2006

Retrotransposon mediated genomic fluidity in the human and chimpanzee lineages

Kyudong Han
Louisiana State University and Agricultural and Mechanical College, khan3@lsu.edu

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_dissertations

Recommended Citation
Han, Kyudong, "Retrotransposon mediated genomic fluidity in the human and chimpanzee lineages" (2006). LSU Doctoral Dissertations. 807.
https://digitalcommons.lsu.edu/gradschool_dissertations/807

This Dissertation is brought to you for free and open access by the Graduate School at LSU Digital Commons. It has been accepted for inclusion in LSU Doctoral Dissertations by an authorized graduate school editor of LSU Digital Commons. For more information, please contact gradetd@lsu.edu.
RETROTRANSPOSON MEDIATED GENOMIC FLUIDITY
IN THE HUMAN AND CHIMPANZEE LINEAGES

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

Kyudong Han
B.S., Dankook University, 2000
M.S., Dankook University, 2002
December 2006
ACKNOWLEDGEMENTS

I would like to thank my major professor, Dr. Mark A. Batzer for his intellectual mentorship and support and the members of my dissertation committee Drs. Joomyeong Kim, Stephania Cormier, and Craig Hart for their scientific advice and encouragement during the course of my research. I would like to acknowledge the following collaborators for their contributions to various aspects of this research. For his contribution to chapters two, three, four, and five: Dr. Richard Cordaux. For their contribution to chapter two: Dr. Jinchuan Xing and Hui Wang. For their contribution to chapters four and five: Dr. Ping Liang, Jianxin Wang, Shurjo Sen, and Jungnam Lee. In addition, I would like to thank Drs. Dale Hedges and Pauline Callinan for their critical reading of this manuscript and fruitful discussion during the course of my research.

I would like to thank my father, mother, and sisters for their love and encouragement.

Lastly, I would like to express my gratitude to my wife, Jungnam, for her love and devotion.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ................................................................................................................................. ii

LIST OF TABLES .................................................................................................................................................. iv

LIST OF FIGURES ................................................................................................................................................ v

ABSTRACT .......................................................................................................................................................... vii

CHAPTER ONE: BACKGROUND ........................................................................................................................ 1

CHAPTER TWO: UNDER THE GENOMIC RADAR: THE STEALTH MODEL OF ALU AMPLIFICATION ................................................................. 14

CHAPTER THREE: DIFFERENT EVOLUTIONARY FATES OF RECENTLY INTEGRATED HUMAN AND CHIMPANZEE LINE-1 RETROTRANSPOSONS .............................................................................................................. 42

CHAPTER FOUR: GENOMIC REARRANGEMENTS BY LINE-1 INSERTION-MEDIATED DELETION IN THE HUMAN AND CHIMPANZEE LINEAGES .................................................................................................................. 69

CHAPTER FIVE: HUMAN GENOMIC DELETIONS MEDIATED BY RECOMBINATION BETWEEN ALU ELEMENTS ......................................................... 104

CHAPTER SIX: SUMMARY ................................................................................................................................... 138

APPENDIX A: LETTERS OF PERMISSION ........................................................................................................ 143

VITA .................................................................................................................................................................. 146
LIST OF TABLES

2.1 *Alu*Yb loci identified in this study ................................................................. 18

2.2 Primate phylogenetic PCR panel results for chimpanzee *Alu*Yb elements .......... 19

3.1 Subfamily classification and age estimates for species-specific L1 elements .......... 46

3.2 Insertions and deletions in the coding region of 5 full-length chimpanzee-specific L1 elements ........................................................................................................ 57

4.1 Structural summary of L1 insertion-mediated deletions ........................................ 78

4.2 L1 insertion-mediated deletion frequency and polymorphism levels within the human and chimpanzee lineages ......................................................................... 81

5.1 Summary of human-specific ARMD events .......................................................... 106

5.2 Genomic DNA sequences deleted by ARMD ......................................................... 119
LIST OF FIGURES

2.1 Polymorphic $Alu_{Yb8}$ elements in the common chimpanzee genome .................................. 20
2.2 Lineage specific $Alu_{Yb8}$ insertions ................................................................. 21
2.3 Sequence analysis of the Pan1 locus ................................................................. 23
2.4 Gene conversion of $Alu_{Yb}$ elements ............................................................... 24
2.5 Median-joining network of the human-specific $Alu_{Yb}$ elements ................. 28
2.6 Putative evolutionary scenario for the $Alu_{Yb}$ lineage ........................................... 31
3.1 Median-joining network of L1 subfamilies ...................................................... 45
3.2 Phylogenetic tree of 5’ UTR consensus sequence of L1CS, L1HS and L1PA subfamilies .......................................................... 50
3.3 Species-specific L1 insertions ............................................................................. 53
3.4 Size distribution of species-specific L1 elements ............................................. 55
3.5 Analysis of genomic environment adjacent to species-specific L1 elements ...... 59
4.1 L1 insertion-mediated deletion in the human genome ....................................... 73
4.2 Endonuclease cleavage site preferences for the L1IMDs ...................................... 75
4.3 Median-joining network of the L1 elements associated with L1IMD .............. 77
4.4 Size distribution of the L1IMDs ......................................................................... 80
4.5 Models for the creation of L1IMDs ..................................................................... 88
4.6 Models for the formation of deletion associated atypical L1 elements .......... 92
5.1 ARMD in the human genome ............................................................................. 107
5.2 Density of ARMD events (red lines) and all $Alu$ insertions (blue lines) on individual human chromosomes ....................................................... 109
5.3 $Alu$ subfamily composition in ARMD events ................................................ 110
5.4 Size distribution of human-specific ARMD events, displayed in 100-bp bin sizes .... 112
5.5 Four different types of the recombination between Alu elements .................................. 114

5.6 Recombination window between Alu elements and percentage frequencies of breakage (during recombination) at different positions along an Alu consensus sequence ......................................................................................................................... 115

5.7 Density of ARMD events (red lines) and RefSeq genes (blue lines) on individual human chromosomes ................................................................. 117

5.8 Computational data mining for human lineage-specific ARMD loci .......................... 128
ABSTRACT

LINE-1s (Long interspersed elements or L1s) and Alus are highly successful non-long terminal repeat retrotransposons with copy numbers of ~520,000 and >1 million within the human genome, respectively. They are associated with human genetic variation and genomic rearrangement. Although they are abundant throughout primate genomes, their propagation strategy remains poorly understood. The recently released human and chimpanzee draft genome sequences provide the opportunity to compare the human genome with the chimpanzee genome. Thus, we were able to assess how these elements expanded in primate genomes and how they create genomic instability during their integration into the host genome as well as subsequent post-insertion recombination between elements. To understand the expansion of Alu elements, we first analyzed the evolutionary history of the AluYb lineage which is one of most active Alu lineages in the human genome. We suggest that the evolutionary success of Alu elements is driven at least in part by “stealth driver” elements that maintain low retrotransposition activity over extended periods of time and occasionally produce short-lived hyperactive copies responsible for the formation and remarkable expansion of Alu elements within the genome. Second, we conducted a detailed characterization of chimpanzee-specific L1 subfamily diversity. Our results showed that L1 elements have experienced different evolutionary fates in humans and chimpanzees lineages. These differential evolutionary paths may be the result of random variation or the product of competition between L1 subfamily lineages. Third, we report 50 deletion events in human and chimpanzee genomes directly linked to the insertion of L1 elements, resulting in the loss of ~18 kb of human genomic sequence and ~15 kb of chimpanzee genomic sequence. This study provides the basis for developing models of the mechanisms for small and large L1 insertion-mediated deletions. Fourth, we analyzed the magnitude of Alu
recombination-mediated deletions in the human lineage subsequent to the human-chimpanzee divergence. We identified 492 human-specific deletions (for a total of ~400 kb of sequence) attributable to this process. The majority of the deletions coincide with known or predicted genes, which implicates this process in creating a substantial portion of the genomic differences between humans and chimpanzees.
CHAPTER ONE:

BACKGROUND
Recent genome sequencing projects have consistently shown that the non-coding DNA occupies much more of the genome than authentic coding DNA in eukaryotic genomes (Lander et al. 2001; Waterston et al. 2002; CSAC 2005). The majority of this ‘junk’ DNA is comprised of transposable elements. These elements account for half the total length of the human genome (Lander et al. 2001). Recent studies of active transposable elements suggest that they can alter gene expression towards either harmful or beneficial effects, especially when an active gene is targeted (Deininger and Batzer 1999; Moran et al. 1999). These elements are associated with genomic fluidity via de novo insertions, insertion-mediated deletions, and recombination events (Deininger and Batzer 1999; Gilbert et al. 2002; Symer et al. 2002; Callinan et al. 2005; Han et al. 2005; Sen et al. 2006; Xing et al. 2006). The picture now emerging from the literature demonstrates that these transposable elements, so called ‘junk’ DNA, have various functions and play a significant and dynamic role in the process of the evolution of genomes.

Transposable elements, which are mobile segments of genetic material, were first discovered by Barbara McClintock in her study of the variegated color pattern in maize (McClintock 1956). In mammals, transposable elements are broadly divided into two categories, transposons and retrotransposons, based on their manner of mobilization. DNA transposons move in a “cut and paste” mechanism and are currently active in the genomes of bacteria, plants, and insects (Mizuuchi 1992; Lander et al. 2001). However, retrotransposons mobilize to a new location in the genome via an RNA intermediate, thereby duplicating the element (Luan et al. 1993; Feng et al. 1996; Moran et al. 1996). As a result of this “copy and paste” mechanism, retrotransposons accumulate much faster in the genome and have had a major impact on genomic architecture and fluidity (Deininger and Batzer 2002).
Long interspersed elements-1 (LINE-1 or L1) are the most successful autonomous retrotransposons in mammals. A full-length functional L1 element is about 6 kb in length and contains a 5’ untranslated region (UTR) bearing an internal RNA polymerase II promoter, two non-overlapping open reading frames (ORF1 and ORF2), which are separated by a ~60 bp-long intergenic spacer, and a 3’ UTR ending in a poly(A) tail (Kazazian and Moran 1998). ORF1 encodes an RNA-binding protein that has nucleic acid chaperone activity in vitro, and ORF2 encodes both reverse transcriptase and endonuclease activities (Mathias et al. 1991; Feng et al. 1996; Kolosha and Martin 1997). L1 elements propagate through an RNA intermediate in a process known as retrotransposition, which is thought to occur by a mechanism termed target primed reverse transcription; the insertion process typically results in 7-20 bp-long target site duplications (TSDs) flanking each side of the L1 element (Fanning and Singer 1987; Luan et al. 1993).

By comparison, short interspersed elements (SINEs) are a class of non-autonomous retrotransposons, which are thought to utilize the L1 enzymatic machinery for transposition. They are typically <500 bp long, contain internal RNA polymerase III promoters, and encode no proteins (Okada 1991; Batzer and Deininger 2002; Kajikawa and Okada 2002; Dewannieux et al. 2003). As the most successful SINEs in primate genomes, Alu elements have enjoyed remarkable proliferation during the primate radiation over the past 65 million years and have expanded to more than one million copies in the human genome (Lander et al. 2001; Batzer and Deininger 2002). The full-length of an Alu sequence is ~300 bp, depending on the length of the poly(A) tail, and was ancestrally derived from the 7SL RNA gene. Each Alu element is a dimer-like structure; the 3’ half has an additional 31 bp insertion relative to the 5’ half. The Alu element
includes internal promoters for RNA polymerase III, a poly(A) tail, and TSDs that are considered a hallmark of retrotransposition (Deininger et al. 1981; Ullu and Tschudi 1984).

Despite considerable progress in the understanding of the biology and distribution of Alu elements throughout primate taxa (Salem et al. 2003; Singer et al. 2003; Carter et al. 2004; Hedges et al. 2004; Otieno et al. 2004; Roos et al. 2004; Ray et al. In press), a great deal of uncertainty still remains concerning their strategy for survival. It has been generally accepted that the amplification of most Alu elements occurs through a small number of long-lived high activity “master” genes (Deininger et al. 1992; Deininger and Batzer 1993), although there is considerable debate as to the details of this amplification strategy (Matera et al. 1990; Schmid 1993; Batzer et al. 1995; Cordaux et al. 2004; Price et al. 2004). In this model, the mutations accumulated in the “master” genes are inherited by the copies they produced and consequently, a series of hierarchical Alu subfamilies that share the novel diagnostic mutation(s) are generated (Slagel et al. 1987; Willard et al. 1987; Britten et al. 1988; Deininger and Slagel 1988; Jurka and Smith 1988; Deininger et al. 1992; Batzer and Deininger 2002). The evolutionary history of the AluYa5 lineage, one of the most active human Alu lineages, suggests that the “founder” gene of this Alu lineage existed long before the major expansion of the lineage within the human genome (Leefflang et al. 1993). Contrary to the prediction of the “master” gene model, it has maintained low retrotransposition activity and this “founder” gene itself may not be directly responsible for the propagation of the recent human AluYa5 elements (Shaikh and Deininger 1996). Thus, these studies suggest that the expansion of Alu elements may follow a more complex propagation mechanism. Unfortunately, aside from the AluYa5 lineage, little data exist concerning the evolutionary origin of other Alu subfamilies in humans, making it difficult to assess the evolutionary significance of the results reported in the original AluYa5 lineage studies. In chapter
two, to gain additional insight into Alu subfamily propagation, we reconstructed the evolutionary history of the AluYb lineage, one of the largest and most active Alu lineages in the human genome (Jurka 1993; Carter et al. 2004) and proposed an improved “stealth model” of Alu amplification in primate lineage (Han et al. 2005).

L1 elements account for ~17% of the human genome (Lander et al. 2001). The L1 family emerged around 120 million years (myrs) ago (Smit et al. 1995; Khan et al. 2006) and is still actively expanding in humans, as demonstrated by the existence of highly polymorphic L1 elements in human populations (Sheen et al. 2000; Myers et al. 2002; Badge et al. 2003; Boissinot et al. 2004; Seleme et al. 2006; Wang et al. 2006) and de novo L1 insertions responsible for genetic disorders (Chen et al. 2005). The detection of several hundred species-specific L1 insertions in both the human and chimpanzee genomes further supports the recent mobilization of this family of retrotransposons (CSAC 2005; Mills et al. 2006). Contrary to the non-autonomous Alu retrotransposons in which different subfamilies are capable of concomitant expansions (Batzer and Deininger 2002; Xing et al. 2004; Hedges et al. 2005), a single line of successive L1 subfamilies has amplified within the past 40 myrs in the primate lineage leading to humans (Khan et al. 2006). The most recently evolved, human-specific L1 subfamilies have been well characterized (Boissinot et al. 2000; Myers et al. 2002; Ovchinnikov et al. 2002; Salem et al. 2003; Boissinot et al. 2004) and the recent completion of the chimpanzee genome sequence (CSAC 2005) facilitates comparisons of the recent patterns of diversity and evolution of L1 subfamilies since the divergence of human and chimpanzee, ~6 million years ago (Goodman et al. 1998). In chapter three, we characterized chimpanzee-specific L1 subfamily diversity, in detail, as a comparison with their human-specific counterparts. Our results indicate that L1
elements have experienced drastically different evolutionary fates in humans and chimpanzees within the past ~6 myrs.

Both mammalian cell culture assays and previous genomic analyses have implicated L1s as agents in complex genomic rearrangements. Mechanisms of L1-mediated genomic instability include (i) unequal homologous recombination between L1 elements (Burwinkel and Kilimann 1998; Ostertag and Kazazian 2001); (ii) generation of interstitial (> 3 Kb) deletions in the target sequence (Gilbert et al. 2002; Symer et al. 2002; Han et al. 2005) and (iii) transduction of varying amounts of 3’ flanking sequence along with the L1 itself during retrotransposition (Pickeral et al. 2000). The last process is also a mechanism for L1-mediated exon shuffling (Moran et al. 1999; Goodier et al. 2000; Pickeral et al. 2000). The L1 enzymatic machinery may also be utilized during pseudogene processing and Alu element mobilization (Wei et al. 2001; Dewannieux et al. 2003). In chapter four, we identified species-specific L1 insertion-mediated deletion candidates via computational screening and wet-bench experimental verification, resulting in the removal of ~18 Kb of human genomic sequence and ~15 Kb of chimpanzee genomic sequence within the past ~6 myrs. We also propose mechanisms to explain the correlation of L1 insertion size with the size of the deletion it causes and suggest models for the formation of truncation/inversion structures during L1 integration processes associated with target site deletions.

In addition to classic retrotransposition-associated insertion mutations, Alu elements can create genomic instability by the deletion of host DNA sequences during their integration into the genome and by creating genomic deletions associated with intrachromosomal and interchromosomal recombination events (Deininger and Batzer 1999; Callinan et al. 2005). Multiple features predispose Alu elements to successful recombination, including their proximity
in the genome (one insertion every 3 kb on average), the high GC content of their sequence, and
the remarkable sequence similarity among Alu subfamilies of widely different ages. Overall, the
recombinogenic nature of these elements is reflected in the various forms of cancer and genetic
disorders associated with Alu-mediated recombination events (Myerowitz and Hogikyan 1987;
1998; Deininger and Batzer 1999; Hattori et al. 1999; Rohlfs et al. 2000; Batzer and Deininger
2002). In chapter five, we identified ~400 kb of human-specific Alu recombination-mediated
deletions (ARMDs), the distribution of which is biased toward gene-dense regions of the
genome, which raises the possibility that ARMD may have played a role in the divergence of
humans and chimpanzees (Sen et al. 2006). About 60% of the ARMDs are located in genes; and,
in at least three instances, exons have been deleted in human genes relative to their chimpanzee
orthologs. The nature of the altered genes suggests that ARMD might have played a role in
shaping the unique traits of the human and chimpanzee lineages.

References


Batzer, M. A., C. M. Rubin, U. Hellmann-Blumberg, M. Alegria-Hartman, E. P. Leeflang, J. D.
and insertion polymorphism in two small subfamilies of recently amplified human Alu

Boissinot, S., P. Chevret and A. V. Furano (2000). "L1 (LINE-1) retrotransposon evolution and


CHAPTER TWO:

UNDER THE GENOMIC RADAR: THE STEALTH MODEL OF \textit{ALU} AMPLIFICATION* 

*Reprinted by permission of Genome Research
Introduction

As the most successful SINEs (Short INterspersed Elements) in primate genomes, *Alu* elements have enjoyed remarkable proliferation during the primate radiation and have expanded to more than one million copies in the human genome (Lander et al. 2001; Batzer and Deininger 2002). Despite considerable progress in the understanding of their biology and distribution throughout primate taxa (Salem et al. 2003b; Singer et al. 2003; Carter et al. 2004; Hedges et al. 2004; Otieno et al. 2004; Roos et al. 2004; Ray et al. 2005), a great deal of uncertainty still remains concerning their strategy for survival. It has been generally accepted that the amplification of most *Alu* elements occurs through a small number of long-lived high activity “master” genes (Deininger et al. 1992; Deininger and Batzer 1993), although there is considerable debate as to the details of this amplification strategy (Matera et al. 1990; Schmid 1993; Batzer et al. 1995; Cordaux et al. 2004; Price et al. 2004). In this model, the mutations accumulated in the “master” genes are inherited by the copies they produced and consequently, a series of hierarchical *Alu* subfamilies that share the novel diagnostic mutation(s) are generated (Slagel et al. 1987; Willard et al. 1987; Britten et al. 1988; Deininger and Slagel 1988; Jurka and Smith 1988; Deininger et al. 1992; Batzer and Deininger 2002). On the other hand, the evolutionary history of the *Alu*Ya5 lineage, one of the most active human *Alu* lineages, suggests that the “founder” gene of this *Alu* lineage existed long before the major expansion of the lineage within the human genome (Leeflang et al. 1993). Contrary to the prediction of the “master” gene model, it has maintained low retrotranspositional activity and this “founder” gene itself may not be directly responsible for the propagation of the recent human *Alu*Ya5 elements (Shaikh and Deininger 1996). Thus, these studies suggest that the expansion of *Alu* elements may follow a more complex propagation mechanism. Unfortunately, aside from the *Alu*Ya5 lineage, little data
exist concerning the evolutionary origin of other Alu subfamilies in humans, making it difficult to assess the evolutionary significance of the results reported in the original AluYa5 lineage studies.

To gain additional insight into Alu subfamily propagation, we reconstructed the evolutionary history of the AluYb lineage, one of the largest and most active Alu lineages in the human genome (Jurka 1993; Carter et al. 2004) that composes approximately 40% of the human-specific Alu elements (Hedges et al. 2004). This lineage, originally termed Sb2, is characterized by a seven-nucleotide duplication involving positions 246 through 252 of the AluY consensus sequence (Jurka 1993; Batzer et al. 1996). The AluYb lineage in the human genome is subdivided into three major subfamilies: AluYb7, Yb8 and Yb9 (Batzer et al. 1996; Roy-Engel et al. 2001; Jurka et al. 2002; Carter et al. 2004) based on diagnostic mutations following the standardized nomenclature for Alu repeats (Batzer et al. 1996). The majority of the human AluYb elements integrated into the genome during the last 3 to 4 million years (myrs) and reached a total copy number of about 2000 elements (Carter et al. 2004). The human diseases caused by de novo AluYb8 insertions suggest that this subfamily is currently actively retrotransposing and a comprehensive analysis of the Yb lineage indicated that about 20% of AluYb elements are polymorphic in human genome (Muratani et al. 1991; Oldridge et al. 1999; Carter et al. 2004). Previous studies suggested that the evolutionary history of the AluYb lineage is much older than its period of major expansion in the human genome, and AluYb elements have also been identified in other non-human primates (Zietkiewicz et al. 1994; Gibbons et al. 2004). Nevertheless, the extent to which these elements are distributed among non-human primate species remains undetermined.

Using both a computational approach and polymerase chain reaction (PCR) display
methodology, we have determined the distribution of AluYb elements in different hominoid genomes. We find that the long term evolutionary history of the AluYb lineage exhibits a pattern that is remarkably similar to that of the AluYa lineage. Thus, the evolution of the AluYa and Yb lineages illustrate a common strategy for Alu element proliferation. We propose a model for the expansion and survival of Alu elements in the primate order.

