Retrotransposon mediated genomic variation in the human and chimpanzee lineages

Jungnam Lee
Louisiana State University and Agricultural and Mechanical College, jlee59@lsu.edu

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

Recommended Citation
https://digitalcommons.lsu.edu/gradschool_dissertations/1572
RETROTRANSPOSON MEDIATED GENOMIC VARIATION
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
Jungnam Lee
B.S., Dankook University, 1999
M.S., Dankook University, 2003
December 2008
ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my major professor, Dr. Mark A. Batzer, for his kind advice, intellectual mentorship, and unending support over the last four years. I thank the members of my dissertation committee, Drs. Joomyeong Kim, Andrew Whitehead, and Fernando Galvez for their invaluable comments and suggestions that helped me during my graduate studies. I also thank the following collaborators for their contributions to this research.

In chapter one, I identified and characterized human- and chimpanzee-specific L1s. In chapter two, I studied Alu recombination-mediated deletion in the chimpanzee genome. In chapter three, I investigated retrotransposon recombination-mediated inversion between the human and chimpanzee genomes. For his contribution to chapters two, three, and four: Dr. Kyudong Han.

For their contribution to chapters two and three: Drs. Richard Cordaux and Ping Liang. For his contribution to chapter four: Dr. Heui-Soo Kim. In addition, I thank all other members of the Batzer laboratory including Jerilyn Walker and T. Joshua Meyer for their critical reading of this manuscript and great comments.

I am grateful to my father, mother, sister, and brothers for their constant love, encouragement, and support. I also thank my husband, Kyudong, for his love and fruitful discussion on my research. Lastly, I thank my lovely son, Andrew, for coming to school with me since he was two-months old and his patience during my graduate studies.
# TABLE OF CONTENTS

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

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

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

ABSTRACT ..................................................................................................................................... vi

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

CHAPTER TWO: DIFFERENT EVOLUTIONARY FATES OF RECENTLY INTEGRATED HUMAN AND CHIMPANZEE LINE-1 RETROTRANSPOSONS ............................................................. 9

CHAPTER THREE: ALU RECOMBINATION-MEDIATED STRUCTURAL DELETIONS IN THE CHIMPANZEE GENOME ................................................................. 36

CHAPTER FOUR: CHROMOSOMAL INVERSIONS BETWEEN THE HUMAN AND CHIMPANZEE LINEAGES CAUSED BY RETROTRANSPOSONS ..................................................... 68

CHAPTER FIVE: SUMMARY ......................................................................................................... 94

APPENDIX: LETTERS OF PERMISSION ................................................................................... 99

VITA ............................................................................................................................................. 101
### LIST OF TABLES

2.1 Subfamily classification and age estimates for species-specific L1 elements ............ 13

2.2 Insertions and deletions in the coding region of 5 full-length chimpanzee-specific L1 elements .......................................................... 24

3.1 Summary of chimpanzee-specific ARMD events .................................................. 39

3.2 Classification of genomic DNA deleted by ARMDs in chimpanzee lineage ............... 42

4.1 Summary of retrotransposon recombination-mediated inversion .............................. 73
LIST OF FIGURES

1.1 Schematic representations of the retrotransposons ........................................................ 3
2.1 Median-joining network of L1 subfamilies ................................................................. 12
2.2 Phylogenetic tree of 5’ UTR consensus sequence of L1Pt, L1Hs and L1PA subfamilies ................................................................. 17
2.3 Species-specific L1 insertions ...................................................................................... 20
2.4 Size distribution of species-specific L1 elements ......................................................... 22
2.5 Analysis of genomic environment adjacent to species-specific L1 elements ............ 26
3.1 Size distribution of chimpanzee-specific ARMD events ............................................. 41
3.2 Alu subfamily composition in ARMD events ............................................................ 44
3.3 Comparison of Alu subfamilies involved in ARMD events ......................................... 44
3.4 Recombination breakpoints during chimpanzee-specific ARMD events ................. 47
3.5 Incomplete lineage sorting and parallel independent ARMD events ....................... 51
3.6 Exonic deletions caused by two ARMD events ......................................................... 58
4.1 The 253 inversion loci between the human and chimpanzee lineages ...................... 72
4.2 Sequence alignment of one recombined and two prerecombined Alu elements involved in an Alu-RMI event ................................................. 74
4.3 Alu and L1 subfamilies involved in RRMI events ...................................................... 75
4.4 Analysis of GC content in flanking regions of RRMI loci .......................................... 79
4.5 RRMI between human and chimpanzee lineages ..................................................... 88
ABSTRACT

LINE-1 (Long INterspersed Element-1 or L1) and Alu elements are important sources of structural variation in primate genomes because they are highly active retrotransposons with copy numbers of ~520,000 and >1.2 million within the human genome, respectively. Although the bulk of these elements have resided in their respective host genomes for a long time, and have thus accumulated random mutations, overall these elements retain high levels of sequence identity among themselves. The presence of many nearly-identical retrotransposons located close to each other (e.g., Alu-Alu or L1-L1 pairs) disposes their host genomes to unequal homologous DNA recombination events that generate genomic deletions and inversions of varying sizes.

Through computational comparisons of the human and chimpanzee genome sequences, and using rhesus macaque and orangutan genome sequences as outgroups, we have identified species-specific genomic variation. In the first analysis, we identified human and chimpanzee-specific L1s and examined their sequence evolution. We show that L1 retrotransposition activity is slightly higher in the human lineage, relative to the chimpanzee lineage, and that L1s have experienced different evolutionary fates in these two lineages, resulting from random variation or competition between L1 subfamily lineages. Next, we analyzed the magnitude of Alu recombination-mediated deletions (ARMDs) in the chimpanzee lineage subsequent to the human-chimpanzee divergence (~6 million years ago). We have identified 663 chimpanzee lineage-specific deletions (involving a total of ~771 kb of genomic sequence) attributable to this process. The RefSeq databases indicate that 13 exons in six genes are annotated as either demonstrably or putatively functional in the human genome, and 299 intronic regions have been deleted through ARMDs in the chimpanzee lineage. In the third analysis, we characterize chromosomal inversion events between the human and chimpanzee genomes caused by inverted
L1-L1 or Alu-Alu pairs. We have identified 49 retrotransposon recombination-mediated inversion (RRMI) loci and, among them, three RRMI loci contain inverted exonic regions in known genes. Therefore, we suggest that L1 and Alu elements have contributed to the genomic and phenotypic diversity between humans and chimpanzees since the divergence of the two species.
CHAPTER ONE:

BACKGROUND
The human lineage diverged from its closest extant lineage, the chimpanzee, approximately six million years ago (Mya) (Goodman et al., 1998). Since then, their genomes have evolved independently from one another, accumulating the genomic differences seen between the two species. The proportion of genomic differences was estimated to be 6.59%, which consists of a 1.52% difference due to base substitution and a 5.07% difference due to the presence of indels (Wetterbom et al., 2006). There are many mechanisms responsible for creating these base substitutions and indels. Among them, mobile elements are considered a major factor accelerating the genomic divergence between the two species. Although mobile elements account for ~45% of the human genome (Deininger et al., 2003; Lander et al., 2001a), most of these elements are inactive and incapable of further retrotransposition. The mobile elements that are older than six million years were inherited from the common ancestor of humans and chimpanzees, and thus they rarely contribute significantly to the genomic variation found between these lineages. However, these elements can still contribute to genomic variation between the two lineages through action of recombination. In contrast, a small fraction of mobile elements retain the ability to retrotranspose, and these elements contribute to the genomic divergence between humans and chimpanzees through \textit{de novo} insertions and insertion-mediated deletions. Through these three general mechanisms, \textit{de novo} insertion, insertion-mediated deletion, and the post-insertion recombination, mobile elements can alter the expression patterns of genes or cause genomic rearrangements in their host genomes (Callinan et al., 2005; Han et al., 2005; Sen et al., 2006).

Mobile elements were first discovered in the maize genome by Barbara McClintock (McClintock, 1956). They are largely divided into two groups, DNA transposons and retrotransposons, based on their propagation method. DNA transposons propagate using a “cut
and paste” method (Mizuuchi, 1992), whereas retrotransposons amplify in a “copy and paste”
mechanism and integrate into a new genomic region via an RNA intermediate (Luan et al.,
1993). Therefore, retrotransposons typically accumulate much faster in their host genomes than
do DNA transposons. Retrotransposons are categorized into two groups: LTR (Long Terminal
Repeat) and non-LTR retrotransposons. LTR retrotransposons carry long terminal repeats on
both ends while non-LTR retrotransposons do not. Among the non-LTR retrotransposons, L1
and Alu elements are the most ubiquitous in primate genomes. They have independently
proliferated in their respective hosts since the divergence of the human and chimpanzee lineages,
and account for ~17% and ~11% of the human genome, respectively (Smit et al., 1995).

Figure 1.1. Schematic representations of the retrotransposons. Retrotransposons are divided
into two groups based on the capability to encode enzymes (e.g., endonuclease and reverse
transcriptase). Autonomous elements contain ORFs which encode the enzymes needed for their
mobilization while non-autonomous elements do not contain the ORFs.
L1s 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) separated by a ~60 bp intergenic spacer, and a 3’ UTR ending in a poly(A) tail (Kazazian and Moran, 1998). ORF1 encodes an RNA-binding protein that has shown nucleic acid chaperone activity in vitro, and ORF2 encodes both reverse transcriptase and endonuclease activities (Feng et al., 1996; Kolosha and Martin, 1997; Mathias et al., 1991). 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). The L1 family emerged around 120 Mya (Khan et al., 2006; Smit et al., 1995) and is still actively expanding in humans, as demonstrated by the existence of highly polymorphic L1 elements in human populations (Badge et al., 2003; Boissinot et al., 2004; Myers et al., 2002; Seleme et al., 2006; Sheen et al., 2000; Wang et al., 2006) as well as the de novo L1 insertions found to be responsible for some genetic disorders (Chen et al., 2005). L1 subfamilies can be 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 (e.g., humans and apes) within the past 25 million years, named L1PA1 to L1PA5 (Boissinot et al., 2000; Khan et al., 2006; Lander et al., 2001a; Smit et al., 1995). In chapter two, we identified human- and chimpanzee-specific L1 elements that inserted into their host genomes after the divergence of human and chimpanzee lineages. We characterized the L1 elements based on their subfamily, size, insertion site preferences, and ability to retrotranspose. In addition, we compared sequence evolution patterns of human-
specific L1s with those of chimpanzee-specific L1s and described the factors that led to these differences.

*Alu* elements, a family of short interspersed elements (SINEs), emerged ~65 Mya and have successfully proliferated in primate genomes with > 1.2 million copies. This element originated from the 7SL RNA gene during an early stage of the primate radiation. *Alu* elements are ~300bp long and have a dimeric structure, consisting of left and right monomers. The left monomer contains internal promoters (i.e. A and B boxes) for polymerase III to initiate *Alu* transcription. Unlike L1s, *Alu* elements are non-autonomous, and thus must borrow the enzymatic machinery necessary for retrotransposition from L1s. Thanks to this shared machinery, *Alu* elements retrotranspose via target primed reverse transcription and each element is usually accompanied by TSD, just as with L1s. The *Alu* family consists of a number of subfamilies, that maintain a high level of sequence identity among themselves (70%-99.7%) and which are capable of concomitant expansions (Batzer and Deininger, 2002; Hedges et al., 2005; Xing et al., 2004). Due to the high sequence identity between *Alu* elements combined with their high copy number in the genome, they are often involved in non-allelic homologous recombination. Such recombination in host genome can cause species-specific local genomic instability and has been reported as the major source of genomic disorders (Deininger and Batzer, 1999; Shaw and Lupski, 2004). In chapter three, we analyze genomic deletions caused by post-insertion recombination between *Alu* elements in the chimpanzee lineage. We computationally identified *Alu* recombination mediated deletion (ARMD) events in the chimpanzee genome and experimentally verified them. Then, we estimated the magnitude of chimpanzee genome deleted by ARMD events and compared chimpanzee-specific ARMD with human-specific ARMD to find out the overall impact of ARMD on the genomic divergence.
between these two species.

