analyses of the human genome have shown that L1 elements have reached a copy number in excess of 500,000 and comprise some 17% of the genomic sequence (Lander *et al.* 2001). Numerous studies indicate that some subclasses of L1 element are still actively expanding by retrotransposition in extant human genomes (Ostertag *et al.* 2001). Retrotranspositionally active L1 elements are approximately 6 kb in length, as shown in Figure 4.1. Evidence suggests that L1 elements have orchestrated large-scale alterations in the genomic architecture of human beings, as they are the major source of reverse transcriptase, upon which other retrotransposable elements and processed pseudogenes have amplified (Ostertag *et al.* 2001). As a result, L1 elements are both directly and indirectly responsible for the vast majority of retrotransposable element-derived variation and disease within the human genome. The propagation of L1 has resulted in disease-causing *de novo* insertions within genes, many of which disrupt exons or alter RNA splicing in the mutant alleles. In addition, the 500,000 L1 elements in the human genome provide long regions of sequence identity that represent numerous sites for unequal homologous recombination and mutation. Despite their vast numbers

and retrotransposition activity, L1 elements are directly responsible for less than 20% of all retrotransposable element-related human diseases, even though experimental evidence suggests that L1s demonstrate a *cis* preference for their own replication machinery, see review (Ostertag *et al.* 2001). The paucity of disease-causing L1 insertions may stem from L1 AT-rich insertion preference, essentially sidestepping the sensitive coding regions of the genome, or perhaps new L1 insertions are subject to appreciable amounts of negative selection because of their size. Additionally, distant L1 spacing may mean that recombination between L1 elements would induce fatal genetic damage and be eliminated. Due to the paucity of disease-causing L1 recombination events, we will not cover this particular mechanism here. Instead, we will focus on what is currently known concerning L1 retrotransposition, retrotransposition-mediated genomic deletion and 3' transduction and their contribution to human diseases.

L1 Retrotransposition

Newly inserted L1 elements have induced disease in sixteen separate documented cases and the vast majority of these elements belong to one of the youngest L1 subfamilies, termed Ta. The L1 Ta subfamily is approximately 2 million years old and shows a high level of polymorphism (insertion presence/absence) in diverse human populations (Myers *et al.* 2002).

In 2001, a comprehensive study of newly inserted L1 elements and related diseases was published (Ostertag *et al.* 2001). The data gathered in this study indicated that nine out of the thirteen disease-causing L1 insertions discovered up until that time disrupted sex-linked genes, namely Factor VIII, Dystrophin or CYBB (Ostertag *et al.* 2001). This observation suggests that some genes are hotspots for mobile element integration, or that the ensuing genic damage was easily detected due to their genomic position on the X chromosome, i.e. through ascertainment bias. Since the review in 2001 (Ostertag *et al.* 2001), three new cases of L1 induced X-linked

genetic disease have been discovered. The first case describes an L1 insertion into the RPS6KA3 gene causing Coffin-Lowry syndrome (van den Hurk *et al.* 2003). Second, a disruption of intronic splicing through an L1 insertion into the CHM gene causing choroideremia (Martinez-Garay *et al.* 2003), and finally, a case of hemophilia B induced by L1 disruption of the Factor IX gene (Mukherjee *et al.* 2004).

L1 disease-causing insertions have been mapped to both the exons and introns of genes. Most exonic L1 integrations are presumably lethal due to the introduction of premature stop codons and are likely eliminated from the population. However, nine instances of exonic integration have resulted in phenotypically tolerable diseases in humans. Some intronic L1 insertions may also be lethal, but some studies have documented the existence of tolerable intronic insertions (Ostertag *et al.* 2001). L1 elements have recently been shown to reduce mRNA transcript levels due to their presence within introns (Han *et al.* 2004). This phenomenon is related to the inefficiency of RNA polymerase II to transcribe through L1 elements (Han *et al.* 2004). Researchers suggest that L1 elements may act as "molecular rheostats" by directly altering gene expression in this way (Han *et al.* 2004). Another study also recently demonstrated that RNA polymerase II transcription of L1 elements is adversely affected due to multiple termination and polyadenylation signals along the length of the L1 element (Perepelitsa-Belancio *et al.* 2003). It was proposed that premature RNA polymerase II termination could be a way that L1 elements limit their damage to host genomes (Perepelitsa-Belancio *et al.* 2003). At the same time, it would also mean that the stalling of polymerase molecules along L1 sequence would increase the negative impact of L1 insertions into genes (Perepelitsa-Belancio *et al.* 2003). Intergenic insertions of L1 may also alter gene expression throughout the human genome. L1 elements possess one RNA polymerase II promoter on their sense strand and another on their

anti-sense strand that have been implicated in the enhancement of some genes (Factor IX and apolipoprotein Lp(a) genes) and in the formation of chimeric mRNA transcripts (Ostertag *et al.* 2001). Given the high insertion polymorphism levels of young L1 elements within the human genome, intronic and intergenic insertions could profoundly influence gene expression on both the individual and population level.

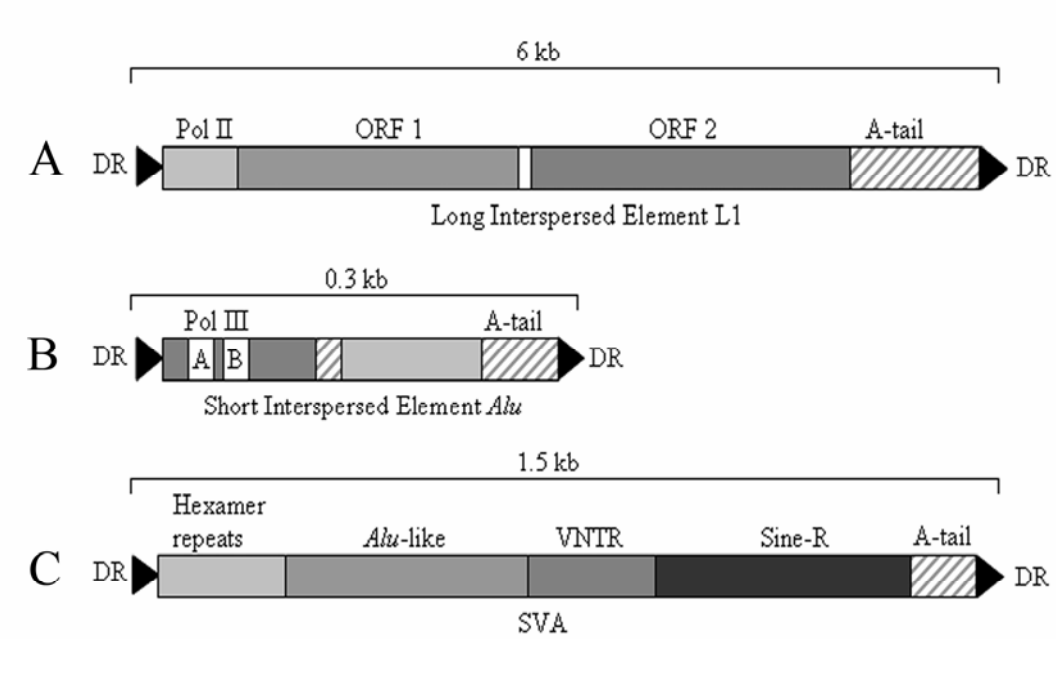

Figure 4.1 Active Retrotransposons within the Human Genome
A. **Long INterspersed Element L1.** L1s are approximately 6 kb long and possess a 5' UTR, in addition to a RNA polymerase II promoter. Full-length elements encode two open reading frames that produce a reverse transcriptase and endonuclease, as well as an RNA binding protein. Each L1 element has a 3' UTR, an oligo-dA tail and is flanked by direct repeat sequences (DR).
B. **Short INterspersed Element *Alu*.** *Alu* SINEs are approximately 300 bp long and comprise two arms separated by a middle A-rich tract. They possess an RNA polymerase III promoter (A and B box), in addition to a variable length oligo-dA rich tail. *Alu* elements are flanked by short direct repeats (DR).
C. **SVA.** Full-length SVA elements are approximately 1.5 kb long, and are composed of several repeat elements: a $CCCTC_n$ hexamer repeat, an anti-sense *Alu*, a variable number of tandem repeats, and a SINE-R element. SVA elements posses an oligo dA-rich tail and are flanked by short direct repeats (DR).
***** not drawn to scale

L1 Retrotransposition-Mediated Deletion

L1 retrotransposition-mediated deletion was first reported in 2002, where L1 integrations within cultured human cells resulted in target site deletions spanning from 1 bp to 70,000 bp at a rate of about 10% (Gilbert *et al.* 2002, Kazazian *et al.* 2002, Symer *et al.* 2002). These studies hinted at the vast impact that L1 retrotransposition-mediated deletion may have had on primate genomes. If 10% of the L1 retrotranspositions induced deletions, then over 5,000 L1 retrotranspositions would be responsible for eliminating megabases of primate genomic DNA.

Retrotransposition events that resulted in deleted target site DNA were found to possess atypical characteristics, including a lack of target site duplications (TSDs), non-canonical L1 EN (endonuclease) nick sites and sometimes the absence of an oligo-dA rich tail, see (Gilbert *et al.* 2002, Morrish *et al.* 2002, Symer *et al.* 2002). Researchers proposed two models, based on evidence from *in vitro* retrotransposition studies, to help explain the mechanism for the insertion-deletion events. The first model proposed that L1 EN nicking variation on the top strand could account for TSD-less L1 element structure, in addition to genomic deletion at the site of insertion (Gilbert *et al.* 2002, Morrish *et al.* 2002, Symer *et al.* 2002). The second mechanism suggested that L1 reverse transcriptase could initiate TPRT from existing breaks in the genome, not depending on L1 EN for the initial nick (Morrish *et al.* 2002). Recently, a third model was formulated to explain the mechanism of retrotransposition-mediated deletion, named promiscuous TPRT (pTPRT) (Callinan *et al.* In Press.). This model states that a retrotransposable element RNA transcript may hybridize to a region of genomic DNA downstream of a genomic break in order to initiate TPRT. The displaced single stranded DNA is removed through enzymatic degradation or by mechanical force, in order to create the target site deletion.