Results

AluYb Elements in the Common Chimpanzee Genome

To determine the evolutionary history of the AluYb lineage, we first computationally retrieved all of the AluYb elements from the first draft of the common chimpanzee (Pan troglodytes) genomic sequence (panTro1 Nov. 2003 assembly). A total of twelve AluYb elements were identified and subjected to PCR amplification on a common chimpanzee population panel and a separate primate panel composed of human and eight additional non-human primates (see Materials and Methods). The number of AluYb elements recovered from the chimpanzee draft sequence is in good agreement with a previous study (Gibbons et al. 2004). Detailed information on each locus including primer sequences, annealing temperature, PCR product sizes and chromosomal locations are shown in Table 2.1. Among the twelve elements identified, ten belong to the AluYb8 subfamily while the other two are non AluYb7/8/9 elements (Table 2.2). Out of the ten AluYb8 elements, four loci were specific to the common chimpanzee lineage including three elements that were polymorphic within the chimpanzee population panel (Figure 2.1, Figure 2.2B). Five elements were shared between the pygmy chimpanzee and common chimpanzee lineage (Figure 2.2C) and one was present within human, chimpanzee and gorilla genomes. These results suggest that the initial expansion of the AluYb8 subfamily predates the divergence of gorillas and humans/chimpanzees, which is thought to have occurred.
Table 2.1. *Alu* Yb loci identified in this study

<table>
<thead>
<tr>
<th>Name</th>
<th>Forward primer sequence (5'-3')</th>
<th>Reverse primer sequence (5'-3')</th>
<th>Chr. Loc. in human</th>
<th>Direct repeat</th>
<th>A. T.</th>
<th>Product size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>in human</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pan1*</td>
<td>ACCAAAATGCAGTCTCTTGTGTT</td>
<td>CTTTTTCTTTATGCGCATTTTCT</td>
<td>5 : 171753413</td>
<td>AAAACATTCCTC</td>
<td>55</td>
<td>892</td>
</tr>
<tr>
<td>Pan2*</td>
<td>CTTGGCCTATTGATATTITCTCT</td>
<td>GCCTCAGAAGAGTTTTTGTGT</td>
<td>12 : 42243314</td>
<td>AAAACTCCCTCTGAG</td>
<td>58.6</td>
<td>481</td>
</tr>
<tr>
<td>Pan3*</td>
<td>CCGATATCATGCATTITCCATA</td>
<td>TGAGCAAGAAAGACATGATTGAA</td>
<td>3 : 155790154</td>
<td>AAAATAATAAACCA</td>
<td>60</td>
<td>728</td>
</tr>
<tr>
<td>Pan4*</td>
<td>AGGGCAATTACTATATGGTCAGGAG</td>
<td>TTGTCCAGTCTCTTACAATGAGACCA</td>
<td>18 : 66254955</td>
<td>ACAATAGAACATTCCTCT</td>
<td>55</td>
<td>403</td>
</tr>
<tr>
<td>Pan5*</td>
<td>GCTCTCATTCTGCTCTCTATT</td>
<td>TCTGGAATTTAAAATTCACC</td>
<td>10 : 97871078</td>
<td>AAAACAGCAAGT</td>
<td>55</td>
<td>504</td>
</tr>
<tr>
<td>Pan6*</td>
<td>GGCATTTTTAGCTTCTTTATGAGC</td>
<td>CATGCTAGACATGAGAACAACCA</td>
<td>2 : 183429337</td>
<td>GAAATAGTCTCTCTGCT</td>
<td>60</td>
<td>749</td>
</tr>
<tr>
<td>Pan7*</td>
<td>CCGACGTCTTCTTCTCTCTCTGA</td>
<td>TCAGCAGACAAATAAGGAACAGAG</td>
<td>3 : 49401325</td>
<td>AAAAGAAAACGAGTCAC</td>
<td>55</td>
<td>502</td>
</tr>
<tr>
<td>Pan8*</td>
<td>CCCCATGTGATCCCTTTCCCTTTA</td>
<td>ATGCAGTCTTTAGTGATTTATGAGA</td>
<td>8 : 101149947</td>
<td>AAAAGAGACAGGAAA</td>
<td>55</td>
<td>834</td>
</tr>
<tr>
<td>Pan9*</td>
<td>CAAGGCCCCAGACCTTGTAGTATTG</td>
<td>TCCAGAAGTTAGATGAGTCCAGA</td>
<td>14 : 52808642</td>
<td>AAAAAGATGATT</td>
<td>55</td>
<td>434</td>
</tr>
<tr>
<td>Pan10*</td>
<td>TGTGCGAGATTTTATGATGACTT</td>
<td>TCCATGCTCTTGAGGAGAAATGAGA</td>
<td>4 : 72958275</td>
<td>AAAAAGATCATCTG</td>
<td>55</td>
<td>500</td>
</tr>
<tr>
<td>Pan11*</td>
<td>AGGAGCCTCTTCTAGAAATCCCA</td>
<td>AAAAGATCAGAGAGTACGACCAGAG</td>
<td>2 : 214354697</td>
<td>AAAAGAGGAGGAAAAC</td>
<td>60</td>
<td>521</td>
</tr>
<tr>
<td>Pan12*</td>
<td>GGTGCTGCTTATGAGGTATTCCATG</td>
<td>TATTTCCCAATCTACAGAAGAAGA</td>
<td>X : 93270745</td>
<td>AAAAATTAGCCGGGCA</td>
<td>60</td>
<td>871</td>
</tr>
<tr>
<td>Gorilla7#</td>
<td>TGAAGAAATTTGAGAAGGAAGAGA</td>
<td>ACTGCAAGATGTCGTCACTACAT</td>
<td>13 : 3266822</td>
<td>AAAAGAAATGACTCAACAG</td>
<td>55</td>
<td>583</td>
</tr>
<tr>
<td>Gorilla8#</td>
<td>GCACCTACATCAGTTTCAGTTA</td>
<td>TCGAGCATCTCTTCTCTGCA</td>
<td>2 : 229645741</td>
<td>AAAAAAGAGGCGAAGG</td>
<td>60</td>
<td>540</td>
</tr>
<tr>
<td>Gorilla16#</td>
<td>AATCTTCTGGGTAGGTGGAAT</td>
<td>CACAAGCTACACCTAATGGAGAAT</td>
<td>X : 24004038</td>
<td>AAAAGATGAGTGTGCTCA</td>
<td>55</td>
<td>531</td>
</tr>
<tr>
<td>Gorilla19#</td>
<td>CAAGTTGCTTATGATGAGGTGTG</td>
<td>GAGGCCCTAAATGATAGCTAGGGG</td>
<td>7 : 10262937</td>
<td>AAAAGAGTAGATCCTT</td>
<td>59.5</td>
<td>597</td>
</tr>
<tr>
<td>Gorilla21#</td>
<td>AACAAGAGATGCTAGAAAGAACCAAT</td>
<td>CGGAGTTGGACACATTTTCTTTT</td>
<td>1 : 51666132</td>
<td>AAAAGAAAGAGGAGGA</td>
<td>55</td>
<td>749</td>
</tr>
<tr>
<td>Pygmy4#</td>
<td>GCAGTTTTTTCCATTTGCTCTA</td>
<td>TTTGAGCTTTTTTTTGGGCTTT</td>
<td>2 : 130058820</td>
<td>AAAACTTACTATA</td>
<td>55</td>
<td>600</td>
</tr>
</tbody>
</table>

a. *: retrieved from chimpanzee genome draft sequence; #: identified using PCR display methodology.
b. Chromosomal location.
c. Annealing temperature.
Table 2.2. Primate phylogenetic PCR panel results for chimpanzee *Alu*Yb elements

<table>
<thead>
<tr>
<th>Name</th>
<th>Chimpanzee Panel</th>
<th>Human (HeLa)</th>
<th>Pygmy chimpanzee</th>
<th>Gorilla</th>
<th>Orangutan</th>
<th>Gibbon</th>
<th>Siamang</th>
<th>Green Monkey</th>
<th>Owl Monkey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pan1</td>
<td>Fixed Present (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>Polymorphic (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Pan2</td>
<td>Fixed Present (<em>Alu</em>Yb6)</td>
<td>+ (<em>Alu</em>Yb6)</td>
<td>+ (<em>Alu</em>Yb6)</td>
<td>+ (<em>Alu</em>Yb6)</td>
<td>+ (<em>Alu</em>Yb6)</td>
<td>+ (<em>Alu</em>Yb6)</td>
<td>+ (<em>Alu</em>Sg)</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Pan3</td>
<td>Fixed Present (<em>Alu</em>Yb)</td>
<td>+ (<em>Alu</em>Yb)</td>
<td>+ (<em>Alu</em>Yb)</td>
<td>+ (<em>Alu</em>Y)</td>
<td>+ (<em>Alu</em>Y)</td>
<td>+ (<em>Alu</em>Y)</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Pan4</td>
<td>Fixed Present (<em>Alu</em>Yb8)</td>
<td>+ (<em>Alu</em>Yb8)</td>
<td>+ (<em>Alu</em>Yb8)</td>
<td>+ (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Pan5</td>
<td>Polymorphic (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Pan6</td>
<td>Fixed Present (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>+ (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Pan7</td>
<td>Fixed Present (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>+ (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Pan8</td>
<td>Fixed Present (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Pan9</td>
<td>Fixed Present (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>+ (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Pan10</td>
<td>Fixed Present (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>+ (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Pan11</td>
<td>Polymorphic (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Pan12</td>
<td>Polymorphic (<em>Alu</em>Yb8)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

+/-: Presence/Absence of *Alu* element; x: no amplification
Figure 2.1. Polymorphic AluYb8 elements in the common chimpanzee genome. Gel chromatographs of two loci are shown. The locus name is shown on the right of the picture. The product sizes for filled and empty alleles (pre-integration size) are indicated on the left of the picture. The DNA panel is composed of twelve unrelated common chimpanzee individuals and other primate species. The template used in each reaction is listed on the top of the picture.

~7 myrs ago (Goodman et al. 1998). In addition, two non-AluYb7/8/9 elements displayed PCR amplicon sizes consistent with the presence of an Alu element in all the hominoid primates we tested.

To confirm that the PCR products were derived from authentic AluYb elements, all of the filled amplicons were cloned and sequenced. This sequencing effort revealed additional insight into the evolutionary history of the following four loci: Pan1, Pan2, Pan3 and Pan4. At the Pan1 locus, an AluYb8 element was inserted into the chimpanzee lineage (pygmy and common) after the divergence of humans and chimps. However, the pygmy chimpanzee and common
**Figure 2.2. Lineage specific AluYb8 insertions.** Four examples of gel chromatographs are shown. The locus designation is shown on the right of the picture while the product sizes for filled and empty alleles (pre-integration size) are indicated on the left of the picture. The DNA template used in each reaction is listed on the top of the picture. (A) Pygmy chimpanzee specific AluYb8 insertion. (B) Common chimpanzee specific AluYb8 insertion. (C) Chimpanzee lineage specific AluYb8 insertion. (D) Gorilla specific AluYb8 insertion.
chimpanzee showed both filled and empty size amplicons in their PCR amplifications, suggesting the presence of an Alu insertion polymorphism in both species (Figure 2.3A). Sequence analysis of the locus showed that the AluYb8 element inserted immediately upstream of an AluSx element in the same orientation and that the integration site is partially shared (three copies of direct repeats). More recently, a non-homologous recombination or intra-chromosomal recombination event in the common chimpanzee genome generated a hybrid Alu element which is composed of the first half of the older AluSx element and the second half of the newly inserted AluYb8 element. In contrast, the smaller allele amplified in other primates appeared to be the pre-integration site of the AluYb8 element (Figure 2.3B).

DNA sequence analysis of the Pan2 and Pan3 loci resulted in the recovery of two gene conversion events. The Pan2 locus appears to be the oldest Alu element that we recovered. PCR analysis of the locus showed the presence of an Alu element or filled allele in all the hominoid primates we examined. The green monkey also showed a filled size amplicon, but the amplicon was slightly smaller than the predicted size (Figure 2.4A). DNA sequence analysis showed that all hominoid primates possessed an AluYb element at the Pan2 locus with the AluYb lineage diagnostic duplication and five additional mutations that characterize the AluYb8 consensus sequence. However, an AluSg element was found in the orthologous locus in green monkey. This result suggests that the AluYb6 element (missing two diagnostic mutations compared to the AluYb8 consensus) at the Pan2 locus was introduced into the genome via a gene conversion event after the divergence of Old World monkeys and hominoids, but before the radiation of the hominoid primates. Thus, the AluYb6 element at the Pan2 locus is 18 to 25 million years old (Goodman et al. 1998). Given the sequence identity between the AluYb and the AluSg consensus sequences, we estimate that the starting point of the gene conversion event was located within
Figure 2.3. Sequence analysis of the Pan1 locus. (A) The gel chromatographs of PCR amplification results are shown. The template used in each lane is listed on the top of the gel picture. The product sizes for filled and empty sites (pre-integration size) are indicated on the left of the picture. (B) Schematic diagrams for the possible evolutionary scenarios. Light blue triangles denote the amplicons with an \textit{Alu}Yb8 insertion; orange crosses denote the pre-integration products and the yellow star denotes the recombination product in the common chimpanzee genome. Flanking sequences are shown as green boxes; target site direct repeats are shown in red and pink boxes. \textit{Alu} elements are shown as arrows and the direction of arrow indicates the orientation (5’->3’) with the head of the arrow denoting the end of the \textit{Alu} elements.
the first 75bps of the *Alu* element and that the 3’ terminus was located between positions #267 and #310.

The Pan3 locus also contains a gene conversion event (Figure 2.4B). Siamang, gibbon and orangutan possess an *Alu* element that does not contain the *Alu*Yb lineage diagnostic duplication and has the highest sequence identity to the *Alu*Y consensus. However, the *Alu* elements in the gorilla, chimpanzee and human genomes belong to the *Alu*Yb lineage, as indicated by the presence of the *Alu*Yb diagnostic duplication in these elements. This indicates

![Figure 2.4. Gene conversion of *Alu*Yb elements.](image)

Gel chromatographs of PCR products derived from a phylogenetic analysis of the Pan2 locus (A) and Pan3 locus (B) are shown on the left. The DNA template used in each lane is shown on the top of the gel picture. The product sizes for filled and empty alleles (pre-integration size) are indicated on the left of the picture. The schematic diagrams depict the potential evolutionary scenarios on the right. Flanking sequences are shown as green boxes; target site direct repeats are shown in red boxes. *Alu* elements are shown as arrows and the direction of arrow indicates the orientation (5’->3’) with the head of the arrow denoting the end of the *Alu* elements.
that a gene conversion event occurred after the divergence of gorillas and orangutans but prior to the divergence of gorillas, humans and chimpanzees. During this process, a small proportion of the ancestral \textit{AluY} element has been converted to an \textit{AluYb} like sequence including the \textit{AluYb} diagnostic duplication and another adjacent tightly linked diagnostic mutation. The gene converted region could be as small as 8 bp (251-259) or as long as 30 bp (237-267) in this case. Thus, the overall gene conversion rate (2/12) we observed here seems to be much higher than in previous studies (Maeda \textit{et al.} 1988; Kass \textit{et al.} 1995; Roy-Engel \textit{et al.} 2002; Salem \textit{et al.} 2003a). However, this is not surprising since the difference may be due to the small sample size in the current study or reflect the longer evolutionary time period that each element may have been subjected to gene conversion.

The Pan4 locus contains the oldest \textit{AluYb8} element we identified. Interestingly, the Pan4 \textit{AluYb8} element in the human genome only has one CpG mutation (G->A) at position #5 compared to the \textit{AluYb8} consensus sequence, while the chimpanzee and gorilla \textit{AluYb8} elements at the orthologous Pan4 locus each have accumulated 5 species specific mutations compared to the \textit{AluYb8} consensus sequence. Since no mutations are shared by the human, chimpanzee and gorilla \textit{AluYb8} elements at the Pan4 locus, we believe that the ancestral \textit{AluYb8} element at the Pan4 locus was a canonical \textit{AluYb8} element. Using the BLAT program (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start), we identified three additional \textit{AluYb8} elements in the human genome with 100\% sequence identity to the Pan4 element. This result suggests that the Pan4 element may still be retrotranspositionally active in the human genome, although we can not completely rule out the possibility that these four elements independently mutated at position #5. Multiple alignments of the Pan1, Pan2, Pan3 and Pan4 loci are available on our website under publications (http://batzerlab.lsu.edu).
AluYb Insertions in Other Primate Genomes

To further investigate the propagation of the AluYb lineage, we next searched for the presence of AluYb elements in additional non-human primate genomes. Since no complete draft genomic sequences are available other than human and common chimpanzee, we employed a modified PCR display method to identify AluYb elements from other primate genomes (see Materials and Methods). The display procedure was performed for four hominoid primates: pygmy chimpanzee, gorilla, orangutan and gibbon; three Old World monkeys: green monkey, rhesus monkey and pig-tailed macaque. In addition, this approach was also applied to the common chimpanzee to identify additional AluYb elements absent from the common chimpanzee genomic sequence. Two restriction enzymes were used for every template and a minimum of 72 colonies were sequenced for each species. For the common chimpanzee, a total of seven AluYb elements were retrieved, all of which had previously been identified from the draft sequence. This suggested that our method involving two restriction enzymes yielded a ~60% coverage (7/12) of the AluYb elements in the genome. One pygmy chimpanzee specific and five gorilla specific AluYb8 insertions were identified and confirmed by a PCR analysis of non-human primates (Figure 2.2A, D) and DNA sequencing. No AluYb element was recovered from the orangutan genome using the PCR display approach. The only locus identified within the gibbon genome was Pan2, which had been previously identified in the chimpanzee genome using the computational approach. Despite at least two trials for each species, no AluYb elements were identified in the three Old World monkeys (Green Monkey, Rhesus Monkey and Pig-tailed Macaque) examined. All of the new AluYb loci identified by PCR display are listed in Table 2.1.

Age Estimates for the AluYb8 Insertions in Chimpanzee and Gorilla

To estimate the average age of AluYb8 elements in chimpanzee genome, CpG and non-
CpG mutation densities were calculated for ten chimpanzee specific \textit{Alu}Yb8 elements as reported previously (Xing \textit{et al.} 2004) using the chimpanzee \textit{Alu}Yb8 consensus. The ten elements contained a total of nine non-CpG mutations out of 2420 nucleotides and ten CpG mutations out of 460 CpG nucleotides. The mutation densities were 0.37\% \pm 0.30\% (average \pm standard deviation) and 2.17\% \pm 2.29\% for the non-CpG nucleotides and CpG nucleotides, respectively. Using a neutral mutation rate of 0.0015/site/myr for non-CpG sites and a mutation rate of 0.0090/site/myr for CpG sites (Xing \textit{et al.} 2004), the average non-CpG and CpG mutation densities yield age estimates of 2.48 \pm 2.03 and 2.42 \pm 2.55 myrs, respectively. For the gorilla specific \textit{Alu}Yb8 elements, a total of five elements were analyzed. The mutation densities were 0.17\% \pm 0.23\% and 3.48\% \pm 3.30\% for the non-CpG and CpG sites, yielding age estimates of 1.10 \pm 1.51 and 3.86 \pm 3.66 myrs, respectively.

\textbf{Phylogenetic Analysis of \textit{Alu}Yb Elements in the Human Genome}

The presence of \textit{Alu}Yb8 elements in different primate genomes suggests that the origin of the \textit{Alu}Yb8 subfamily may be much older than its major expansion in humans. In addition, the presence of \textit{Alu}Yb8 but not \textit{Alu}Yb7 or \textit{Alu}Yb9 elements in the chimpanzee genome suggests that the \textit{Alu}Yb8 subfamily may be the ancestral component of the human \textit{Alu}Yb lineage and would therefore predate the other two major human \textit{Alu}Yb subfamilies (i.e. \textit{Alu}Yb7 and \textit{Alu}Yb9). To test this hypothesis, we examined the phylogenetic relationships of the different components of the \textit{Alu}Yb lineage in the human genome, using a median-joining network approach (Cordaux \textit{et al.} 2004), as implemented in the software NETWORK 4.1 (Bandelt \textit{et al.} 1999) available at http://www.fluxus-engineering.com/sharenet.htm. From the \textit{Alu}Yb elements identified by Carter \textit{et al.} (2004) in the human genome, we removed truncated elements and members of the \textit{Alu}Yb7, \textit{Alu}Yb8 and \textit{Alu}Yb9 subfamilies, leaving 36 previously unclassified \textit{Alu}Yb elements. After
deleting the middle A-rich region and poly-A tail of the elements, a network of the *Alu*Yb lineage was reconstructed using the 36 non-*Alu*Yb7/8/9 elements and the consensus sequences of the *Alu*Yb7, *Alu*Yb8 and *Alu*Yb9 subfamilies. Collectively, these three subfamilies comprise ~2000 copies. A preliminary analysis suggested that several nucleotide positions may have mutated more than once. Thus, the *Alu*Yb network was calculated with these putative hyper-variable positions down weighted to 1 (positions #64, #98 and #144) or 5 (position #174 and

![Median-joining network of the human-specific AluYb elements.](image)

**Figure 2.5. Median-joining network of the human-specific *Alu*Yb elements.**

The network of the *Alu*Yb lineage was reconstructed using the 36 non Yb7/8/9 elements and the consensus sequences of the Yb7, Yb8 and Yb9 subfamilies as representatives of these three subfamilies. Black circles denote sequence types. Reconstructed nodes are identified as empty circles. The size of circles indicates the number of *Alu* loci with this sequence type while arbitrary sizes were chosen for the Yb7/8/9 nodes to represent the relative sizes of the three subfamilies. Lines denote substitution steps, with a one-step distance being indicated in the lower-right corner. Broken lines indicate that the length of the branch is not drawn to scale.
position #211), while other positions were given a weight of 10. The resulting network (Figure 2.5) shows that the AluYb8 subfamily occupies a central position in the network. In addition, the AluYb8 node is associated with the highest number of direct branches (9), as compared to the AluYb7 and AluYb9 nodes (4 branches each). Finally, 72% of the non AluYb7/8/9 elements are more closely related to AluYb8 than to AluYb7 or AluYb9. Taken together, these results are strongly suggestive that the AluYb8 elements are ancestral to the other AluYb subfamilies in the human genome (Posada and Crandall 2001; Cordaux et al. 2004)

**Discussion**

**The Origin of the AluYb Lineage**

The AluYb lineage is one of the most active Alu lineages in the human genome with an estimated copy number of ~2,000 (Carter et al. 2004). To obtain further insight into the origin of the AluYb lineage in the primate order, we analyzed the draft sequence of the common chimpanzee genome and identified twelve AluYb insertions, ten of which are members of the AluYb8 subfamily, while the other two are non-Yb7/8/9 elements. The presence of an AluYb element at the Pan2 locus within siamang and gibbon genomes suggested the AluYb lineage originated before the divergence of all hominoid primates. However, no AluYb elements have been identified in multiple old world monkey genomes, thus the origin of the AluYb lineage was after the divergence of Old World monkeys and apes. These results place the origin of the AluYb lineage at the early stage of the hominoid evolution, about 18 to 25 million years ago (Goodman et al. 1998). It is worth noting that although the AluYb6 insertion at the Pan2 locus is the oldest AluYb element we identified, it was most likely generated via a gene conversion event. Therefore, it may not be the founder gene of the AluYb lineage but rather an early offspring of the AluYb founder gene, which was subsequently lost in extant primates.
The Pan4 AluYb element appears to have been fixed in the human, chimpanzee and gorilla genomes and contains all eight diagnostic mutations that characterize the AluYb8 consensus sequence. This suggests that within the 10 myrs or so after the AluYb lineage initially arose, it was not very active in terms of retrotransposition, if at all. However, this lineage retained its retrotranspositional potential during this extended period of time. Furthermore, our results suggest that in gorilla and chimpanzee genomes, the copy number of AluYb elements is two orders of magnitude lower than that in human. Therefore, over several million years following the insertion of the AluYb8 at the Pan4 locus, the AluYb lineage still retained a very low retrotransposition activity (~1.5 fixed copies per myr in the chimpanzee genome) until the major expansion of the AluYb lineage in the human genome within the past 3-4 myrs (Carter et al. 2004; Xing et al. 2004).

The “Stealth Driver” Model of Alu Evolution

The long evolutionary history of AluYb lineage leads to the conclusion that the AluYb lineage has remained in the genome with little or no retrotransposition activity for an extended period of time while retaining the ability to generate an appreciable number of new copies later in a species-specific manner (Figure 2.6). This scenario is different from the classic “master” gene model in which a “master” gene is defined as an element that is highly active over long periods of time (Deininger et al. 1992). In general, the amplification dynamics of the AluYb lineage show a striking similarity to that of the AluYa5 lineage (Leeflang et al. 1993; Shaikh and Deininger 1996). Although the existence of low activity Alu source genes has previously been suggested for the AluYa5 subfamily (Shaikh and Deininger 1996), here we provide evidence that low retrotransposition activity Alu source genes should be recognized as a major factor driving Alu expansion and evolution.
Figure 2.6. Putative evolutionary scenario for the *Alu*Yb lineage. A schematic diagram of the hominid primates is shown with the approximate time scale shown on the bottom in million years. The blue lines indicate the expansion of the *Alu*Yb lineage and the thickness of the lines represents its relative retrotransposition activity. The estimated copy number of *Alu*Yb elements in various primates is shown after their names. The blue triangle represents the estimated integration time period of *Alu*Yb founder gene and star represents the estimated integration time period of the oldest known *Alu*Yb8 element (Pan4).