As described above, homologous recombination between Alu elements can cause deletions within their host genome. The two Alu elements involved in each ARMD event are oriented in the same direction. However, when two Alu elements are oriented in opposite direction to one another, recombination may still occur, but this result in a chromosomal inversion rather than a deletion. Like Alu elements, L1s retain high level of sequence identity among themselves and exist at high copy numbers in the human and chimpanzee genomes. This suggests that L1s also have the potential to cause chromosomal inversions through recombination between oppositely oriented L1 pairs. In chapter four, we characterized chromosomal inversions between the human and chimpanzee genomes caused by both L1 and Alu elements. Chromosomal inversions have been implicated as potential drivers that led to the speciation of humans and chimpanzees from their common ancestor. These events also cause genomic variation within each species. This study delineates impact of L1 and Alu elements in generating this type of genomic variation in the human and chimpanzee lineages.

References


CHAPTER TWO:

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

*Reprinted by permission of *Gene
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 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 (Feng et al., 1996; Kolosha and Martin, 1997; Mathias et al., 1991). 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., 2001a). The L1 family emerged around 120 million years (myrs) ago (Khan et al., 2006; Smit et al., 1995) and is still actively expanding in humans, as demonstrated by the existence of highly polymorphic L1 elements in human populations (Badge et al., 2003; Boissinot et al., 2004; Myers et al., 2002; Seleme et al., 2006; Sheen et al., 2000; 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; Mathews et al., 2003; Mills et al., 2006a). Contrary to the non-autonomous Alu retrotransposons in which different subfamilies are capable of concomitant expansions (Batzer and Deininger, 2002; Hedges et al., 2005; Xing et al., 2004), 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 can be 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 (Boissinot et al., 2000; Khan et al., 2006; Lander et al., 2001a; Smit et al., 1995). The most recently evolved, Homo sapiens-specific (Hs) L1 subfamilies have been well characterized (Boissinot et al., 2000; Boissinot et al., 2004; Myers et al., 2002; Ovchinnikov et al., 2002; Salem et al., 2003a) 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 Pan troglodytes-specific (Pt) L1 elements have previously been published (CSAC, 2005; Mills et al., 2006a). Here, we report a detailed characterization of Pt 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 Pt 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; Mathews et al., 2003; Mills et al., 2006a). 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 Pt L1 elements. While this represents more than half of all Hs L1 elements identified, it barely accounts for one fifth of all Pt elements, suggesting that Pt 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 Pt for human and chimpanzee, respectively, and we use the RepeatMasker subfamily assignment for shared L1 subfamilies (Table 2.1 and Figure 2.1). 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 L1Pt-2A is the oldest (A) L1 subfamily belonging to P. troglodytes-specific (Pt) subfamily.

**Figure 2.1.** Median-joining network of L1 subfamilies. The network was reconstructed using Hs and Pt L1 elements. Empty circles represent Hs L1 subfamilies. Filled circles represent Pt 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 (L1Pt) shadow boxes.
Table 2.1. 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;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Chimp (%)</td>
</tr>
<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;b&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;b&lt;/sup&gt;</td>
<td>0% (0/49)</td>
<td>26.6</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;b&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;b&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;b&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;b&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;b&lt;/sup&gt;</td>
<td>7% (5/67)</td>
<td>27.1</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;b&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></td>
<td>9% (1/11)</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;c&lt;/sup&gt;</td>
<td>0</td>
</tr>
<tr>
<td>L1Hs-Ta0</td>
<td>L1Hs-Ta0</td>
<td>2.7 ± 0.2</td>
<td></td>
<td>38.0</td>
</tr>
<tr>
<td>L1Hs-Ta1</td>
<td>L1Hs-Ta1</td>
<td>1.9 ± 0.2</td>
<td>45%&lt;sup&gt;c&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>L1Pt-1A</td>
<td>L1PA2</td>
<td>6.2 ± 0.8</td>
<td></td>
<td>30% (3/10)</td>
</tr>
<tr>
<td>L1 Pt -1B</td>
<td>L1PA2</td>
<td>3.9 ± 0.5</td>
<td></td>
<td>15.0</td>
</tr>
<tr>
<td>L1 Pt -2A</td>
<td>L1PA2</td>
<td>4.7 ± 0.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 Pt -2B</td>
<td>L1PA2</td>
<td>2.9 ± 0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 Pt -2C</td>
<td>L1PA2</td>
<td>2.9 ± 0.4</td>
<td>80% (8/10)</td>
<td>27.5</td>
</tr>
<tr>
<td>L1 Pt -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></td>
<td>3.8</td>
</tr>
</tbody>
</table>

<sup>a</sup> Based on 1,000 Hs and 207 Pt L1 elements.

<sup>b</sup> Estimated from both Hs and Pt L1 elements.

<sup>c</sup> Data from Salem et al. (2003) and Myers et al. (2002).
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 Pt 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 2.1). 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 2.1), 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 2.1). However, since only about one fifth of all Pt 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 Pt elements are not shared with human (Table 2.1). Given that our human sample includes 1,000 L1 copies, it is very unlikely that these subfamilies would appear to be Pt 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 Pt 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 2.1, and Supplemental Figure 2.1 in Batzer Laboratory Web site; http://batzerlab.lsu.edu). This network, rooted with the older L1PA3 consensus sequence, shows the global sequential order in which the successive L1 subfamilies arose (Figure 2.1). Moreover, the ages estimated independently for individual subfamilies based on within-subfamily sequence diversity are in complete agreement with this phylogenetic structure (Figure 2.1 and Table 2.1). 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 Pt subfamilies belong to two independent L1 lineages, termed L1Pt-1 and L1Pt-2 (Figure 2.1 and Table 2.1), 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 L1Pt lineages we identified (i.e. L1Pt-1 and L1Pt-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 L1Pt subfamily lineages are highly similar to each other (Figure 2.2) and to the L1Hs and L1PA2 5’ UTRs. More generally, both L1Pt 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 L1Pt-2 lineage may have had an advantage over the L1Pt-1 lineage. Indeed, not only is the L1Pt-2 lineage represented by twice as many copies as the L1Pt-1 lineage, but three of the four L1Pt-2 subfamilies are 2-3 myrs-old, whereas the youngest L1Pt-1 subfamily is ~4 myrs-old (Table 2.1). Interestingly, we identified two full-length L1Pt-2 copies with intact ORF1 and ORF2, while L1Pt-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 (Dewannieux et al., 2003;
Figure 2.2. Phylogenetic tree of 5' UTR consensus sequence of L1Pt, 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 L1Pt lineages in a pink shadow box were generated in this study. Bootstrap values (%) are shown above each branch.

Wei et al., 2001), 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 L1Pt copies have been inactivated so far and they both happen to belong to the L1Pt-2 lineage) or because a selective process is acting to specifically preserve the integrity of the ORFs of these two particular L1Pt-2 copies. It is worthy to note here that although competition is a
plausible explanation for the differential evolutionary successes of the L1Pt-1 and L1Pt-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 Pt elements were genotyped in 12 chimpanzees. As expected (Hedges et al., 2005), polymorphism levels decreased with subfamily ages (Table 2.1). 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 (Salem et al., 2005; Xing et al., 2003).

The comparison between Pt and Hs L1 subfamilies of similar ages indicates that the polymorphism levels of Pt 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 2.1). These results are consistent with those observed for Hs and Pt *Alu* elements, that also showed that the polymorphism levels of Pt *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 2.1, 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 2.1). 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 2.1) 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 2.1) were present in gorilla but absent from either humans or chimpanzees in our panel (Figure 2.3). 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 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
Figure 2.3. 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.

literature (Hedges et al., 2004; Ray et al., 2006b; Salem et al., 2003b). 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 2.1). 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 Pt L1 elements on these chromosomes. On average, Hs L1
elements were about fourfold longer than Pt L1 elements (i.e. 2,533 vs. 641 bp; Figure 2.4). 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 Pt L1 elements (Boissinot et al., 2000; Boissinot et al., 2004; Mills et al.,
2006a; Myers et al., 2002) (Figure 2.4). By contrast, ~86% (89/103) of Pt L1 elements are
shorter than 1 kb vs. only ~48% (66/138) of Hs L1 elements (Figure 2.4). Therefore, Pt L1
elements appear to be more severely truncated than their Hs counterparts. The reason for such
structural differences between Hs and Pt 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 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
Figure 2.4. 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 Pt L1 elements on chromosomes 1 and 21 are grouped in 500 bp bins.

of selection in chimpanzees than in humans, given that the chimpanzee effective population size is higher than that of humans (Fischer et al., 2004; Graur and Li, 2000) 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 Pt 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 (Martin et al., 2005; Szak et al., 2002). 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 Pt 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 Pt L1 elements located on chromosome 1 and 21 possessed intact ORFs. Given this result, we extended our investigation of full-length Pt L1 elements to the whole chimpanzee genome. We identified a total of 19 full-length Pt 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 2.2). In most cases, those insertions were located in homopolymeric tracts (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
Table 2.2. 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></th>
<th>Deletions</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 bp</td>
<td>2 bp</td>
<td>3 bp</td>
<td>1 bp</td>
<td>2 bp</td>
<td>3 bp</td>
<td>4 bp</td>
<td>6 bp</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number in chimpanzee genome sequence (Nov. 2003 freeze)</td>
<td>56</td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number confirmed by DNA sequencing in this study</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

al., 2006a). To test this hypothesis, we selected 5 full-length Pt 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 2.2).

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 Pt 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., 2003).

Two of the intact chimpanzee L1 elements belong to the subfamily lineage L1Pt-2B and three are L1PA2 members. As discussed above (see section “Comparison of 5’ UTR sequences”), this may contribute to explain why the L1Pt-2 subfamily lineage seems to have been more successful than the L1Pt-1 lineage in recent chimpanzee evolution.
Genomic Distribution of Human and Chimpanzee L1 Insertions

To test whether Hs and Pt 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 Pt L1 elements had very similar GC content distributions, both being skewed towards AT-rich regions of the genome (Figure 2.5A). Indeed, 74% (102/138) and 83% (86/103) of Hs and Pt 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., 2001a). We also compared the gene density of 1 Mb flanking genomic sequence each side of L1 elements. Again, we found that Hs and Pt L1 elements had similar gene density distributions, skewed towards gene-poor regions of the genomes (Figure 2.5B). These results are not unexpected, however, since there is a positive correlation between GC content and gene density (Lander et al., 2001a; Versteeg et al., 2003).

To investigate global polymorphism levels of Hs and Pt L1 elements regardless of subfamily affiliation, we randomly selected 31 Hs and 31 Pt 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 Pt 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.