A recent survey of L1 disease-causing insertions reported two instances of retrotransposition-mediated deletion in humans: a 1 bp deletion in the DMD gene and another 6 bp deletion in the FCMD gene that resulted in Duchene muscular dystrophy and Fukuyama-type congenital muscular dystrophy, respectively (Kondo-Iida *et al.* 1999, Narita *et al.* 1993). In both cases, the disease phenotype resulted from the L1 element insertion, rather than through deletion of genomic sequence at the target site. These two cases are among only six other published *in vivo* examples of L1 retrotransposition-mediated deletion in the human genome to date (Ho *et al.* 2005, Vincent *et al.* 2003). Further research is underway at this time to determine the frequency of L1 retrotransposition-mediated deletion in the native human genome and its resultant impact on genomic instability and evolution.

L1-Mediated 3' Transduction

A decade ago, a mechanism was detected by which L1 alters the primate genome. It was termed 3' transduction (Holmes *et al.* 1994). The discovery of 3' transduction coincided with the insertion of L1 into the dystrophin gene, manifesting muscular dystrophy in a single human individual (Holmes *et al.* 1994). Since then, cell based studies have documented the ability of L1 elements to shuffle genomic DNA, including exons, using this mechanism, see (Moran *et al.* 1999). During 3' transduction, a read-through transcript of the L1 element transcribes flanking genomic material downstream by virtue of a weak L1 termination and poly-adenylation signal. Transduction of adjacent genomic DNA by L1 elements may result in the creation of new exons and in the alteration of gene expression through promoter and enhancer shuffling.

Computational analyses have indicated that L1-mediated transduction of genomic material may occur at a rate of one in every five L1 retrotransposition events and that approximately 1% of the human haploid genome may have arisen by this mechanism (Goodier *et*

al. 2000). In some instances, due to the severe truncation of L1 elements upon reverse transcription, it is possible that the transduced sequence will not reside adjacent to its L1 element thereby artificially reducing estimates of the impact that 3' transduction has had on the architecture of the human genome.

Non-Autonomous Retrotransposons and Disease

Alu Elements

The *Alu* family represents an enormously successful lineage of retrotransposons, whose origin and amplification coincided with the radiation of primates some 65 million years ago (Batzer *et al.* 2002). *Alu* elements are non-autonomous retrotransposons that mobilize in a copy and paste fashion. They are approximately 300 bp long and comprise two nearly identical arms separated by a middle A-rich tract, in addition to a 3' oligo dA-rich tail (Figure 4.1). Recent data suggest that only a fraction of *Alu* elements, termed source genes, are retrotranspositionally competent and responsible for producing over one million *Alu* copies within the primate order (Batzer *et al.* 2002). Although the exact characteristics of a source gene are unclear, *Alu* element age, RNA polymerase III promoter integrity and the length and homogeneity of the oligo-dA rich tail are considered major factors influencing retrotransposition potential (Batzer *et al.* 2002). *Alu* elements have continued to mobilize throughout the evolution of primates, as evidenced by human lineage-specific elements. These elements are absent from orthologous loci in non-human primates and exhibit high levels of polymorphism with respect to their insertion presence and absence in different human individuals. Recent estimates of *Alu* insertion numbers in the human lineage (~7000-9000) suggest that *Alu* elements are amplifying at a rate of one new insert approximately every 15-20 births, see (Deininger *et al.* 1993) for theory. Thus, it is not

surprising that recent *Alu* retrotransposition events have given rise to a number of human diseases.

Alu elements are known to create genetic instability and disease in a number of different ways. We will deal with each mechanism in turn and assess the prevalence, importance and resultant impact on the integrity of the human genome.

Alu Retrotransposition

From a review of current literature, 25 newly integrated *Alu* elements have been determined to induce disease states in human beings. Approximately eleven of the *Alu* elements integrated within introns and either caused partial intron retention within the mature mRNA through *Alu* exonization, or exon skipping (Ferlini *et al.* 1998, Ganguly *et al.* 2003, Knebelmann *et al.* 1995, Lev-Maor *et al.* 2003, Mitchell *et al.* 1991, Ostertag *et al.* 2001, Vervoort *et al.* 1998). A study by Lev-Maor *et al*. described the process of *Alu* exonization in a 2003 study, where the retention of anti-sense *Alu* elements within the mature mRNA transcript was attributed to the introduction of new splice sites from the *Alu* sequence (Lev-Maor *et al.* 2003). One recent study has proposed that exonized *Alu* elements are almost exclusively alternatively spliced, and that '*Alu*ternative' splicing is accountable for producing variable exonic transcripts in over 5% of genes (Kreahling *et al.* 2004). The retention of *Alu* elements within mRNA transcripts could contribute to subtle differences in gene expression between individuals and populations.

Alu repeats are rarely found within the coding regions of genes, as this may disrupt the gene's function. However, twelve exon insertion events have been described in the literature, see review (Ostertag *et al.* 2001). Since the publication of that review in 2001, two other studies have reported *Alu* integration into exons as the cause of genetic disease. In the first case, a young *Alu*Ya5 element inserted into codon 650 of the renal chloride channel gene, CLCN5,

resulting in Dent's disease, a cause of renal failure (Claverie-Martin *et al.* 2003). The second study reports a case of hemophilia A as a direct result of *Alu* integration into exon 14 of the Factor VIII gene (Sukarova *et al.* 2001). The total number of *Alu* retrotransposition insertions (both intronic and exonic) contributing to disease phenotypes within the human lineage equals 25. The total number of mutations in the Human Mutation Genetic Database (http://archive.uwcm.ac.uk/uwcm/mg/hgmd0.html) currently exceeds 44,000, as of January 2005). Therefore, *Alu* element insertional disruption accounts for 0.05% of all human mutations. However, only non-lethal mutations that cause observable phenotypes will be captured by this statistic. *Alu* insertions that are lethal and those that cause only mild phenotypes will be missed and thereby underestimate the true number of detrimental *Alu* insertions.

Alu-Alu Recombination

Alu-Alu unequal homologous recombination usually involves crossover between evolutionarily older elements within the genome, see (Deininger *et al.* 1999). *Alu* elements appear to possess particular characteristics that make them prone to recombination. These are: (1) the relatively close proximity of *Alu* elements within the genome, making most recombination events tolerable. (2) The sequence identity of *Alu* elements (greater than 75%, on average), which promotes efficient base pairing during crossover. (3) The vast number of *Alu* elements that create numerous identical DNA stretches, increasing the probability for recombination. (4) A chi-like motif within the *Alu* sequence that may stimulate recombination. Since 1999, approximately 25 new *Alu-Alu* recombination events have been linked to human disease. This makes the updated contribution of *Alu-Alu* recombination (both germline and somatic) to human genetic disease 0.17% (74/44,000).

Alu elements have also been linked to the presence of gene-rich segmental duplications within the human genome (Bailey *et al.* 2003). Given that 5-6% of the human genome sequence was created through segmental duplication events, *Alu-Alu* recombination may have contributed significantly to altered gene expression and species evolution (Bailey *et al.* 2003). In addition, mobile element recombination may occur in regions devoid of genes and still impact gene expression (Balemans *et al.* 2002). The fact that gene expression can be altered by the recombination of non-coding DNA is especially interesting since it is estimated that over 40 polymorphic *Alu-Alu* recombination events exist within humans (unpublished data). *Alu-Alu* recombination may therefore play a significant role in determining individual- and population-specific disease susceptibility.

Novel Mechanisms of *Alu*-Mediated Genomic Instability

Two novel mechanisms of *Alu*-associated genomic instability have recently been reported, *Alu* retrotransposition-mediated deletion (Callinan *et al.* In Press.) and gene conversion-mediated deletion (Salem *et al.* 2003). Both mechanisms involve the retrotransposition of a new *Alu* element coupled to the deletion of genomic material at the target integration site. *Alu* retrotransposition-mediated deletion involves the integration of an *Alu* cDNA transcript at a new site in the genome, similar to the retrotransposition-mediated deletion mechanism of L1. Gene conversion-mediated deletion involves the non-reciprocal conversion of an older *Alu* element into a younger *Alu* element. Due to the retrotransposition activity of *Alu* elements within humans over the last five million years, numerous chances have arisen for both types of deletion-inducing events.

A recent study of retrotransposition-mediated deletion determined that approximately 9,000 bases of human DNA have been deleted through this process (Callinan *et al.* In Press.). In

one instance, a 1002 bp deletion caused the functional loss of a retroviral transforming gene, *c-rel,* within the human lineage (Callinan *et al.* In Press.). Research indicates that *c-rel* may have important roles in regulating cell proliferation and differentiation (Bishop 1982). If the entire primate order is taken into account, approximately one megabase of DNA may have been deleted through *Alu* retrotransposition-mediated deletion since *Alu* elements evolved 65 million years ago.

Gene conversion-mediated deletion events have yet to be studied in such detail, although preliminary data suggest this mechanism could be as prevalent, if not more, than retrotransposition-mediated deletion (unpublished). The first published example of exonic disruption mediated by gene-conversion deletion occurred in the CMAH gene in humans. The deletion event encompassed a 92 bp exon encoding CMP-*N*-acetylneuraminic acid hydroxylase. The partial deletion of CMAH induced a biochemical difference in a sialic acid cell surface receptor between humans and non-human primates. Only two other examples of gene conversion-mediated deletion have been reported to date, and arise from the young *Alu*Yg6 and Yb8 subfamilies (Carter *et al.* 2004, Salem *et al.* 2003). Given the fact that *Alu* elements tend to reside in gene rich regions, gene conversion-mediated deletion by young *Alu* family members may be responsible for the deletion of other exonic or regulatory regions within the human genome.