We propose a model of *Alu* evolution and retrotransposition in which the low activity *Alu* elements are termed “stealth drivers”. In contrast to “master” genes, “stealth drivers” are not responsible for generating the majority of new *Alu* copies, but rather for maintaining genomic retrotransposition capacity over extended periods of time. By generating new *Alu* copies at a slow rate, a stealth driver may generate some daughter elements that are capable of much higher retrotransposition rates. These hyperactive daughter elements may act as “master” genes for the amplification of *Alu* subfamilies and are responsible for producing the majority of the subfamily
members. In addition, the new “master” genes may also give birth to additional retrotranspositionally active Alu copies that serve as secondary “master” genes or sprouts and also contribute to the expansion of the subfamily (Deininger and Batzer 1993). This view is supported by the fact that recently integrated human Alu subfamilies typically contain ~15% of such secondary “master” genes in addition to the primary subfamily “master” gene (Cordaux et al. 2004). While the highly active “master” genes which are produced from the stealth driver would be deleterious and generally be subject to negative selection, the low activity stealth driver itself will not be subject to such selection and would allow the lineage to persist for extended periods of time. Furthermore, both stealth driver and its daughter elements will also generate new elements that exhibit low levels of retrotransposition activity, effectively becoming new "stealth driver" elements. In fact, it is possible that the continuation of a lineage over extended periods of time may depend on the production of new driver elements.

But the question remains as to why, after persisting relatively quietly for millions of years, a lineage can show a sudden increase in numbers, as appears to be the case for AluYa and AluYb lineages. Under the "stealth driver" model, the master or stealth locus does not need to be "turned on", as one or more such drivers has been active during the entire history of the lineage. There are multiple scenarios which may account for the sudden expansion of the Alu lineage. In the scenario that we favor, periods of rapid expansion may be related to the ability of highly active daughter elements to escape selection at the population level and consequently produce more progeny. The change in the efficiency of natural selection in weeding out overactive elements may be related to population bottlenecks or other demographic factors (Hedges et al. 2004). Alternatively, periodic increases in element numbers may simply be due to the stochastic nature by which active daughter elements are produced. Yet another possibility is that,
contingent on the relative abundance of available L1 retrotransposition machinery, there may be fluctuation in *Alu* expansion rates (Dewannieux et al. 2003). Less likely is the possibility that molecular host defense mechanisms, which were previously suppressing the activity of *Alu* elements, failed for some reason during these periods.

In the *Alu*Yb lineage, multiple lines of evidence suggest that the *Alu*Yb8 element at the Pan4 locus may be a recent stealth driver gene in human specific *Alu* retrotransposition. First, the Pan4 element is the oldest *Alu*Yb8 element we identified; second, it only accumulated one point mutation over the last 7 myrs; third, the presence of three identical human *Alu*Yb8 elements is consistent with recent low levels of retrotranspositional activity of this element. On the other hand, since the Pan4 *Alu*Yb8 element has accumulated five mutations in the chimpanzee and gorilla genome, and there are no other *Alu*Yb elements in these genomes with the same mutations, the *Alu*Yb8 element at Pan4 locus is unlikely to be the current driver gene in the chimpanzee and gorilla genomes. The reasons why the Pan4 *Alu*Yb8 element may be a stealth driver in human but not chimpanzee and gorilla are unclear. However, there is a striking correlation between sequence similarity to the *Alu*Yb8 consensus sequence and total number of *Alu*Yb copies in these species, raising the possibility of mutational inactivation of the Pan4 *Alu*Yb8 element in chimpanzee and gorilla, but not in the human lineage.

Previous studies have shown that the amplification of the *Alu* family reached its peak about 30 myrs ago and subsequently underwent retrotranspositional quiescence (Shen et al. 1991; Britten 1994). Although we still do not know the underlying mechanisms for the retrotranspositional quiescence, the “stealth driver” model may explain why the *Alu* lineage has been subjected to periods of retrotranspositional quiescence interspersed with episodic bursts of amplification as suggested by the accumulation of at least 5,000 human specific *Alu* elements.
since the human-chimpanzee divergence (Carroll et al. 2001; Batzer and Deininger 2002; Xing et al. 2003; Carter et al. 2004; Otieno et al. 2004). The AluYa and AluYb lineages that comprise more than 60% of the human-specific Alu elements may be just two successful examples of this strategy. Similar patterns of amplification have also been observed in the retrotransposition of rodent SINE family, ID, (Kim et al. 1994) and the rodent LINE (Long INterspersed Element) family, Lx (Pascale et al. 1993). In addition to these evolutionary observations, there is also experimental evidence that indicates that some varieties of mobile elements evolve strategies to attenuate their own activity. In Alu, the acquisition of a second monomer to form its dimeric structure has been linked to decreased retrotranspositional activity (Li and Schmid 2004). In vitro modifications to currently active L1 elements can produce orders of magnitude increases in the L1 amplification rate (Han and Boeke 2004). In addition, cryptic polyadenylation sites throughout the L1 sequence may serve to quell the number of full length, retrotransposition competent L1 copies generated (Perepelitsa-Belancio and Deininger 2003). Taken together, this evidence suggests that the stealth driver model may not be a unique feature of the Alu family itself, but rather be one variant of a common survival strategy for SINE and LINE elements. More generally, the ability of mobile elements to maintain low to moderate levels of amplification activity, rather than more rapid duplication rates, may be a common feature of long-lived, successful families of transposons.

Materials and Methods

Computational Identification of AluYb Elements

A 31bp (TGCGCCACTGCAGTCCGCAGTCCGGCCTGGG) oligonucleotide that included the AluYb lineage specific duplication, was used to screen the common chimpanzee genome draft sequences (panTro1 Nov. 2003 assembly) using the Basic Local Alignment Search
Tool (BLAST) program available at http://www.ensembl.org/multi/blastview (Altschul et al. 1990). All Alu elements that have the diagnostic seven base pair duplication were selected and extracted along with one thousand base pairs of unique DNA sequence adjacent to both ends of the elements. The program RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) was then used to annotate all known repeat elements within the DNA sequence. Flanking oligonucleotide primers for the PCR amplification of each Alu element were then designed using Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The primers were subsequently screened against the GenBank NR database using BLAST queries to determine if they resided in unique DNA sequences and would only amplify the Alu elements of interest.

**Identification of AluYb Elements with PCR Display**

The Alu element PCR display methodology has been reported previously (Ray et al. 2005). Using this approach, 500ng of genomic DNA was partially digested using restriction endonucleases NdeI or MseI as recommended by the manufacturer (New England Biolabs, Beverly, MA) in 120 μl reactions. Digestion products were then ligated with double stranded linkers and amplified and “displayed” using the primer LNP (5’-GAATTCGTCAACATAGCAT TTCT-3’) and an AluYb-specific primer (5’-GGCCGGACTGCGGACT-3’) to acquire partial Alu sequences and the accompanying flanking unique sequences from each template. Since the Alu sequence in the amplicon is about 300bp long, the PCR products were then purified by BD CHROMAS SPIN™ -400 columns (BD Biosciences) to select the fragments larger than 400bp so that enough unique flanking sequence can be obtained to locate the orthologous sequences in the draft sequence of the human genome. A second round amplification was performed using the LNP oligonucleotide and a second nested AluYb-specific primer (5’-AATCTCGGCTCACTGC
AAGCTCCGCT -3’) to increase the specificity of the amplicons. The second round PCR products were separated on a 2% agarose gel and fragments larger than 400bp were excised and extracted from the gel using the Wizard gel purification kit (Promega). The purified products were then cloned into the TOPO-TA cloning vector (Invitrogen). At least seventy-two clones were randomly isolated from each template and DNA sequences were determined from both strands using chain termination sequencing (Sanger et al. 1977) on an ABI 3100 automated DNA sequencer.

After obtaining the sequences, the BLAST-Like Alignment Tool (BLAT) program (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start) (Kent 2002) was used to compare the resulting partial AluYb8 element and its adjacent flanking sequence with human draft sequences to identify orthologous sequences. Using a combination of the unique flanking sequence and the orthologous human DNA sequences, oligonucleotide primers were designed around each newly identified AluYb element as outlined above.

PCR Analysis of AluYb Elements

All of the AluYb loci were screened on a panel composed of human genomic DNA (HeLa cell line ATCC CCL-2) and DNA samples from the following non-human primate species: *Pan troglodytes* (common chimpanzee), *Pan paniscus* (bonobo or pygmy chimpanzee), *Gorilla gorilla* (lowland gorilla), *Pongo pygmaeus* (orangutan), *Hylobates syndactylus* (siamang), *Hylobates lar* (white handed gibbon), *Chlorocebus aethiops sabaeus* (green monkey) and *Aotus trivirgatus* (owl monkey). The non-human primate DNA is available as a primate phylogenetic panel (PRP00001) from the Coriell Institute for Medical Research. The chimpanzee specific AluYb loci were also screened for insertion presence/absence using a
common chimpanzee population panel composed of twelve unrelated individuals of unknown geographic origin, which was provided by the Southwest Foundation for Biomedical Research.

PCR amplification of each locus was performed in 25 µl reactions using 10-50 ng of target DNA, 200 nM of each oligonucleotide primer, 200 µM dNTP’s in 50 mM KCl, 1.5 mM MgCl₂, 10 mM Tris-HCl (pH 8.4) and 2.5 units Taq DNA polymerase. Each sample was subjected to an initial denaturation step of 94° C for 150 seconds, followed by 32 cycles of one minute of denaturation at 94° C, one minute of annealing at optimal annealing temperature, one minute of extension at 72° C, followed by a final extension step at 72° C for ten minutes. Resulting PCR products were fractionated on a 2% agarose gel with 0.25 µg of ethidium bromide and visualized using UV fluorescence.

DNA Sequence Analysis

To confirm the presence of AluYb elements, all PCR products suggesting the presence of an Alu element were gel purified using the Wizard gel purification kit (Promega). Purified PCR products were then cloned into vectors using the TOPO TA cloning kit (Invitrogen) and sequenced using chain termination sequencing (Sanger et al. 1977) on an Applied Biosystems 3100 automated DNA sequencer. All clones were sequenced in both directions to confirm the sequence. The DNA sequences generated in this study are available in the GenBank under accession numbers AY791249 to AY791290.

References


Britten, R. J. (1994). "Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago." Proc Natl Acad Sci U S A 91(13): 6148-50.


CHAPTER THREE:

DIFFERENT EVOLUTIONARY FATES OF RECENTLY INTEGRATED HUMAN AND CHIMPANZEE LINE-1 RETROTRANSPOSONS
Introduction

Long interspersed elements-1 (LINE-1 or L1) are the most successful autonomous retrotransposons in mammals. A full-length functional L1 element is about 6 kb in length and contains a 5’ untranslated region (UTR) bearing an internal RNA polymerase II promoter, two non-overlapping open reading frames (ORF1 and ORF2), which are separated by an ~60 bp-long intergenic spacer, and a 3’ UTR ending in a poly(A) tail (Kazazian and Moran 1998). ORF1 encodes an RNA-binding protein that has nucleic acid chaperone activity in vitro, and ORF2 encodes both reverse transcriptase and endonuclease activities (Mathias et al. 1991; Feng et al. 1996; Kolosha and Martin 1997). L1 elements propagate through an RNA intermediate in a process known as retrotransposition, which is thought to occur by a mechanism termed target primed reverse transcription; the insertion process typically results in 7-20 bp-long target site duplications flanking each side of the L1 element (Fanning and Singer 1987; Luan et al. 1993).

With >500,000 copies, L1 elements account for ~17% of the human genome (Lander et al. 2001). The L1 family emerged around 120 million years (myrs) ago (Smit et al. 1995; Khan et al. 2006) and is still actively expanding in humans, as demonstrated by the existence of highly polymorphic L1 elements in human populations (Sheen et al. 2000; Myers et al. 2002; Badge et al. 2003; Boissinot et al. 2004; Seleme et al. 2006; Wang et al. 2006) and de novo L1 insertions responsible for genetic disorders (Chen et al. 2005). The detection of several hundred species-specific L1 insertions in both the human and chimpanzee genomes further supports the recent mobilization of this family of retrotransposons (CSAC 2005; Mills et al. 2006). Contrary to the non-autonomous Alu retrotransposons in which different subfamilies are capable of concomitant expansions (Batzer and Deininger 2002; Xing et al. 2004; Hedges et al. 2005), a single line of successive L1 subfamilies has amplified within the past 40 myrs in the primate lineage leading to
humans (Khan et al. 2006). L1 subfamilies are distinguished by diagnostic substitutions that are shared by all members of any given subfamily. For example, five subfamilies are thought to have amplified in hominoid primates (i.e., humans and apes) within the past 25 myrs, named L1PA1 to L1PA5 (Smit et al. 1995; Boissinot et al. 2000; Lander et al. 2001; Khan et al. 2006). The most recently evolved, human-specific (HS) L1 subfamilies have been well characterized (Boissinot et al. 2000; Myers et al. 2002; Ovchinnikov et al. 2002; Salem et al. 2003; Boissinot et al. 2004) and the recent completion of the chimpanzee genome sequence (CSAC 2005) facilitates comparisons of the recent patterns of diversity and evolution of L1 subfamilies since the divergence of human and chimpanzee, ~6 million years ago (Goodman et al. 1998). Global overviews of HS and chimpanzee-specific (CS) L1 elements have previously been published (CSAC 2005; Mills et al. 2006). Here, we report a detailed characterization of CS L1 subfamily diversity and a comparison with their HS counterparts. Our results indicate that L1 elements have experienced drastically different evolutionary fates in humans and chimpanzees within the past ~6 myrs.

**Results and Discussion**

**L1 Elements and Nomenclature Used in This Study**

Our comparison of the human and chimpanzee genome sequences resulted in the identification of 1,835 HS and 1,190 CS L1 elements. These figures compare favorably with previous estimates, considering the differences in the computational methodologies and requirements for validation of candidate loci used in the different studies (CSAC 2005; Mills et al. 2006). Because L1 elements are often truncated or rearranged (Smit et al. 1995; Szak et al. 2002), we based our analyses of L1 subfamily diversity and relationships on 864 bp-long sequences encompassing the last 665 bp of ORF2 and the entire 3’ UTR, to maximize the
number of elements included in the analyses. This approach resulted in the inclusion of 1,000 HS and 207 CS L1 elements. While this represents more than half of all HS L1 elements identified, it barely accounts for one fifth of all CS elements, suggesting that CS L1 elements tend to be more severely truncated than HS L1 elements (see below).

In the following text, we refer to species-specific L1 subfamilies as HS and CS for human and chimpanzee, respectively, and we use the RepeatMasker subfamily assignment for shared L1 subfamilies (Table 3.1 and Figure 3.1). Each subfamily name is further identified by an Arabic numeral indicating the L1 subfamily lineage to which it belongs, followed by an upper-case letter identifying the subfamily within the sequential lineage (lower case-letters are also added for isolated subfamilies outside of the sequential lineage). Upper- and lower-case letters follow the Latin alphabet, starting from the oldest subfamily in the lineage. For example, subfamily L1CS-2A is the oldest (A) L1 subfamily belonging to the chimpanzee-specific (CS) subfamily

![Figure 3.1. Median-joining network of L1 subfamilies.](image)

The network was reconstructed using HS and CS L1 elements. Empty circles represent HS L1 subfamilies. Filled circles represent CS L1 subfamilies. Lineage names and ages of some nodes are shown. The lines represent substitution steps, with a one-step distance indicated in the bottom-right corner. The network is also classified as yellow (L1PA3 and L1PA2-1A), pink (L1PA2), green (L1HS) and blue (L1CS) shadow boxes.
Table 3.1. Subfamily classification and age estimates for species-specific L1 elements

<table>
<thead>
<tr>
<th>Classification in present study</th>
<th>RepeatMasker classification</th>
<th>Age ± SD (myrs)</th>
<th>Polymorphism level</th>
<th>Proportion of species-specific L1 elements&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L1 subfamilies shared by human, chimpanzee and gorilla, but not orangutan</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA3-1A</td>
<td>L1PA3</td>
<td>12.7 ± 0.8&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA3-1Aa</td>
<td>L1PA3</td>
<td>12.2 ± 0.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA3-1B</td>
<td>L1PA3/L1PA2</td>
<td>12.6 ± 0.6&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0% (0/49)</td>
<td>26.6% 25.6%</td>
</tr>
<tr>
<td>L1PA3-1Ba</td>
<td>L1PA3/L1PA2</td>
<td>10.3 ± 0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA3-1Bb</td>
<td>L1PA3/L1PA2</td>
<td>10.2 ± 0.5&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA2-1A</td>
<td>L1PA2</td>
<td>9.0 ± 0.4&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L1 subfamilies shared by human and chimpanzee, but not gorilla</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA2-1B</td>
<td>L1PA2</td>
<td>7.6 ± 0.5&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA2-1C</td>
<td>L1PA2</td>
<td>8.0 ± 0.5&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA2-1D</td>
<td>L1PA2</td>
<td>7.9 ± 0.5&lt;sup&gt;a&lt;/sup&gt;</td>
<td>7% (5/67)</td>
<td>27.1% 34.8%</td>
</tr>
<tr>
<td>L1PA2-1Da</td>
<td>L1PA2</td>
<td>7.8 ± 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA2-1Db</td>
<td>L1PA2</td>
<td>6.0 ± 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1PA2-1E</td>
<td>L1PA2</td>
<td>6.5 ± 0.4&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Human-specific L1 subfamilies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1HS-1A</td>
<td>L1PA2</td>
<td>5.7 ± 0.8</td>
<td>9% (1/11)</td>
<td></td>
</tr>
<tr>
<td>L1HS-1B</td>
<td>L1PA2</td>
<td>4.4 ± 0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1HS-preTa</td>
<td>L1HS-preTa</td>
<td>3.1 ± 0.3</td>
<td>14%&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0% 38.0%</td>
</tr>
<tr>
<td>L1HS-Ta0</td>
<td>L1HS-Ta0</td>
<td>2.7 ± 0.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1HS-Ta1</td>
<td>L1HS-Ta1</td>
<td>1.9 ± 0.2</td>
<td>45%&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
</tr>
<tr>
<td><strong>Chimpanzee-specific L1 subfamilies</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1CS-1A</td>
<td>L1PA2</td>
<td>6.2 ± 0.8</td>
<td>30% (3/10)</td>
<td>15.0%</td>
</tr>
<tr>
<td>L1CS-1B</td>
<td>L1PA2</td>
<td>3.9 ± 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1CS-2A</td>
<td>L1PA2</td>
<td>4.7 ± 0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1CS-2B</td>
<td>L1PA2</td>
<td>2.9 ± 0.3</td>
<td>80% (8/10)</td>
<td>27.5%</td>
</tr>
<tr>
<td>L1CS-2C</td>
<td>L1PA2</td>
<td>2.9 ± 0.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1CS-2D</td>
<td>L1PA2</td>
<td>2.4 ± 0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Others</strong></td>
<td></td>
<td></td>
<td>3.8%</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

<sup>a</sup> Estimated from both HS and CS L1 elements.
<sup>b</sup> Data from Salem et al. (2003) and Myers et al. (2002).
<sup>c</sup> Based on 1,000 HS and 207 CS L1 elements.
lineage 2. Subfamily L1PA2-1D is the fourth oldest (D) L1PA2 subfamily belonging to the subfamily lineage 1 shared between human and chimpanzee. Subfamily L1PA2-1Da is the oldest isolated subfamily (a) stemming from L1PA2-1D. Throughout the manuscript we use the designations commonly employed in the literature for the previously characterized HS subfamilies PreTa, Ta0 and Ta1 (Skowronski et al. 1988), which could also be referred to as L1HS-1C, L1HS-1D and L1HS-1E, respectively, according to the terminology applied to the other L1 subfamilies.

**L1 Subfamily Diversity**

We arbitrarily set the minimum number of elements to form a subfamily as 1% of all species-specific elements examined, or 10 HS and 2 CS L1 elements. Using this criterion, we could assign greater than 98% of all species-specific L1 elements to 17 human subfamilies containing 10-131 copies and 14 chimpanzee subfamilies containing 5-27 copies (Table 3.1). By extrapolation to total genome size, these figures imply that at least 20-30 copies of each subfamily are present in their respective genomes.

With respect to human subfamilies, we recovered the previously identified preTa, Ta0 and Ta1 HS subfamilies (Skowronski et al. 1988), that account for 31.5% of all HS L1 elements. All other HS L1 elements were assigned to the older L1PA2 or L1PA3 subfamilies by RepeatMasker. Interestingly, although we analyzed species-specific L1 elements, eight subfamilies were shared between the human and chimpanzee genomes, all of which were estimated to be older than 6 myrs (Table 3.1), an age consistent with the human-chimpanzee divergence time (Goodman et al. 1998). These results underscore the important distinction that needs to be made about the species-specific nature of L1 individual copies versus subfamilies. Four additional human L1 subfamilies have ages estimated to be greater than 6 myrs, but are
apparently not shared with chimpanzee (Table 3.1). However, since only about one fifth of all CS L1 elements could be examined, it is conceivable that these four apparently HS L1 subfamilies are actually present in the chimpanzee genome but are truncated to such an extent that they were not recognized or included in our analyses. By contrast, the two remaining human subfamilies also absent from chimpanzee (i.e. L1HS-1A and L1HS-1B) have estimated ages of 4-6 myrs; they are therefore likely true HS subfamilies.

With respect to the 14 L1 subfamilies identified in chimpanzee, beyond the eight subfamilies shared with human, the six other subfamilies that account for 42.5% of all CS elements are not shared with human (Table 3.1). Given that our human sample includes 1,000 L1 copies, it is very unlikely that these subfamilies would appear to be CS as a consequence of not having been sampled from the entire set of HS L1 elements. Moreover, these six subfamilies are estimated to be 2-6 myrs-old, therefore postdating the human-chimpanzee divergence time (Goodman et al. 1998). Therefore we believe they are true CS L1 subfamilies.

**Phylogenetic Relationships of L1 Subfamilies**

To reconstruct the relationships among the different L1 subfamilies identified in human and chimpanzee, we applied the median-joining network method (Bandelt et al. 1999; Cordaux et al. 2004) using the consensus sequences of each L1 subfamily (Figure 3.1, and Supplemental Figure 3.1 in Batzer Laboratory Web site). This network, rooted with the older L1PA3 consensus sequence, shows the global sequential order in which the successive L1 subfamilies arose (Figure 3.1). Moreover, the ages estimated independently for individual subfamilies based on within-subfamily sequence diversity are in complete agreement with this phylogenetic structure (Figure 3.1 and Table 3.1). In particular, the sequential order observed for the subfamilies shared between human and chimpanzee, and HS subfamilies is in perfect
agreement with previous studies (Boissinot et al. 2000; Khan et al. 2006). In sharp contrast with the human L1 subfamily single-lineage structure, the 6 CS subfamilies belong to two independent L1 lineages, termed L1CS-1 and L1CS-2 (Figure 3.1 and Table 3.1), which encompass two and four subfamilies, respectively.