Conclusions

Our analyses indicate that L1 elements have had very different evolutionary dynamics in
Figure 2.5. 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.
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 (Cordaux and Batzer, 2006; Hedges et al., 2004) 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.
Within the chimpanzee genome, two Pt 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 Pt 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 Pt 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 (*P. 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 (*P. 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 2.1 from the Batzer Laboratory Web site (http://batzerlab.lsu.edu).
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 2.2 in Batzer Laboratory Web site; http://batzerlab.lsu.edu, (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 Pt 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 (Pascale et al., 1993; Voliva et al., 1984). The software MEGA 3.1 (Kumar et al., 2004) was used to build neighbor-joining trees of the 5'UTR consensus sequences of two Pt 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/geece.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 THREE:

ALU RECOMBINATION-MEDIATED STRUCTURAL DELETIONS IN THE CHIMPANZEE GENOME*

*Reprinted by permission of *PLoS Genetics
Introduction

Mobile elements are a major source of genetic diversity in mammals (Batzer and Deininger, 2002; Deininger and Batzer, 2002). Alu elements, a family of short interspersed elements (SINEs), emerged ~65 million y ago (Mya) and have successfully proliferated in primate genomes with >1.2 million copies (Batzer and Deininger, 2002; CSAC, 2005; Lander et al., 2001a; RMGSAC, 2007). Alu elements consist of a left monomer and a right monomer (Batzer and Deininger, 2002; Quentin, 1992). Each of these monomers independently evolved from 7SL-RNA (Kriegs et al., 2007) and subsequently fused into the dimeric Alu element in the primate lineage (Quentin, 1992). Alu elements are known to be associated with primate-specific genomic alterations by several mechanisms, including de novo insertion, insertion-mediated deletion, and unequal recombination between Alu elements (Callinan et al., 2005; Deininger and Batzer, 1999; Han et al., 2005; Sen et al., 2006). The Alu family consists of a number of subfamilies, which maintain high sequence identity among themselves (70%–99.7%) (Britten et al., 1988; Jurka and Smith, 1988; Schmid and Maraia, 1992; Slagel et al., 1987).

Mispairing between two Alu elements has been shown to be a frequent cause of deletion or duplication in the host genome (Deininger and Batzer, 1999; Hackenberg et al., 2005; Sen et al., 2006). A recent study of human-specific Alu recombination-mediated deletion (ARMD) reported a significant number of events associated with Alu elements (Sen et al., 2006). An ARMD may arise through either interchromosomal recombination by mismatch of sister or nonsister chromatids during meiosis (Chance et al., 1994) or by intrachromosomal recombination between two Alu elements on the same chromosome. Previously, Sen et al. (2006) found 492 human-specific ARMD events responsible for ~400 kb of deleted genomic sequence in the human lineage (Sen et al., 2006). Here, we report 663 chimpanzee-specific ARMD events.
identified from comparative analysis of the chimpanzee and human genomes. The chimpanzee-specific ARMD events deleted a total of ~771 kb of genomic sequence in chimpanzees, including exonic deletions in six genes, sometime after the divergence of the human and chimpanzee lineages (~6 Mya). ARMD events in the chimpanzee genome have generated large deletions (up to ~32 kb) relative to human-specific ARMD events. Taking deletions in both the human and chimpanzee lineages into account, we suggest that ARMD events may have contributed to genomic and phenotypic diversity between humans and chimpanzees.

**Results**

**A Genome-Wide Analysis of Chimpanzee-Specific ARMD Events**

To investigate chimpanzee-specific ARMD loci, we first computationally compared the chimpanzee (panTro1) and human (hg17) genome reference sequences. A total of 1,538 ARMD candidates were initially retrieved using panTro1. These loci were converted to panTro2 (March 2006), which, due to the better quality of the sequence assembly, allowed us to eliminate a number of loci that mimicked authentic ARMD loci. Through a comparison of panTro1 and panTro2, we discarded 258 of the 1,538 loci (Table 3.1). The remaining 1,280 loci were manually inspected using the repetitive DNA annotation utility RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker). In terms of local sequence architecture, human-specific mobile element insertions between two preexisting adjacent *Alu* elements could be computationally confused with a chimpanzee-specific deletion. Because the consensus sequences of the human-specific mobile elements (e.g., *AluYb8*, *AluYa5*, SVA, and L1Hs) have been well established in RepeatMasker, we were able to identify and eliminate from our analysis 189 human-specific insertion loci, including processed pseudogenes. The remaining 1,091 candidate ARMD loci were inspected using triple alignments of human (hg18),
Table 3.1. Summary of chimpanzee-specific ARMD events

<table>
<thead>
<tr>
<th>Classification</th>
<th>Number of loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computationally predicted deletion loci</td>
<td>1538</td>
</tr>
<tr>
<td>Discarded:</td>
<td></td>
</tr>
<tr>
<td>Discarded after manual inspection</td>
<td>513</td>
</tr>
<tr>
<td>Failed PCR verification†</td>
<td>240</td>
</tr>
<tr>
<td>Wrong assembly in panTro1</td>
<td>258</td>
</tr>
<tr>
<td>Wrong assembly in panTro2</td>
<td>5</td>
</tr>
<tr>
<td>Candidate ARMD events:</td>
<td>1025</td>
</tr>
<tr>
<td>Computational error (Alu insertion in human):</td>
<td>362</td>
</tr>
<tr>
<td>Confirmed by PCR analysis</td>
<td>98</td>
</tr>
<tr>
<td>Analysis based on TSD and/or triple alignment</td>
<td>264</td>
</tr>
<tr>
<td>ARMDs:</td>
<td>663</td>
</tr>
<tr>
<td>Confirmed by PCR analysis</td>
<td>233</td>
</tr>
<tr>
<td>Analysis based on TSD and/or triple alignment</td>
<td>430</td>
</tr>
</tbody>
</table>

† The loci could not be amplified due to the presence of other repeat elements in the flanking sequence.

In this study, we combined computational data mining and wet-bench experimental verification, an approach that is optimal for identifying lineage-specific insertions and deletions (Sen et al., 2006). Whereas Sen et al. (2006) computationally compared the human and chimpanzee genomes, in our analysis, the draft version of the rhesus macaque genome sequence chinpanzee (panTro2), and rhesus macaque (rheMac2) sequences at each locus, and also on the basis of their target site duplication (TSD) structures (see Materials and Methods). After manual inspection, 342 of the candidate ARMD loci were examined by PCR to verify their status as authentic ARMD loci. Finally, combining computational and experimental results, 663 loci were confirmed as bona fide chimpanzee-specific ARMD loci (Table 3.1, and Supplemental Dataset 3.1 in Batzer Laboratory Web site; http://batzerlab.lsu.edu).

In this study, we combined computational data mining and wet-bench experimental verification, an approach that is optimal for identifying lineage-specific insertions and deletions (Sen et al., 2006). Whereas Sen et al. (2006) computationally compared the human and chimpanzee genomes, in our analysis, the draft version of the rhesus macaque genome sequence
was used as an outgroup when filtering computational output for false positives (see Materials and Methods). This allowed us to eliminate 215 candidate ARMD loci prior to wet-bench verification, minimizing the cost and time needed to confirm authentic chimpanzee-specific ARMD events, as compared with the previous human-specific ARMD study.

**Genomic Deletion through Chimpanzee-Specific ARMD Events**

Since the human-chimpanzee divergence ~6 Mya, chimpanzee-specific ARMD events have occurred 1.3 times as often as their human-specific counterparts (663 chimpanzee-specific versus 492 human-specific events). The total amount of genomic DNA deleted by ARMD events from the chimpanzee genome is estimated to be 771,497 bp. However, when we consider that the average indel divergence between the human and chimpanzee genomes has been estimated at 5.07% (Wetterbom et al., 2006), the precise amount of DNA deleted through ARMDs in the chimpanzee genome could be anywhere between ~733 and ~811 kb (±5.07% of ~771 kb). The size distribution of DNA sequences deleted through chimpanzee-specific ARMD events ranged from 111 to 31,861 bp, with 1,164 bp average and 615 bp median ARMD sizes. Similar to the pattern observed in human-specific ARMD events (Sen et al., 2006), a histogram of the size distribution of chimpanzee-specific ARMDs is skewed toward deletions of shorter size, with ~68% (449 of 663) of the deletion events shorter than 1 kb (Figure 3.1). As expected, about 70% of the deleted genomic DNA sequences are composed of repetitive elements (Table 3.2), of which *Alu* element sequences account for ~64% (338 kb of 528 kb). Interestingly, the amount of sequence deleted through the ARMD process from the chimpanzee genome is twice as much as that from the human genome during the same period of time. Ten chimpanzee-specific ARMD events were found to have each deleted >7.3 kb of sequence (Figure 3.1); ARMD sizes this large were not observed in the human-specific study. Among these, the largest deleted sequence is
31,861 bp in length, within which only the *SLC9A3P2* pseudogene and two intergenic regions are found in the ancestral sequence (i.e., human ortholog). To examine the possible effects of the removal of ancestral genomic sequences during the 663 chimpanzee lineage-specific ARMD events, we retrieved the pre-recombination sequences (i.e., unaltered orthologs) from the human genome. About 46% (305 of 663) of the ARMD events were located within known or predicted RefSeq genes (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606), and five
Table 3.2. Classification of genomic DNA deleted by ARMDs in chimpanzee lineage

<table>
<thead>
<tr>
<th>Classification</th>
<th>Amount (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alu†</td>
<td>338,489</td>
</tr>
<tr>
<td>MIR</td>
<td>11,527</td>
</tr>
<tr>
<td>L1</td>
<td>82,872</td>
</tr>
<tr>
<td>L2</td>
<td>10,663</td>
</tr>
<tr>
<td>L3</td>
<td>1,135</td>
</tr>
<tr>
<td>LTR</td>
<td>48,650</td>
</tr>
<tr>
<td>MER1</td>
<td>7,638</td>
</tr>
<tr>
<td>MER2</td>
<td>9,336</td>
</tr>
<tr>
<td>Other DNA repeats</td>
<td>5,385</td>
</tr>
<tr>
<td>RNA repeats</td>
<td>229</td>
</tr>
<tr>
<td>Simple repeats</td>
<td>9,174</td>
</tr>
<tr>
<td>Satellite repeats</td>
<td>2,908</td>
</tr>
<tr>
<td>Unique DNA</td>
<td>243,491</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>771,497</strong></td>
</tr>
</tbody>
</table>

† Includes truncated Alu elements.

ARMD events generated 13 exonic deletions in six genes annotated as either demonstrably or putatively functional in the human genome. Among them, two ARMD events deleted exons from demonstrably functional genes in the *NBR2* (neighbor for *BRCA1* [breast cancer 1] gene 2) and *HTR3D* (5-hydroxytryptamine [serotonin] receptor 3 family member D) genes. While no alternative pre-mRNA spliced forms exist for the *NBR2* gene, the *HTR3D* gene shows three alternative pre-mRNA spliced forms in the human according to the ECR Browser (http://ecrbrowser.dcode.org). Among them, one of the *HTR3D* isoforms does not contain exon 3, which was deleted from the chimpanzee genome. Thus, chimpanzees could produce a similar protein to the *HTR3D* isoform mentioned above, because the ARMD event deleted the entire exon 3 and portions of some introns in the chimpanzee genome. However, we cannot rule out that the ARMD event has produced cryptic splicing sites causing either nonfunctionalization or
subfunctionalization of HTR3D. The remaining three chimpanzee ARMD events generated exonic deletions in four putative human genes of unknown function (LOC339766, LOC127295, LOC729351, and LOC645203).

To further analyze the genomic sequences lost due to the ARMD process in the chimpanzee genome, we used the National Center for Biotechnology Information’s (NCBI) UniGene utility (http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene) to look at the orthologous loci in the human genome, which contained sequences that would have been present in the chimpanzee genome if the ARMD events had not occurred. UniGene indicated that 164 ARMD events had caused deletions of coding sequence on the basis of expressed sequence tags (ESTs), although this number decreased to 94 when a high threshold indicating protein similarities (≥98% ProtEST) was selected (Supplemental Table 3.1 from the Batzer Laboratory Web site; http://batzerlab.lsu.edu). This number is much higher than the exonic deletions in six genes generated by ARMD events reported above when RefSeq annotation was used instead.