SVA Elements

The SVA element is the least well-documented retrotransposon residing within the human genome. First reported in 1994, SVA elements are a composite retrotransposon consisting of a SINE-R element, a variable number of tandem repeats (VNTR) section and an *Alu* component, all contained within direct repeats (Figure 4.1), see (Ostertag *et al.* 2003). A

recent computational study of SVA elements indicated that there are approximately 1,750-3,500 SVA elements in the human haploid genome, substantially fewer than other retrotransposons such as *Alu* and L1. Low nucleotide sequence divergences within the SVA family suggest that their small number may be the result of their recent proliferation and origin, rather than low retrotranspositional activity. SVA retrotransposition has been verified from studies documenting their involvement in the induction of disease states. Previous research has revealed the presence of a SVA-mediated transduction within the α-spectrin gene (SPTA1) (Ostertag *et al.* 2003). Two other cases of disease-causing SVA insertions have also been reported. The first describes a SVA insertion into an intron of the btk gene, resulting in immunodeficiency X-linked agammaglobulinemia (XLA)(Ostertag *et al.* 2003). The second case was reported as a cause for Fukuyama-type congenital muscular dystrophy, following disruption of the futukin gene, see review (Ostertag *et al.* 2003).

Collectively, L1, *Alu* and SVA retrotransposable elements are responsible for 0.27% (118/44,000) of all human mutations discovered to date. They introduce genetic variation, and disease, on occasion, to human beings via an array of interesting mechanisms. Although researchers in the field of human genetics have explored the major mutational mechanisms of retrotransposable elements, their overall contribution to genomic diversity remains to be quantified.

References

Bailey, J. A.*, et al.* (2003). "An Alu transposition model for the origin and expansion of human segmental duplications." Am J Hum Genet 73 (4): 823-34.

Balemans, W.*, et al.* (2002). "Identification of a 52 kb deletion downstream of the SOST gene in patients with van Buchem disease." J Med Genet 39 (2): 91-7.

Batzer, M. A. and Deininger, P. L. (2002). "Alu repeats and human genomic diversity." Nat Rev Genet 3 (5): 370-9.

Boeke, J. D. (1997). "LINEs and Alus--the polyA connection." Nat Genet 16 (1): 6-7.

Callinan, P. A., *et al.* (In Press). "Alu retrotransposition-mediated deletion." Journal of Molecular Biology.

Carter, A. B., *et al.* (2004). "Genome wide analysis of the human Yb lineage." Human Genomics 1 167-168.

Chu, W. M., *et al.* (1998). "Potential Alu function: regulation of the activity of double-stranded RNA-activated kinase PKR." Mol Cell Biol 18 (1): 58-68.

Claverie-Martin, F., *et al.* (2003). "De novo insertion of an Alu sequence in the coding region of the CLCN5 gene results in Dent's disease." Hum Genet 113 (6): 480-5.

Deininger, P. L. and Batzer, M. A. (1999). "Alu repeats and human disease." Mol Genet Metab 67 (3): 183-93.

Ferlini, A., *et al.* (1998). "A novel Alu-like element rearranged in the dystrophin gene causes a splicing mutation in a family with X-linked dilated cardiomyopathy." Am J Hum Genet 63 (2): 436-46.

Ganguly, A., *et al.* (2003). "Exon skipping caused by an intronic insertion of a young Alu Yb9 element leads to severe hemophilia A." Hum Genet 113 (4): 348-52.

Gilbert, N., *et al.* (2002). "Genomic deletions created upon LINE-1 retrotransposition." Cell 110 (3): 315-25.

Goodier, J. L., *et al.* (2000). "Transduction of 3'-flanking sequences is common in L1 retrotransposition." Hum Mol Genet 9 (4): 653-7.

Han, J. S., *et al.* (2004). "Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes." Nature 429 (6989): 268-74.

Ho, H. J., *et al.* (2005). "Straightening out the LINEs: LINE-1 orthologous loci." Genomics 85 (2): 201-7.

Holmes, S. E., *et al.* (1994). "A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion." Nat Genet 7 (2): 143-8.

Jelinek, W. R., *et al.* (1980). "Ubiquitous, interspersed repeated sequences in mammalian genomes." Proc Natl Acad Sci U S A 77 (3): 1398-402.

Kaminker, J. S., *et al.* (2002). "The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective." Genome Biol 3 (12): RESEARCH0084.

Kazazian, H. H., Jr. and Goodier, J. L. (2002). "LINE drive. retrotransposition and genome instability." Cell 110 (3): 277-80.

Kleckner, N. (1981). "Transposable elements in prokaryotes." Annu Rev Genet 15 341-404.

Knebelmann, B., *et al.* (1995). "Splice-mediated insertion of an Alu sequence in the COL4A3 mRNA causing autosomal recessive Alport syndrome." Hum Mol Genet 4 (4): 675-9.

Kondo-Iida, E., *et al.* (1999). "Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD)." Hum Mol Genet 8 (12): 2303-9.

Kreahling, J. and Graveley, B. R. (2004). "The origins and implications of Aluternative splicing." Trends Genet 20 (1): 1-4.

Lander, E. S., *et al.* (2001). "Initial sequencing and analysis of the human genome." Nature 409 (6822): 860-921.

Lev-Maor, G., *et al.* (2003). "The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons." Science 300 (5623): 1288-91.

Martinez-Garay, I., *et al.* (2003). "Intronic L1 insertion and F268S, novel mutations in RPS6KA3 (RSK2) causing Coffin-Lowry syndrome." Clin Genet 64 (6): 491-6.

Mitchell, G. A., *et al.* (1991). "Splice-mediated insertion of an Alu sequence inactivates ornithine delta-aminotransferase: a role for Alu elements in human mutation." Proc Natl Acad Sci U S A 88 (3): 815-9.

Moran, J. V., *et al.* (1999). "Exon shuffling by L1 retrotransposition." Science 283 (5407): 1530-4.

Morrish, T. A., *et al.* (2002). "DNA repair mediated by endonuclease-independent LINE-1 retrotransposition." Nat Genet 31 (2): 159-65.

Mukherjee, S., *et al.* (2004). "Molecular pathology of haemophilia B: identification of five novel mutations including a LINE 1 insertion in Indian patients." Haemophilia 10 (3): 259-63.

Myers, J. S., *et al.* (2002). "A comprehensive analysis of recently integrated human Ta L1 elements." Am J Hum Genet 71 (2): 312-26.

Narita, N., *et al.* (1993). "Insertion of a 5' truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy." J Clin Invest 91 (5): 1862-7.

Orgel, L. E. and Crick, F. H. (1980). "Selfish DNA: the ultimate parasite." Nature 284 (5757): 604-7.

Ostertag, E. M., *et al.* (2003). "SVA elements are nonautonomous retrotransposons that cause disease in humans." Am J Hum Genet 73 (6): 1444-51.

Ostertag, E. M. and Kazazian, H. H., Jr. (2001). "Biology of mammalian L1 retrotransposons." Annu Rev Genet 35 501-38.

Perepelitsa-Belancio, V. and Deininger, P. (2003). "RNA truncation by premature polyadenylation attenuates human mobile element activity." Nat Genet 35 (4): 363-6.

Salem, A. H., *et al.* (2003). "Recently integrated Alu elements and human genomic diversity." Mol Biol Evol 20 (8): 1349-61.

Sukarova, E., *et al.* (2001). "An Alu insert as the cause of a severe form of hemophilia A." Acta Haematol 106 (3): 126-9.

Symer, D. E., *et al.* (2002). "Human l1 retrotransposition is associated with genetic instability in vivo." Cell 110 (3): 327-38.

van den Hurk, J. A., *et al.* (2003). "Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon." Hum Genet 113 (3): 268-75.

Vervoort, R., *et al.* (1998). "A mutation (IVS8+0.6kbdelTC) creating a new donor splice site activates a cryptic exon in an Alu-element in intron 8 of the human beta-glucuronidase gene." Hum Genet 103 (6): 686-93.

Vincent, B. J., *et al.* (2003). "Following the LINEs: an analysis of primate genomic variation at human-specific LINE-1 insertion sites." Mol Biol Evol 20 (8): 1338-48.

Wessler, S. R. (2001). "Plant transposable elements. A hard act to follow." Plant Physiol 125 (1): 149-51.

CHAPTER 5:

CONCLUSIONS

Mobile elements contribute great dynamism to the genomes they inhabit by introducing genetic variation. An excellent example is the *Alu* element of the primate order, the most abundant retrotransposable element with over one million copies per haploid genome. Alongside other retrotransposable elements, L1 and SVA, *Alu* elements are useful as fossils for studies of diversity, species identification, and evolution due to their high polymorphism levels and homoplasy-free characteristics. The contribution of *Alu*, L1 and SVA elements to genetic instability within primate genomes is also of considerable interest given their proposed roles in disease causation.

Chapter 2 analyzed the *Alu* element distribution and diversity on the human sex chromosomes. Our analysis computationally ascertained 344 sex chromosome-specific *Alu* elements, 225 of which were empirically tested for insertion presence/absence by polymerase chain reaction. Our results showed that insertion bias on the human sex chromosomes was subfamily specific and not endemic to the young *Alu* element subfamilies studied as a whole. We concluded that the distribution of different classes of mobile elements on the sex chromosomes is the result of a number of complex processes such as mobilization mechanism and integration site preferences that are mobile element specific. We found that recently integrated *Alu* subfamily members on the X and Y chromosomes exhibited reduced polymorphism as compared to their autosomal counterparts. We determined that *Alu* element age did not contribute to the low polymorphism, but that lower effective population size and reduced recombination on the X and Y chromosomes could contribute to the polymorphism levels found. Our study has identified sixteen additional polymorphic sex-linked markers that will prove useful in future research studies of human identification, diversity and evolution.

The genomic instability detected in the genomes of human and common chimpanzee in chapter 3 introduces a new type of genetic variation mediated by mobile elements in primate genomes. Using computational methods supported by wet bench experimentation, it was determined that *in vivo Alu* retrotransposition-mediated deletion was responsible for 33 deletion events within human and chimpanzee over the last 5 million years of evolution. An extrapolation of the observed *Alu* retrotransposition rate 0.21-0.28% to the entire primate order suggests that during the course of primate evolution, *Alu* retrotransposition may have contributed to over 3000 deletion events, deleting approximately 900 kb of DNA in the process. We believe that *Alu* retrotransposition-mediated deletion could have influenced genome evolution and ultimately, speciation within the primate order.