**Comparison of 5’ UTR Sequences**

It has recently been proposed that the number of retrotransposition-active L1 lineages at a given period of primate evolution is correlated with the extent of 5’ UTR sequence variation among subfamilies (Khan et al. 2006). Therefore, we analyzed the 5’ UTR sequences of the two L1CS lineages we identified (i.e. L1CS-1 and L1CS-2) in conjunction with the 5’UTR of other L1 subfamilies (i.e. L1HS and L1PA2-13). Our results indicate that the 5’UTRs of both L1CS subfamily lineages are highly similar to each other (Figure 3.2) and to the L1HS and L1PA2 5’ UTRs. More generally, both L1CS subfamily lineages fall within the cluster of L1 subfamilies which have been sharing a common 5’UTR presumably recruited ~40 myrs ago (Khan et al. 2006). The presence of two L1 subfamily lineages with similar 5’UTRs in the chimpanzee genome suggests that they might be (or might have been recently) competing with each other for the same transcription factors (Khan et al. 2006). If so, two lines of evidence suggest that the L1CS-2 lineage may have had an advantage over the L1CS-1 lineage. Indeed, not only is the L1CS-2 lineage represented by twice as many copies as the L1CS-1 lineage, but three of the four L1CS-2 subfamilies are 2-3 myrs-old, whereas the youngest L1CS-1 subfamily is ~4 myrs-old (Table 3.1). Interestingly, we identified two full-length L1CS-2 copies with intact ORF1 and ORF2, while L1CS-1 does not possess any detectable full-length copy with intact ORFs (i.e., putatively retrotransposition-competent) in the chimpanzee genome reference sequence (see below). Because L1 retrotransposition molecules exhibit strong cis-preference...
Figure 3.2. Phylogenetic tree of 5' UTR consensus sequence of L1CS, L1HS and L1PA subfamilies. This neighbor joining tree is built by using 5’ UTR consensus sequences, based on observed number of nucleotide differences. The 5’ UTR consensus sequences of L1HS and L1PA2-13 families in light yellow and blue shadow boxes were from Khan et al. (2006) and the sequences of two L1CS lineages in a pink shadow box were generated in this study. Bootstrap values (%) are shown above each branch.

(Wei et al. 2001; Dewannieux et al. 2003), the differential number of retrotransposition-competent L1 copies among lineages may provide an advantage in the putative competition among L1 lineages. However, it is currently unknown whether the preservation of ORFs in some L1 copies is only the result of chance (i.e., because of the stochastic occurrence of ORF-disrupting mutations, all but two full-length L1CS copies have been inactivated so far and they both happen to belong to the L1CS-2 lineage) or because a selective process is acting to specifically preserve the integrity of the ORFs of these two particular L1CS-2 copies. It is
worthy to note here that although competition is a plausible explanation for the differential evolutionary successes of the L1CS-1 and L1CS-2 lineages, random chance alone could have led to the same evolutionary outcome.

**Insertion Polymorphism Levels of L1 Subfamilies**

To estimate the polymorphism levels (i.e., the proportion of polymorphic elements for insertion presence/absence) associated with the different L1 subfamilies, we analyzed a total of 147 L1 elements from the different subfamilies using locus-specific PCR reactions. Eighty two HS elements were genotyped in 80 humans and 65 CS elements were genotyped in 12 chimpanzees. As expected (Hedges et al. 2005), polymorphism levels decreased with subfamily ages (Table 3.1). For example, 45-80% of L1 elements belonging to subfamilies younger than ~3 myrs are polymorphic, and 9-30% of L1 elements are polymorphic in subfamilies that are estimated to be ~3-6 myrs-old. By contrast, in ~6-8 myrs-old subfamilies, only 7% of the L1 elements are polymorphic, and in subfamilies older than ~9 myrs, no elements are polymorphic. This result is consistent with the polymorphism levels observed for Alu subfamilies of similar ages, in which Alu subfamilies older than ~10 myrs, for example, virtually lack polymorphic elements (Xing et al. 2003; Salem et al. 2005).

The comparison between CS and HS L1 subfamilies of similar ages indicates that the polymorphism levels of CS subfamilies is about twice as high as that of HS subfamilies, e.g., 80% vs. 45% for <3 myrs-old L1 subfamilies and 30% vs. 9-14% for 3-6 myrs-old L1 subfamilies (Table 3.1). These results are consistent with those observed for HS and CS Alu elements, that also showed that the polymorphism levels of CS Alu subfamilies is about twice as high as that of HS Alu subfamilies (Hedges et al. 2004).
**Comparisons with Gorilla and Orangutan**

As shown in Table 3.1, several L1 subfamilies exhibit ages predating the human-chimpanzee divergence ~6 myrs ago (Goodman *et al.* 1998), based on subfamily sequence diversity. In fact, the oldest L1 subfamilies containing species-specific elements are estimated to be about twice as old as the human-chimpanzee divergence time (Table 3.1). To investigate whether these represent L1 subfamilies that have been producing new copies over extended periods of time or if the L1 elements have inserted prior to the human-chimpanzee divergence but were lost in either species (for example as a result of lineage sorting events), we genotyped the 147 L1 elements described in the previous section in gorilla and orangutan. None of the 147 elements were present in the orangutan genome. This result is consistent with the fact that the oldest L1 subfamilies examined are ~12 myrs-old (Table 3.1) and thus they postdate the divergence of orangutans and the ancestor of gorillas, chimpanzees and humans, estimated to have taken place ~14 myrs ago (Goodman *et al.* 1998). By contrast, 16 out of 49 L1 elements belonging to the 6 oldest subfamilies examined (*i.e.*, ~9-12 myrs-old, Table 3.1) were present in gorilla but absent from either humans or chimpanzees in our panel (Figure 3.3). DNA sequence analysis of the PCR products derived from these L1 elements showed that they are shared between gorilla and either human or chimpanzee and are identical-by-descent rather than derived from parallel, independent insertion events.

Because these elements belong to L1 subfamilies, which have presumably expanded before the divergence of gorillas and the ancestor of humans and chimpanzees, it is not unexpected that some elements are shared with gorilla. One explanation for this phylogenetic distribution is that the L1 elements inserted prior to the divergence of the three species and were still polymorphic at the time of speciation. As a result, some elements have become fixed in
Figure 3.3. Species-specific L1 insertions. Agarose gel chromatographs derived from the analysis of two loci are shown. The DNA template used in each lane is shown at top. The product sizes for filled and empty alleles are displayed at the left or right. (A) An HS L1 insertion. (B) Lineage sorting of an L1 insertion.

some species while being lost in others; many examples illustrating this process of lineage sorting of mobile element insertion polymorphisms involving closely related species exist in the literature (Salem et al. 2003; Hedges et al. 2004; Ray et al. 2006). It is likely that most individual copies of the shared L1 subfamilies are also shared by the different primate species, but since our analyses were designed to detect L1 elements differentially inserted between human and chimpanzee, shared L1 elements would not be recovered.

By contrast, none of the 98 L1 elements belonging to 8 myrs-old or younger L1 subfamilies was present in the gorilla genome. Therefore, our data suggest that the divergence of gorillas and the ancestor of humans and chimpanzees occurred ~8-9 myrs ago, corresponding to the time window between the oldest L1 subfamilies shared by human and chimpanzee to the exclusion of gorilla (L1PA2-1B/C/D) and the youngest L1 subfamily shared by human,
chimpanzee and gorilla (L1PA2-1A) (Table 3.1). Our results therefore suggest that the successive speciation events leading to the human, chimpanzee and gorilla lineages occurred within a restricted period of time, consistent with previous studies (Goodman et al. 1998). Such limited time periods between speciation events are particularly prone to lineage sorting of genetic variants because polymorphic L1 loci at the time of speciation can be independently fixed or lost in each species, as exemplified by the analysis of retrotransposon insertions among African cichlid fish species which are thought to have experienced a radiation several myrs ago (Takahashi et al. 2001; Terai et al. 2003).

**Structural Comparison of Human and Chimpanzee L1 Insertions**

To investigate structural differences between L1 insertions that are differentially inserted in human and chimpanzee, we focused on the comparison of the genomic sequences of human and chimpanzee chromosomes 1 and 21 (using the new chimpanzee chromosome designation). We identified 138 HS and 103 CS L1 elements on these chromosomes. On average, HS L1 elements were about fourfold longer than CS L1 elements (i.e. 2,533 vs. 641 bp; Figure 3.4). This sharp difference is explained by the fact that ~30% (41/138) of HS L1 elements were full-length vs. only ~2% (2/103) of CS L1 elements (Boissinot et al. 2000; Myers et al. 2002; Boissinot et al. 2004; Mills et al. 2006) (Figure 3.4). By contrast, ~86% (89/103) of CS L1 elements are shorter than 1 kb vs. only ~48% (66/138) of HS L1 elements (Figure 3.4). Therefore, CS L1 elements appear to be more severely truncated than their HS counterparts. The reason for such structural differences between HS and CS L1 elements is currently unknown. We cannot presently exclude the possibility that this observation is the result of lower genome coverage or sequence quality available for the chimpanzee genome as compared to the highly refined human genome draft sequence. It is also possible that one or several biological processes
Figure 3.4. Size distribution of species-specific L1 elements. A comparison of the sizes of species-specific L1 insertions from chromosomes 1 and 21 are shown. The HS and CS L1 elements on chromosomes 1 and 21 are grouped in 500 bp bins.

are responsible for these differences. For example, assuming that full-length or relatively long L1 elements are more deleterious than severely truncated elements (Boissinot et al. 2001), the size differences observed between chimpanzee and human L1 elements could be explained by a higher efficiency of selection in chimpanzees than in humans, given that the chimpanzee effective population size is higher than that of humans (Graur and Li 2000; Fischer et al. 2004) and that the efficiency of selection theoretically increases with effective population size (Graur and Li 2000). An alternative explanation might be that, due to innovations in the host or L1 biology, L1 elements have become less adept at integrating themselves into the chimpanzee genome.

Among the truncated L1 elements inserted on chromosomes 1 and 21, 29% (28/97) and 21% (21/101) of the HS and CS L1 elements, respectively, showed 5’ inversions. The inverted L1 elements were grouped into three classes, according to the structure of the junctions between the two inverted segments: deletion, overlap and precise join, as previously described (Szak et
Examination of the junctions showed that 57% (16/28) and 43% (12/28) of truncated HS L1 elements belonged to the deletion and overlap class, respectively. By comparison, 81% (17/21), 14% (3/21) and 5% (1/21) of the truncated CS elements belonged to the deletion, overlap and precise join classes. Hence, the deletion class of inverted L1 elements was the most frequent in chimpanzee, similar to what has been reported in human and mouse (Gilbert et al. 2002; Gilbert et al. 2005; Martin et al. 2005).

Next, we examined the coding sequence of full-length L1 elements to investigate whether they are intact and thus encode putatively functional proteins required for retrotransposition. We found that 32 out of 41 full-length HS L1 elements inserted on chromosomes 1 and 21 contained substitutions introducing premature stop codons within ORF1 or ORF2, while 9 elements encoded putatively functional proteins. Given that chromosome 1 and 21 represent ~9% of the entire human genome, we would predict that ~100 (9/9%) intact L1 elements exist in the human genome. This figure is very close to the ~90 human retrotransposition-competent L1 elements previously identified in a genome-wide analysis (Brouha et al. 2003). The similarity between the two values suggests that the features of L1 elements inserted on chromosomes 1 and 21 constitute a good approximation of genome-wide patterns of L1 diversity. By contrast with humans, none of the full-length CS L1 elements located on chromosome 1 and 21 possessed intact ORFs. Given this result, we extended our investigation of full-length CS L1 elements to the whole chimpanzee genome. We identified a total of 19 full-length CS L1 elements genome-wide, one of which contained an Alu element inserted in ORF1. However, again, none of the L1 elements was apparently intact. Strikingly, the chimpanzee L1 elements showed a frequent occurrence of 1 or 2 bp insertions responsible for frameshifts and the introduction of premature stop codons (Table 3.2). In most cases, those insertions were located in homopolymeric tracts.
Table 3.2. Insertions and deletions in the coding region of 5 full-length chimpanzee-specific L1 elements

<table>
<thead>
<tr>
<th>Size</th>
<th>Insertions</th>
<th>Deletions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 bp</td>
<td>2 bp</td>
</tr>
<tr>
<td>Number in chimpanzee genome sequence (Nov. 2003 freeze)</td>
<td>56</td>
<td>8</td>
</tr>
<tr>
<td>Number confirmed by DNA sequencing in this study</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

(e.g., presence of four T nucleotides in a row in one copy with a frameshift, whereas the consensus of all other L1 sequences examined would possess only three T nucleotides preserving the ORF). These results suggest that at least some of these insertions may not be authentic, for example resulting from sequencing errors in the draft sequence of the chimpanzee sequence used in this study (Mills et al. 2006). To test this hypothesis, we selected 5 full-length CS L1 elements and resequenced them using DNA from the chimpanzee individual analyzed in the chimpanzee genome project, known as Clint (CSAC 2005). None of the 64 insertions of 1 or 2 bp present in the chimpanzee genome reference sequence (Nov. 2003 freeze) were found in our sequence analysis (Table 3.2). By contrast, the single 3-bp insertion detected in the reference sequence was confirmed as an authentic event. It turns out that this insertion introduced a codon that did not disrupt the ORF of the L1 element. In addition, all but one deletion sequenced (7/8) were confirmed as authentic events. These results suggest that small insertions are likely to be artifacts whereas most small deletions appear to be authentic. Therefore, we reanalyzed the 19 full-length CS L1 elements computationally after removing all 1 or 2-bp insertions. Using this approach, we identified five intact L1 elements in the chimpanzee genome, that is considerably lower than the ~90 retrotransposition-competent L1 elements identified in the human genome (Brouha et al.)
Two of the intact chimpanzee L1 elements belong to the subfamily lineage L1CS-2B and three are L1PA2 members. As discussed above (see section “Comparison of 5’ UTR sequences”), this may contribute to explain why the L1CS-2 subfamily lineage seems to have been more successful than the L1CS-1 lineage in recent chimpanzee evolution.

Genomic Distribution of Human and Chimpanzee L1 Insertions

To test whether HS and CS L1 elements inserted in genomic regions with similar properties, we analyzed the GC content and gene density of genomic regions flanking the L1 elements inserted on chromosomes 1 and 21. We examined the GC content of 20 kb flanking genomic sequence each side of the L1 elements. The results showed that HS and CS L1 elements had very similar GC content distributions, both being skewed towards AT-rich regions of the genome (Figure 3.5A). Indeed, 74% (102/138) and 83% (86/103) of HS and CS L1 elements, respectively are found in AT-rich regions (defined as regions with GC content less than the 41% genome-wide average), whereas, in comparison, 58% of the human genome consists of AT-rich regions (Lander et al. 2001). We also compared the gene density of 1 Mb flanking genomic sequence each side of L1 elements. Again, we found that HS and CS L1 elements had similar gene density distributions, skewed towards gene-poor regions of the genomes (Figure 3.5B). These results are not unexpected, however, since there is a positive correlation between GC content and gene density (Lander et al. 2001; Versteeg et al. 2003).

To investigate global polymorphism levels of HS and CS L1 elements regardless of subfamily affiliation, we randomly selected 31 HS and 31 CS L1 elements located on chromosomes 1 and 21 and genotyped them in our relevant human or chimpanzee population panels. We found that 10% (3/31) and 23% (7/31) of the HS and CS L1 elements, respectively, were polymorphic. Hence, consistent with the L1 subfamily-specific polymorphism results (see
above) and previously reported Alu element results (Hedges et al. 2004), the global L1 insertion polymorphism level is about twice as high in chimpanzees as in humans.

**Figure 3.5. Analysis of genomic environment adjacent to species-specific L1 elements.** (A) **Analysis of GC content.** The vertical axis represents the number of L1 loci within each GC bin. The highest frequency of species-specific L1 loci is shown in 35-39% GC bin from both the human and chimpanzee genomes. (B) **Number of genes flanking L1 elements.** The distribution of the number of genes flanking species-specific L1 elements is similar in both the human and chimpanzee genomes.
Conclusions

Our analyses indicate that L1 elements have had very different evolutionary dynamics in the chimpanzee and human genomes, within the past ~6 myrs. Although the species-specific L1 copy numbers are on the same order in both species (1,200-2,000 copies; this study (CSAC 2005)), the number of retrotransposition-competent elements appears to be much higher in the human genome than in the chimpanzee genome. Nevertheless, in the human genome, only a subset of all retrotransposition-competent L1 elements may be responsible for most L1 insertions (Brouha et al. 2003; Seleme et al. 2006), indicating that the total number of apparently intact L1 elements in a genome is not necessarily predictive of the overall L1 activity. Interestingly, we identified two recent lineages of L1 subfamilies in the chimpanzee genome. The two lineages seem to have coexisted for several myrs, but only one shows evidence of expansion within the past three myrs. This lineage contains twice as many copies as the other lineage and we identified two retrotransposition-competent L1 elements belonging to this most recently active lineage in the chimpanzee genome, whereas no retrotransposition-competent L1 element can be identified in the other, apparently less active lineage. If the differential evolutionary dynamics of these two L1 subfamily lineages is not the result of chance, our results suggest that the coexistence of several L1 lineages might be unstable (Khan et al. 2006), and that a situation of competition between two L1 subfamily lineages may be resolved in a very short evolutionary period of time, perhaps on the order of just a few myrs. Our data suggest that speciation events and associated host demographic changes (Hedges et al. 2004; Cordaux and Batzer 2006) may facilitate the coexistence of multiple L1 subfamily lineages within species. Therefore, cases of coexistence of multiple L1 subfamily lineages may have been quite common during evolution. However, if this situation is evolutionarily unstable and quickly leads to the loss of activity of
one of the lineages, then it would appear on a large evolutionary time scale as though all or most L1 subfamilies in one species belong to one major lineage of subfamilies, as previously reported (Khan et al. 2006). Within the chimpanzee genome, two CS L1 subfamily lineages can be unambiguously detected, presumably because of the short evolutionary time-depth involved. Therefore, the chimpanzee genome constitutes an excellent model in which to further analyze the evolutionary dynamics of L1 retrotransposons.

**Materials and Methods**

**Computational Identification of L1 Elements**

We identified all L1 elements with complete 3’end sequences in the human genome (hg16, UCSC July 2003 freeze) by Basic Local Alignment Search Tool (BLAST) querying the genome with the 3’-most 50 bp preceding the poly-A tail of the L1 consensus sequence. This strategy yielded ~110,000 candidate elements, corresponding to the most recent fraction of all L1 elements inserted in the human genome. Next, 300 bp-long sequences covering each L1 3’-end and 100 bp of flanking sequence immediately downstream the poly-A tail were extracted. The exact terminus of the poly-A tails in these L1 sequences was determined by a BLAST search with the 50 bp L1 consensus sequence to which a tract of 100 adenosines was added. The sequences were used as queries for BLAST searches against the chimpanzee genome sequence (UCSC Nov. 2003 freeze). Queries with matches limited to the 100 bp L1 3’ end flanking regions in human were collected as candidates representing the orthologous pre-integration sites of the human L1 insertions. Then, we extracted the 800-bp region centered at the chimpanzee pre-integration site, along with the human L1 insertion and 400 bp upstream and downstream flanking sequence. To reduce false positives, pairs of chimpanzee and human non-L1 genomic sequences were required to exhibit >95% identity over their entire length. This resulted in 1,989
candidate HS L1 insertions. The procedure was repeated by reversing the order of the human and chimpanzee genome sequences to identify candidate CS L1 insertions, resulting in the recovery of 1,207 loci. All candidate loci were subsequently subjected to manual verification, yielding a total of 1,835 HS and 1,190 CS L1 elements.

**PCR Amplification and DNA Sequencing**

Cell lines used to isolate DNA samples were as follows: human (*Homo sapiens*) HeLa (American Type Culture Collection [ATCC] number CCL2), common chimpanzee Clint (*Pan troglodytes*; cell line NS06006B), gorilla (*Gorilla gorilla*; cell line AG05251) and orangutan (*Pongo pygmaeus*; cell line ATCC CR6301). DNA samples from 20 European, 20 African American and 20 Asian human individuals isolated from peripheral blood lymphocytes were available from previous studies in our lab, and DNA samples from 20 South American individuals were obtained from the Coriell Institute for Medical Research. A common chimpanzee (*Pan troglodytes*) population panel composed of 12 unrelated individuals of unknown geographic origin was obtained from the Southwest Foundation for Biomedical Research.

Oligonucleotide primers for the PCR amplification of L1 elements were designed using the software Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). PCR amplification of each locus was performed in 25 µl reactions using 10-50 ng DNA, 200 nM of each oligonucleotide primer, 200 µM dNTPs in 50mM KCl, 1.5 mM MgCl2, 10mM Tris-HCl (pH 8.4) and 2.5 U Taq DNA polymerase. Each sample was subjected to an initial denaturation step of 5 min at 95°C, followed by 35 cycles of PCR at 1 min of denaturation at 95°C, 1 min at the annealing temperature, 1 min of extension at 72°C, followed by a final extension step of 10 min at 72°C. The resulting products were loaded on 2% agarose gels, stained with ethidium
bromide, and visualized using UV fluorescence. Detailed conditions for all PCR assays designed in this study are available in Supplemental Table 3.1 from the Batzer Laboratory Web site.

Individual PCR products were purified from the gels using the Wizard® gel purification kit (Promega) and cloned into vectors using the TOPO-TA Cloning® kit (Invitrogen), according to the manufacturer’s instructions. DNA sequencing was performed using chain termination sequencing on an Applied Biosystems 3100 automated DNA sequencer. The DNA sequences from this study have been deposited in GenBank under accession numbers DQ375560-DQ375750.

PCR amplification of 5 full-length L1 loci was performed in 50 µl reactions using 200 ng DNA, 300 nM of each oligonucleotide primer, 200 µM dNTPs, 1mM MgSO₄, 2% DMSO, and 2 U KOD Hifi DNA polymerase (Novagen). Each sample was subjected to heating for 2 min at 94°C to activate the polymerase, followed by 35 cycles of PCR at 15 sec of denaturation at 94°C, 30 sec of annealing at 60°C, 5 min of extension at 72°C. The PCR products were purified using the Wizard® PCR clean-up system (Promega). DNA sequencing was completed using 26 L1 internal primers (Supplemental Table 3.2 in Batzer Laboratory Web site, (Seleme et al. 2006)). These DNA sequences have been deposited in GenBank under accession numbers DQ456866-DQ456870.

**Data Analyses**

We aligned 864 bp corresponding to ORF2 3’ end and entire 3’ UTR (excluding the G₄TG₆AG₆AG₃ repeat exhibiting variable length among sequences) of 1,000 HS and 207 CS L1 elements, using the software BioEdit v.7.0 (Hall 1999). L1 subfamily consensus sequences were generated based on putative diagnostic substitutions using the module MegAlign available in the package DNASTar. The relationships among the subfamilies were reconstructed using a median-
joining network (Bandelt et al. 1999; Cordaux et al. 2004), as implemented in the software NETWORK 4.111 (http://www.fluxus-engineering.com/sharenet.htm). The age of the subfamilies were calculated with NETWORK, based on the divergence among all the copies of each subfamily. We used a nucleotide mutation rate of 0.15% per site per myr (Miyamoto et al. 1987), assuming that L1 elements accumulate mutations at the neutral rate after their insertion (Voliva et al. 1984; Pascale et al. 1993). The software MEGA 3.1 (Kumar et al. 2004) was used to build neighbor-joining trees of the 5’UTR consensus sequences of two CS L1 subfamily lineages and other L1 subfamilies (L1HS and L1PA2-13; (Khan et al. 2006)), based on the observed number of nucleotide differences and Kimura 2-parameters distance. Support for the branching patterns was evaluated based on 1,000 bootstrap replicates.