Structural Features of ARMD Events

Ten different Alu subfamilies are associated with chimpanzee-specific ARMD events: AluJo, AluJb, AluSx, AluSq, AluSp, AluSg, AluSg1, AluSc, AluY, and AluYd8. Their composition and ratio in chimpanzee-specific ARMD events are remarkably similar to those in human-specific ARMD events (Figure 3.2). The Alu subfamily analysis shows that the number of elements from each Alu subfamily involved in the ARMD process is proportional to the genome-wide copy number of each Alu subfamily in the chimpanzee genome. For example, the AluS subfamily has contributed the most to chimpanzee-specific ARMD events because it is the most successful Alu subfamily in the primate genome in terms of copy number. However, we found one exception to this rule; the AluJ subfamily is more ubiquitous than the AluY subfamily in both
Figure 3.2. *Alu* subfamily composition in ARMD events. Proportion of all *Alu* elements involved in chimpanzee- and human-specific ARMD events (red and blue bars, respectively) that belong to each *Alu* subfamily as noted.

Figure 3.3. Comparison of *Alu* subfamilies involved in ARMD events. Proportion of *Alu* elements involved in chimpanzee-specific (red bars) and human-specific (blue bars) ARMD events versus proportion of total *Alu* elements in each subfamily in the chimpanzee genome (gray bars).
the chimpanzee and human genomes (Figure 3.3), but more members of the $AluY$ subfamily were found to be involved in the ARMD process. The major expansion of the $AluJ$ subfamily in primate genomes occurred ~60 Mya, whereas the $AluY$ subfamily expanded only ~24 Mya (Jurka and Smith, 1988; Price et al., 2004; Shen et al., 1991). On the basis of these ages, the individual members of the $AluJ$ subfamily have likely accumulated more point mutations than those of the $AluY$ subfamily. As a result, $AluY$ copies have more sequence identity among them than do the $AluJ$ copies, which results in increased involvement in ARMD events. In addition, we investigated intra-$Alu$ subfamily recombination-mediated deletions for both the $AluJ$ and $AluY$ subfamilies. Of the 103 events involving at least one $AluJ$ element in the ARMD event, only 15 (14.6%) involved recombination between two $AluJ$ elements. The $AluY$ subfamily shows a higher rate of intra-subfamily recombination than the $AluJ$ subfamily, with 219 loci in which at least one $AluY$ element was involved in the recombination event, and 57 (26%) that were between two $AluY$ elements. This suggests that the rate of recombination between $AluY$ elements is 1.8 times higher than that between $AluJ$ elements. Taken together, this suggests that, in addition to the copy number of each $Alu$ subfamily, the level of sequence identity between the individual $Alu$ elements in the genome is also an important variable influencing ARMD events.

From a mechanistic viewpoint, four different types of recombination may occur between two $Alu$ elements. An $Alu$ element consists of left and right monomers. In the first type, comprising about 88% (583 of 663) of the ARMD events in our study, the recombination occurred between the same monomers of the two $Alu$ elements. A second type of recombination occurred between two $Alu$ elements in which one had previously integrated into the middle of the other. Such insertions are commonly found in both the chimpanzee and human genomes because each $Alu$ element bears two endonuclease cleavage sites ($5'$-TTTT/A-3') between its two
monomers. About 8% (51 of 663) of the ARMD events in the chimpanzee genome are products of this second type of recombination. The third type of recombination, seen in 25 of the 663 events (~4%), involved recombination between the left and right monomers on two separate \textit{Alu} elements. The last type occurred between oppositely oriented \textit{Alu} elements. Instances of this type of ARMD are very rare, found only in four of the 663 cases (0.6%). This style of recombination is likely to be uncommon because the stretch of sequence identity between two \textit{Alu} elements oriented in opposite directions to one another is too short to frequently generate unequal homologous recombination. Instead, these two \textit{Alu} elements are more likely to cause \textit{Alu} recombination-mediated inversions or A-to-I RNA editing through the posttranscriptional modification of RNA sequences (Athanasiadis et al., 2004).

**Analysis of the ARMD “Hotspots”**

To analyze the frequency of recombination at different positions along the length of the \textit{Alu} elements (which we refer to as “recombination breakpoints”) at our ARMD loci, we aligned the two intact human \textit{Alu} elements involved in each recombination event with the single chimeric \textit{Alu} element from the chimpanzee genome (Supplemental Figure 3.1 from the Batzer Laboratory Web site; http://batzerlab.lsu.edu). The windows between the two \textit{Alu} elements range in size from 1 to 116 bp, with a mean of 20 bp and a mode of 22 bp. In general, the ARMD loci generated by intra-\textit{Alu} subfamily recombination, as well as the recombination events between relatively young \textit{Alu} elements, show longer stretches of sequence identity than others. Through this analysis, we identified a recombination “hotspot” on the \textit{Alu} consensus sequence (5’-TGTAATCCAGCAGTTGGGAGG-3’), located between positions 24 and 45 (Figure 3.4). This recombination hotspot is congruent with previous studies of gene rearrangements in the human LDL-receptor gene involving \textit{Alu} elements (Rudiger et al., 1995), and with the pattern of
Figure 3.4. Recombination breakpoints during chimpanzee-specific ARMD events. Percentage of ARMD events found to have breakpoints at different positions along an Alu consensus sequence. The “hotspot” region is represented by a conserved 22-bp nucleotide sequence found in 634 ARMD loci (the first and second types of ARMD events) using WebLogo analysis (http://weblogo.berkeley.edu). The dashed line represents the average percentage (0.0035%) of breakpoints across the entire length of the Alu consensus sequence.

recombination found in the 492 human-specific ARMD events (Sen et al., 2006). Of these studies, the former suggested that the hotspot sequence (therein called the “core sequence”) might induce genetic recombination because it subsumes the prokaryotic chi sequence (the pentanucleotide motif CCAGC), which is known to stimulate recBC-dependent recombination (Stahl, 1979). We searched for and found the CCAGC motif at four places (positions 31–35, 85–89, 166–170, and 251–255) along the Alu consensus sequences. The percentages of breakpoints found at these positions are 0.00886%, 0.00336%, 0.00406%, and 0.00372%, respectively. Among these, the percentages of breakpoints found at the latter three positions are similar to the
average percentage of breakpoints across the entire length of the Alu elements (0.0035%) in our ARMD events. The only spot where the motif is found that showed a substantially higher percentage of breakpoints is the one located at positions 31–35, which is within our proposed hotspot. Therefore, this motif may invoke, but does not seem to be essential for the generation of ARMD events.

Interestingly, the 22-bp hotspot sequence contains no CpG dinucleotides. These CpG dinucleotides have been shown to mutate approximately six times faster than other dinucleotides in Alu elements (Xing et al., 2004) due to cytosine methylation and subsequent deamination (Bird, 1980). In addition, when we aligned the consensus sequences of the 10 different Alu subfamilies involved in ARMDs, we found that the hotspot sequence is located within the longest stretch of their conserved regions. Furthermore, using the software utility WebLogo (Crooks et al., 2004), we confirmed that this 22-bp sequence is the most conserved region among Alu elements involved in ARMD events (Figure 3.4). Therefore, the recombination hotspot that we have identified, by virtue of having an increased level of conservation among the Alu subfamilies involved in the ARMDs in our study, has potentially allowed frequent recombination between Alu repeats from different Alu subfamilies to occur.

Genomic Environment of ARMD Events

Most Alu elements located in the primate genomes that have been sequenced (e.g., human, chimpanzee, and rhesus macaque) exist in high-GC content regions (CSAC, 2005; Lander et al., 2001a; RMGSAC, 2007), and also have high GC content (an average of ~62.7%). Moreover, it has also been previously reported that human-specific ARMD events preferentially occur in areas of high GC content (~45% GC content, on average) (Sen et al., 2006). To analyze the genomic environment of chimpanzee-specific ARMD events, we estimated the GC content of
20 kb (±10 kb in either direction) of neighboring sequence for each ARMD locus. Our results indicate that the chimpanzee-specific ARMDs are similar to human-specific ARMDs in having a tendency to occur in GC rich regions (45.2% GC content, on average). This preference is correlated with the distribution of Alu elements involved in ARMDs (Figure 3.3) because the genomic distribution of ARMD events would in effect have an a priori dependence on the preferred locations of Alu elements after insertion of the different Alu subfamilies. About 74% of chimpanzee-specific ARMDs are associated with the older Alu subfamilies, AluJ and AluS. Although young Alu subfamilies are found in AT-rich, gene-poor regions, the older Alu subfamilies are most often found in GC-rich, gene-rich regions (CSAC, 2005). This could account for the preferential occurrence of ARMD events in GC-rich regions. Moreover, the local rate of genomic recombination has been shown to be positively correlated with GC content (Fullerton et al., 2001), which may further explain the observed distribution of ARMD events. About 44% of genomic DNA deleted through ARMD events were Alu sequences in the human ortholog. This could indicate that regions of high local Alu element density within chromosomes are more likely to provide increased opportunities for local recombination, a trend previously noticed during analysis of the global genomic distribution of human lineage-specific ARMD events (Sen et al., 2006).

To further characterize the genomic environment of chimpanzee-specific ARMD events, we estimated the gene density of the genomic regions flanking each chimeric Alu element resulting from the process by extracting 4 Mb of flanking genomic sequences (±2 Mb in either direction), and counting the number of known or predicted chimpanzee RefSeq genes. The gene density of the flanking regions of chimpanzee-specific ARMD events is estimated to be, on average, one gene per 60.7 kb, which is similar to that of human-specific ARMD events (one
gene per 66 kb). This indicates that the global distribution of chimpanzee-specific ARMD events is biased towards gene-rich regions, since the global average gene density in the chimpanzee genome is approximately one gene per 112 kb. To test for any relationship between the size of an ARMD and its flanking gene density or GC content, we performed a correlation test. While the $r$-values for both tests were negative, as would be expected given the danger of large deletions in gene-rich areas, the low $p$-values indicate that no significant correlation exists between the two variables in either test (gene density: $r = -0.028; p = 0.472$; GC content: $r = -0.065; p = 0.095$).

**Chimpanzee-Specific ARMD Polymorphism**

In order to estimate the polymorphism rates in chimpanzees, we analyzed and amplified a total of 50 chimpanzee-specific ARMD loci on a panel composed of genomic DNA from 12 unrelated chimpanzee individuals (see Materials and Methods). Our results show that the polymorphism level of chimpanzee-specific ARMDs (28%) is about two times higher than the polymorphism rate of human-specific ARMD events (15%) (Sen et al., 2006), which is in general agreement with the polymorphism levels from previous studies of chimpanzee- or human-specific retrotransposons (e.g., Alu and L1 elements) (Hedges et al., 2004; Lee et al., 2007).