As an extension of the study into *Alu* retrotransposition-mediated deletion in chapter 3, chapter 4 reviewed the current literature concerning retrotransposable elements and disease within humans. We estimated that 0.27% of all currently known human disease mutations were due to the activity of retrotransposons within human genomes. A number of different mechanisms by which genome alteration occurs were identified. It was concluded that although researchers in the field of human genetics have discovered many mutational mechanisms for retrotransposable elements, their contribution to genetic variation within humans is still being resolved.

Overall, it is clear that LINE, SINE and SVA elements are responsible for introducing a tremendous amount of genetic variation within the primate order. By detailing the location and distribution of these elements, we are able to assess the retrotransposition dynamics of mobile elements. Through analysis of their insertion polymorphism levels, we are able to build a picture of the population dynamics associated with the species in which they reside. Through utilization

of comparative genomics techniques, the variation that *Alu* elements have introduced into primate genomes through retrotranspositional activity can be elucidated.

APPENDIX A:

LETTERS OF PERMISSION

01 February 2005 Our ref: HG/HDN/FEB05/J007

Ms Pauline Callinan
Louisiana State University
Email: Pcalli1@lsu.edu

Dear Ms Callinan

GENE, Vol 317, No 1 – 2, 2003, pp 103 – 110, P A Callinan et al, "Comprehensive…"

As per your letter dated 25th January 2005, 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

Helen Gainford
Rights Manager

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

ELSEVIER

3 March 2005 Our ref: HG/QC/MAR05/J025

Pauline Callinan
Louisiana State University
pcalli1@lsu.edu

Dear Ms. Callinan

JOURNAL OF MOLECULAR BIOLOGY, In Press, 2005, Callinan et al, "Alu Retrotransposition-mediated…"

As per your email date 18 February 2005, 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

Helen Gainford
Rights Manager

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

APPENDIX B:

SUPPLEMENTARY DATA TO CHAPTER 2

Name	Accession	Location	5' Primer sequence (5'-3')	3' Primer sequence (5'-3')	AT[1]	Human Diversity[2]	Product Size	
							Filled	Empty
Ya5420	AC004823	chrX:116284524-116400496	AAACATTAGGCCACCCTTCC	GGCAGCATGTGGAGTATGG	63	FP	426	102
Ya5DP4	AC017047	chrX:4670075-4850396	AACACCTCTGATGTAGCTTATG	CTAGGCCACCATTAAGCCAA	55	LF	649	334
Ya5DP2	AC074035	chrX:2646878-2836432	GTAACCAACAGCCTGATTTTGA	GACCTGCCATTTTCTAAGAAGCTAT	60	FP	462	172
Ya5DP69	AF047825	chrX:129328529-129413663	AATAAATTGCTTGCATGGGG	TCACAGGAGCCACCTCTTCT	55	FP	500	182
Ya5NBC118	AC005913	chrX:29824239-29971362	AATACGTGTGTCTGTGTGTATATGTTT	TGCATACCTTCCCAGAGATAATG	60	FP	533	235
Ya5DP16	AL121577	chrX:36904840-37080370	CTGACTGCTATGTCACAGCTACTTC	GGGGATATGTGAATGTGTATATGTG	60	FP	454	176
Ya5DP92	AF002992	chrX:155813783-155917819	ACAGGAGTCCATGTCAAGGG	TCAGGGTTTATGATCCAGGC	55	FP	447	119
Ya5 491	U69730	chrX:9810906-9875672	ACATGAATGTGCCATTGGTT	CAAGAAGGCAGCTGTCCTAGA	55	IF	435	96
Ya5NBC103	AL034408	chrX:62513993-62643841	ACTCTCTCTCCTACATCACTGACTTCTC	GTAAGCTTTGAGTTCAGAGGACAGATA	58	FP	556	237
Ya5DP8	AC005859	chrX:11177501-11380379	AGAAAGGGCGCTTACACTGA	CCATAGCTTTACAGGGGTGC	55	FP	494	168
Ya5DP60	AL035067	chrX:110968801-111103018	AGGATTGGGTCTACTGTGCAA	GGAATTATCAAATGAAAAAGCCA	55	FP	460	131
Ya5DP3	AC023104	chrX:4095243-4260035	ATCTTGAGAATCTCTACCAC	TCCTCTGGATTTCAGGGTTG	55	HF	487	162
Ya5NBC66	AC006210	chrX:26126751-26312398	ATGGTAATTTCCCTCATTTGTCA	GTAATGTCCTCCATTGTTCATTTG	61	FP	448	115
Ya5DP10	AC009858	chrX:16660990-16840489	CAAAGCCCTCAGATACTGAAA	TTGGCCATTCATTTTCTTCC	55	FP	390	68
Ya5NBC362	AL050308	chrX:142956655-143169738	CAAGTTTGTTGGCATAGAGGTG	ATCAATCCAGGAGCCGTTTT	60	FP	506	187
Ya5a2DP1	AL035423	chrX:130859858-130999951	CACAACAAAGTACTGCAAAGAGT	CTTTGTTTTCTGATTTTGGAAGG	55	HF	939	615
Ya5DP91	AF274857	chrX:155080500-155220669	CACCTCCCCTTCCCTTAAAA	GGGGGAATAAAAATCTCCAGG	55	FP	472	150
Ya5NBC34	AL031575	chrX:28407821-28485259	CACTCTGATACTTATCTCTGTGCCTGTAT	TGAGAGACATCAAACCAGAAATCC	60	FP	494	150
Ya5NBC313	AL121823	chrX:89292879-89478034	CACTTGCCATTGACTCCAAA	GGCTGGGTTGTGTGAGTTCT	60	FP	481	174
Ya5DP74	AL390879	chrX:137836321-138008600	CAGAAGCACAGAGGAAAGGG	AACCTGCATTACGGGCTATG	55	FP	1040	716
Ya5DP65	AL512286	chrX:119941032-120032906	CAGGCTGACCACACAATCAT	GCTACAAGGGAAAACTGGCA	55	FP	456	159

(table cont.)

Ya5DP15	AL451103	chrX:34868434-35043817	CAGGCTTGCACAAATATCCA	TTATATGAAGCACATTGAAGAAATG	55	FP	445	139
Ya5NBC326	AL133500	chrX:70223216-70424625	CCAAGAGACCACTTCCTATTTCA	AATGGGGGAGAGGACAGTCT	60	FP	539	216
Ya5 489	Z81364	chrX:130766117-130842210	CCATTATGACCAGTTGTGTGTTG	CCGGCCAAAAGCATTGTA	55	FP	433	115
Ya5 467	Z92844	chrX:42519788-42671585	CCCCTCCTCAGTTTTTGGAT	GGCTTAATAGCCAAGAGAGTGC	60	FP	400	85
Ya5 417	AF067122	chrX:155561893-155628434	CCTTCCCATAAACCCACTGA	CCAAAATTTGCTCCATGTTG	55	FP	441	121
Ya5NBC344	AL109853	chrX:132277087-132383551	CGTGAGAAAGCATAGGCAAC	TCCTTTCCTTATGCCTGCAA	60	FP	472	158
Ya5DP13	AC004470	chrX:21230949-21438905	CTATAGAGCCAAGCCTGATACTCTG	GTATGGGGAATGTGACAAGGAG	60	HF	430	141
Ya5DP18	AF241732	chrX:38416627-38459556	CTCAGTGTTCCCTCCTCTGG	ATGCGCTATGTCTTTTTGGG	55	FP	879	554
Ya5NBC80	AL590410	chrX:54568403-54757014	CTCTCCTGTGTCCATACTTCTT	CTGGCATGGAGATTTCTTAC	60	FP	368	47
Ya5DP88	AC005731	chrX:151553784-151697727	CTGAACCAAACTGGAAGGGA	GATTCACGTTGCACTTTTACCA	55	FP	490	175
Ya5DP5	AC019219	chrX:6134097-6314114	TATATGGGTAAAGATCCAAAGCAAGG	AGAATAATGCCTTAGCATTCAGCAG	60	FP	293	115
Ya5DP62	AL049591	chrX:114555491-114677890	GAATGAATGCAATGCCTAAGGT	AACCTATCTAGGGAGACCAGCAG	60	HF	410	115
Ya5DP77	AL356785	chrX:140674109-140839680	GAAGGATGATCTCTCCTTAC	TGCAAGGAGAGTTGGCATAA	55	HF	620	298
Ya5DP86	AL109654	chrX:148555591-148737740	GAGTAGTGTACATGAGGGGTTAT	AGGGCTGAGACAGTGTCTTC	55	FP	657	327
Ya5DP76	AL353788	chrX:138017665-138180403	GCAAATGTTCATTAAGAAAGCTGA	ATGGATTTTTGCTCTGCCC	55	FP	485	163
Ya5 455	AC002368	chrX:151258956-151583771	GCAACTTTCCCATGTTTTCC	TGGATGCAAGGTCTAAATTCG	55	FP	416	114
Ya5NBC170	Z94722	chrX:92120551-92227389	GCAAGACCTGTGTGTATGCTTAAAT	GAGAGTACACGAAAATACAGGCTTT	60	FP	521	195
Ya5 425	AL022166	chrX:54807015-54936240	GCACAGACAAGCTGCTCAAG	GAAGCCTGGCATGGAGATT	60	FP	431	110
Ya5DP53	AL359641	chrX:98554165-98729296	GCCAGGAACAGACAAGGTGT	TTGCCTTTTGGTGTTGTTCA	55	FP	490	177
Ya5DP40	AL031116	chrX:86290983-86441140	GCCTCATCCTGTACCATACTCC	TCCCACACTATTCTGATTTCTTCTT	55	FP	482	161
Ya5DP52	AL390027	chrX:98223595-98423785	GCCTGAGATGTGGGAGTAAAC	CAGCCTTCAAACTTGCACCT	55	FP	423	293
Ya5NBC37	AC002476	chrX:120184952-120332053	GCTTGAGGTTTTCATACTACTCTTATCTTT	ACTGTATAAGCATTTTCCTCTTTATCTTTC	60	IF	497	184
Ya5DP61	AL121878	chrX:114065698-114188586	GCTTTCTGCAGCAAAACTCA	CAGATGGCAAGAGCCTGAA	55	FP	684	370
Ya5NBC98	AL049591	chrX:114555491-114677890	TATAGCTAGTAAATGGTAGAGCCAGGA	CTGTCTAAGATAGTGATTGGACCTACTATG	55	HF	504	209
Ya5DP84	AL445258	chrX:147855595-148031077	GGAGCTGCAGGAGTTGTCTT	CCAGGAGCAGGAGAGAACAA	55	FP	496	173
Ya5 477	Z92844	chrX:42519788-42671585	GGCTTAATAGCCAAGAGAGTGC	AACCCCTCCTCAGTTTTTGG	55	FP	400	87
Ya5DP70	AL023799	chrX:130812222-130905926	GGGGAATGAGAGGGAAATGT	AAGACAGCCAAAATTCAGTTAAAAA	55	FP	1190	868

(table cont.)