For flanking sequence GC content analysis, we used the BLAST-Like Alignment Tool (BLAT) server (http://genome.ucsc.edu/cgi-bin/hgBlat) to extract 20 kb of flanking sequence in either direction of each L1 element examined, after excluding 100 bp downstream of the polyadenylation signal to prevent bias towards excessive adenosine residues. The percentage of GC nucleotides in the flanking sequence of each L1 element was calculated using the EMBOSS GeeCee server (http://bioweb.pasteur.fr/seqanal/interfaces/geeccee.html). For the gene density analysis, we counted the number of genes within 2 Mb sequences surrounding the 5’ and 3’ ends of each L1 element examined.

References


CHAPTER FOUR:

GENOMIC REARRANGEMENTS BY LINE-1 INSERTION-MEDIATED DELETION IN THE HUMAN AND CHIMPANZEE LINEAGES*

*Reprinted by permission of Nucleic Acids Research
Introduction

Long INterspersed Elements (LINE-1s or L1s) are abundant non-LTR retrotransposons in mammalian genomes and comprise ~17% of the human genome (Lander et al. 2001). They have reached copy numbers of about 520,000 (Lander et al. 2001; Ostertag and Kazazian 2001) and have expanded over the past 100-150 million years (Smit et al. 1995). In their full-length state, they are capable of autonomous retrotransposition through an RNA intermediate. However, ~ 99.8% of extant L1s in the human genome are retrotransposition-defective (Sassaman et al. 1997), either due to point mutations or larger changes such as 5’ truncations, 5’ inversions or other internal rearrangements (Kazazian and Moran 1998; Ostertag and Kazazian 2001; Gilbert et al. 2002; Myers et al. 2002). While extant human L1-derived elements have an average size of 900 bp for all L1 copies (Lander et al. 2001), an active full-length L1 element is about 6 Kb in length, and encodes two open reading frames (ORFs) separated by a 63 bp spacer region. The first L1-encoded protein, ORF1p, is a 40 kDa RNA-binding protein, while the second, ORF2p, is a 150 kDa protein with both endonuclease (EN) and reverse transcriptase (RT) activities (Mathias et al. 1991; Feng et al. 1996). The two ORFs are preceded by a 5’ untranslated region (5’-UTR), which contains an internal promoter for RNA polymerase II, and are followed by a 3’ UTR ending in a poly(A) tail. The L1-encoded proteins predominantly exhibit cis-preference, transposing the same RNA that encoded them (Wei et al. 2001; Dewannieux et al. 2003).

The number of full-length retrotransposition-competent L1 elements that are currently estimated to be propagating in the human genome, however, is much lower than the total number of insertions, with estimates varying between 60 and 100 elements (Sassaman et al. 1997; Kazazian and Goodier 2002; Brouha et al. 2003). The mobilization of L1 elements is based on a mechanism termed target-primed reverse transcription (TPRT) which provides useful landmarks.
for the identification of L1 insertion (Luan et al. 1993). During this process, a single-strand nick in the genomic DNA is made by the L1 EN at the 5’-TTTT/A-3’ consensus cleavage site (Feng et al. 1996; Jurka 1997; Cost and Boeke 1998; Morrish et al. 2002) on the antisense strand, after which the L1 RNA transcript anneals by its poly(A) tail to the cleavage site and primes reverse transcription. After the synthesis of the complementary DNA copy and its covalent attachment to the target DNA, second strand synthesis occurs using the first strand as a template. Single-stranded regions remaining in the target DNA at either end are filled in to create target site duplications (TSDs), structural hallmarks of the TPRT process which have been used in the computational location of L1 insertions (Szak et al. 2002). However, in situations where L1 integration results in the deletion of portions of target DNA, TSDs may not be formed, and a number of studies have reported L1 insertions without TSDs of any length (Morrish et al. 2002; Gilbert et al. 2005).

Both mammalian cell culture assays and previous genomic analyses have implicated L1s as agents in complex genomic rearrangements. Mechanisms of L1-mediated genomic instability include (i) unequal homologous recombination between L1 elements (Burwinkel and Kilimann 1998; Ostertag and Kazazian 2001); (ii) generation of interstitial (> 3 Kb) deletions in the target sequence (Gilbert et al. 2002; Symer et al. 2002) and (iii) transduction of varying amounts of 3’ flanking sequence along with the L1 itself during retrotransposition (Pickerel et al. 2000). The last process is also a mechanism for L1-mediated exon shuffling (Moran et al. 1999; Goodier et al. 2000; Pickerel et al. 2000). The L1 enzymatic machinery may also be utilized during pseudogene processing and Alu element mobilization (Wei et al. 2001; Dewannieux et al. 2003).

Previous analyses of genomic deletions created upon L1 retrotransposition in human DNA have almost exclusively relied on cell culture assays and described de novo L1
retrotransposition events associated with target site deletions (Gilbert et al. 2002; Symer et al. 2002). Large interstitial deletions, ranging up to 71 Kb, have been reported as one of the consequences of L1 retrotransposition (Gilbert et al. 2002). However, the artificially constructed L1 insertion cassettes utilized in these assays permit the recovery of large and full-length L1 insertions only, and the extent of genomic deletion identified in these analyses may not represent the actual extent of existing deletions associated with L1 insertions in the human genome. The recent completion of the draft chimpanzee genome sequence (PanTro1; Nov. 2003 freeze) provides the first opportunity to locate and quantify in an evolutionary framework existing human-specific and chimpanzee-specific L1 insertion-mediated deletions (L1IMDs). In this study, we identified species-specific L1IMD candidates via computational screening of the draft genomic sequences of Homo sapiens and Pan troglodytes and confirmed them experimentally. We find that L1 insertions are directly responsible for the removal of ~18 Kb of human genomic sequence and ~15 Kb of chimpanzee genomic sequence within the past 4-6 million years and may have generated over 11,000 deletion events during the radiation of the primate order, resulting in the removal of up to 7.5 Mb of DNA in the process. We also propose mechanisms to explain the correlation of L1 insertion size with the size of the deletion it causes and suggest models for the formation of truncation/inversion structures during L1 integration processes associated with target site deletions.

Results

A Genome-Wide Analysis of Human- and Chimpanzee-Specific L1IMDs

To locate L1IMD loci in the human and common chimpanzee lineages, we first compared data from the draft human and common chimpanzee genomic sequences. We computationally detected 30 human-specific and 33 chimpanzee-specific L1 insertion candidates associated with
extra (non-homologous) sequences at the orthologous loci in the other genome. PCR display and manual inspection of the DNA sequences resulted in the exclusion of four human loci and six chimpanzee loci as false positives for L1IMD. These cases were due to stretches of Ns in the chimpanzee genome assembly (corresponding to unsequenced regions) or species-specific Alu element insertions in the 5’ end of the loci, leading to partial mismatches at the orthologous locus in the other species, one of the prerequisites in our computational approach to identify candidate L1IMD loci. This resulted in the validation of 26 and 27 L1IMDs identified from the human and chimpanzee genomes, respectively. PCR analysis of all but one (LH4) L1IMD loci in five primate species showed that all the L1IMDs were specific to the species from which they were identified (Figure 4.1). Locus LH4 could not be amplified due to the presence of other repeat elements in the flanking sequence. However, on the basis of (i) the 99.5% similarity of the L1 element inserted at this locus to the consensus sequence of the human-specific L1Hs subfamily.

Figure 4.1. L1 insertion-mediated deletion in the human genome. (A) Gel chromatographs of PCR products from a phylogenetic analysis of the human-specific L1IMD are shown. The DNA template used in each lane is shown at top. The product sizes for filled and empty alleles are indicated at the left. (B) The schematic diagrams depict the insertion of the L1 element (orange boxes) and the deletion of genomic DNA (blue boxes). Flanking unique DNA sequences are shown as light blue boxes.
and (ii) the presence of extra (non-homologous) genomic sequence at this locus in the common chimpanzee genome, the L1 insertion and associated deletion at locus LH4 were included in our dataset of human-specific genomic deletions directly associated with L1 insertion.

Because the L1 elements associated with L1IMD were not flanked by TSDs, the only possible hallmark of TPRT in our L1IMD events was the presence of L1 EN cleavage sites. To confirm that the deletions observed in the human and chimpanzee genomes were generated during the process of L1 insertion rather than prior to (and therefore independently of) the L1 insertion, we looked for L1 EN cleavage motifs in our L1IMD loci and divided the loci into categories based on the number of differences with the 5’-TTTT/A-3’ consensus L1 EN cleavage site (Jurka 1997; Boeke and Devine 1998; Morrish et al. 2002). For each locus, we compared the sequence corresponding to the insertion site predicted to the consensus EN cleavage motif to see if it was L1 EN-generated or not. To conservatively exclude ‘false’ cleavage motifs arising from post-insertion mutations mimicking the L1 EN consensus cleavage sequence, we down-weighted the number of transition differences with the consensus EN cleavage motif by a factor 0.5 because transitions in the cleavage site that conserve the homopurine or homopyrimidine runs are generally better tolerated by the EN than transversions (Cost et al. 2001). Additionally, we further down-weighted transitions by a second factor 0.5, because of their more frequent occurrence than transversions in GC-poor regions (Nachman and Crowell 2000). In both humans and chimpanzees, the frequency spectra of the integration site preferences showed unimodal distributions with modes at 0.5 differences from the consensus sequence 5’-TTTT/A-3’ (Figure 4.2). The L1 EN site preference of our L1IMDs is thus very similar to that of L1-Ta subfamily elements (n = 282) identified in a previous study (Morrish et al. 2002). However, three of the 53 loci (LH11, LH12 and LC6) identified computationally as L1IMD candidates had cleavage sites
Figure 4.2. Endonuclease cleavage site preferences for the L1IMDs. The number of differences from the consensus L1 endonuclease cleavage site (TTTT/A) are shown after down-weighting transitions. The data are analyzed for (A) The L1-Ta subfamily elements identified in Morrish et al. (2002); (B) Human lineage specific L1 insertions (LH11 and LH12 excluded as number of differences \(\geq 2.5\)); (C) Chimpanzee lineage specific L1 insertions (LC6 excluded as number of differences \(\geq 2.5\))
substantially differing from the consensus by four or more substitutions while the maximum number of substitutions observed in the L1-Ta subfamily is three (Figure 4.2), hence casting doubt on the use of EN during insertion of these elements. We believe that these deletions are the products of EN independent insertions similar to those reported in previous cell culture assays (Morrish et al. 2002). To be conservative, these three elements were removed from the analyses, resulting in a final dataset of 24 and 26 L1IMD loci in the human and chimpanzee genomes, respectively, with deletions produced unambiguously by an L1 EN-dependent mechanism.

Characteristics of the L1 Insertions Associated with L1IMDs

The L1 insertions in our study ranged in size from 61 to 5174 bp. Of the 24 human L1 insertions, eight belonged to the L1Hs subfamily according to RepeatMasker, 14 to L1PA2 and 2 could not be confidently assigned to any subfamily. As to the 26 chimpanzee L1 insertions, 23 belonged to the L1PA2 subfamily, one to L1PA5 while two could not be confidently assigned to any subfamily. Median-joining network analysis (Figure 4.3) of the L1 elements in our study, using substitutions at the 4 key subfamily-diagnostic sequence positions (i.e., bp 5930-5932 and 6015 in the 3’ UTR of the full-length L1 consensus sequence) shows that the chronological order in evolutionary time (from youngest to oldest) of the L1 elements in our study is Ta (ACA/G) - PreTa (ACG/G) - ACG/A - GCG/A or AAG/A - GCG/G - L1PA2 (GAG/A). This evolutionary order is consistent with previous analyses of L1 insertions utilizing other phylogenetic approaches such as neighbor-joining, maximum-likelihood and maximum parsimony analyses (Ovchinnikov et al. 2002).

All the elements were 5’ truncated to different degrees (Ovchinnikov et al. 2001), with most having their 5’ start position located in the 3’ UTR of the consensus full-length L1.3 reference sequence (Dombroski et al. 1993) (Table 4.1). The size distribution of the L1 insertions
Figure 4.3. Median-joining network of the L1 elements associated with L1IMD. Empty circles denote human-specific L1 elements. Filled circles denote chimpanzee-specific L1 elements. The size of circles indicates the number of L1 loci with that sequence type. The lines denote substitution steps, with a one-step distance indicated in the top-left corner. The subfamily-specific diagnostic sequence positions (corresponding to positions 5930-5932 and 6015 in the 3’ UTR of the full-length L1 consensus sequence) are specified below each relevant node.
Table 4.1. Structural summary of L1 insertion-mediated deletions

<table>
<thead>
<tr>
<th>Feature</th>
<th>Human</th>
<th>Chimpanzee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full-length L1 insertions</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5’ truncated L1 insertions</td>
<td>24</td>
<td>26</td>
</tr>
<tr>
<td>Internal rearrangements</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Non-inverted</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5’t truncation/inversions</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>With TSDs of any length</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total L1 size (bp)</td>
<td>31,617</td>
<td>25,031</td>
</tr>
<tr>
<td>Mean of L1 size (bp)</td>
<td>1322</td>
<td>963</td>
</tr>
<tr>
<td>Total deletion size (bp)</td>
<td>17,671</td>
<td>14,923</td>
</tr>
<tr>
<td>Mean of deletion size (bp)</td>
<td>736</td>
<td>574</td>
</tr>
<tr>
<td>Median of deletion size (bp)</td>
<td>21</td>
<td>73</td>
</tr>
</tbody>
</table>

is similar to that obtained in a previous human cell culture assay of L1-mediated genomic instability (Symer et al. 2002). As to chromosomal distribution, the majority of the L1IMDs were located on chromosomes 1 to 12, which probably relates to both the larger size of these chromosomes and their higher density of truncated (3’ intact) L1 insertions (Szak et al. 2002).

Four human-specific L1 insertions (at loci LH17, LH19, LH26 and LH31) showed the presence of partially duplicated or internally rearranged L1 segments, suggesting either an atypical structure for the particular L1 insertion or two independent L1 insertions into the same locus during a relatively short time. Given the size of the human genome (~3300 Mb), two L1 insertions occurring at exactly the same location four times in 24 human loci is very improbable.
considering that there have been no instances of L1 element insertion homoplasy ever reported (Salem et al. 2003; Ho et al. 2005; Salem et al. 2005). Loci LH17 and LH 31 each consist of two L1PA2 segments in the same orientation with 300 bp and 286 bp gaps between the two segments, respectively, relative to the L1PA2 consensus sequence. These loci probably represent single L1 insertion events associated with internal deletions. The other two loci, LH19 and LH26, each consist of two identical L1PA2 segments in tandem, with 53 bp and 189 bp stretches respectively being repeated in the same orientation without any intervening region. Two chimpanzee loci (LC26 and LC27) also presumably resulted from 5' truncation/inversion events, with overlapping junctions between the inverted segments (Szak et al. 2002).

The poly(A) tails of the L1 inserts ranged in length from 2 to 64 bases, with similar averages of 19 bases in humans and 21 bases in chimpanzees. Our value for the average poly(A) tail lengths for human L1 insertions is thus much lower than those from two previous cell culture assays of de novo L1 retrotransposition in HeLa cells, that reported averages of ~60 residues (Gilbert et al. 2002) and 88±27 residues (Symer et al. 2002). Furthermore, the 23 bp average length of the poly(A) tail among members of the youngest L1Hs subfamily was slightly higher than the 16 bp average for the older L1PA2 subfamily elements. Our data thus suggest the occurrence of post-insertional shortening of poly(A) tails over time, possibly due to replication slippage (Ovchinnikov et al. 2001; Roy-Engel et al. 2002). While the poly(A) tails in the de novo insertions identified in the aforementioned studies are exclusive runs of adenosine residues, the tails of the L1s identified in our study show considerable patterning and incidence of other nucleotide residues, with TA\textsubscript{1\(n\)} being the most common pattern (six cases in the chimpanzee L1s and four cases in human L1s), which corroborates the findings of Szak et al. (2002). We found no significant correlation between the size of the poly(A) tail and the size of the L1 insertion in
our dataset \((r = 0.12, P = 0.84)\).

**Characteristics of the L1IMDs**

L1IMD events resulted in the deletion of 17,671 nucleotides from the human genome and 14,921 nucleotides from the chimpanzee genome (Table 4.1). The size distribution of the deletions (Figure 4.4) showed a strong bias towards the smaller sizes, with 50% of the chimpanzee L1IMDs and 58% of the human L1IMDs showing sizes of <200 bp. However, both human and chimpanzee events were also characterized by 20-30% of L1IMDs longer than 1 Kb. These observations were further reflected by the medians of the L1IMD sizes being an order of magnitude smaller than the average L1IMD size in both human and chimpanzee (Table 4.1). The L1IMD loci in our study in both human and chimpanzee lineages showed significant \((P < 0.05\) in both species) positive correlations between the size of the L1IMD and the size of its associated L1 insertion.

**Figure 4.4. Size distribution of the L1IMDs.** The size distribution of all the L1IMD events identified in the human and chimpanzee lineages is displayed in 500bp intervals or bins.
L1IMD Polymorphism

To estimate the level of polymorphism associated with human-specific L1IMD loci, we amplified them in 80 individuals from four geographically diverse populations. In all, five out of 23 loci (~22%) were polymorphic (Table 4.2), three of which contained L1Hs elements and two contained L1PA2 elements. Within our common chimpanzee panel of 12 individuals, four out of 26 loci (~15%) were polymorphic (Table 4.2), three of which contained L1PA2 elements and one contained a L1PA5 element. Overall, this indicates that human L1IMDs are associated with slightly higher polymorphism rates than their chimpanzee counterparts. These results contrast with those obtained for Alu retrotransposition-mediated deletions (ARDs) (Callinan et al. 2005) and Alu insertions (Hedges et al. 2004) in the context of human/chimpanzee comparisons, in which the polymorphism rates were found to be about twice as high in chimpanzee as in human. These data could be indicative of a slowdown of L1 retrotransposition within the chimpanzee lineage as compared to the human lineage.

Table 4.2. L1 insertion-mediated deletion frequency and polymorphism levels within the human and chimpanzee lineages

<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Chimpanzee</th>
<th>Human to Chimpanzee ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total observed L1IMDs</td>
<td>24</td>
<td>26</td>
<td>0.92</td>
</tr>
<tr>
<td>PCR amplified</td>
<td>23</td>
<td>26</td>
<td>-</td>
</tr>
<tr>
<td>Fixed present</td>
<td>18</td>
<td>22</td>
<td>-</td>
</tr>
<tr>
<td>Polymorphic loci</td>
<td>5</td>
<td>4</td>
<td>-</td>
</tr>
<tr>
<td>Polymorphic fraction</td>
<td>0.22</td>
<td>0.15</td>
<td>1.41</td>
</tr>
<tr>
<td>Adjusted polymorphic loci</td>
<td>10</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Adjusted number of L1IMDs</td>
<td>29</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>
Genomic Environment of L1IMDs

Contrary to non-autonomous Alu elements, L1s seem to have a preference for GC-poor regions of the genome (Ovchinnikov et al. 2001; Boissinot et al. 2004), which may be a consequence of either the L1 EN site preference (Cost and Boeke 1998) or of faster removal of L1s from GC–rich regions (Boissinot et al. 2001). To analyze whether L1 insertions causing deletions in the target sequence behaved differently from typical insertions, we analyzed GC content of 40 Kb of the flanking sequences (20 Kb each from the 5’ and 3’ ends) of the L1IMDs. Because poly(A) tails are shortened over time by the combined effects of mutation and replication slippage (Ovchinnikov et al. 2001) causing the presence of ‘fossil’ poly(A) tails in the 3’ flanking sequence, we avoided bias towards excessive adenosine residues by counting 20 Kb at the 3’ end after excluding 100 bp from the end of the poly-adenylation signal (AATAAA) of the L1 inserts. The mean GC content for the flanking regions of the human-specific and chimpanzee-specific L1IMDs was 38% and 39%, respectively. Compared to the ~42% average GC content of the draft human and chimpanzee genomes (Lander et al. 2001; Watanabe et al. 2004), L1IMD loci thus seem to be concentrated in AT-rich areas of the genome. Remarkably, ARDs in the human and chimpanzee genomes also show a preference for AT-rich locations (Callinan et al. 2005). The reduced GC content (~36%) around the eight youngest human L1 elements belonging to the L1Hs subfamily in our dataset (LH4, LH15, LH17, LH19, LH20, LH22, LH23, LH24) is consistent with previous findings (Boissinot et al. 2004).

To further characterize the genomic context in which L1IMDs occur, we calculated known and predicted gene densities in 4, 2 and 0.5 Mb windows lying immediately 5’ and 3’ to the L1IMDs (see supplementary data for gene counts in Batzer Laboratory Web site). Our results indicate that L1IMDs are concentrated in regions of low gene density (i.e. 1 gene per ~200 Kb,
which contrasts with the human genomic average of 1 gene per ~100 Kb) (IHGSC 2004). To test whether the size of the L1 insertions at L1IMD loci showed any relation to its surrounding gene density, we performed correlation tests for each window size (4, 2 and 0.5 Mb) in both chimpanzee and human. While we found no significant correlation \((-0.16 < r < 0.34, P > 0.05\) in all cases), the \(r\)-value itself was negative in five out of six tests, opening the possibility that analysis of a larger dataset of L1 insertions may show a trend towards shorter L1 insertions in gene-rich areas of the genome. Because the chimpanzee LC23 locus was located in an unusually gene dense region in the short arm of chromosome 9 \((i.e. 1 \text{ gene per } ~30 \text{ Kb})\), we repeated our correlation tests involving chimpanzee loci including and excluding this locus. However, the results were similar.

To characterize L1 insertions causing deletions within genes, we analyzed the 14 L1IMD loci (ten in human and four in chimpanzee) that were located within the introns of known or predicted genes. Eight of these were in collinear orientation with the gene transcript, while six were in antisense orientation. The average length of the L1 insertions within introns was considerably lower than the average L1 insertion length observed at non-intron L1IMD loci in both human and chimpanzee \( (849 \text{ vs. } 1601 \text{ bp and } 474 \text{ vs. } 1053 \text{ bp, respectively}) \). These 47% and 55% reductions, respectively, might indicate that smaller L1 insertions are better tolerated than longer ones within the introns of genes.

**Discussion**

The role of Alu and L1 retrotransposons in the creation of genomic instability is no longer a matter of dispute (Gilbert et al. 2002; Kazazian and Goodier 2002; Symer et al. 2002; Callinan et al. 2005). While extensive cell culture analyses have documented in detail the types and prevalence of genomic rearrangements by L1 insertion \textit{in vitro}, the possibility remains that \textit{in
vivo, evolutionary factors such as selection, variation in the number of actively retrotransposing elements and differences in effective population size (Boissinot et al. 2001; Hedges et al. 2004) may substantially impact the spectrum of these rearrangements. To test the latter, we made use of the genome sequence of our closest living relative, the common chimpanzee (Pan troglodytes), and performed a human/chimpanzee comparison of L1IMD events.

**Evolutionary Levels of L1IMD**

The previous cell culture analyses of Symer et al. (2002) and Gilbert et al. (2002), have both reported the presence of large (> 3 Kb) deletions associated with L1 retrotransposition, with one candidate in Gilbert et al. (2002) even deleting at least 24 Kb and possibly as much as 71 Kb of target sequence. However, such massive deletions are very unlikely to persist in the population because of the likelihood that such events would delete regions of the genome required for survival and thus would subsequently be removed by selection. Consistent with this view, we find that the vast majority of L1IMDs with some degree of evolutionary success are shorter than a few hundred bases in both the human and chimpanzee lineages. In fact, the total amount of lineage specific deleted sequences through L1IMD in the latest draft of the human genome is estimated to be only ~17.7 Kb, corresponding to an average deletion rate of ~3.5 Kb per haploid genome per million years (Myrs) within the ~5 Myrs since the divergence of humans and chimpanzees (Goodman et al. 1998; Chen and Li 2001). The rate of deletion in the chimpanzee genome is also similar at ~3 Kb per haploid genome per Myrs.