**Incomplete Lineage Sorting and Parallel Independent ARMDs**

About 32% of the ARMD candidates were found to have ambiguous TSD structures and a triple alignment that proved too complex to assign ARMD status to the locus solely on the basis of our computational output. These loci were verified experimentally using PCR (see Materials and Methods) to determine the authenticity of the chimpanzee-specific ARMDs and identify false positives in the computational data, which were usually caused by human-specific Alu insertions. However, 16 ambiguous loci were identified at which human-specific Alu
Figure 3.5. Incomplete lineage sorting and parallel independent ARMD events. The DNA template used in each reaction is listed on top of the gel chromatograph (M, 100-bp ladder; H, human; C, chimpanzee; G, gorilla; O, orangutan). The large and small sizes of PCR products indicate two Alu elements and one Alu element, respectively. The thunderbolts represent recombination events between two Alu elements, causing ARMDs. Possible scenarios that explain the observed chromatograph: (A) chimpanzee-specific ARMDs, (B) incomplete lineage sorting of an ARMD events, and (C) parallel independent ARMD events.

insertions were not present. In 11 of these loci, the human and gorilla genomes appear to have two Alu elements, while the chimpanzee and orangutan genomes have only one element at the orthologous position. DNA sequence analysis of the PCR products classified five of these 11 loci as chimpanzee-specific ARMDs, with the second of the two recombining Alu elements having integrated into the host genome after the divergence of orangutan and the common ancestor of humans, chimpanzees, and gorillas (Figure 3.5A). Four out of the 11 loci show a pattern consistent with incomplete lineage sorting, in which the ARMD event occurred before the divergence of great apes and was still polymorphic at the time of speciation. Subsequently, the
chimeric Alu elements produced by these ARMD events became fixed in the chimpanzee and orangutan lineages while the two original Alu elements involved in the ARMDs were fixed in the human and gorilla genomes (Figure 3.5B). Incomplete lineage sorting has been reported in cases of retrotransposon insertion polymorphism involving closely related species (Lee et al., 2007; Ray et al., 2006a). In cases where the time between any genomic event and a subsequent speciation is very short, incomplete lineage sorting can easily occur. The remaining two of the 11 ambiguous loci were identified as parallel independent ARMD events in separate primate genomes by aligning the pre-recombination sequence and chimeric Alu elements (Figure 3.5C). These events suggest that orthologous loci may experience two independent lineage-specific ARMDs at different times (i.e., chimpanzee-specific ARMDs and orangutan-specific ARMDs).

In contrast, PCR analysis of the remaining five ambiguous loci (from the 16 referred to above) showed that humans and orangutans have two Alu elements, whereas chimpanzees and gorillas have only one at the orthologous position. Of these five loci, three showed a pattern suggesting incomplete lineage sorting events, while the other two were parallel independent ARMDs. For one of the loci displaying a parallel independent ARMD event, the structural characteristics of the two chimeric Alu elements resulting from independent recombination events are clearly different between the chimpanzee and gorilla genomes. The 574-bp chimpanzee genomic deletion occurred between the left monomer on the first Alu and the right monomer on the second Alu, whereas the 708-bp genomic deletion in the gorilla happened between the two left monomers of the two Alu elements.

These results indicate that at least ~0.9% of chimpanzee-specific ARMD loci (2 of 233 loci which were analyzed by PCR) are shared by the gorilla genome and another ~0.9% are shared by the orangutan genome, due to parallel independent ARMDs at two different time
points in two separate primate genomes. As such, the presence of independently occurring ARMD events in both the human and chimpanzee genomes could lead to false negative events being missed during the previous analysis done by Sen et al. (2006), although the frequency of such false negatives is likely to be very low. In addition, we believe that the human orthologs of the chimpanzee-specific ARMD loci represent sites predisposed for potential future ARMDs in the human genome that could generate human lineage-specific rearrangements and genetic disorders. Identifying putative ARMD hotspot genomic regions is not surprising based upon the frequency of \textit{Alu}-mediated recombination events that have given rise to mutations in a number of different loci, including the \textit{LDLR} and \textit{MLL1} genes (Deininger and Batzer, 1999; Hess, 2004; Lehrman et al., 1985; Purandare and Patel, 1997).

\textbf{Discussion}

\textbf{Differential Level of Lineage-Specific ARMD Events}

Despite the high level of overall similarity between their genomes, humans and chimpanzees have subtly different genomic landscapes because of alterations such as insertions, deletions, inversions, and duplications after their divergence from a common ancestral primate (Bailey and Eichler, 2006; Callinan et al., 2005; Cheng et al., 2005; Deininger and Batzer, 1999; Han et al., 2005; Sen et al., 2006). Although from a mechanistic viewpoint, the chimpanzee-specific ARMD events are similar to the human-specific ones, the total number and size of deletions are substantially different between the two lineages. One reason for the observed differences between these two lineage-specific ARMD patterns may be the increased genetic diversity of the chimpanzee population as compared to the human population, which is known to have experienced a significant reduction in its effective population size after the divergence of humans and chimpanzees (Chen and Li, 2001), leading to a consequent reduction in genetic
diversity. These results are supported by the higher polymorphism level for chimpanzee-specific ARMDs than human-specific ARMDs.

**Balance of Chimpanzee Genome Size**

*Alu* elements as well as other retrotransposons can contribute to the size expansion of primate genomes by increasing their copy numbers and causing homology-mediated segmental duplications (Bailey et al., 2003; Liu et al., 2003; Petrov, 2001). However, the retrotransposon-mediated increase in genome size is not unilateral, because several processes such as retrotransposon-mediated deletions and recombination-mediated deletions concurrently act in the opposite direction, causing reduction in genome size as well (Callinan et al., 2005; Han et al., 2005; Sen et al., 2006). Retrotransposon-mediated negative control of genome size has been well documented in plants such as *Arabidopsis* and rice (Devos et al., 2002; Ma et al., 2004).

In this study, we analyzed the contribution of ARMDs to genome size regulation in the chimpanzee genome by estimating an *Alu*-mediated sequence turnover rate, which is the amount of sequence increase caused by chimpanzee-specific *Alu* insertions relative to the amount of reduction by the chimpanzee-specific ARMD process. The copy number of chimpanzee-specific *Alu* elements (i.e., those that inserted after the divergence of human and chimpanzee) is ~2,340, accounting for ~700 kb of inserted sequence in the chimpanzee lineage (CSAC, 2005), while the amount of sequence deleted by chimpanzee-specific ARMDs is ~771 kb. Therefore, within the past ~6 million y, the genome size of chimpanzees has not expanded but rather has contracted by ~71 kb, when considering the combined effects of *Alu* retrotransposition and recombination-mediated deletion (i.e., the *Alu*-mediated sequence turnover rate is more than 100% in the chimpanzee genome). This observation suggests that ARMD events efficiently counteract genomic expansion caused by novel *Alu* inserts in the chimpanzee genome when compared to the
human genome. A previous analysis of human-specific ARMD events indicates that the Alu-mediated sequence turnover rate is ~20% in the human genome (Sen et al., 2006). This significantly different turnover rate between the two species could be explained by differences in the tempo of Alu amplification (i.e., higher Alu retrotransposition activity in the human genome) and rates of ARMD events (i.e., higher ARMD activity in the chimpanzee genome). Ultimately, it is worth noting that at least in the chimpanzee lineage, concurrent Alu insertion/ARMD mechanisms have balanced the gain and loss of sequences during Alu-mediated genomic alterations.

**Retrotransposition of Chimeric Alu Elements**

To investigate whether chimeric Alu elements are able to retrotranspose in the chimpanzee genome, we tried to find progeny of the 663 chimpanzee-specific chimeric Alu elements using the BLAST-Like Alignment Tool (BLAT) program (http://genome.ucsc.edu/cgi-bin/hgBlat). However, we failed to recover any such elements in the chimpanzee genome for one or more of a number of reasons. First, Alu elements involved in ARMD events are expected to be relatively old (i.e., more than 6 million y) because our comparative analysis detects only ARMD events involving Alu elements that were inserted into the genome before the divergence of humans and chimpanzees. Therefore, most of the ARMD-associated Alu elements probably lost their ability to retrotranspose before the Alu–Alu recombination process. In reality, the contribution of chimpanzee-specific young Alu elements to the ARMD process may be extremely limited due to their low copy number (~2,000 copies) in the chimpanzee genome (CSAC, 2005). Indeed, ARMD events generated by the relatively young AluY subfamilies account for 0.19% of the total AluY elements in the chimpanzee genome. Second, only a few source genes are responsible for new Alu subfamily amplification through retrotransposition.
Although some *Alu* subfamilies (e.g., *Alu*Yc1) are still active in the chimpanzee genome (CSAC, 2005; Hedges et al., 2004), it is improbable that their source gene(s) are involved in the *Alu--Alu* recombination events. Similarly during an earlier analysis (Sen et al., 2006), we investigated the retrotransposition ability of 492 human-specific ARMD-generated chimeric *Alu* elements and were unable to recover their progeny as well.

**ARMD as an Endogenous Process Affecting Human and Chimpanzee Variation**

Recently, the genomic relationship and genetic divergence between the human and chimpanzee genomes have been the subjects of extensive comparative genomic analyses on the basis of their respective draft genome sequences (Cheng et al., 2005; CSAC, 2005; Feuk et al., 2005; Mills et al., 2006a; Mills et al., 2006b). However, these studies have not focused on *Alu*-mediated genomic deletions in the chimpanzee lineage, aside from the 14 *Alu* retrotransposition-mediated deletions reported previously (Callinan et al., 2005).

Thus, our study forms the first comprehensive analysis of recombination-mediated genomic alteration by *Alu* elements in a nonhuman primate (chimpanzee) lineage. We found 305 chimpanzee-specific deletions within protein-coding genes as annotated by the RefSeq gene annotation database, 299 genes from which introns were deleted, and six genes in which thirteen exons were deleted. Remarkably, two chimpanzee-specific ARMD events deleted exons from genes demonstrably functional in the human lineage (*NBR2* and *HTR3D*), providing direct proof that the ARMD process contributes to creating phenotypic differences between humans and chimpanzees. The *NBR2* gene is located near the *BRCA1* gene on Chromosome 17, which is responsible for tumor repressor activity in the human genome, and shares a common promoter for transcription, forming a bidirectional transcriptional unit with *BRCA1*. Although the complete *NBR2* cDNA sequence is ~1.3 kb, it has a short open reading frame (112 amino acids), and is
subject to nonsense-mediated decay (Jin et al., 2004; Xu et al., 1997). In humans, this gene is suppressed by a non–tissue-specific protein complex that binds to its first intron (i.e., the 18-bp repressor element) (Suen et al., 2005). However, in the chimpanzee lineage, an ARMD event occurred between the third intron and the 3′ flanking region, causing an exonic deletion (Figure 3.6A). Thus, this ARMD event could potentially inhibit NBR2 gene expression in the chimpanzee genome, regardless of whether or not the repressor element is present. Although the exonic deletion of the NBR2 gene has been independently reported through a comparative analysis of cancer genes between the human and chimpanzee genomes, the previous analysis did not report what caused this genetic difference between human and chimpanzee genomes (Puente et al., 2006). Our study of chimpanzee-specific ARMDs illuminates the underlying molecular mechanism for this deletion.

A chimpanzee-specific ARMD event also deleted the first coding exon of HTR3D, a functional gene in humans (Figure 3.6B). This gene belongs to the 5-HT3 serotonin receptor-like gene family, which has been recently characterized (Niesler et al., 2003). The 5-HT3D subunit is not a functional receptor on its own (i.e., a homomeric receptor), but when it binds to the 5-HT3A subunit to form the heteroligomeric receptor, 5-HT, maximum response is significantly increased as compared to the homomeric 5-HT3A receptor (Niesler et al., 2007). HTR3D is primarily expressed in the gastrointestinal tract (Niesler et al., 2007), where serotonin is synthesized extensively (Kobayashi et al., 1991). We speculate that the exonic deletion in this gene caused by the chimpanzee-specific ARMD event may lead to a reduction in serotonin levels in the chimpanzee lineage, and thus have an impact on physiological variation between the human and chimpanzee lineages.