Ya5DP12	AC017058	chrX:19068390-19241039	GGGTTGATTTAGTGGCCCTT	TCCTTTCAGATTTTCGTGGG	55	FP	374	59
Ya5DP97	AC011142	chrX:12380392-12557081	TACTATATCCCCCATGCCCA	ACTTGGTCCTCTCTCCAGCA	55	FP	1075	749
Ya5DP59	AL360224	chrX:109503420-109660581	TAGAGAATGAGGGTGGCTGG	TCGTGACCTTAGCACATGGA	55	FP	472	158
Ya5NBC99	AL031312	chrX:146122637-146208640	TATACACACACACACAGAGAATGACTG	CCTGACTCGAAAGTACTGTTTTCTAAG	55	FP	515	198
Ya5DP22	AL590223	chrX:47743014-47959685	TCTAAACCTGCCCTAGCTAGATACC	TCCTTTCTCAAAACTGCTTTCC	60	FP	516	190
Ya5DP56	Z70051	chrX:104660637-104705312	TGAAGATGTTTCTCTCCCCAG	AGTGGAAGAGAAAGGGTGGG	55	FP	487	374
Ya5DP68	AL391002	chrX:126496085-126581721	TGATTTCACTATGAAACCCACTC	TGAAGGACTCAAAATTTTCCAC	55	FP	405	89
Ya5DP66	AC002377	chrX:120825392-120967170	TGGACTGCTATCTCACGCTG	TTGGTTTTCTGGCAAGTTCC	55	FP	938	624
Ya5DP41	AL137015	chrX:86883045-86982571	TGGAGACATGAATACATTTTAGACA	CCAACAGATTTCACTTTTTGCTT	60	X/Y	464	149
Ya5DP83	AL445258	chrX:147855595-148031077	TGGATTAAATACAGGCAGAAAGC	TGCAGCAAAGATCTTCCAGA	55	FP	478	164
Ya5DP6	AC073533	chrX:6458416-6640471	TGGGTGTTTGCATCAAGAAA	GCAGGCAGAGAGGACAGGTA	55	FP	731	412
Ya5DP44	AC004072	chrX:90436734-90607391	TGTCATCTTTATCTGCCTTGGA	ACGGAGATTCTGCTTCAACAA	55	X/Y	398	89
Ya5 466	AC002377	chrX:120960081-121101859	TGTCTTACAACTCCCCACTCAA	CCTGGCTCTTCCAAGTTAGG	60	FP	426	94
Ya5DP34	AL359885	chrX:79179019-79255815	TTAGGTCACCTCTCCCTTGC	CAAGTGTCTGCAAAAAGGCA	55	FP	1131	800
Ya5DP82	AL512285	chrX:146753057-146823003	TTTAAAAACATAACCCAGTTGAAAA	CACCCATTAATTCACTACCCAA	55	FP	1084	785
Ya5DP54	AL355593	chrX:98735910-98903974	TTTAAAGAAAGCCTGTGATGGA	AAATGAATTGGCCCACCTTT	55	FP	493	178
Ya5DP57	Z83850	chrX:105136491-105269471	TTACCTCAACAGTGACATAACAGCA	ATAGTGAAGCAGAGAACTGTTGGTT	60	HF	652	349
Ya8BGK21	AC016678	chrY:18083142-18225923	AATATCCACCAAGAACAGAAGCTTTAG	AATCTTTGACTAGGCCCTGTAAGTT				
Yb8DP1	AC079824	chrX:29704853-29824238	TCACCAATTATCCTCCTCCA	CGAGATGAATAAACACTGCACA	60	FP	442	235
Yb8DP2	AL049643	chrX:32572391-32691085	TCCTTTTATAAATTGGACAGAAAGC	TTCAAATGTCCAGCCAATTG	60	IF	400	48
Yb8DP3	AC022212	chrX:38096933-38284245	TTGTATTCCAGGGATCAGGC	GGGAGCCTGGGATTTTAGAG	60	FP	465	111
Yb8DP4	AC091810	chrX:39109332-39209804	TGGACTCCCACTGAGATGTG	ACTCACCCGCTAATTGTGCT	60	FP	499	145
Yb8DP5	AL023875	chrX:41894031-42016355	CCTTAATTTTGTTTCCCGCA	TTCACAGCTGGATCAGTTCAA	60	FP	451	102
Yb8DP7	AL034370	chrX:43613478-43733422	AAATGGTGGAAAAGATGCCA	CCCATCACAACTGTACCCAA	60	FP	485	119
Yb8DP8	AF196779	chrX:49459890-49643885	GAACTTAGAGAGAGCTAGTC	GTGCATCTTAGTATGAACTC	62	FP	673	358
Yb8DP9	AC078991	chrX:3366309-3536127	GAGACAGAGGCTACATGTGA	AACAGCAAATGAAATCGCCT	60	FP	1039	692
Yb8DP12	Z82211	chrX:56385934-56518162	ATGGACATCTCTGGTACGGC	CTAATTCCCCTGGCTGCATA	55	FP	489	151

(table cont.)

Yb8DP13	AL158016	chrX:65925564-65996226	TAGGTTCATGAAGGCAAGGG	TGTCAATTAGAAGGCCTGGG	55	FP	479	258
Yb8DP18	Z98255	chrX:74382876-74552873	CAGTCTGTCTTCAGACCAGA	AGAAATGAATTAACGTGGC	62	FP	1026	626
Yb8DP22	AL358796	chrX:71193981-71539035	CTGGGGAAACAGACATAGTC	ACTTAGTGGACCTTCGTGGA	59	FP	727	485
Yb8DP25	AL591431	chrX:78222054-78373070	TGATGGGCATCACTGAAATC	CATTCTTAATGGGCCAATTTCT	60	FP	482	137
Yb8DP27	AL590031	chrX:78671333-78816485	TCATGCTGGAAAGGGCTATT	GCTTCCCACCTGAGCTAACA	60	FP	433	79
Yb8DP36	AL590043	chrX:94963848-95106616	AGTCAGTGACACCCACATGC	TGATGGAAGGATTTAAGCCAA	55	FP	500	142
Yb8DP38	AC003048	chrX:8164628-8205708	TACTGAGGCCATCGAGGAAC	CTCTCCTCACATCCCCGTAT	58	FP	491	145
Yb8DP39	AC002349	chrX:9399852-9559714	TGCAGATCTTATCAGCACATTG	ATTCATCCACCATCAGGGAA	55	FP	454	89
Yb8DP42	AC002449	chrX:113337879-113511645	GAAACCCAGTTTCACCATTTG	CAATGCATCTGTACCATGCTA	55	FP	670	318
Yb8DP43	AC005000	chrX:114817798-114925111	CCAAGGCAATCAATTTAGCC	TTCAAGATGCAGTCACTCGG	55	FP	897	544
Yb8DP44	AL357562	chrX:121846492-121975456	TTCATGTGGGCTTTTTGTGA	CAGCAAATTGTTCACAGTCCA	55	FP	471	123
Yb8DP45	AC002981	chrX:10814208-10967775	CCATCAATACATCGCTGGAA	TGTTCACCACCTTTCAACCA	62	FP	478	135
Yb8DP49	Ac002422	chrX:129115374-129275464	GACTAGGGGTTTGTGCCAGA	TCCCCCATTTCTGTTGTTGT	57	HF	459	138
Yb8DP51	AL138745	chrX:129973729-130197972	GCTTGCAACCTTACTGCCTC	GACAAAGCCTGAAGCCACTT	60	FP	414	68
Yb8DP52	AL022162	chrX:130258976-130259910	TGGGGGCACTTTACTAGGAT	CCACAGCTGGAGAACACTGA	60	FP	399	51
Yb8DP55	AL034400	chrX:133947057-134088818	GTGCTGCTGTAGCATTGCAT	GAAAGAACAGAGAACAGCCCA	60	FP	488	134
Yb8DP56	Z97196	chrX:134723484-134812365	AGACACCATCTGTGGGAAGG	ATTAAGGGCACTGTGCAACC	60	FP	461	120
Yb8DP58	AL390879	chrX:137999986-138172265	GTGGATGCCATTTTGGCTAC	TCCTTCATAGCCCGTAATGC	60	FP	494	161
Yb8DP59	AL022576	chrX:138442263-138579373	CTTGTGGGGACAACACTCCT	CTTCCTTCCACAGCCATTGT	60	FP	829	469
Yb8DP61	AL356785	chrX:140831401-140996972	GAGTAGCTACGTAAATACCC	TCCACACTTCATTCAAAGCC	59	FP	523	176
Yb8DP63	AL109653	chrX:147856085-148017425	CCCCTTCCTCTCACATAGCA	TTTATTCCCCCATTCCACAA	60	FP	1180	830
Yb8DP64	AC079383	chrX:12531947-12683222	CGTTTTCTATTTCCCACCACA	CCAACATTTTTCCTCCAAGG	55	FP	318	74
Yb8DP65	AC002524	chrX:13210194-13412733	CAGCTAGGCCTTGGAGATCA	TGCAAGCCAAATGAAAGAAA	55	FP	472	127
Yb8DP68	AF030876	chrX:157681205-157793960	CAAAGTCCTGTTGCGTACCTC	GCTGATGGCTACAACCCTGT	55	FP	953	630
Yb8DP70	AC078993	chrX:15756369-15970369	TTTGAATCAATATGTATATGGTGGA	CAGTTCCCATGACTTGGCTT	55	FP	437	71
Yb8DP76	AL592043.	chrX:33755847-33940359	GAGGCTAATATCAGCAAGCCA	TGTTTCAGCCAAAGAATGGA	60	FP	477	146
Yb8DP79	AL035088-AC016681	chrX:107155092-107301449, chrY:5852850-5921375	AGATTTCCAGAGGGAGCCAT	TTTCAACAGTCTTCTTTCGCA	60	X/Y	428	96

(table cont.)