To estimate the number of human-specific L1 insertions, we reasoned that all human-specific L1 elements belong to only 3 subfamilies (L1Ta, L1preTa and L1PA2) (Myers et al. 2002; Salem et al. 2003; Furano et al. 2004). Given that both empirical (Boissinot et al. 2004) and theoretical (Hedges et al. 2004) evidence suggests that the analysis of a single genome
results in the recovery of only ~50% of all polymorphic elements in a subfamily, we estimated each L1 subfamily copy number as the sum of the number of fixed elements and twice the number of polymorphic elements detected in the human genome reference sequence. This resulted in a total of ~5800 L1 elements for these three subfamilies. However, not all of these L1 elements are specific to humans (Buzdin et al. 2003). Using the method of identification of human-specific L1 insertions from Buzdin et al. (2003), we conclude that ~1300 L1 elements have inserted in the human genome since the human/chimpanzee divergence. Given that L1 elements in the human genome have an average size of ~1 Kb (Lander et al. 2001), we calculate that the insertion of L1 elements within the past 5 Myrs resulted in the addition of ~1.3 Mb of sequence to the human genome. This is two orders of magnitude higher than the ~18 Kb length of sequence deleted in the same period by L1IMDs. On a larger time scale, assuming that ~2.2% of L1 insertions are associated with L1IMD in primates (29/1300 in humans) and the median deletion size of 21 bp from the L1IMD events in our study, the ~520,000 L1 elements that inserted in primate genomes were responsible for the deletion of a minimum of ~240 Kb of DNA sequences. However, if we perform the same calculation using the average L1IMD size of 655 bp, then almost 7.5 Mb of primate genomic DNA would have been deleted during the retrotransposition of L1 elements. It is also interesting to note that ~520 Mb (520,000 L1 elements with an average size of 1 Kb) of sequence has been added to the genome by the insertion of L1s in the same time period. This is reflective of the ongoing process of renewal of genomic sequences through the retrotransposition process.

**Chronological Framework of L1IMD Events**

We were able to place our L1IMD events in a chronological framework on the basis of (i) the results of the median-joining network analysis (Figure 4.3); (ii) the observation that about
two thirds of the human-specific L1IMDs are caused by L1PA2 insertions vs. about one third caused by members of the younger L1Hs subfamily; and (iii) only 20% of the chimpanzee-specific L1IMD events were specific to the common chimpanzee and 80% are shared with the pygmy chimpanzee. Taken together, these results suggest that L1IMD events in the human genome may have occurred to a large extent soon after the human/chimpanzee divergence when the L1PA2 subfamily was active, although they may be continuing to accumulate, as suggested by the non-trivial contribution of L1Hs members. In the chimpanzee lineage as well, the majority of L1IMDs is older than 1-2 Myrs, which corresponds to the divergence time of common and pygmy chimpanzees (Goodman et al. 1998; Chen and Li 2001). However, these observations may, at least partly, be influenced by the overrepresentation of older insertions within genomic sequences (i.e. younger events are more likely to be polymorphic than older events and could remain undetected when a small number of individuals were sequenced). Nevertheless, the fact that 23 out of 26 L1IMDs in the common chimpanzee involve L1PA2 elements suggests that the L1PA2 subfamily may still be actively undergoing retrotransposition in the chimpanzee lineage.

Interestingly, among the chimpanzee-specific L1IMDs, we found an ancient L1PA5 element (LC8) that was polymorphic. The L1PA5 subfamily is ~25 Myrs old (Furano et al. 2004). We excluded the possibility of polymorphism being maintained by balancing selection acting on this locus because of the low gene density in its vicinity. It is worthy to note that Bennett et al. (2004) also recently identified four polymorphic old AluS elements and one L1PA3 polymorphism. Therefore, this suggests that at least some copies of older L1 retrotransposon subfamilies can retain the ability of retrotransposition for extended periods of time similar to Alu elements (Han et al. 2005). Alternatively, it is possible that these polymorphisms have been maintained over a very long period of time by chance. Although this is expected to happen very
rarely, it may not be surprising to find a few such cases in view of the hundreds of thousands of L1 and Alu elements (Britten 1994; Batzer and Deininger 2002; Furano et al. 2004) that have inserted during primate evolution. However, we favor the former explanation in the case of the polymorphic L1PA5 element we detected, because DNA sequencing of the locus showed that the L1PA5 insert was specific to the chimpanzee lineage and absent from all other primate genomes we examined.

**Different Mechanisms May Exist for Different Deletion Sizes**

The sizes of the L1IMDs we identified are in general agreement with the size range of similar deletions (13 deletion events ranging from 2 bp-14 Kb) identified in a recent study of L1 retrotransposition in cell culture (Gilbert et al. 2005). However, our sample size for L1IMDs is substantially larger. Very large deletions like those seen in cell culture analyses (Gilbert et al. 2002; Symer et al. 2002) did not appear in our study, presumably because they are more likely to have been removed from the populations rapidly due to their deleterious nature (especially if they were located in gene-rich regions). Interestingly, in both the human and chimpanzee datasets, we noticed a tendency for the deletions to be either very short (*i.e.* < 100 bp) or, to a lesser extent, relatively large (> 1 Kb), which possibly indicates the concomitant action of two different mechanisms of L1IMD acting on different scales. This dichotomy in deletion sizes was also observed by Gilbert et al. (2002), and our data would seem to fit their general models for small and large L1IMD events, to which we propose further extensions to better explain some of the L1 structures that are unique to our study. In general, small deletions may be caused by the creation of 5’ overhangs by top strand cleavage being inexactly opposed to bottom strand cleavage in an upstream direction, with subsequent 5’-3’ exonuclease activity on both the exposed 5’ ends (Figure 4.5A). By contrast, larger deletions may be explained if the nascent L1
Figure 4.5. Models for the creation of L1IMDs. (A) Formation of small deletions. 5’ overhangs created by inexact cleavage of the top strand by the L1 EN are subject to 5’-3’ exonuclease activity, that removes small single-stranded stretches from both the plus and minus strands (dotted light blue lines), which would otherwise have been the templates for the formation of TSDs. Subsequent ligation of the L1 cDNA to the upstream minus strand sequence and plus strand sequence synthesis by cellular enzymes results in the creation of small deletions and an L1 insertion without TSDs. (B) Formation of large deletions. For any preexisting double strand break that has a 3’ overhang (red) for base pairing of the L1 cDNA (blue), a longer cDNA transcript is more likely to contain a stretch of sequence that has adequate complementary bases for annealing (pink) than a shorter one. Subsequent recombinational repair would remove a large segment of the target sequence, extending downstream to the original integration site (dotted black line) and resulting in an L1 insertion without TSDs.
cDNA invades a double-strand break with a 3’ overhang located upstream to the initial integration site (Figure 4.5B), with gap repair removing the intervening single-stranded segment and causing a large deletion (Gilbert et al. 2002; Gilbert et al. 2005). Additionally, we suggest that large deletions could result if palindromic stretches downstream of the original site of integration, mechanically or enzymatically held in single-strand conformation during the physical integration of the L1 DNA, formed hairpin loops which were subsequently removed by repair enzymes. Remarkably, a similar pattern of deletion size differences (small or large) also characterizes the deletions caused in the target sequence by the retrotransposition of Alu elements (Callinan et al. 2005). Taken together, the data from genomic deletions caused by L1 and Alu retrotransposon insertions are consistent with the view that two different mechanisms underlie the deletions of small and large stretches of target sequence, especially as both Alu retrotransposition-mediated deletions (Callinan et al. 2005) and the L1IMDs in our study are whole-genome analyses that should represent the comprehensive picture of such deletions.

**A Model for Correlation between Insert Size and Deletion Size**

In both our human and chimpanzee data sets, we noted a significant positive correlation between the size of the L1 insertion and the size of the deletion caused thereupon. In the extension of the model of Gilbert et al. (2002) described above for the creation of large deletions, we propose a probability-based mechanism to further explain the observed correlation (Figure 4.5B). Our model assumes that given the prior presence of a 3’ overhang in the double-strand break (which is a necessary prerequisite for the occurrence of the deletion by this mechanism) a longer segment of newly transcribed minus strand L1 cDNA is more likely to contain the adequate number of complementary bases (and thus be able to bind with sufficient strength) than a shorter segment. A longer stretch of complementarity than expected by chance between the end
of the L1 cDNA and the region surrounding the 5’ end of the L1 insertion in the ancestral (pre-insertion) sequence would provide support for this model. To quantify this parameter, we located (a) the 5’ start position of the L1 insertions with respect to the L1.3 consensus sequences and; (b) the site corresponding to the 5’ start position of the human-specific L1 insertions in the chimpanzee genomic sequence and *vice versa*. Next, we isolated 15 bp stretches of sequence in the 5’ direction from both these locations in the L1.3 consensus sequence and the genomic sequences, respectively, and aligned them. In all the 12 L1IMD loci that had large deletions corresponding to large L1 insertions (both sizes above 500 bp), we found between 27% and 53% complementary bases, which would indicate that potential binding sites were present in all the cases (see supplementary data for alignments in Batzer Laboratory Web site). Additionally, in seven out of the 15 loci, the first two (LH28, LH30, LC4, LC31) to three (LH17, LH27, LC29) bases in the 3’ end of the alignments were complementary. This further indicates that these bases could have been utilized for binding between the L1 transcript and the target sequence. Recent computational analyses of the 5’ junctions of young L1 insertions in the human genome (Zingler et al. 2005) suggest that microhomology-mediated end-joining is the likely mechanism for 5’-end attachment during the retrotransposition of 5’truncated L1 elements. Thus, our results support this hypothesis and indicate that longer L1 cDNA strands, because of the higher probability of possessing such microhomology with the pre-integration site, are better suited to the creation of longer genomic deletions by bridging double strand breaks. The presence of two double strand breaks (one at the original integration site and one upstream of it) would also lessen the chance of mechanical obstruction to the annealing of the L1 cDNA across the potential deleted region. We note that as proposed in Gilbert et al. (2005), the site of integration is very likely to be a “host/parasite battleground”, with the L1 cDNA trying to finish
reverse transcription and the host enzymatic machinery opposing it. Given the odds against the simultaneous occurrence of L1 insertion reaching comparatively near full-length and the presence of a double-strand break with a 3’ overhang conducive to binding, the lower number of large deletions corresponding to large insertions (6/26 in chimpanzee and 6/24 in human) lends support to our model.

**Rearrangements within the L1 Elements Associated with L1IMD.**

Six of the L1IMD loci were also characterized by rearrangements within the sequence of the L1 insertion, resulting in atypical L1 structures. Of these, two were both 5’ truncated and partially inverted (LC26 and LC27) while the other four (LH17, LH19, LH26 and LH31) were 5’ truncated non-inverted L1 elements that showed internal rearrangements. Previous cell culture studies have also shown that L1 rearrangements can occur during the process of retrotransposition (Gilbert *et al.* 2002; Gilbert *et al.* 2005). In our study, the presence of the homologous sequence from the respective closest ancestors allowed us to confirm that these loci did not have prior insertions of endogenous L1 elements at the pre-integration sites. The probability of two independent L1 insertions into the same locus after the human-chimpanzee divergence is extremely small, given the large size of the human and chimpanzee genomes and the estimated number of L1 insertions specific to these lineages (e.g. ~1300 in humans), which leads us to suggest that mechanistic processes led to the generation of these particular structures during the retrotransposition events. Of the non-inverted atypical L1 elements, LH19 and LH26 are strong candidates for gene duplication, with portions of the L1.3 consensus sequence repeated in parallel orientation without any intervening region (53 and 189 bp, respectively). LH17 and LH31 were 5’ truncated L1 insertions that showed two stretches of the consensus L1.3 sequence with a gap of ~300 bp in between them. We propose a novel mechanism for this
Figure 4.6. Models for the formation of deletion associated atypical L1 elements. (A) Formation of a non-inverted atypical L1 insertion resulting in a large deletion. The L1 mRNA (green) forms a loop, with microhomology stretches within its sequence annealing to each other. The resulting L1 cDNA (blue) has an internal breakpoint (orange) where a stretch of the consensus sequence (complementary to the loop) is missing. Arrows show the orientation of the two parts of the L1 insertion. (B) Formation of a 5’ truncation/inversion resulting in a large deletion. Annealing of the L1 mRNA (green) to a complementary sequence in the 3’ overhang of a preexisting double-strand break leads to the transcription of a second stretch (purple) apart from the original cDNA (blue). Subsequently, both dissociate from the mRNA and form an ‘inversion junction’ (circled in red). Recombinational repair removes the stretch of DNA between the double strand break and the original site of integration. Plus-strand synthesis results in a 5’ truncated L1 with the inverted portion being reverse complementary to the consensus sequence. Arrows show the orientation of the L1 segments in the inversion.
structure, by which stretches of microhomology within the L1.3 consensus sequence might have led to the L1 mRNA looping back on itself (Figure 4.6A), resulting in the formation of an L1 insertion with the characteristic structure observed and an associated deletion of target site DNA. The presence of at least one such 8 bp homologous stretch was visually confirmed by us in both the cases.

With respect to the 5’ truncation/inversions in our study (LC26 and LC27), a mechanism termed ‘twin priming’ has been suggested for the creation of such structures during L1 retrotransposition (Ostertag and Kazazian 2001). However, the existing model does not incorporate the possibility of creation of large deletions during this process. To provide a possible explanation for the large deletions caused at these loci (2973 and 1175 bp, respectively), we suggest a ‘modified twin priming’ model, whereby a stretch of complementarity between the extended L1 mRNA and a 3’ overhang formed at a preexisting double strand break would lead to a second site of priming on the mRNA (Figure 4.6B). Subsequently, dissociation of the two newly synthesized cDNA segments from the mRNA and the formation of an ‘inversion junction’, followed by double strand synthesis, would lead to the removal of the intervening DNA (between the original site of TPRT and the double strand break) with the formation of a rearranged L1 element with the truncation/inversion structure observed.

**Conclusion**

In conclusion, our study demonstrates that L1IMDs are not restricted to transformed cells but are also a feature of *in vivo* insertions as well, and that this process has been active in causing deletions in both the human and chimpanzee lineages. Our *in vivo* evolutionary analysis and prior *in vitro* cell culture studies of deletions caused by L1 retrotransposition provide pictures that differ at first sight, but can be reconciled by evolutionary factors. While 16-25% of L1
insertions identified in the cell culture studies cause deletions at the target site (Gilbert et al. 2002; Symer et al. 2002; Gilbert et al. 2005), only ~2.2% of existing human-specific L1 insertions seem to be directly linked to genomic deletions [compared to 0.2-0.4% for Alu elements (Callinan et al. 2005)]. As the currently available chimpanzee assembly covers ~95% of the genome sequence while the human genome sequence is considered to be “finished” (UCSC genome database), our human-chimpanzee comparison probably recovered most species-specific L1IMD events. A slight underestimation due to different levels of completion of the human and chimpanzee genome sequences could not account for the ~10-fold difference between in vivo and in vitro L1IMD rates. The difference in the rate of L1IMD estimated from cell culture-based analyses and genome-based analyses may more likely reflect the differences in the number of these events that are tolerated in the genome after natural selection has occurred. Thus, our study validates the use of cell culture retrotransposition assays as surrogate models to deduce the underlying mechanisms for these complex genomic rearrangements.

The extent of genomic deletion is reduced compared to the amount of sequence inserted by the L1 retrotransposition process. In addition evidence from our study indicates that many large L1IMDs such as those identified in cell culture assays do not persist in the primate lineage over time. We propose new mechanisms for the creation of some of the specific L1 structures reported in our analysis. Most of the existing human-specific deletions appear to have taken place soon after the divergence of the human and chimpanzee lineages. The atypical L1 elements created during the deletion process could also be sources for new L1 subfamilies in both the human and chimpanzee lineages (Saxton and Martin 1998; Gilbert et al. 2002).
Materials and Methods

Computational Analysis

To identify L1IMD candidate loci in the human genome, we first identified all L1 elements that have intact 3’ sequence in the July 2003 freeze of the human genome (hg16: UCSC genome database at http://genome.ucsc.edu/ENCODE/) by querying the genome sequence with the 50 bp of the 3’-end of the L1 consensus sequence (excluding the poly(A) tail), using the command line version of the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990). The BLAST output file was then processed by a set of in-house Perl programs to extract entries that contain matches with at least 96% sequence similarity to the query sequence over at least 40 bp, resulting in a total of 49,791 L1 entries. Using a cutoff value of 96% similarity ensured that the most recent L1 inserts (including human-specific events) were selected for further analysis. For each entry, 400 bp of sequence downstream of the start of the query (including the match to the query sequence, the poly(A) tail and the 3’-end flanking sequence) were extracted from the human genome sequence. The exact start of the 3’-end flanking sequences was determined for each entry by aligning it with the 50 bp L1 consensus sequence used as the initial query, with which a stretch of 100 adenosines was now included to simulate the poly(A) tail. The 3’ sequence immediately flanking the L1 element identified for each entry was then used as a query to search the chimpanzee genome (PanTro1; Nov. 2003 freeze). If the best match started immediately after the poly(A) tail, the locus was considered to be a human-specific L1 insertion and the start of the matching region was considered to be the insertion site in the human genome. For each identified locus, we extracted 1000 bp and 100 bp of sequence in the 5’ and 3’ regions of the pre-insertion site, respectively, from the chimpanzee genome. The 5’ chimpanzee sequences were then used to query the human genome. If a 1000 bp chimpanzee
sequence only matched the human sequence at its 5’ end, the unmatched sequence at the 3’ end was considered as a L1IMD candidate in the human genome. In cases where there was no match in the entire 1000 bp of the query sequence, the 5’ flanking sequences from the chimpanzee genome were progressively extended until a good partial match at the 5’ end could be identified in the human sequence. These cases were considered to represent deletions that were close to or longer than 1000 bp.

Chimpanzee L1IMD candidates were identified by reversing the query and target genomes and using the same approach as described above. All candidate loci were then subjected to manual verification, resulting in a total of 30 and 33 putative L1IMDs in the human and chimpanzee genomes, respectively.

**PCR Amplification and DNA Sequence Analysis**

To experimentally verify the L1IMD candidate loci, flanking oligonucleotide primers were designed using the primer design software Primer3 (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The primers were subsequently screened against the GenBank NR and HTGS databases using BLAST queries to determine if they resided in unique DNA sequences. Detailed information for each locus including primer sequences, annealing temperature, PCR product sizes and chromosomal locations can be found in the “Publications” section of our website (http://batzerlab.lsu.edu).

PCR amplification of each locus was performed in 25 µl reactions using 10-50 ng DNA, 200 nM of each oligonucleotide primer, 200 µM dNTP’s in 50 mM KCl, 1.5 mM MgCl2, 10 mM Tris-HCl (pH 8.4) and 2.5 units Taq DNA polymerase. Reactions were subjected to an initial denaturation step of 94° C for four minutes, followed by 32 cycles of one minute of denaturation at 94° C, one minute of annealing at optimal annealing temperature and one minute
of extension at 72° C, followed by a final extension step at 72° C for ten minutes on a Biorad™ iCycler thermocycler. Resulting PCR products were separated on 2% agarose gels, stained with ethidium bromide and visualized using UV fluorescence.

Individual PCR products were purified from the gels using the Wizard® gel purification kit (Promega) and cloned into vectors using the TOPO-TA Cloning® kit (Invitrogen). For each sample, three colonies were randomly selected and sequenced on an Applied Biosystems AB3100 automated DNA sequencer using chain termination sequencing (Sanger et al. 1977). All clones were sequenced in both directions using M13 forward and reverse primers to confirm the sequence, analyzed using the Seqman™ program in the DNASTAR suite and aligned using the BioEdit sequence alignment software package (http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

For each locus, this procedure was applied to one individual from each of five different primate species, including Homo sapiens (HeLa cell line ATCC CCL-2), Pan troglodytes (common chimpanzee; cell line AG06939B), Pan paniscus (bonobo or pygmy chimpanzee; cell line AG05253B), Gorilla gorilla (Western lowland gorilla; cell line AG05251) and Pongo pygmaeus (orangutan; cell line ATCC CR6301). The DNA sequences from this study are available in GenBank under accession numbers DQ017967-DQ018078.

Polymorphism Analysis

To evaluate the extent of polymorphism associated with the validated L1IMD loci, each locus was further amplified in the genomes of 80 humans (20 individuals from each of four populations, see below) and 12 unrelated common chimpanzees, following the PCR protocol described above. Our human population panel was composed of DNA from African-American, European and Asian populations (isolated from peripheral blood lymphocytes) available from previous studies in our lab and South American population DNA (HD17 and HD18) purchased
from the Coriell Institute for Medical Research. The common chimpanzee population panel was prepared from genomic DNA of twelve unrelated individuals of unknown geographic origin and subspecies affiliation, which was provided by the Southwest Foundation for Biomedical Research.

**Phylogenetic Analysis of L1IMDs**

To examine the phylogenetic relationships of the human and chimpanzee L1 elements identified in this study, we constructed a median-joining network (Cordaux *et al.* 2004; Han *et al.* 2005) using the software NETWORK ver. 4.1.1.0 (Bandelt *et al.* 1999) available at http://www.fluxus-engineering.com/sharenet.htm. The network was generated using a 94 bp stretch corresponding to positions 5930-6023 in the 3’ end consensus sequence of the L1Hs and L1PA2 reference sequences obtained from the RepeatMasker database. Elements LC9 and LH29 had to be excluded from this analysis because of truncations in the region analyzed.

**Analysis of Flanking Sequences**

For GC content analysis, we used the BLAST-Like Alignment Tool (BLAT) server (Kent 2002) available at http://genome.ucsc.edu/cgi-bin/hgBlat to isolate 20 Kb of flanking sequence in either direction from the reference human and chimpanzee draft sequences after adjustment at the 3’ end to prevent bias towards excessive adenosine residues (see results). We used the EMBOSS GeeCee server (http://emboss.sourceforge.net/apps/geecce.html) to calculate GC percentages. To characterize the gene-frequency neighborhoods of the L1IMDs, we pinpointed exact chromosomal location of the L1 insertions with BLAT, and then used the NCBI MapViewer interface (http://www.ncbi.nlm.nih.gov/mapview/) to map all known genes within 4, 2 and 0.5 Mb windows surrounding the 5’ and 3’ ends of the L1IMDs.
References


Britten, R. J. (1994). "Evidence that most human Alu sequences were inserted in a process that ceased about 30 million years ago." Proc Natl Acad Sci U S A 91(13): 6148-50.


CHAPTER FIVE:

HUMAN GENOMIC DELETIONS MEDIATED BY RECOMBINATION BETWEEN ALU ELEMENTS*

*Reprinted by permission of American Journal of Human Genetics
Introduction

With a copy number of >1 million, Alu elements are one of the most successful non-LTR retrotransposon families in the human genome (Lander et al. 2001). In addition to classic retrotransposition-associated insertion mutations, Alu elements can create genomic instability by the deletion of host DNA sequences during their integration into the genome and by creating genomic deletions associated with intrachromosomal and interchromosomal recombination events (Deininger and Batzer 1999; Callinan et al. 2005). Multiple features predispose Alu elements to successful recombination, including their proximity in the genome (one insertion every 3 kb on average), the high GC content of their sequence (~62.7%), and the remarkable sequence similarity (70-100%) among Alu subfamilies of widely different ages. Overall, the recombinogenic nature of these elements is reflected in the various forms of cancer and genetic disorders associated with Alu-mediated recombination events (Myerowitz and Hogikyan 1987; Huang et al. 1989; Marshall et al. 1996; Rothberg et al. 1997; Levran et al. 1998; Tvrdik et al. 1998; Deininger and Batzer 1999; Hattori et al. 1999; Rohlfis et al. 2000; Batzer and Deininger 2002).