The analyses using the RefSeq and UniGene annotations (see Results) indicate that
Figure 3.6. Exonic deletions caused by two ARMD events. Black arrows represent the direction of transcription, and gray and black boxes indicate the noncoding exons and coding exons, respectively. Green and purple arrows indicate elements from two different Alu subfamilies, and dual-color arrows indicate chimeric Alus generated by ARMD events (map is not drawn to scale). (A) An exonic deletion within the NBR2 gene. The AluSg and AluY elements are located within the third intron and the 3’ flanking sequence, respectively, in the human genome. The exon4 sequence is deleted due to an ARMD event in the chimpanzee lineage. (B) An exonic deletion within the HTR3D gene. The AluSx and AluSq elements are located within the second and third introns, respectively, in the human genome. The exon3 sequence, which is includes the initiation codon ATG, is deleted due to an ARMD event in the chimpanzee lineage.

ARMD events could have affected the expression of many genes. Moreover, intronic or intergenic deletions caused by ARMD events may also affect the levels of gene expression in both the human and chimpanzee genomes through alteration of splicing patterns and loss of transcription factor binding sites, further contributing to the divergence of the human and chimpanzee lineages. Additional studies of the functional genomics of the genes altered in both
human and chimpanzee ARMD events will be instructive and provide new insight into the genetic and phenotypic differences between the two species.

**Conclusion**

Retrotransposon-mediated genomic rearrangement could be one of the major factors responsible for the lineage-specific changes in genomes that ultimately lead to speciation. Comparative investigations of the ARMD events apparent between the human and chimpanzee genomes indicate that this process plays an important role in the biological differences between humans and chimpanzees, and provides a reliable record of lineage-specific evolutionary histories due to the nearly homoplasy-free nature of these mutations. Moreover, in the chimpanzee lineage, the chimpanzee-specific ARMD process has completely counteracted the genomic expansion caused by new Alu inserts since the divergence of the chimpanzee and human lineages. The existence of parallel independent ARMD events found at the orthologous loci of some of the 663 chimpanzee-specific ARMD events suggest that other chimpanzee-specific ARMD orthologs in humans may be predisposed to undergo recombination between the two Alu elements in the future. These ARMD orthologous loci may be sites of unstable structure in humans as well as other apes, because they still preserve the pre-recombination structure that has proven itself susceptible to unequal recombination in the chimpanzee lineage.

**Materials and Methods**

**Computational Search and Manual Inspection of Chimpanzee-Specific ARMD Loci**

To computationally screen the chimpanzee genome for potential ARMD loci, we used a technique previously described by Sen et al. (2006) in a study of human lineage-specific ARMD events, with the distinction that, for this analysis, the query and target genomes were reversed. In summary, we extracted 400 bp of 5′ and 3′ flanking sequence for all chimpanzee Alu elements
(PanTro1; November 2003 freeze) and joined the two 400 bp sequences to form a single “query” sequence. A best match for each query sequence was determined by using BLAT (Kent, 2002) against the reference human genome (hg17; May 2004 freeze). Then, the sequence in the human genome (the “hit”) found between the orthologs of the two 400 bp stretches of the query was extracted and aligned with the chimpanzee Alu element sequence initially used to design the query (the “query Alu”) using a local installation of the NCBI bl2seq utility.

One hallmark of de novo Alu insertion is the presence of TSDs flanking each side of the Alu element, generated by the target-site primed reverse transcription process (Cost and Boeke, 1998; Deininger and Batzer, 2002; Jurka, 1997; Luan et al., 1993). However, the single chimeric Alu element created by an ARMD event lacks matching TSD structures in the chimpanzee because it is comprised of fragments from a pair of Alu elements with mutually unique TSDs at the orthologous ancestral locus (Sen et al., 2006). If a potential ARMD locus exhibited the structures of a valid ARMD as described by Sen et al. (2006), we accepted the computational detection as an authentic ARMD locus. In addition, we used the BLAT software utility (Kent, 2002) to compare the human, chimpanzee, and rhesus macaque genomes at each potential ARMD locus. If the two Alu elements in the human genome that are considered to be the pre-recombination Alu elements for an ARMD locus are shared with the rhesus macaque genome at orthologous loci, despite the presence or absence of TSDs, the single Alu element remaining at the orthologous chimpanzee locus is most likely a chimeric element generated an ARMD event. On the basis of these features, we manually inspected 1,538 potential ARMD loci retrieved by the computational data analysis. However, some loci displayed ambiguous TSD structure or remained ambiguous after analysis using the triple alignment. These loci were subjected to PCR analysis and, if necessary, DNA sequencing in order to confirm or eliminate each as being
products of bona fide ARMD events.

**PCR Amplification and DNA Sequence Analysis**

PCR analysis was performed using four different primate species as templates. The cell lines used to isolate DNA samples corresponding the primate species are as follows: human (*Homo sapiens*) HeLa (CCL2; American Type Culture Collection [ATCC], http://atcc.org), common chimpanzee “Clint” (*Pan troglodytes*; NS06006B), gorilla (*Gorilla gorilla*; AG05251) and orangutan (*Pongo pygmaeus*; AG05252A). To evaluate polymorphism rates, we amplified 50 randomly selected ARMD loci on a common chimpanzee population panel composed of 12 unrelated individuals of unknown geographic origin obtained from the Southwest Foundation for Biomedical Research (San Antonio, Texas).

Oligonucleotide primers for the PCR amplification of ARMD events were designed using the Primer3 utility (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The sequences of the oligonucleotide primers, annealing temperatures, and PCR product sizes are shown in Supplemental Table 3.2 from the Batzer Laboratory Web site (http://batzerlab.lsu.edu). Each PCR amplification was performed in 25-µl reactions using 10–50 ng 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 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, and 1 min of extension at 72 °C, followed by a final extension step of 10 min at 72 °C. PCR amplicons were loaded on 1%–2% agarose gels, depending on the amplicon sizes, stained with ethidium bromide, and visualized using UV fluorescence. In cases where the expected size of the PCR product was greater than 1.5 kb, *iTaq* (Bio-Rad, http://www.bio-rad.com) or *Ex Taq* polymerase (TaKaRa, http://www.takara-bio.com) were used, following the
manufacturer’s suggested protocols.

When necessary, individual PCR amplicons were gel purified using the Wizard gel purification kit (Promega, http://www.promega.com) and cloned into vectors using the TOPO-TA Cloning kit (Invitrogen, http://www.invitrogen.com) according to the manufacturer’s instructions. DNA sequencing was performed using dideoxy chain-termination sequencing (Sanger et al., 1977) on an Applied Biosystems ABI3130XL automated DNA sequencer (Applied Biosystems, http://www.appliedbiosystems.com). Raw sequence reads were assembled using DNASTAR’s Seqman program in the Lasergene version 5.0 software package (http://www.dnastar.com).

Analysis of Flanking Sequences

For each chimpanzee-specific ARMD locus, 10 kb of flanking sequence upstream and downstream were collected using a combination of in-house Perl scripts and the nibFrag utility bundled with the BLAT software package. The GC content of the flanking regions of each ARMD locus was calculated by analyzing the combined 20 kb of flanking sequence using another in-house Perl script, which excluded Ns from the analysis. Gene density around individual ARMD loci was estimated using the NCBI Map Viewer utility, run on Build 2.1 of the Pan troglodytes genome (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9598). The neighboring 2 Mb of sequence 5′ and 3′ to each chimeric chimpanzee Alu element was analyzed, and the number of genes found within this combined 4 Mb were noted. All computer programs used are available from the authors upon request.

Accession Numbers

The gorilla and orangutan DNA sequences generated during the course of this study have been deposited in GenBank (http://www.ncbi.nlm.nih.gov/Genbank) under accession numbers
EF682150–EF682182. The GenBank accession numbers for the three HTR3D isoforms discussed in this article are NM_182537, BC101090, and AJ437318.

**References**


CHAPTER FOUR:

CHROMOSOMAL INVERSIONS BETWEEN THE HUMAN AND CHIMPANZEE LINEAGES CAUSED BY RETROTRANSPOSONS
Introduction

Mobile elements make up ~45% of the human genome (Deininger et al., 2003). Among them are L1 and Alu elements, that have been active since well before the divergence of the human and chimpanzee lineages, and remain active in their host genomes. These two elements mobilize via a “copy and paste” mechanism and integrate into new genomic regions by means of an RNA intermediate (Luan et al., 1993). A full-length functional L1 element is about 6 kb in length and able to code for enzymes which are required for L1 retrotransposition, making the L1 an autonomous element (Mathias et al., 1991). By contrast, the Alu element is 300 bp long and does not encode the means of its own retrotransposition, instead borrowing the enzymatic machinery of the L1 elements for its propagation (Deininger et al., 1992; Weiner, 2000), making it a non-autonomous mobile element. Although L1 elements contribute the most to the genome in terms of total size, Alu elements are the most successful mobile element family in terms of copy number, reaching a copy number of ~1.2 million in the human genome (Lander et al., 2001b).

L1 and Alu elements have played an important role in shaping their host genomes. They can alter gene expression patterns and cause chromosomal rearrangements through various mechanisms including novel insertion, insertion-mediated deletion, and unequal homologous recombination between elements (Callinan et al., 2005; Han et al., 2005; Sen et al., 2006). Sequence identity between two retrotransposons of the same type (e.g., Alu-Alu and L1-L1) can lead to non-allelic homologous recombination between them, that subsequently results in chromosomal rearrangements such as duplications, deletions, translocations, and inversions (Bailey and Eichler, 2006; Feuk et al., 2005; Han et al., 2007; Sen et al., 2006). Such recombination can cause species-specific local genomic instability and has been reported as a
major source of genomic disorders (Shaw and Lupski, 2004).

Inverted \textit{Alu} and L1 pairs (i.e., two \textit{Alu} elements or two L1 elements inserted in opposite orientations along a chromosome) have caused chromosomal rearrangements in their host genomes through several mechanisms including large inverted duplications, translocations, inversions, and deletions (Li and Bray, 1993; Lobachev et al., 2000; Narayanan et al., 2006). Due to their sequence similarity, they have the ability to form a hairpin structure in single-stranded DNA or a cruciform structure in double-stranded DNA (Lobachev et al., 2000; Nag and Kurst, 1997; Zheng and Sinden, 1988). These structures can potentially block progression of the replication fork and cause intra- or inter-molecular template switching of DNA polymerase between the inverted elements (Lobachev et al., 2002; Lobachev et al., 2000). In reality, inverted \textit{Alu} pairs cause a 1000-fold increase in homologous recombination (Lobachev et al., 2000). Here, we report for the first time a genome-wide analysis of \textit{retrotransposon recombination-mediated inversion} (RRMI), causing genomic and subsequently phenotypic differences between humans and chimpanzees. The previously reported mechanism, \textit{Alu recombination-mediated deletion} (ARMD), alters or interrupts gene function through the deletion of intronic and exonic regions. By contrast, RRMI usually does not cause any change in genome size. Instead, it could alter the structure of genes or transcription of genes by inverting intron or exon sequences and introducing alternative gene splicing sites. Through the comparison of human and chimpanzee draft genome sequences (CSAC, 2005; Lander et al., 2001b), we identified 49 RRMI loci, 28 of which were human-specific inversions and 21 were chimpanzee-specific inversions. Among them, 53% of the RRMI occurred within genic regions. Interestingly, we found that three RRMI events caused alteration of exonic regions in known genes with ten RRMIs that are polymorphic within a species. These findings suggest that recombination between inverted L1 and \textit{Alu} pairs might
have generated genomic variation within a species as well as between species.