Yb8DP80	AL137065	chrX:107787396-107906706	CCATGATCATTTCCCTGACC	CCTGTCTGTTCTGCTTCTTTGG	57	FP	458	126
Yb8DP81	AC008162	chrX:120517329-120638169	CAGTTTCCTGGGTCCTGTGT	CAAGGCTTCCAGCTTAGGAA	57	FP	460	128
Yb8NBC8	Z98950	chrX:143336947-143460502	AAGAAAACTGATGGGGAAAG	CCAACTAGAGAAACGGAGAA	60	FP	599	198
Yb8NBC30	Z95124	chrX:84348492-84423053	TTGCCTTGGATGGCATATCT	AAATGGCCGGAGTAAGTCCT	55	IF	497	194
Yb8NBC38	AC002367	chrX:27624355-27772954	CGAGAGAAAGGGGTAGAAAGC	AATGCCTTCCAAGGACATCTT	60	FP	480	311
Yb8NBC62	AL031368	chrX:28485260-28629149	TGCCACACATTGTTCTAGGC	TGCCAACTATTGGAGGAGATG	45	FP	548	307
Yb8NBC75	Z68328	chrX:104956504-105000946	CCCACTGTGTTTATTGTTCC	GCTAAAGTACCCAGACCAAG	60	FP	519	200
Yb8NBC102	AL049591	chrX:114555491-114677890	TATAGCTAGTAAATGGTAGAGCCAGGA	CTGTCTAAGATAGTGATTGGACCTACTATG	60	HF	504	209
Yb8NBC133	Z84470	chrX:74641008-74790533	GCCATTGATCCCACAGAAAT	GCTGTGAATTCGTTGGTCCT	55	FP	536	232
Yb8NBC170	AL109653	chrX:147856085-148017425	TCCCCAAAGAAGGAGAGACA	TTCCCCCATTCCACAATTTA	60	FP	599	275
Yb8NBC221	AL034370	chrX:43613478-43733422	AATTCAAGCCAATGAACCAC	TCAGTGCTCTGAAGAAGCTCA	60	FP	431	97
Yb8NBC239	AF031078	chrX:157681205-157793960	TTGCTGACAGATCAGGGATG	TCCCCCTTCAAACCTATTCC	55	FP	730	419
Yb8NBC242	AC002349	chrX:9399852-9559714	ATCCACCATCAGGGAATCAA	TGCAGATCTTATCAGCACATTG	60	FP	450	117
Yb8NBC246	AC002981	chrX:10814208-10967775	CACCACCTTTCAACCAGGAA	ATCGCTGGAATGTGGTTCTC	60	FP	464	149
Yb8NBC247	AC002366	chrX:10014142-10273343	GCAGCACAAAGTAGTGGTTGG	TGCACCCACTTGATATGCTT	60	FP	551	259
Yb8NBC256	Z73986	chrX:100506131-100636835	CCCACAATTTCCACTTCAGG	GCATTGCTTCCCTTCTATTTC	55	FP	503	24
Yb8NBC269	AC091810-AF241734	chrX:39109332-39209804, chrX:38989413-39109331	CACGCTTAACCTCTACCACCA	TGGACTCCCACTGAGATGTG	60	FP	587	261
Yb8NBC483	AC012078	chrX:88400703-88612982, chrY:3556128-3732799	GGCCAAGAGCATTCCAAAAT	GCCAATTGGTCAGGGTACAA	58	X/Y	744	422
Yb8NBC578	AL159988	chrX:146926640-147038418	TTTTTGCAGATGCTTCCCTA	CCCTTGATCCAGATGTGATG	55	IF	380	72
Yb8NBC594	AC087225	chrX:158577659-158680667	AGCAGGTGGTTAGGTCTTGG	CAGGGGGAGGGAACATTAAC	60	FP	428	103
Yb8NBC613	AL158201	chrX:66488109-66630063	GTCGCTTCACCTTGCACTTT	CAATCTGTGAAGGCTGAGGA	55	IF	459	124
Yb8NBC634	AL390840	chrX:92693201-92890811	AACAGAAAGGCATCATTTGC	GGGGGCATTTATTACTGCTT	55	IF	420	95
Yb9DP1	AL050305	chrX:32824774-32964031	TGACGACAAAGCACAAGGAC	TGGGGAGAATTTTACAAAACTAGG	60	FP	499	165
Yb9DP10	AC002477	chrX:119582373-119706467	CCAATTCCACAAAGGCAAAT	TTAGCTGCCTGACACGTCC	62	FP	1144	825
Yb9DP13	AF277315	chrX:158097366-158244593	ATGGAAACTGCACAGAGAGG	CTCTCTGGGCAGACCACG	62	FP	620	531
Yb9NBC251	AC002477	chrX:119582373-119706467	CGGCCCTGATATGTCTTTGA	TCCACAAAGGCAAATGGATA	60	FP	838	500
Yc1DP2	AL353136	chrX:64692940-64885444	GGCCTATATTGCTATCACGCA	TTTTCTCTCAGGTTCTCTGTAAACT	60	FP	1050	721

(table cont.)

Yc1DP4	AL357752	chrX:68485370-68664236	AAACATGGGAGGGAGGAAAG	GCTCAGAAACTCCCAACCAG	60	FP	486	318
Yc1DP5	AL121601	chrX:123991202-124124592	CAACCAGAGATCTTAAAATGTGA	TCAGCGTGAGAGCCCATATT	60	FP	452	330
Yc1DP7	AL031054	chrX:144887772-145086787	GACCCCAAAGGTTCAAGTCA	GCATGCCCACTAGCAGTGTA	60	FP	1072	731
Yc1DP8	AJ239323	chrX:50201890-50304742	CAATTTCCTGGCATTTGGAG	TTCAAGATGCAGTCACTCGG	60	FP	345	62
Yc1DP10	AJ239320	chrX:69939181-70231674	CACTTTTTCTTATTTGGCCCAG	ATGGGCAATTCAATGTTTCC	60	FP	428	65
Yc1DP11	Z75741	chrX:128441659-128443464	AACCTCACATTTTCCAAAGGTA	TCTTGCTTCCTGAGTCGGTT	60	FP	691	380
Yc1DP13	AL137013	chrX:73134712-73280970	AGGCCTCAAAGTTTAGGGGA	ATCAAAGGGGAATACTGGGG	60	FP	424	338
Yc1DP14	AL049643	chrX:32572391-32691085	CCACTGCAGGCAGGATTATT	GCATGCCTGATTCCACACTA	60	FP	480	314
Yc1DP16	Z86061	chrX:95243418-95361328	AGCATGCAAGGAAAGGGATA	TTCTCAGTTTCCAATCTTAGGGA	60	FP	486	134
Yc1DP18	Z98046	chrX:53964263-54042043	CAAGGTTTGGGTTCTGCTGT	CATGGACACAGTGGTGAAGG	60	FP	412	81
Yc1DP21	AL589872	chrX:53255422-53447504	CTTGAAGCTGCTCAGTAAGG	TAGCCATATCCACACA	60	FP	567	240
Yc1DP22	AL049562	chrX:128109596-128200796	GCAAAACTTTGCGCTAATCC	ATGGGAAGCTTTCCCTGACT	60	FP	746	415
Yc1DP24	AL158819	chrX:54387419-54562331	GGGGAAATGGGCCTAGTAAA	AATCACCTTAACGCCACAGC	60	FP	470	142
Yc1DP26	AL096861	chrX:150067750-150197435	TGCAATAAAGAGTGTTCCTCTCC	CCCAAACTTGGTAGGTGAAAA	60	FP	482	147
Yc1DP27	Z83823	chrX:125012174-125121452	TCACGTCTCTCCTTTGCTCA	CTCTGGAAGCCTGCTATTGG	60	FP	1072	775
Yc1DP30	AL591431	chrX:78222054-78373070	TGCCTTACCCAATACACATTT	AAGGCAAAAGTCCATAAAGCA	60	FP	498	172
Yc1DP32	AL365179	chrX:61340404-61521254	CCAAAGGAGGTGGCTACTCA	GCACCCTGGTGAGAAATTGT	60	FP	422	73
Yc1DP34	AL356317	chrX:62409559-62514092	TGGATCTGCTATCAGAATGGAC	TTTGTGCAAAATAGGACCCTT	60	FP	499	194
Yc1DP35	AL031319	chrX:109958712-110057481	GCCTTGGGCTGCTATCATAA	GGGCAGAATAACGCAAGATT	60	FP	500	185
Yc1DP38	AL359854	chrX:61176831-61340403	CCAAAGGAGGTGGCTACTCA	GCACCCTGGTGAGAAATTGT	60	FP	423	113
Yc1DP39	AC073614	chrX:25176210-25306010	CCAACAGACAGCTTTCCACA	CAAGTCGAGGTTCTCCCTCA	60	FP	498	200
Yd3JX170	AC005000	chrX:114817798-114925111	GTGATTGCTACTGCTTTTTGCTT	ACCTGATGAACATTTTAGGAACC	60	FP	570	255
Yd3JX757	AL139396	chrX:52597320-52775770	CATTAGAAATCAGAATGGCTTCG	CTTGGTTTATTCCTTTGCTATGC	60	FP	549	250
Yd3JX437	AL034412	chrX:46070143-46177191	TGGTGTACCTTAGTCCAAAGACC	TTTGCATCTCAGAACTTTTTCCT	60	IF	547	235
Yd3JX545	U73479	chrX:20177044-20213072	AGGTTATGAAAGGGTCTGCTTTT	GATATTTGGACACACACACCTAAA	60	FP	680	355
Yd3JXD75	AJ239320	chrX:69939181-70231674	TGTACTTGCCCCATCTTCTGTAT	TATTCTGAAAATCTTGGGGGTGT	60	FP	546	226
Yd6JX284	AL591591	chrX:32998640-33102756	TTTCCTGATGGAAGCAGTGTATT	TGTTAGCATAATTGATCCCAAAAT	60	FP	517	200

(table cont.)