However, clinical studies of isolated disease-causing deletions, although useful from a medical viewpoint and in demonstrating the existence of Alu Recombination-Mediated Deletions (ARMDs), do not adequately depict the overall contribution of this process to the architecture of the genome and the associated impact on gene function. The availability of a genome sequence for the common chimpanzee (Pan troglodytes), the closest evolutionary relative of the human lineage (CSAC 2005), has allowed us to perform a comparative genomic assessment of the extent of ARMD in the human genome over the past ~6 million years, since the divergence of the human and chimpanzee lineages (Miyamoto et al. 1987; Wildman et al. 2003). In this study,
we identified ~400 kb of human-specific ARMD, the distribution of which is biased toward gene-dense regions of the genome, which raises the possibility that ARMD may have played a role in the divergence of humans and chimpanzees. About 60% of the ARMDs are located in genes, and, in at least three instances, exons have been deleted in human genes relative to their chimpanzee orthologs. The nature of the altered genes suggests that ARMD might have played a role in shaping the unique traits of the human and chimpanzee lineages. Mechanistically, we characterized the physical aspects of the deletion process and proposed different models for ARMD.

**Results**

**A Whole-Genome Analysis of Human-Specific ARMD Events**

To identify putative ARMD loci, we first computationally compared the human and chimpanzee genomes. Subsequently, we manually inspected and, if needed, experimentally verified individual loci. Of the 1332 computationally predicted deletions that we initially recovered, 461 were discarded after manual inspection (Table 5.1). The causes for rejection of computationally predicted ARMD loci were: (a) insertion of an *Alu* or other retroelement at the

<table>
<thead>
<tr>
<th>Classification</th>
<th>No. of loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computationally predicted deletion loci</td>
<td>1322</td>
</tr>
<tr>
<td>Discarded after manual inspection</td>
<td>461</td>
</tr>
<tr>
<td>Candidate ARMD events:</td>
<td>871</td>
</tr>
<tr>
<td>False-positive events (<em>Alu</em> insertion in chimpanzee):</td>
<td>379</td>
</tr>
<tr>
<td>Confirmed by PCR analysis</td>
<td>189</td>
</tr>
<tr>
<td>Analysis based on TSD structure</td>
<td>190</td>
</tr>
<tr>
<td>ARMDs:</td>
<td>492</td>
</tr>
<tr>
<td>Confirmed by PCR analysis</td>
<td>163</td>
</tr>
<tr>
<td>Analysis based on TSD structure</td>
<td>329</td>
</tr>
</tbody>
</table>
orthologous chimpanzee locus, which leads to the presence of sequence that the computer erroneously assumed to be deleted in the human genome (38 cases), (b) authentic deletion products in the human genome that were not products of \textit{Alu-Alu} recombination (211 cases), and (c) computational errors in alignment of the human and chimpanzee genomes (212 cases). On the basis of sequence architecture, the remaining 871 loci represented putative ARMD events in the human lineage. All of these loci were further manually inspected and were analyzed, for comparison of the ancestral predeletion and human postdeletion states, by use of a TSD-based strategy as described below (see Materials and Methods). In addition, we experimentally verified the authenticity of 352 candidate ARMD loci by PCR (Table 5.1 and Figure 5.1). To be conservative, we discarded all loci in which an alternative mechanism (\textit{e.g.}, random genomic

\textbf{Figure 5.1. ARMD in the human genome.} Individual ARMD candidate loci amplified by PCR. (A) Agarose-gel chromatograph of PCR products derived from an authentic human-specific ARMD event. (B) Agarose-gel chromatograph of PCR products derived from an ARMD false positive event (\textit{Alu} insertion in chimpanzee). The DNA templates used in each reaction are shown above the chromatographs.

107
deletion), distinct from ARMD, could have produced the deletion. Specifically, ARMD events can be distinguished from random genomic deletions occurring at Alu insertion sites because an ARMD event reconstitutes an uninterrupted chimeric Alu element (i.e., with no internal deletion), whereas the probability of this happening through chance alone (as would be the case with a random deletion) is remote. Indeed, the probability of two ~280-bp Alu elements breaking by chance at a homologous site is only 1 in ~80,000 (1 in 280 × 1 in 280). Hence, although we cannot formally exclude the possibility that a few random deletions may precisely mimic the ARMD process, we believe the overall impact of these nonauthentic events on our estimates would be minimal.

The manual verification of the 871 loci resulted in a final dataset of 492 ARMD events spanning the entire human genome (Table 5.1). Nine ARMD loci on the Y chromosome were all located in the pseudoautosomal part of this chromosome and hence were identical copies of deletion loci on the X chromosome. As a result, each event was counted only once during the analysis. In general, the loci analyzed in this study suggest that the combination of computational data mining and experimental validation is the “gold standard” when conducting comparative genomic searches for lineage-specific deletions. As we observed during the course of this study, lineage-specific insertions in one genome stand a risk of being characterized as deletions in the other when only two genomes are compared in a computational analysis. In our analysis, we minimized the chances of including such events by using three other hominoid genomes as controls during experimental verification of the events.

**Extent of Genomic Deletion and Size Distribution of ARMD Events**

The number of ARMD events is positively correlated with the number of Alu elements present on each chromosome ($r = 0.69$; $P < .0005$). This is expected, since physical proximity
between repetitive elements strongly predisposes them to recombination (Inoue and Lupski 2002). Simultaneous mapping of ARMD loci and all *Alu* insertions on each chromosome highlights the tendency for deletions to cluster with regions of high local *Alu* density (Figure 5.2). Additionally, sequence analysis of the *Alu* elements involved in ARMD events indicates that the number of elements from each *Alu* subfamily (Figure 5.3) is proportional to their genomewide copy number (Batzer and Deininger 2002), with no bias observed for elements from

![Figure 5.2](image_url)

*Figure 5.2. Density of ARMD events (red lines) and all *Alu* insertions (blue lines) on individual human chromosomes.*
Figure 5.3. *Alu* subfamily composition in ARMD events. (A) Proportion of *Alu* elements involved in ARMD events (unblackened bar) versus total number of *Alu* elements (blackened bar) for each subfamily. (B) Subfamily ratios of upstream and downstream *Alu* elements involved in ARMD events (unblackened and blackened bars, respectively).
older subfamilies (such as *AluJ*) that would have had more time for recombination because of their age. This implies that *Alu* elements throughout the genome have similar chances of recombining with each other, as opposed to a mechanism of preferential recombination between members of an individual subfamily, and that proximity between the elements is the major factor involved in the process. Additional evidence supporting this position comes from the fact that ~40% (197 of 492) of ARMD events result from inter-*Alu* subfamily recombinations. However, within this context, the amount of sequence identity between the two elements at a locus also appears to be proportional to their chances of successful recombination, since young *AluY* elements are over-represented at ARMD loci compared with their total number in the genome, whereas the opposite is true for older, highly diverged *AluJ* elements.

The total amount of genomic sequence deleted by this process in the human lineage (i.e., after the human-chimpanzee divergence ~6 million years ago) is estimated to be 396,420 bp. This is probably a conservative estimate, since our comparative analysis of the human and chimpanzee genomes detects ARMD events only between *Alu* elements that were inserted before the human-chimpanzee divergence. Therefore, it would miss ARMD loci involving newly inserted human-specific *Alu* elements (Carter *et al.* 2004; Otieno *et al.* 2004). However, the contribution of human-specific *Alu* elements to ARMD is probably relatively limited, given that there are only ~7000 such insertions (CSAC 2005), as compared with >1 million *Alu* elements shared between the human and chimpanzee genomes.

The ARMDs range in size between 101 and 7255 bp, with an average size of ~806 bp. A histogram of the size frequency distribution of ARMDs reveals a skew towards shorter ARMD sizes, with ~75% (368 of 492) of the deletions shorter than 1 kb (Figure 5.4). Thus, the median ARMD length of 468 bp better represents the most common size category. However, in terms of
Figure 5.4. Size distribution of human-specific ARMD events, displayed in 100-bp bin sizes.
total genomic sequence deleted, the ~25% ARMD events >1 kb were responsible for ~62% (245,263 of 396,420 bp) of the total sequence deleted. Our computational analyses did not return any ARMD loci with deletions <100 bp. Strictly speaking, Alu-Alu recombination elements should not cause deletions of <300 bp (i.e., the length of a complete Alu element), because, even if the recombining elements were immediately adjacent to each other, this would be the smallest possible amount of sequence deleted. However, the individual left and right monomers of the dimeric Alu element can freely exist in the genome, and these types of elements are accounted for in our study. This resulted in the ability of our study to detect deletions smaller than the expected minimum of ~300 bp.

**Structural Characteristics of ARMD Events**

Pairs of Alu elements that recombined to cause human genomic deletions were in parallel orientation in almost all cases (490 of 492). Most probably, this is a direct consequence of the increased length of hybridization available from this arrangement, as the parallel orientation would allow for homology over longer stretches between pairs of Alu elements located on the homologous chromosomes during recombination. Analysis of the Alu trios at each locus (i.e., two pre-ARMD Alu elements in chimpanzee and one postdeletion element in human) suggests four possible recombination mechanisms. Of these, unequal recombination between adjacent Alu elements on homologous chromosomes (Figure 5.5A, left panel) accounts for ~74% (366 of 492) of the deletions, whereas the other three putative mechanisms were less frequent (Figure 5.5B-5.5D). Our study captures both intrachromosomal (Figure 5.5A, right panel) and interchromosomal (Figure 5.5A, left panel) recombination-mediated deletions.

For each deletion, we located the points on the Alu consensus sequence where the two intact chimpanzee Alu elements involved in the recombination were broken and subsequently
Figure 5.5. Four different types of the recombination between Alu elements. Black and gray lines represent flaking and intervening regions, respectively. Dotted red circles denote recombining regions, and red and pink arrows represent TSDs of the two elements, respectively. (A) Interchromosomal (left) and intrachromosomal (right) recombination between two Alu elements (light blue and green). (B) Recombination between two Alu elements, one of which previously inserted into the other (L and R indicate left and right Alu monomers). (C) Recombination between left and right Alu monomers on two different elements. (D) Recombination between oppositely oriented Alu elements (only two cases observed).
Figure 5.6. Recombination window between \textit{Alu} elements and percentage frequencies of breakage (during recombination) at different positions along an \textit{Alu} consensus sequence. The structure of a typical \textit{Alu} element is shown in the lower panel. The length of the \textit{Alu} consensus sequence is \textasciitilde 282 bp, excluding the 3’ poly(A) tail. The element consists of left (light blue) and right (purple) monomers. The left monomer contains an RNA polymerase III promoter (green boxes A and B). TSDs (red boxes), usually between 7-20 bp long, are created at each end during the \textit{Alu} insertion process.

attached to each other to form the resulting single human \textit{Alu} element. Plotting the frequency distribution of recombination breakpoints at different positions on the \textit{Alu} consensus sequence revealed a recombination “hotspot” encompassing positions 21-48 (Figure 5.6), which is consistent with an earlier study based on a smaller dataset (Rudiger \textit{et al.} 1995). To uncover the reasons underlying the observed “adhesive” nature of this part of the \textit{Alu} element, we aligned the consensus sequences of 10 \textit{Alu} subfamilies (\textit{Alu}Jo, \textit{Alu}Jb, \textit{Alu}Sx, \textit{Alu}Sp, \textit{Alu}Sq, \textit{Alu}Sg, \textit{Alu}Sg1, \textit{Alu}Sc, \textit{Alu}Y, and \textit{Alu}Yd8) and analyzed the levels of conservation and GC content of regions that tended to recombine at frequencies exceeding the mean (0.08) across all positions in our ARMD events. This analysis indicated that both parameters were substantially higher in these
regions than in the rest of the *Alu* sequence, with the major inferred recombination hotspot referred to above showing >60% GC (as compared to the ~62.7% average GC content for the 10 *Alu* consensus sequences) and complete conservation across all subfamilies. Although these factors may be responsible for higher recombination frequencies in this region, other reasons are also plausible, such as the location of this stretch near the L1 endonuclease cleavage site at the 5’ end of the *Alu* element, which make it closer to putative breakage sites during the recombination process.

**Genomic Environment of ARMD Events**

*Alu* elements in the human genome show a preference for high GC content areas, except for the most recently integrated subfamilies (Lander *et al.* 2001; Cordaux *et al.* 2006). However, since only a fraction (984 of ~1.2 million) of the total number of *Alu* insertions is associated with the ARMD process, it may well be that, in this respect, the deletions themselves behave differently from the *Alu* family as a whole. To characterize the sequence context in which ARMD events occur, we calculated the percentage GC content in 20-kb windows of flanking sequence centered on the ARMD loci. Compared with previous analyses of *Alu* and L1 insertion-mediated (as opposed to postinsertional recombination-mediated) genomic deletions (Callinan *et al.* 2005; Han *et al.* 2005), which are preferentially localized in low-GC content neighborhoods (~38% GC), ARMD events tend to occur in high-GC content regions (~45% GC content on average). This is also substantially higher than the ~41% global average GC content of the human genome (Lander *et al.* 2001). Since high-GC content areas of the genome also show higher gene density (Lander *et al.* 2001; IHGSC 2004), we analyzed 4 Mb (2 Mb in either direction) of sequence flanking ARMD events, for the presence of known and predicted human RefSeq genes. We found the gene density around ARMD events to be, on average, one gene per
66 kb, which, as expected, is higher than the global average gene density (approximately one gene per 150 kb) (IHGSC 2004) and the average gene density in the vicinity of L1 insertion-mediated deletions (approximately one gene per 200 kb) (Han et al. 2005). Thus, ARMD events seem to be concentrated in gene-rich regions of the human genome. The tendency for clustering of ARMD events and genes becomes even more apparent when their densities are plotted side by side on each chromosome (Figure 5.7). Interestingly, the neighboring GC content showed a

![Figure 5.7. Density of ARMD events (red lines) and RefSeq genes (blue lines) on individual human chromosomes.](image-url)
significant negative correlation with the deletion size ($r = -0.17; P < .0001$).

About 45% (219 of 492) of ARMD events were located within known or predicted human RefSeq genes, and an additional ~15% (76 of 492) were in intergenic regions of the human genome but were located within predicted chimpanzee genes. Since ≤25% of the human genome represents currently known genes (including both exon and intron sequences) (Venter et al. 2001; IHGSC 2004; Sakharkar et al. 2004), the relative density of ARMD events within genic regions is remarkably high. This would indicate that, a priori, the probability of this process interfering with gene function is higher than the two retrotransposon insertion-mediated deletion mechanisms mentioned above. To test this hypothesis, we extracted the ancestral prerecombination sequence at each ARMD locus (i.e., the sequence present in the chimpanzee genome but deleted in the human genome) and analyzed its location in the chimpanzee genome to see whether it mapped to a protein-coding region. In three instances, the ARMD event deleted an entire exon from a gene that is functional in the chimpanzee genome. To confirm that these three ARMD loci did not represent assembly errors, we resequenced them in the human and chimpanzee genomes. One of the three genes, LOC471177 is a model chimpanzee gene similar to the human CHRNA9 gene (MIM 605116), a member of the ligand-gated ionic channel family that is associated with cochlea hair cell development (Lustig and Peng 2002). Of the other two, LOC452742 is similar to the human model gene LOC440141 (which encodes the mitochondrial ribosomal protein S31), and LOC471116 encodes a hypothetical protein with a conserved high-molecular weight glutenin subunit.

**Characteristics of the Genomic Sequences Lost during ARMD**

Previous analyses have suggested that recombination may be responsible for the bias towards high-GC content areas observed for Alu elements in the human genome (Brookfield
If so, one would expect that ARMD events preferentially remove low-GC content sequence, consequently causing a shift in the opposite direction. However, simulation results revealed that the GC content of both RSNA and RSG (41.9% and 41.4%, respectively) were significantly lower than the ~45.4% GC content of the observed deleted sequences (\(P\) value < .00001 in both cases). Moreover, the RSNA and RSG \textit{Alu} contents (20.6% and 11.4%, respectively) also had significantly lower values when compared to the \textit{Alu} content of the observed deleted sequences (27.0%; \(P < .0001\), compared with both RSNA and RSG). In addition to \textit{Alu} elements, repetitive DNA from elements of other families, for a total of 86,442 bp, was removed by ARMD (Table 5.2).

### Table 5.2. Genomic DNA sequences deleted by ARMD

<table>
<thead>
<tr>
<th>Classification</th>
<th>Amount (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Alu}\textsuperscript{a}</td>
<td>192,102</td>
</tr>
<tr>
<td>MIR</td>
<td>4780</td>
</tr>
<tr>
<td>7SL RNA</td>
<td>306</td>
</tr>
<tr>
<td>L1</td>
<td>41,491</td>
</tr>
<tr>
<td>L2</td>
<td>7312</td>
</tr>
<tr>
<td>L3</td>
<td>163</td>
</tr>
<tr>
<td>LTR</td>
<td>23,336</td>
</tr>
<tr>
<td>MER1</td>
<td>3575</td>
</tr>
<tr>
<td>MER2</td>
<td>2555</td>
</tr>
<tr>
<td>Other DNA repeat elements</td>
<td>669</td>
</tr>
<tr>
<td>Simple repeat</td>
<td>2255</td>
</tr>
<tr>
<td>Nonrepetitive DNA</td>
<td>117,876</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>396,420</strong></td>
</tr>
</tbody>
</table>

\( \textsuperscript{a} \) Including truncated \textit{Alu} elements.
Discussion

Role of the ARMD Process in Human Genome Evolution

Retrotransposons such as Alu elements are associated with size expansion in primate genomes (Petrov 2001; Liu et al. 2003). This is a consequence of their increasing copy number and also an indirect result of their implication in homology-mediated segmental duplications (Bailey et al. 2003). For example, the high retrotransposition activity of the Alu family in the human lineage has been responsible for the addition of ~2.1 Mb to the human genome within the past ~6 million years (Hedges et al. 2004; CSAC 2005). In this context, our study provides the first comprehensive assessment of a postretrotransposition process that has had an appreciable impact on the dynamics of human genome-size evolution. Previous in vivo evolutionary analyses have characterized human and chimpanzee genomic deletions generated on Alu and L1 insertion (Callinan et al. 2005; Han et al. 2005). However, the combined extent of human-specific deletion attributable to these mechanisms is an order of magnitude lower than that resulting from ARMD (~30 Kb for Alu and L1 insertion-mediated deletions combined, vs. ~400 kb for ARMD alone). The relative amounts of sequence inserted (by Alu retrotransposition) and deleted (by ARMD) imply an Alu-mediated sequence turnover rate of ~20% (i.e., ~400-kb deleted sequence vs. ~2.1-Mb inserted sequence) in the human genome within the past ~6 million years. This indicates that ARMD is capable of mitigating, at least partially, the increase in genome size caused by new retrotransposon insertions.

The scope of retrotransposon-mediated reduction of genome size further broadens when we consider that L1 elements (another mobile DNA family) are capable of creating deletions by a recombination process analogous to ARMD (Burwinkel and Kilimann 1998; Bailey et al. 2003). The higher average distance between L1 insertions in the human genome (one element
per 6.3 kb) (Lander *et al.* 2001) as well as the lower GC content of L1 elements (~43%, excluding the poly(A) tail) (Dombroski *et al.* 1993) may be contributing factors to the paucity of L1-mediated recombination events as compared to ARMD events. Even so, the greater length of L1 elements (~6 kb vs. ~300 bp for Alu elements) (Dombroski *et al.* 1993) and their high copy number (~520,000 elements) (Lander *et al.* 2001) still indicate that this family may represent another source of retrotransposon recombination-mediated deletions in the human genome. However, a broader comparative genomic study of such retrotransposon recombination-mediated deletion mechanisms in both the human and chimpanzee lineages is needed before the comprehensive role of transposable elements in primate genome-size evolution can be determined. In this respect, at least in the case of plants, studies have already shown that the genome of *Arabidopsis thaliana* uses recombination-mediated deletion to counterbalance genome expansion, which may be one of the reasons for its remarkably compact size (Devos *et al.* 2002).

Recent analyses of human-genome variation have emphasized the importance of deletions in creating genetic diversity among humans (Iafrate *et al.* 2004; Conrad *et al.* 2006; Hinds *et al.* 2006; McCarroll *et al.* 2006). Our results offer insight into one of the mechanisms that may contribute to the creation of such deletions. Interestingly, the majority of the deletion variants identified in the recent studies cited above (McCarroll *et al.* 2005; Conrad *et al.* 2006; Hinds *et al.* 2006) are polymorphic between human individuals or populations. Although their contribution to between-individual genetic diversity is undisputed, the persistence of these deletions over evolutionary time cannot be taken for granted. By contrast, the deletions reported in our study have a low polymorphism rate (15%) among the 80 diverse human genomes we genotyped. This may represent the difference in the comparative timescales of these between-
human genomic deletion variants (Conrad et al. 2006; Hinds et al. 2006) and our human-chimpanzee comparison. In an earlier analysis (Han et al. 2005), we showed that only a fraction of the deletions caused by in vitro L1 retrotransposition (Gilbert et al. 2002; Symer et al. 2002; Gilbert et al. 2005) persist in the human genome over evolutionary time. Additionally, comparative genomic studies across a range of organisms indicate that genomic deletions that ultimately reach fixation tend to be smaller than those detected before any selective force operates (i.e., in cell culture analyses) (Gregory 2004). Analogous to this situation, ARMD events (which had a median length of 468 bp) were, in general, smaller than the deletion variants characterized by the recent studies of human-genome variation, which had a range of 1-745 kb (McCarroll et al. 2005; Conrad et al. 2006; Hinds et al. 2006). Since our study focuses on a longer evolutionary time scale and would preferentially capture those ARMD events that have not been selected against, it is possible that the deletions we detected represent the smaller evolutionary remainder of a group of older and perhaps larger deletions.

**ARMD as an Agent in Human-Chimpanzee Divergence**

The human and chimpanzee genomes are characterized by only ~1.4% divergence at the nucleotide-sequence level (Ebersberger et al. 2002; Watanabe et al. 2004; CSAC 2005; Newman et al. 2005). With the completion of the draft chimpanzee genome, the focus has shifted to identifying differences rather than locating similarities. Regarding actual genetic change, although a comprehensive assessment of protein-coding portions of the chimpanzee genome is not yet available, functional classes of genes that are under accelerated evolution in one lineage or the other have been characterized by recent studies (Clark et al. 2003; Dorus et al. 2004).