**Results**

**A Whole-Genome Scan for Inversion Events between Human and Chimpanzee Lineages**

To identify potential inversion loci between human and chimpanzee lineages, we computationally compared human with chimpanzee genome reference sequences. We initially obtained a total of 6887 inversion candidates ranging in size from 27 bp to 47.3 Mb and discarded 986 loci whose human chromosomal positions were unknown or random. The remaining 5902 loci were subjected to flanking sequence analysis as described in the materials and methods section. Among them, 3055 loci were categorized as false positives for inversions between the human and chimpanzee genomes. Our computational methodology excluded these loci due to a failure of UCSC’s liftOver utility to find the orthologous positions between the two species. These failures result from species specific-genomic deletions, duplications, or splits, after their removal, a total of 2847 loci were collected as candidate inversion loci.

These loci were then subjected to manual inspection. Sequence disagreement between human and chimpanzee genome sequences resulting from the unsequenced regions of the chimpanzee genome and genomic defragmentation (Giordano et al., 2007) significantly reduced our ability to find the inversion breakpoints, especially when the sequence disagreement occurred in the genomic regions where an inversion began or ended. As such, many inversion events may have been eliminated from our data set even though likely to be authentic. Intrachromosomal duplications in which the duplicate is inserted in the reverse direction (inverted duplication) are likely to be a major source of false positives for this analysis. To identify and eliminate them from our data set, we used human inversion sequence as a query for BLAT against human genome sequence. A false positive showed two highest score hits in the
BLAT results, corresponding to the query sequence and the inverted duplication sequence (+ and -), respectively. We removed these false positive inversion loci from our data and finally confirmed 253 inversion events (Figure 4.1) whose inversion breakpoints are able to be characterized.

Figure 4.1. The 253 inversion loci between the human and chimpanzee lineages. Blue and red circles indicate Alu-RMI and L1-RMI events, respectively. All inversions except for those caused by RRMI are indicated by green circles. The karyotype images were created using the idiographica webtool (Kin and Ono 2007).
Breakpoint Examination for RRMI

To characterize inversion breakpoints, we retrieved human flanking sequence of the 253 inversion loci and used them, combined with the putative inversion sequence, as queries for BLAT searches against the chimpanzee genome sequence (panTro2). The flanking regions, as expected, matched between human and chimpanzee genomes. However, the inverted region stood out clearly, allowing the beginning and end of each inversion, the breakpoints, to be identified.

To identify RRMI events, we examined whether L1s or Alu elements spanned the two inversion breakpoints of each inversion event, and whether or not their orientation was opposite to one another. For example, Alu-RMI occurs when two Alu elements are found spanning the two breakpoints of an inversion and are oriented in opposite directions along the chromosome. We found 49 RRMI loci (28 Alu-RMI and 21 L1-RMI, Table 4.1, Supplemental Table 4.1 in Batzer Laboratory Web site; http://batzerlab.lsu.edu) out of the 253 inversion events. Intriguingly, 63 of the remaining 204 inversions were also associated with an L1 or Alu element. For these, however, the retrotransposon spanned only one of the two breakpoints, while the other breakpoint was located independently of repetitive elements. One possible explanation for these

<table>
<thead>
<tr>
<th>Retrotransposon-RMI</th>
<th>Human-specific inversion</th>
<th>Chimpanzee-specific inversion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><em>Alu</em>-RMI</td>
<td><em>L1</em>-RMI</td>
</tr>
<tr>
<td>Total events †</td>
<td>14 (3)</td>
<td>13 (1)</td>
</tr>
<tr>
<td>Total inversion size (bp)</td>
<td>27078</td>
<td>185831</td>
</tr>
<tr>
<td>Average of inversion size (bp)</td>
<td>1934</td>
<td>14294</td>
</tr>
</tbody>
</table>

† The numbers within the parentheses indicate the numbers of RRMI which are accompanied by the deletion of partial inverted sequence.
loci is that microhomology between the retrotransposon and the genomic region where the other inversion breakpoint occurs induced the recombination event responsible for the inversion.

When an inversion occurs, the retrotransposons spanning the inversion breakpoints recombine, becoming chimeric elements consisting of the front portion of one element and the back portion of the other. To further characterize the inversion breakpoints of the RRMI loci, we aligned the two ancestral, pre-recombined retrotransposons (e.g., AluSg and AluSx) with one of the recombined retrotransposons for each RRMI locus (Figure 4.2). These alignments allowed more precise determination of where the breakpoints occurred within each element. The inversion breakpoints on L1 and Alu elements were found to be evenly distributed, which indicates that no recombination hotspot exists for these retrotransposons regarding inversion events between the human and chimpanzee genomes.

Figure 4.2. Sequence alignment of one recombined and two prerecombined Alu elements involved in an Alu-RMI event. The recombined (chimeric) Alu element and two prerecombined Alu elements that contributed to its formation are showed in order. Identical nucleotides shared among elements are indicated by dots. Otherwise, differences are shown with letters. The recombination breakpoint for this event is located in the yellow box.
RRMI Characterization

As described above, we examined the ancestral state of each RRMI locus using three methodologies. Among the 49 RRMI loci, 27 loci were human-specific inversions whereas 22 loci were chimpanzee-specific inversions. We grouped them into L1-RMI and Alu-RMI depending on the type of retrotransposon that spanned the inversion breakpoints. As shown in Table 4.1, the 49 loci contained 21 L1-RMIs and 28 Alu-RMIs.

We further investigated the subfamilies of L1 and Alu elements involved in the inversion events. The analysis of Alu subfamilies showed that the number of elements from each Alu subfamily involved in Alu-RMI is proportional to their genome-wide copy number (Figure 4.3).

Figure 4.3. Alu and L1 subfamilies involved in RRMI events. The proportion of Alu elements involved in Alu-RMI events (blue bars) and the total number of Alu elements for each subfamily (black bars) are compared in the left side. The proportion of LINEs involved in L1-RMI events (red bars) and the total number of LINEs for each subfamily (gray bars) are compared in the right side.
This result implies that the elements with high copy number are more frequently subjected the recombination than are elements with lower copy numbers. However, more members of the *AluY* subfamily are involved in the *Alu*-RMI events than those of the *AluJ* subfamily, even though the *AluJ* subfamily has a higher copy number than the *AluY* subfamily in the human and chimpanzee genomes. It is useful to note here that the *AluY* subfamily is younger and, therefore, its members tend to have more sequence identity with one another, relative to the *AluJ* subfamily. This suggests that, along with copy number, a high level of sequence identity is also important in the recombination between the two *Alu* elements. This finding is consistent with the patterns described in studies of species-specific ARMD (Han et al., 2007; Sen et al., 2006). As shown in Figure 4.3, the analysis of L1 subfamilies further supports that sequence identity is an important factor affecting the frequency of recombination between these elements.

The RRMI loci range in size from 166 bp to 81,189 bp with an average and a median size of 5364 bp and 1452 bp, respectively. A majority of *Alu*-RMI loci are responsible for the inversions whose sizes are shorter than 1 kb. In contrast, more than half of L1-RMIs are longer than 10 kb. Interestingly, the average size of human-specific inversions is three times longer than that of chimpanzee-specific inversions. We tested the correlation between the length of elements involved in the inversion event and respective inversion size. This analysis showed a statistically significant positive correlation between the two variables \( r = 0.578; p<0.0001 \), and suggests that the larger the number of nucleotides capable of base pairing between the two elements the larger the inversion is likely to be.

**RRMI Polymorphism**

Through PCR assays, we verified the integrity of 33 RRMI loci and excluded one chimpanzee-specific inversion locus resulting from sequence assembly error in the chimpanzee
genome reference sequence. However, we could not experimentally confirm the remaining loci because they contained a high density of repetitive elements, that inhibit PCR amplification of their respective genomic regions (Batzer et al., 1991). Additionally, we estimated the polymorphism levels of Alu-RMI loci using PCR assay. Nine human-specific Alu-RMIs were genotyped in 80 diverse humans (20 individuals from each of four populations, composed of African-American, European, Asian, and South American individuals) and seven chimpanzee-specific Alu-RMIs were genotyped in 12 unrelated common chimpanzees. Among them, we identified three human-specific Alu-RMI polymorphic loci whose minor allele frequencies were 0.6 %, 1.3 %, and 43%, respectively. Of the three polymorphic loci, the last has been independently reported through an inversion analysis between the human and chimpanzee genomes (Feuk et al., 2005). By contrast, only one chimpanzee-specific Alu-RMI was found to be polymorphic, and its minor allele frequency was 25%.

Ninety polymorphic inversion loci between the human genome project assembly and the Venter genome sequence were previously reported (Levy et al., 2007). Intriguingly, six of the human-specific RRMI loci in our data are found in this data set. Thus, it could be stated that at least nine human-specific RRMI loci including the three loci above contribute to genomic variation within the human population. In addition, two of the nine inversion loci show evidence of inverted exonic regions in two known genes, DOCK3 and USP40. DOCK3 plays an important role in the engulfment of apoptotic cells and in the migration of cells (de Silva et al., 2003), while USP40 encodes an ubiquitin-specific peptidase 40 that is related to Parkinson disease (Li et al., 2006). Based on these results, we suggest that RRMI events are one source of genomic and phenotypic variation in the human population.
RRMI and the Divergence of Humans and Chimpanzees

Any given inversion locus could be polymorphic within a species but fixed between species. Thus, 27 human-specific RRMIIs and 22 chimpanzee-specific RRMIIs independently shape their respective genomes, accelerating the genomic divergence between the two species. Our results show that 26 inversions occurred in genic regions while 23 occurred in intergenic regions. Three chimpanzee-specific events are responsible for the inversion of exonic regions in predicted genes, as annotated by the N-SCAN gene prediction tool (Gross and Brent, 2006). In addition, one human-specific inversion involves an exon of the isoform of the JMJD5 gene (AK310885), which is a putative histone lysine demethylase. Inversions neighboring exons or introns could significantly impact gene function, either by disrupting the gene itself or by generating alternative splice sites or altering gene regulatory networks. Although 23 RRMI events are located in intergenic regions, they could also affect gene expression by locating upstream or on the gene regulatory regions. The effect of RRMI on their host genome is ongoing, leading to continued genomic variation between and within the human and chimpanzee species.

Environmental Characterization of RRMI

To estimate the GC content of the genomic regions neighboring the RRMI loci, we extracted 20 kb of flanking sequences (±10 kb in either direction) for each RRMI which does not include the inverted sequence. For this test, we analyzed L1-RMI loci and Alu-RMI loci separately because L1s tend to occur in low GC genomic regions while Alu insertions preferentially occur in high GC regions (Lander et al., 2001b; Lee et al., 2007). As expected, most of L1-RMI loci were located in GC-poor regions (~39% GC content, on average) while most of Alu-RMIs were found in relatively GC-rich regions (~44% GC content, on average) (Figure 4.4). It was recently reported that young Alu elements are more ubiquitous in AT rich
Figure 4.4. Analysis of GC content in flanking regions of RRMI loci. The vertical axis represents the relative frequency of RRMI loci within each GC bin. Black bars and blue bars indicate \textit{Alu}-RMD and L1-RMD events, respectively.

regions of the human genome (Cordaux et al., 2006). Nonetheless, our results showed that seven out of eight inversion events caused by the \textit{Alu}Y subfamily occurred in genomic regions with GC contents higher than 41\%, the genome-wide average (Lander et al., 2001b).

We estimated the gene density of the genomic regions flanking RRMI loci by counting the number of known or predicted genes in the 4 Mb of the flanking sequences (±2 Mb in either direction). The gene density of the regions neighboring \textit{Alu}-RMI loci is estimated to be one gene per 60 kb, on average. This estimate of the gene density is congruent with the gene density of the flanking regions of ARMD loci in the human and chimpanzee genomes (Han et al., 2007; Sen et al., 2006). This is an expected result because \textit{Alu}-RMI and ARMD events both result from the same mechanism, recombination between \textit{Alu} elements. In contrast, the gene density of the regions neighboring L1-RMI loci is estimated to be one gene per 98 kb which is similar to the
global average gene density in the human genome (one gene per 94 kb). Despite the fact that L1-RMI events were located, on average, in less gene-dense regions of the genome than their Alu-RMI counterparts, we found that five out of the six RRMI events that caused the inversion of exonic regions within known and predicted genes were L1-RMI events.