Yd6JX56	AC079173	chrX:3673291-3838308	ATACTTACCATTGCCTCGTCCTT	ATGTCATGATCGGCTAGTTCTTG	60	FP	530	216
Ya5a2AD3	AC006371	chrY:15065526-15267679	TGGGGAAATCGATGATTTAAGA	AAGACAACGCACAATACCTTTGA	55	X/Y	421	117
Ya5AD585	AC006983-AC024067	chrY:24548626-24728770,chrY:27613358-27721503	TAAAATATTGCAAGGGGATGA	CCAGGTCTGTGTCTTATTTTCTTT	56	FP	867	536
Ya5AD586	AC006983-AC006338-AC010088-AC025735	chrY:24548626-24728770,chrY:26134406-26321043,chrY:24321000-24428486,chrY:25944031-26029302	ACGCAGAACCTGAAATTGTGATT	ACCATGCATAAATAGTGCCAACT	60	FP	524	181
Ya5AD588	AC026061	chrY:22174780-22194121	TGAGCGTCTAATGTGTTAATGAAA	CAAATACTTCAGCCTTGTCAAGAA	60	FP	500	193
Ya5AD589	AC010086	chrY:22595725-22766459	TGCACATACTGCTATTGATG	TGGCTATGCTTTCTTCATCT	55	FP	549	232
Ya5AD591	AC073893	chrY:25211889-25276138	TTGTATTAAAGCCCGTAAAATGG	AAGAATTATCTAGGACAGCTTTGG	55	FP	544	223
Ya5AD592	AC007965-AC007359-AC016752-AC008175	chrY:24895138-25061373,chrY:23324934-23425360,chrY:24895138-25061373,chrY:23742819-23947855	CATCGTGATGGTCTAGATTTCTTT	TTAAGGCATCGGATTCTTTCT	55	X/Y	685	268
Ya5AD593	AC024067	chrY:27613358-27721503	AATTAAAAGCACCCCCAAGA	CTCACCTTCTCTGCTTAACAAAA	60	FP	543	227
Ya5AD594	AC010153-AC016728-AC006983-AC006338-AC010088-	chrY:25840084-25944030,chrY:26321044-26472895,chrY:24548626-24728770,chrY:26134406-26321043,chrY:24321000-24428486,	TGTTTCAGAGAGGACAGAAA	AGTGATTGCCTTGACATAGT	55	X/Y	459	148
Ya5AD595	AC025735	chrY:25944031-26029302	ACGCAGAACCTGAAATTGTGATT	AACCATGCATAAATAGTGCCAAC	60	X/Y	524	182
Ya5AD597	AC023274-AC006328	chrY:25351695-25489176,chrY:26636925-26814493	GTTTGCTCAAGCCCATTAAA	TAAATGTATCCTGGCACCAT	55	X/Y	434	115
Ya5AD598	AC023274-AC007562	chrY:25351695-25489176,chrY:26814494-26951370	AACGCCAAACACAATGACAA	TTTGGCTGCATGAATGTGTT	55	X/Y	592	277
Ya5AD600	AC010094-AC002509	chrY:3732800-3851035,chrX:88482028-88624126	AAAACAGCACAACGTTTTAT	TCTCAAAGCTCTAGGTTAGTTGA	60	FP	396	293
Ya5AD601	AC009491	chrY:8539647-8680380	AGTGGAAGCCATAAAACAAA	ACATAATCCAAGCATGATCC	60	FP	398	299
Ya5AD602	AC006040	chrY:2500001-2686304	CCCAAACCAAAACTGTTACT	TTTGTTCCTGCAGTCAATCT	60	FP	492	291
Ya5AD603	AC006376	chrY:14752949-14924755	TGAGGGAAGAACATTAAGGCATA	AGGTAAGCCAGATCCAGTTTTTA	60	FP	508	189
Ya5AD604	AC010723	chrY:15580278-15754497	AGCTGAAAGAGGACATCAAT	TGATATTCACCAGGGATTCT	55	FP	489	159
Ya5AD606	AC019060	chrY:4618247-4734841	TCTAAGGCAAACATGAGCTT	GAACATCTTAGAGCCTTCAAA	55	X/Y	1038	374
Ya5AD607	AC010977-AC009491-AL121881	chrY:5716765-5852849,chrY:8539647-8680380,chrX:142771104-142956654	AACATCAATTTGAAAACCTAGA	TGAGGAACAAAGGTTTTGAC	55	X/Y	472	141
Ya5AD608	Z95703	chrX:143720946-143847097	ATGAAAACTGTTCAGGGAGATATT	TGGTTAATATCCTGAAGGCAAAA	55	X/Y	629	314
Ya5AD609	AC015978	chrY:18788855-18967434	TTGGAAAGTACACCATAACCACA	GCCCTACTTGTCCATTTTTCAAT	60	FP	505	184
Ya5AD610	AC068541-AC007379	chrY:19834150-19871515,chrY:20027673-20201554	GATGCATGGATGATACAATTT	TGCTCAAGCCCTTTATTATT	55	FP	549	303

(table cont.)

Ya5AD611	AC010133	chrY:20609301-20761174	ATACCTGGAGCTTTTTGTCA	CACGCATAGTCACAAGTTTT	55	FP	551	228
Ya5AD612	AC010889	chrY:20958342-21138265	ACGATTTTCAGAGTTGAAGC	AACTCTTATTTGGAGGGACA	55	FP	542	231
Ya5AD613	AC006998	chrY:16704663-16848722	GGAAACTTAAAGGAAAGGCACAT	CAAATCTTAAGAAAGCCAGTGGA	55	FP	710	400
Ya5AD614	AC016678	chrY:18083142-18225923	TCAGAGAAAATCAAGAAATGC	GAGTGAAAAGGGTGAAAATG	55	FP	549	204
Ya5AD615	AC006999	chrY:18504136-18616813	TTGCACATTTCTTGTTTTCCA	AAATGTGGGGAAATTGGTTT	57	FP	879	549
Ya5AD617	AC007967	chrY:8680381-8867727	ACATGTATACACATAAGTACATGTG	AATGCCAATTATCCTGACTT	55	FP	472	169
Ya5NBC9	AC006382 AC005704	chrY:16848723-17011332 chrX:5295540-5394572	CTTCCCTAGGATTTAAGTCACCATAA AGAC	TTTTCAACTTGTAACTGTAGAGGACAGG AC	60	X/Y	415	102
Ya5NBC153	AC005820	chrY:14465010-14615919	CCAATCTGGGAATTATGACAAGTAG	CTTCAGACTTCTGCTTGATTTCTTC	60	FP	496	186
Ya5NBC155	AC006565	chrY:14420131-14465009	TGTCAATATCAGACAGATCCATGAG	ACTTCCAACTATGTGGTCAGTTTTG	60	X/Y	505	182
Ya5NBC156	AC002531	chrY:14120145-14316044	TGTGGTAAGTGTAGTTTCAAAAGAGT TT	TAATCTCTGGACTGGAAACATAAAA	55	FP	480	148
Ya5NBC172	AC006371	chrY:15065526-15267679	CCAAACGTAAGATTGAGTGG	AGTGGTGTTCTCGGTATTTC	55	FP	473	155
Ya5NBC174	AC006462	chrY:17011333-17151126	TCACTCTTTGTCTTGCTGACTACAG	GCTATAGCTTCTATTTACGGGGAAT	55	FP	526	206
Ya5NBC218	AC006989	chrY:16294804-16452269	AGCCCAACATCTGGTTTTGT	TCCAGTCTCGTGTAAAATAGCTTG	55	FP	445	109
Ya5NBC219	AC006989	chrY:16294804-16452270	CCTGGCAACCACCATTCTAC	AAACCTGGAGGGCATTCTTT	58	FP	445	129
Ya5NBC325	AC009479	chrY:3222117-3377215	CTTCTCTCTCTGAAATGCCAAT	CAGTTGAAAGGTTTGACAATACACC	60	FP	501	184
Ya5NBC413	AC006040	chrY:2500001-2686304	GGGCATTTTCAATCTCTCCA	ATGAAGTTGGAGGGGCAGAG	60	FP	435	119
Ya5NBC503	AC019099	chrY:27901323-28009655	GCTGAAAAGCTGACTGACACC	CAGAAAGGTTTCCCAGTTCG	55	FP	456	156
Ya5NBC508	AC010723	chrY:15580278-15754497	GGTAAAATCCCTCCTTTGAG	GAACTAATTGGGAGAGAGCA	55	FP	405	96
Ya5NBC509	AC010135	chrY:17664290-17841040	TGCTTGTATCAGCAGTCCTCA	CCCTCCATCCATCGAAAAAT	60	FP	390	76
Yb8AD687	AC007320- AC023342	chrY:23555125-23742818, chrY:23425361-23494514	CCAGGAGCTAGGTAATCAACATTT	TGGAAGGGGCAAATAAGAAA	58	FP	622	322
Yb8AD689	AC010723	chrY:15580278-15754497	AAGAATTTGCCAACACAGGTT	TTGTGCACAGGATGATTTGA	60	FP	834	516
Yb8AD690	AC010726	chrY:15782642-15958965	TTAACTAACATGGGCACCAA	AAAAATAGATTGCTCTCCTTCA	55	FP	465	166
Yb8AD693	AC010972	chrY:16532607-16647043	ATGAAATGTCAGCCTGATTC	CTCCCATGAAATGACAAGAT	60	FP	471	122
Yb8AD720	AC025227	chrY:23494515-23555124	TCCTTCTTTGATGGACTTTC	AAGCTATGGTATCAGGGTGA	55	FP	626	314
Yb8AD721	AC012067	chrY:5187228-5351534	TTCTGCCATAGATGAAGGAT	GTATGTGCATGCATCTGTGT	55	FP	533	201
Yb8NBC108	AC010089- AC053490	chrY:26029303-26132458, chrY:24428487-24531718,	TGTCACTTGATTGTCCGCATA	TCAATGGCATCCTGAAAACA	60	FP	550	194
Yb8NBC109	AC006371	chrY:15065526-15267679	GTGCAACTTCAGTTTCTGCTAAGAT	CATGGTTATCTGCAAAGACTATGAC	55	FP	532	212

(table cont.)