In the context of possible events that have altered gene structure or expression between the human and chimpanzee lineages, our study illustrates almost 300 lineage-specific deletions
within protein-coding human or chimpanzee RefSeq genes; it is conceivable that at least some of these ARMD events contributed to phenotypic divergence. Gene shuffling by recombination between Alu elements has already been reported in the human genome (Babcock et al. 2003). Furthermore, in at least two documented instances, Alu elements have caused hominoid lineage-specific exon deletions in functional genes: through an insertion-mediated deletion in the human CMAH gene (Hayakawa et al. 2001) and through ARMD in the human ELN gene (Szabo et al. 1999). In the present study, we show three additional instances in which ARMD has caused the loss of an exon in a human gene, as compared to its chimpanzee ortholog. Of particular interest is the deletion of the fourth exon in the predicted chimpanzee gene LOC471177, which is orthologous to the human CHRNA9 gene. In the human lineage, CHRNA9 is an ionotropic receptor with a probable role in the modulation of auditory stimuli (Glowatzki and Fuchs 2000; Lustig and Peng 2002). Modifications in the function of this gene may lead to a reduction in basilar membrane movement and thus affect the dynamic range of hearing. Although the characterization of the actual gene expression pathways that underlie the differences of humans and chimpanzees has just begun, preliminary data suggest that differences in auditory genes may comprise a subset of the total change (Clark et al. 2003). This is reflected in the fact that the tonal range of normal human speech is probably outside the optimal reception of the chimpanzee auditory system (Martinez et al. 2004). Thus, it is conceivable that CHRNA9 is a member of the group of genes (such as FOXP2 and TECTA) that may be responsible for the unique auditory and olfactory traits that distinguish humans and chimpanzees (Enard et al. 2002; Clark et al. 2003).

Even excluding the three ARMD events listed above that deleted exons, 292 other events located within genes have deleted 229,205 bp of intronic sequence. Although further analysis will be
required for conclusive assignment of specific roles, if any, to the deleted intronic sequences, it is possible that some of them may be associated with alteration of splicing patterns.

**Does ARMD Play a Role in Modifying Alu Distribution?**

Recently integrated or young Alu elements are inserted relatively randomly in the genome; by contrast, older Alu elements are preferentially found in GC-rich areas of the genome (Lander et al. 2001; Cordaux et al. 2006). Both selective and neutral explanations have been offered for this uneven genomic distribution of Alu elements. However, a selective process (Lander et al. 2001) is inconsistent with polymorphism patterns of recently integrated Alu elements (Cordaux et al. 2006). An alternative neutral explanation for the enrichment of Alu elements in GC-rich regions over time involves their preferential loss from GC-poor regions (Brookfield 2001; Lander et al. 2001; Jurka et al. 2004; Hackenberg et al. 2005), a process that might be influenced by ARMD.

However, the high GC content of deleted sequences, along with the preferential occurrence of ARMD events in GC-rich regions, argues against this possibility. To result in the Alu distribution shift, the deletions would need to be much larger in GC-poor than in GC-rich regions (Cordaux et al. 2006). Consistent with this hypothesis, our results indicate that ARMD size is negatively correlated with GC content. However, although ARMD events are significantly larger in GC-poor (i.e., <41% genome average) than in GC-rich (i.e., >41% genome average) regions (~1100 vs. ~700 bp; t test, $P = .0007$), three times as many ARMD events occurred in GC-rich as in GC-poor regions (369 vs. 123). Consequently, the net amount of sequence deleted from GC-poor regions is half that of GC-rich regions (~135 kb vs. ~261 kb). Given that GC-poor regions encompass ~58% of the genome (Lander et al. 2001), it is unlikely that ARMD has played a substantial role in mediating the shift in the Alu distribution towards heavy isochors.
(CSAC 2005). Nevertheless, other types of deletions could contribute more significantly to the yet-unexplained Alu genomic distribution shift.

Interestingly, the results from the simulations we performed suggest that sequences deleted through ARMD contain a statistically significant excess of Alu elements. This implies that the ARMD process may contribute to effective removal of Alu elements from regions in which they have reached high densities. Given the fact that abnormally high Alu density within a particular genomic region would also make it prone to recombination-mediated deletions, this result may reflect a selective force that counteracts the deletion process.

**A Potential Mechanism of Double-Strand Break (DSB) Repair**

Previous analyses have demonstrated the ability of both LTR and non-LTR retrotransposons to cause DSBs in genomic DNA (Zimmerly et al. 1995; Gasior et al. 2006). In particular, the role of the L1 family in the creation and subsequent resolution of DSBs has been extensively analyzed (Gilbert et al. 2005). In vitro, cell-culture studies have shown that homology-directed repair is a major mechanism for patching such breaks and that recombination between repetitive elements is one possible pathway for this process (Richardson and Jasin 2000). Recombination rates are highly increased on artificially induced DSBs in cultured cells, which further implicates this mechanism in “tying up the loose ends” at potentially deleterious DSB loci (Liang et al. 1998).

*In vitro*, a 3:1 excess of recombination deletions versus conservative noncrossover situations was detected in a study of homology-mediated repair at a single predefined DSB locus (Liang et al. 1998). In this context, some of the loci in our study may represent instances of homology-mediated DSB repair, in which the presence of highly conserved Alu sequences on both sides of the break has facilitated its patching. This would be particularly true for loci at
which the deletion would otherwise be selectively neutral, since the act of having repaired a potentially lethal DSB would give it an instant advantage, if only for propagation to the immediately next generation.

**Conclusion**

As high-throughput sequencing techniques become more advanced, the focus of evolutionary studies is shifting more towards genomewide analyses. Our study represents such a situation: we have comprehensively analyzed a major deletion mechanism in the human genome that was previously known only as a result of mutations in isolated disease-causing loci. In view of the fact that deletions are being recognized as an important class of genetic variants that contribute to human diversity and evolution (Conrad *et al*. 2006; Hinds *et al*. 2006; McCarroll *et al*. 2006), ARMD represents one of the major mechanisms for generating such deletions in humans. Moreover, the frequent occurrence of ARMD in gene-rich regions of the genome demonstrates the importance of this process in both biomedical and evolutionary studies. Overall, our results open the field to further studies of deletions caused by recombination between mobile elements and demonstrate one of the possible ways by which the human lineage may have developed a set of unique genetic traits.

**Materials and Methods**

**Computational Data Mining for Identifying Candidate ARMD Loci**

We extracted 400 bp of 5’ and 3’ genomic sequence flanking all human *Alu* elements (Figure 5.8). Next, we joined the two 400-bp stretches to form a single sequence (the “query”). For each query, the best match in the reference chimpanzee genome (PanTro1 [November 2003 freeze]) was identified. Then, the sequence stretch in the chimpanzee genome between the two regions that aligned with the two-400 bp halves of the query (the “hit”) was extracted and
aligned with the human *Alu* sequence initially used to design the query (the “query *Alu*”), by use of a local installation of the National Center for Biotechnology Information Blast 2 Sequences BL2seq utility. Following are the possible alignment results for each sequence pair (see corresponding diagrams in Figure 5.8).

A. There is no match. In this case, an *Alu* insertion-mediated deletion has occurred in the human genome at that locus.

B. There is only one alignment block, and:

B.1. The hit is identical to the query *Alu*. This is shared ancestry of an *Alu* insertion.

B.2. The hit is longer than the query *Alu*, and the extra sequence is entirely composed of a poly(A) tract downstream of the *Alu* sequence. This is a case of extension of the *Alu* poly(A) tail.

B.3. The hit consists of the query *Alu* plus some extra non-poly(A) sequence, and:

B.3a. The extra, non-poly(A) sequence is downstream of the poly(A) tail. This could be a gene conversion event in the chimpanzee genome.

B.3b. The extra, non-poly(A) sequence is upstream of the query *Alu* element or there is extra sequence at both ends. This is a possible *Alu* insertion-mediated deletion event in the human genome.

C. There is more than one alignment blocks, and:

C.1. The beginning and end of the hit match the query *Alu* and the hit is at least 100 bp longer than the query *Alu* sequence (since this size would approximate the expected lower ARMD size limit). This is a candidate ARMD event in the human genome.

C.2. At least one end of the hit has no match to the query *Alu*. This is another possible case for an *Alu* insertion-mediated deletion in the human genome.
Figure 5.8. Computational data mining for human lineage-specific ARMD loci.
(A) No match between query Alu and hit (possible Alu insertion-mediated deletion).
(B.1) Query Alu and hit are identical (shared ancestry of an Alu insertion). (B.2) Hit is longer than query Alu and the extra sequence is a poly(A) tract downstream of the query Alu (extension of the Alu poly(A) tail). (B.3) Hit consists of query Alu plus extra non-poly(A) sequence, and the following. (B.3a) Extra, non-poly(A) sequence is downstream of the query Alu poly(A) tail (may be gene conversion event in the chimpanzee genome). (B.3b) Extra, non-poly(A) sequence is upstream of the query Alu element or there is extra sequence at both ends (possible Alu insertion-mediated deletion event). (C.1) Beginning and end of the hit match query Alu and the hit is at least 100 bp longer than query Alu (candidate human lineage-specific ARMD event). (C.2) At least one end of the hit has no match to query Alu (possible Alu insertion-mediated deletion).

We retained all loci matching case C.1 as pairs of FASTA files (i.e, the orthologous human and chimpanzee sequences). Each human sequence contained the query Alu and its 400-bp flanking sequences on each side, and each chimpanzee sequence contained the entire hit that aligned with the query flanking sequences. All candidate ARMD loci were then manually
inspected and, if necessary, verified by wet bench (PCR) analysis. Orthologous human and chimpanzee sequences for each locus are available from the “Publications” section of the Batzer Laboratory Web site.

**Inspection of Target Site Duplications (TSDs)**

A typical *Alu* insertion is flanked on both sides by identical (or nearly perfect) short, direct repeats (7-20 bp) termed “target-site duplications” (TSDs) (Deininger and Batzer 2002). The single *Alu* element remaining at a human candidate ARMD locus is characterized by the apparent absence of TSDs, since it is composed of fragments from a pair of *Alu* elements with mutually different TSDs, situated at the orthologous ancestral locus (which persists in the chimpanzee genome). This hallmark of the ARMD process offers a direct means of confirming the “chimeric” origin of the human *Alu* element at a deletion locus. Using this property as our basis for verification, we manually inspected all candidate loci returned by the computational analysis. In an unambiguous ARMD event, the TSDs of the two *Alu* elements immediately upstream and downstream of the deleted portion in the chimpanzee genome were perfect matches with the 5’ and 3’ TSDs, respectively, of the orthologous single human *Alu* element. In the next possible scenario, the sequence on any one side of the human *Alu* (upstream or downstream) matched the TSDs of the chimpanzee element on the corresponding side, but the other chimpanzee *Alu* element itself lacked TSDs. However, the sequence immediately flanking this element on the side opposite to the deletion was identical in both human and chimpanzee. In both these cases, we accepted the computational detection as a valid ARMD locus. At loci that showed slight deviations in the sequence architecture from the unambiguous ARMD structures described above (which raise the possibility that one of the two chimpanzee *Alu* elements might be a chimpanzee-specific *Alu* insertion, as opposed to a human-specific ARMD event), we
designed oligonucleotide primers in the nonrepetitive sequences flanking the *Alu* elements in the chimpanzee genome and we experimentally confirmed by PCR (and, where required, by DNA sequencing) that the deletion did exist and was specific to the human genome.

As an additional step to verify the potential ARMD loci that we accepted/rejected based solely on computational identification, we randomly chose two sets of 25 such insertions and deletions and verified them by PCR. Accuracy rates for putative deletion and insertion loci were 100% and 96%, respectively (4% of putative insertions comprising the error were all deletions), confirming the validity of our approach.

**PCR Amplification and DNA Sequence Analysis of ARMD Loci**

We designed oligonucleotide primers using Primer3 software. Detailed information for each locus including primer sequences, annealing temperature and PCR product sizes is available from the “Publications” section of the Batzer Laboratory Web site.

PCR amplification of each locus was performed in 25 μl reactions with 10-50 ng genomic DNA, 200 nM of each oligonucleotide primer, 200 μM dNTPs in 50 mM KCl, 1.5 mM MgCl2, 10 mM Tris-HCl (pH 8.4), and 2.5 units *Taq* DNA polymerase. The conditions for the PCR were an initial denaturation step of 94°C for 4 min; followed by 32 cycles of 1 min of denaturation at 94°C, 1 min of annealing at optimal annealing temperature, and 1 min of extension at 72°C; followed by a final extension step at 72°C for 10 min. PCR amplicons were separated on 2% agarose gels, were stained with ethidium bromide, and were visualized using UV fluorescence.

Individual PCR products were purified from the gels with Wizard gel purification kits (Promega) and were cloned into vectors by use of TOPO-TA Cloning kits (Invitrogen). For each sample, three colonies were randomly selected and sequenced on an Applied Biosystems
ABI3130XL automated DNA sequencer. Each clone was sequenced in both directions with use of M13 forward and reverse primers. The sequence tracks were analyzed using the Seqman program in the DNASTAR suite and were aligned using BioEdit sequence alignment software. Gorilla and orangutan sequences generated during the course of this study have been submitted to GenBank under accession numbers DQ363502-363524.

Loci verified by PCR were screened on a panel of five primate species, including *Homo sapiens* (HeLa; cell line ATCC CCL-2), *P. troglodytes* (common chimpanzee; cell line AG06939B), *Pan paniscus* (bonobo or pygmy chimpanzee; cell line AG05253B), *Gorilla gorilla* (Western lowland gorilla; cell line AG05251) and *Pongo pygmaeus* (orangutan; cell line ATCC CR6301). To evaluate polymorphism rates, we amplified 50 randomly picked ARMD loci on a panel of genomic DNA, from 80 human individuals (20 from each of four populations: African American, South American, European, and Asian) that was available from previous studies in our lab.

**Monte Carlo Simulations of GC and Alu Content**

To test whether the GC and *Alu* contents of the sequences deleted through ARMD differed statistically from the rest of the genome, we performed Monte Carlo simulations comparing the observed deletions to two other sets of sequences. Both these sets comprised randomly extracted sequences equal in number to the observed deletions (492) and mimicked the observed size distribution of ARMD events. The first set was extracted from the regions immediately adjacent to randomly picked *Alu* elements annotated in the reference human genome sequence (called “RSNA”). The second set comprised sequences randomly extracted from the entire genome sequence, with no additional parameters incorporated (called “RSG”). We used 5000 randomized replicates of both sets. For both observed and simulated sets of
sequences, we calculated GC content using in-house Perl scripts, whereas the Alu content was analyzed using a locally installed copy of the RepeatMasker Web server. Additionally, to make our estimate of observed percentage Alu content conservative, we trimmed the deleted sequence at each locus to remove remaining fragments of the two Alu elements that caused the ARMD event.

Statistical significances of the differences in GC and Alu content were based on Z scores obtained by comparing observed values (from the actual set of deleted sequences) with the mean value obtained from the 5000 randomly extracted sequence sets (Hamaker 1978). All computer programs used are available from the authors on request.

References


sequences, one novel base substitution and two tentative hotspot mutations in the hypoxanthine phosphoribosyltransferase (HPRT) gene in five patients with Lesch-Nyhan syndrome." Hum Genet 103(3): 311-8.


CHAPTER SIX:

SUMMARY
Recent human and chimpanzee sequencing projects have revealed that most human and chimpanzee repetitive sequences are derived from transposable elements. These elements have extensive impact on the evolutionary history of both lineages by shaping the genomic landscape. Retrotransposons comprise ~42% of human genome and have mediated genomic fluidity in the host genome during their de novo insertion and during post-retrotransposition processes. In this study, we examined different evolutionary fates of recently integrated retrotransposons in the human and chimpanzee lineages; and analyzed genomic instability that is generated by retrotransposon integration into the genome as well as post-insertion recombination between elements.

In chapter two, we analyzed the evolutionary history of AluYb lineage and proposed a revised model for successful propagation of Alu elements. The amplification of most Alu elements is thought to occur through a limited number of hyperactive “master” genes that produce a high number of copies during long evolutionary periods of time. However, the existence of long-lived, low activity Alu lineages in the human genome suggests a more complex propagation mechanism. Using both computational and wet bench approaches, we reconstructed the evolutionary history of the AluYb lineage, one of the most active Alu lineages in the human genome. We show that the major AluYb lineage expansion in humans is a species-specific event, as non-human primates possess only a handful of AluYb elements. However, the oldest existing AluYb element resided in an orthologous position in all hominoid primate genomes examined, demonstrating that the AluYb lineage originated 18 to 25 million years ago. Thus, the history of the AluYb lineage is characterized by approximately 20 million years of retrotransposition quiescence preceding a major expansion in the human genome within the past a few million years. We suggest that the evolutionary success of the Alu family may be driven at least in part
by “stealth driver” elements that maintain low retrotransposition activity over extended periods of time and occasionally produce short-lived hyperactive copies responsible for the formation and remarkable expansion of Alu elements within the genome.

In chapter three, we reported a detailed characterization of chimpanzee-specific L1 subfamily diversity and a comparison with their human-specific counterparts. Our results indicate that L1 elements have experienced different evolutionary fates in humans and chimpanzees within the past ~6 million years. Although the species-specific L1 copy numbers are on the same order in both species (1,200-2,000 copies), the number of retrotransposition-competent elements appears to be much higher in the human genome than in the chimpanzee genome. Also, while human L1 subfamilies belong to the same lineage, we identified two lineages of recently integrated L1 subfamilies in the chimpanzee genome. The two lineages seem to have coexisted for several million years, but only one shows evidence of expansion within the past three million years. These differential evolutionary paths may be the result of random variation, or the product of competition between L1 subfamily lineages. Our results suggest that the coexistence of several L1 subfamily lineages within a species may be resolved in a very short evolutionary period of time, perhaps in just a few million years. Therefore, the chimpanzee genome constitutes an excellent model in which to analyze the evolutionary dynamics of L1 retrotransposons.

In chapter four, we first identified species-specific L1 insertion-mediated deletion, in vivo and proposed different mechanisms for the deletion. Fifty deletion events in the human and chimpanzee genomes are directly associated with the insertion of L1 elements. Consequently, ~18 kb and ~15 kb of sequence are removed from the human and chimpanzee genome, respectively. Our data suggest that during the primate radiation, L1 insertions may have deleted
up to 7.5 Mb of target genomic sequences. While the results of our *in vivo* analysis differ from those of previous cell culture assays of L1 insertion-mediated deletions in terms of the size and the rate of sequence deletion, evolutionary factors can reconcile the differences. We report a pattern of genomic deletion sizes similar to those created during the genomic retrotransposition of *Alu* elements. Our study provides support for the existence of different mechanisms for small and large L1-mediated deletions, and we present a model for the correlation of L1 element size and the corresponding deletion size. In addition, we show that internal rearrangements can modify L1 structure during retrotransposition events associated with large deletions.

In chapter five, we compare the reference human and chimpanzee genomes to determine the magnitude of *Alu* recombination-mediated deletion process in the human lineage since the human-chimpanzee divergence ~6 million years ago. Combining computational data mining and wet-bench experimental verification, we identified 492 human-specific deletions (totaling ~400 Kb) attributable to this process, making it a significant component of the insertion/deletion spectrum of the human genome. The majority of the deletions (295/492) coincide with known or predicted genes (including three that deleted functional exons as compared to orthologous chimpanzee genes), implicating this process in creating a substantial portion of the genomic differences between humans and chimpanzees. Overall, we find that *Alu* recombination-mediated genomic deletion has had a much higher impact than that reflected by previously identified isolated events, and that it continues to contribute to the dynamic nature of the human genome.

Mobile elements have had a variety of impacts on the genomes that they occupy. Here the underlying expansion of *Alu* and L1 retrotransposons in the human and chimpanzee lineages has been examined. These studies have shed new insight into the amplification mechanism of *Alu* elements and the underlying sequence diversity of L1 elements. The genetic instability mediated
by post-insertion *Alu* element mediated recombination and L1 insertion mediated deletion within the human and chimpanzee lineages has been determined. As a whole these data call into question whether or not retrotransposons should be considered as “junk” DNA in the genomes in which they reside. Rather, mobile elements represent a potent evolutionary force associated with genomic fluidity in the host genome.
APPENDIX A:

LETTERS OF PERMISSION

27 June 2006

Kyudong Han
Louisiana State University, USA

khan3@lsu.edu

Dear Kyudong Han,

Thank you for your request of June 19, 2006. We are pleased to grant you the permission requested. We hereby grant you permission to reprint the article entitled: “Under the genomic radar: the Stealth model of Alu amplification”, which is accepted for publication on Genome Research at January 28, 2005, at no additional charge in your thesis, provided that the source article is suitably acknowledged as a footnote or reference list as follows: "Reprinted from Publication Title, Vol number, Author(s), Title of Article, Pages No., Copyright(Year), with permission from Cold Spring Harbor Laboratory Press.

With kind regards,

Hillary Sussman, Ph.D.
Executive Editor, Genome Research
500 Sunnyside Blvd.
Woodbury, NY 11797
Tel: (516)422-4014
Fax:(516)422-4092
hsussman@cshl.edu
Kyudong Han
Dept of Biological Sciences
Louisiana State University
202 Life Sciences Building
Baton Rouge 70803
USA

Dear Kyudong Han,

Han, ‘Genomic rearrangements by LINE-1 insertion-mediated deletion…’

Thank you for your email dated 20 June, requesting permission to reprint the above material. Our permission is granted without fee to reproduce the material, as you are the original author.

Use of the article is restricted to reproduction in your dissertation, available in print format only, to be used only in the English language. This permission is limited to this particular use and does not allow you to use it elsewhere or in any other format other than specified above.

Please include a credit line in your publication citing full details of the Oxford University Press publication which is the source of the material and by permission of Oxford University Press or the sponsoring society if this is a society journal.

If the credit line or acknowledgement in our publication indicates that material including any illustrations/figures etc was drawn or modified from an earlier source it will be necessary for you to also clear permission with the original publisher. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

Please do not hesitate to contact me if I can be of any further assistance.

Yours sincerely,

Gemma Punits
Journals Rights Assistant
Permission Grant

Kyudong Han
Louisiana State University
202 Life Science Bldg
Baton Rouge, LA 70803

Re: your request dated 6/19/2006

Dear Kyudong Han,

Thank you for your request for permission to use material from the publication(s) of the University of Chicago Press. Permission is granted for use as stated below. Unless specifically granted below, this permission does not allow the use of our material in any other edition or by any additional means of reproduction including (by way of example) motion pictures, sound tapes, electronic formats, and phonograph records; nor does this permission cover book clubs, translations, abridgment, or selections which may be made of the publication. No subsequent use may be made without additional approval. This permission is subject to the following terms:

1. On each copy of the selection full credit must be given to the book or journal, to the author (as well as to the series, editor, or translator, if any), and to the University of Chicago as publisher. In addition, the acknowledgment must include the identical copyright notice as it appears in our publication.
2. This permission does not apply to any part of the selection which is independently copyrighted or which bears a separate source notation. The responsibility for determining the source of the material rests with the prospective publisher of the quoted material.
3. This permission grant for the material listed on the invoice below is provided GRATIS.
4. This permission covers publication of one edition of the work up to 250 copies.
5. Permission granted is non-exclusive and, unless otherwise stated, is valid throughout the world in the English language only.
6. This permission is void if:
   a. More than 10% of your forthcoming work, exclusive of index and bibliography, is composed of University of Chicago Press material.
   b. Publication does not take place within 24 months of the date on this form.
7. This permission includes the use of Braille, large type, or other editions of your work by nonprofit organizations solely for the use by the visually handicapped, provided no fees are charged. The limitations listed in Clause 4, above, do not apply to such use.

Reference

Material Requested


To appear in: a dissertation by Kyudong Han for a degree at Louisiana State University.

Approved by: [Signature] Perry Cartwright, Rights and Permissions

All correspondence on this request must include our reference number (62396).
VITA

Kyudong Han is the son of Junhee Han and Jumsoon Song and is married to Jungnam Lee. He graduated with Bachelor of Science and Master of Science degrees in microbiology from Dankook University in Cheonan, Korea, in August of 2000 and in August of 2002, respectively. He began his doctoral research in the fall of 2003 in the Department of Biological Sciences at Louisiana State University in Baton Rouge, Louisiana, under the direction of Professor Mark A. Batzer. Mr. Han will graduate with the degree of Doctor of Philosophy in December, 2006.