Yb8NBC110	AC006383	chrY:14960516-15065525	AATAGGCTGAATGCCCCAAT	CTAGCATTGCAATCCCTGCTTT	60	X/Y	507	186
Yb8NBC111	AC007320	chrY:23555125-23742818	CCAGTGTCATCATCCAGACTTATTC	TACACACACACACATGCATTCTAAG	60	FP	531	192
Yb8NBC112	AC006999	chrY:18504136-18616813	GCATCTTAACCTAAATACCTGATGC	CAGGGACATAGGGTGTGAGTTACTA	60	FP	503	192
Yb8NBC114	AC004617	chrY:13889626-14035646	GGGTGAGATAGCTTAAGGAAAGAGA	AGATCTTCCCAAGAAGCCTTTC	60	FP	510	164
Yb8NBC160	AC007284 AC016681	chrY:7139521-7310769 chrY:5852850-5921375	CCACACATGGGTACCAGTCC	TTGCTTACCCACAGTCACCTC	60	FP	404	72
Yb8NBC268	AL590492	chrX:91254000-91383356	TGGGGATAGAGGAAGAAGACAA	CCTTTTCATCCAACTACCACTG	60	X/Y	517	188
Yb8NBC496	AC010977	chrY:5716765-5852849	CTGGGATAAAACAAGAGATAACAGG	GGTGTGCAGATTTTTGAGTCAT	60	FP	407	68
Yb8NBC507	AC021107 AC012667	chrY:22887518-23048118 chrY:5351535-5426338	GGCCACGTTCTGTTCTTGTT	TACCGCCTGAACTCCACTTT	53	FP	805	484
Yb8NBC535	AL133274	chrX:90732320-90828381	CTGAATAGAATCAGGGCAACA	CCATCTGGGAATAGTGTGGTG	60	X/Y	482	150
Yb9AD60	AC007678 AC024703	chrY:21877693-21986665 chrY:4241197-4272897	GGAAAACTGAAAGAATCCACACA	TCAGATGCAGGCTTTCTAACTTT	55	FP	439	114
Yb9NBC416	AL162723	chrX:88944751-89173981	GCCTTTTGAAGCTTCTGTCG	TGTTCCTTTGGTTAGGCAGA	59	X/Y	506	187
Yc1AD246	AC010154	chrY:6291830-6346980	TGGGTGGGGCCAAATAAAGAA	TGGGGTTTATTCCTTCAGATGTT	60	FP	589	269
Yc1AD250	AC011751	chrY:17903627-18083141	GGTATGCAAAAAGAAGTGCT	TTCAGATATGTGACCTGCTT	60	FP	472	167
Yc1AD254	AC010877	chrY:14615920-14752948	TGAGCAGAACAGAAAACACA	TGTGTGGCTAGCAAGTTATT	60	FP	445	139
Yc1AD255	AC011302 AC017019	chrY:13382453-13560389 chrY:9394276-9556454	AGCCGTAGTTCACAATGTTT	CACAGGGTGCATATTTTCTT	60	FP	481	154
Yc1NBC28	AC010154	chrY:6291830-6346980	TGGTGAGTTCCTGGTCTTGCTG	TGCTCACTCTTTGGGTCCACAC	60	FP	414	99
Yc1RG 243	AC006998	chrY:16704663-16848722	GGTCTGCTTACCAAATGACTGAG	ACATTCCTGATTCACAGAAGCTC	60	FP	424	136
Yc1RG242	AC007043	chrY:18396934-18504135	GCAGGACACACTTCCTGTTTCT	GTCCAGCACAGAAGAGGAATAAA	60	FP	416	96
Yc1RG244	AC017020	chrY:17266120-17432322	CCTAGAGGATTAGAGTCTGCCCTA	TATCCCCTAAAACTCATGTGTGG	60	FP	459	131
Yd6AD16	AC007247	chrY:7310770-7427357	TGACCCTAAATATACCTTCCA	AGCAACCTTGAGAAGAGTTTT	60	FP	436	127
Yd6AD17	AC007247	chrY:7310770-7427357	TGGATTCTTCCTCTTTTTGG	TTGGCTTCCCTGAGAAAATA	55	FP	575	265

[1]. Annealing temperature.
[2]. Allele frequency was classified as: high frequency polymorphism (HF), intermediate frequency polymorphism (IF), low frequency polymorphism (LF) and fixed present (FP) as previously defined by Carroll et al., 2001.
X/Y indicates a homologous region on the X and Y chromosomes.
Some of the reported *Alu* elements were detected in multiple sequencing contigs suggesting that they are either paralogous elements or the result of sequence assembly artifacts

APPENDIX C:

SUPPLEMENTARY DATA TO CHAPTER 3

HuARD	Subfamily	Chr.	Location	Forward primer	Reverse primer	Annealing temp. ° C	Deleted prod. size	Ancestral prod. size
1	Yb8	10	77318292:77319404	ggcttgttccatgccataac	ttccagtcccagaagtcaca	60	558	323
2	Y	13	73782392:73783502	tgatgtgcaggtctatatttgg	ccacgtggattcatggtcta	55	500	202
3	Yb9	13	76473601:76474684	ccaggttggactgagtcgtt	atggagtgggcaaaattcag	60	399	240
4	Yb8	18	18545981:18547101	caatggaccatctgacagga	cctatttatatgtgggggaaaatcc	55	487	505
5	Yb8	18	39769599:39770621	tgggggtaaaagctgaataa	ctggatttggctttctgcaa	55	400	188
6	Yb9	20	37275442:37276552	cactgtacccagcccacttt	tggagcaatctggaactgaa	60	450	1148
7	Yb8	21	14484889:14485990	tcaattttacctggccctagaa	ggcagggtacagaactgctc	60	596	1493
8	Ya5	22	41655448:41656509	cacaggtgcacaagctcag	aactgaacggcatggagaag	60	599	357
9	Yb8	2	194127466:194128445	r/r	r/r	---	---	---
10	Y	3	87721600:87722703	ggaaggatggatggatggat	atggtgttttgtttcctcctac	55	498	1107
11	Yb	2	212580114:212581054	ggtaaggcgtgacgacaagt	aggcatagttgaccattgacat	60	242	889
12	Yg6	3	135531419:135532443	tcctgtgtcccattttgtga	ccaacctgaccatcattcaa	60	472	556
13	Y	4	31175951:31177004	aatcaacctagttctaagtggtcct	gggaaacagaagtcaagggtaa	55	562	324
14	Ya5	4	81482458:81483562	cttgagagatcctttagatcgcttt	ccatccctactcctggtgaa	55	399	458
15	Yb8	4	180517794:180518910	cttttcttcccacccactca	tttttggatctctggagtgaga	60	624	1867
16	Yb8	5	8147581:8148653	gaggccacagattctgcttc	aacggggcatatttgtgatg	60	394	1650
17	Ya5	6	136332766:136333868	tctaggagataccatttggcatag	tgatgaggaatcaagccttc	55	375	232
18	Ya5	7	43154505:43155579	aaaatcatccccaacccagt	gttgcagaagcttgctgtgt	60	562	305
19	Y	7	83340028:83341136	caatcgtggacaatagttatagcag	aaggagagtttctccattactcg	60	479	323
ChARD								
1	Yc1	12	14762119:14763170	ctagttaccatattctgagcac	cgatggggaagttgtaccag	60.2	479	238
2	Yc1	12	77386007:77386963	gaggtaggccagtgacatcc	cacaaatggacctgaaaccac	65.6	836	692
3	Y	15	36744713:36745651	cctgcatcttttccccctttac	tgtctcccaaatacccagtg	60.2	500	375
4	Y	15	83600102: 83601175	gcaagcaaggattccaaatac	catctttgacccagagattttg	55.9	1294	1078
5	Y	15	100926641:100927546	caggatcaatcagtggagagg	aaagaggaggagggttcagg	65.6	499	431
6	Y	17	45718125:45719236	gacgctctacttgacttatgtgc	ttgtactccccatgattcagc	63.0	468	368
7	Y	20	14335691:14336797	tgatggcgatagtgttggac	cagtttgaacaggaagttggtg	65.0	1193	887

(table cont.)

8	Yc1	3	128808819:128809896	ccatgcccttcttgtttttc	tcttctaagagccagatgcag	58.0	1385	1135
9	Yc1	3	153809148:153810227	tgtgttacatcagggctactg	gctccaccaaaagcatcttc	62.2	1141	891
10	Y	3	166240896:166241963	tgtggtttttctccaggacag	aaacagttcccagaaaaagagg	60.2	429	179
11	Yc1	5	111815878:111816861	ttcctgacttttccctcttctc	cagtgcatcatacagccagac	60.2	698	541
12	Yc1	7	33801384:33802464	tgaatgctcttgtccactgc	agggtgaggaaagattcagg	62.7	801	540
13	Yc1	9	67952742:67953847	cgactaaactgggaatggtg	catttcccagggttaacagg	60.2	420	166
14	Y	X	99346975:99347902	gacatttgagctggttttgg	ccatgacttgctttcagagg	55.9	901	787

VITA

Pauline Ann Callinan was born to Patrick Joseph Callinan and Mary Bridget McGroary on May 12th, 1977 in Bedfordshire, England. Upon graduating from Cardinal Newman Sixth Form College in 1995, she went on to pursue studies in human biology at Loughborough University in Leicestershire, England. Graduating with first class honors in 2000, Pauline relocated to the United States of America to enroll as a doctoral student at Louisiana State University, under the tutelage of Prof. Mark Batzer. Pauline will matriculate her studies in May 2005, when she will receive the degree of Doctor of Philosophy.