GC content is positively correlated with gene density and the local chromosomal recombination rate (Fullerton et al., 2001; Lander et al., 2001b; Payseur and Nachman, 2002). Our results based on GC content, gene density, and frequencies of Alu-RMI and L1-RMI are largely congruent. However, we found one interesting locus that resulted from the recombination between two L2 elements. L2 is an ancestor of L1 and, therefore, presumably inserted to host genome several hundred millions ago (Lovsin et al., 2001). As the time an element resides in a specific genomic locus increases, more nucleotide substitutions accumulate in the elements. This age-related degradation significantly reduces the nucleotide identity between members of L2 subfamily. We investigated this locus in detail and discovered that its flanking sequence has a GC content of 59%. High GC content, as stated above, correlates with high local rates of recombination. Thus, we suggest that the local chromosomal recombination rate may affect the occurrence of chromosomal inversions.

Discussion

Impact of Inversions on the Genomic Variation between Humans and Chimpanzees

Chromosomal rearrangements are thought to be important in the speciation events separating the human from its nearest extant relative, the chimp (Goidts et al., 2005; Rieseberg and Livingstone, 2003). Among them, chromosomal inversions, including nine pericentric inversions, have been considered major drivers in the speciation process (Goidts et al., 2005; Szamalek et al., 2005; Yunis and Prakash, 1982). An inversion results from two breaks on a
single chromosome followed by a reversal of the orientation of the chromosomal segment between the breaks (Jaarola et al., 1998). This mechanism is unlikely to result in insertion and deletion events, and usually does not result in a change in genome size, which makes the identification of these events more difficult. This characteristic, combined with imperfect genome assemblies, makes the estimation of a precise number of inversions between these lineages difficult. As an example, a previous comparative study identified 1576 putative inversions (Feuk et al., 2005), but this data set included a large fraction of false positives, likely resulting from the use of the lower quality early sequence assembly of the chimpanzee genome for comparison with the human genome.

Our study uses a comparison between the highest quality genome assemblies currently available, and identified a total of 323 inversion loci between the human and chimpanzee lineages, regardless of whether they have precise inversion breakpoints. However, this number is likely to be an underestimate because of the method we used to validate candidate inversion events (see Materials and Methods). Large inversions are particularly likely to be eliminated from our data because they are more frequently subjected to species-specific chromosomal rearrangements. This leads to sequence disagreement between inverted and non-inverted sequences, making identification of the original inversion difficult. Our results show that a majority of human-specific inversions reside on chromosomal regions with low local recombination rates. This indicates that inversion events suppress recombination in surrounding regions (Jaarola et al., 1998; Rieseberg and Livingstone, 2003). In addition, half of the inversion events identified involve exonic or intronic regions, that could impact the level of gene expression resulting in phenotypic variation between and within the human and chimpanzee lineages. Thus, we suggest that chromosomal inversions not only contributed to the speciation
events in the human and chimpanzee lineages, but also contribute to their current genomic variation.

**Role of Alu and L1 in Causing Inversion Events**

It has been speculated that mobile elements are one of the factors contributing to chromosomal inversions between the human and chimpanzee lineages. Here, we comprehensively analyzed retrotransposon-mediated inversion between the two species. Among the 253 inversion loci identified, 49 inversions were found to have been caused by inverted L1 or Alu pairs. In addition, 63 of the remaining inversion loci were cases where an L1 or Alu spanned only one of the two inversion breakpoints. Interestingly, one inversion locus contained two L1 elements that were oriented in the same direction, contrary to the expected pattern for RRMI. One possible explanation for this locus is that double-strand breaks (DSBs) occurred within the two L1s, after which the internal sequence was reversely oriented and the breaks repaired. This suggests that L1 and Alu elements could serve as fragile sites that tend to result in chromosomal breaks or gaps leading to inversions (Schmidt et al., 2005). In total, L1 and Alu elements are shown to be responsible for approximately ~44% (112/253 events) of the total inversions between the human and chimpanzee lineages.

Along with retrotransposons, segmental duplications are considered to be major factors contributing to chromosomal inversion (Bailey et al., 2002; Casals and Navarro, 2007; Dennehey et al., 2004; Locke et al., 2003; Schmidt et al., 2005). Sequence identity between the inverted segmental duplications is high enough to cause non-allelic homologous recombination and thus facilitates chromosomal inversion (Schmidt et al., 2005). In addition, the comparison of human and mouse genome sequences showed that the segmental duplications are highly related to chromosomal breakpoints in the inversion areas (Bailey et al., 2004). This finding strongly
supports the relationship between segmental duplications and chromosomal inversions because a chromosomal break is a necessary step in generating an inversion. Interestingly, Alu elements have been suspected as prime candidates to mediate the formation of segmental duplications. This is supported by the fact that the formation of most segmental duplications coincides with the timing of a burst in Alu amplification beginning ~35 million years ago (Shen et al., 1991; Wessler, 2006). Taken together, Alu elements and L1s have a high potential to have mediated the chromosomal inversions observed between the human and chimpanzee lineages.

**Inverted Repeats and Genomic Instability**

L1 and Alu elements are the most abundant mobile elements in the human and chimpanzee genomes (CSAC, 2005; Lander et al., 2001b) and thus L1 and Alu pairs that are inverted in their orientation relative to one another are common throughout the genomes. These inverted repeats have been considered as hotspots in causing chromosomal rearrangements. Base pairing between inverted L1 or Alu pairs can form single-stranded hairpin structures, the formation of which is spontaneous due to the low free energy of the hairpin structure (e.g. the most probable hairpin formation has a $\Delta G$ of -12.4 kcal/mol) (Kelchner and Wendel, 1996). This hairpin structure places the chromosomal regions adjacent to the elements involved into close physical proximity, increasing the likelihood of DSB, recombination, and replication slippage on the regions flanking the stem loop structure. Any DSB could be repaired by non-allelic homologous recombination or non-homologous end joining, resulting in genomic inversions or deletions. In case where recombination between the inverted repeats results in an inversion of the internal sequence, the recombination rate between the inverted repeats is positively related to the size of the repeats but negatively related to the distance between the repeats (Lobachev et al., 1998). Thus, inverted L1 pairs are able to induce the inversion of longer genomic sequences than
inverted Alu pairs, as shown in our results (Table 4.1). Inverted L1 and Alu pairs not only facilitate recombination between themselves, but also increase local recombination rate on their respective chromosomal regions. One previous study reported that inverted repeats increased intrachromosomal and interchromosomal recombinations on their flanking regions 2400-fold and 17000-fold, respectively (Lobachev et al., 1998). In addition, the inverted repeats cause interchromosomal effects by acting as hotspots for mitotic interchromosomal recombination (Gordenin et al., 1993).

During DNA replication, single-stranded DNA can form a secondary structure by allowing base pairing between inverted L1 and Alu pairs, which may predispose DNA polymerase to slip on the replication template, leading to the deletion of some genomic regions. The genomic deletion caused by inverted repeats have been well studied in various organisms, including bacteria, yeast, and human (Gebow et al., 2000; Gordenin et al., 1993; Lobachev et al., 1998). Inverted repeats frequently cause genomic deletions, even during the inversion process resulting from recombination between them. In our data, 12 out of the 49 RRMIs are accompanied by genomic deletions that deleted a portion of the internal sequence and/or the retrotransposon causing the inversion. We extended this examination to the total number of 253 inversion loci identified between humans and chimpanzees and found that ~30% of the inversion events (75/253) involved genomic deletions of variable sizes. Although Alu elements are evenly distributed throughout the genome in terms of their orientation, when Alu pairs whose internal sequence is shorter than 650 bp were counted, two-thirds of the total number of Alu pairs belong to non-inverted Alu pairs in the human genome. However, as the length of the internal sequence increases, the proportions of the non-inverted and inverted Alu pairs become balanced (Stenger et al., 2001). These findings suggest that inverted repeats located close to one another are more
unstable in host genomes.

In conclusion, our study supports that inverted repeats could have played an important role in genome variation between and within the human and chimpanzee lineages. Although the number of inverted L1 and \textit{Alu} pairs is similar between human and chimpanzee, they have shaped different chromosomal regions in independent ways, accelerating genomic variation and subsequent phenotypic variation between the two lineages. In this study, we conducted a genome-wide analysis of RRMI between the human and chimpanzee lineages. However, more detailed studies about other chromosomal rearrangements that may be caused by inverted repeats are required to understand the full extent of their role in chromosomal evolution and speciation.

\textbf{Materials and Methods}

\textbf{Computational Data Mining and Manual Inspection for RRMI Loci}

For the comparison of human and chimpanzee genome reference sequences, we utilized the March 2006 freeze of the human (\textit{Homo sapiens}) genome and the March 2006 freeze of the chimpanzee (\textit{Pan troglodytes}) genome from the University of California Santa Cruz (UCSC). To identify potential RRMI events between the two genomes, we first found all putative inversion loci between them, based on UCSC Table Browser utility, comparing human to chimpanzee genome reference sequences (http://genome.brc.mcw.edu/cgi-bin/hgTables?org=Human&db=hg18&hgsid=2066727&hgta_doMainPage=1). After obtaining the human and chimpanzee genomic positions for each inversion locus, we extracted 15 kb of flanking sequence in either direction of the human genomic position. By using UCSC’s liftOver utility (http://genome.brc.mcw.edu/cgi-bin/hgLiftOver), we obtained the orthologous positions within the chimpanzee genome reference sequence that corresponded to the human flanking sequences. If liftOver failed to return an orthologous position in the chimpanzee genome, the
locus was discarded. The remaining inversion loci were subjected to manual inspection. We extracted the inverted human sequence and 1 kb of flanking sequence in either direction of the inversion. Next, the human sequence was used as a query to search against the chimpanzee genome sequence using UCSC’s BLAST-like alignment tool (BLAT). For each hit in the BLAT search, we retrieved the human and chimpanzee sequences and annotated repeat elements existing in the sequences utilizing RepeatMasker (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) analysis. In the case of authentic inversions between the human and chimpanzee genomes, the RepeatMasker output would show that the order and direction of repetitive elements in the human loci were reversed relative to their chimpanzee counterparts.

**PCR Amplification and DNA Sequencing**

RRMI loci were verified by PCR assay with four different DNA templates including human, chimpanzee, gorilla, and orangutan. Cell lines used to isolate the DNA samples were as follows: *Homo sapiens* (HeLa; ATCC CCL-2), *Pan troglodytes* (common chimpanzee Clint: AG06939B), *Gorilla gorilla* (western lowland gorilla: AG05251), and *Pongo pygmaeus* (orangutan; AG05252A).

Oligonucleotide primers for each RRMI locus were designed using Primer3 software (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi) and then computationally tested utilizing both the Oligonucleotide Properties Calculator (Kibbe, 2007) and UCSC’s In-Silico PCR (http://genome.ucsc.edu/cgi-bin/hgPcr?command=start). The primers were then used to amplify RRMI loci (Supplemental Table 4.2 in Batzer Laboratory Web site; http://batzerlab.lsu.edu). Each PCR amplification was performed in 25 µl reactions with 10–50 ng 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 5 min at 95 °C, followed by 32 cycles of PCR at 15 sec of denaturation at 95 °C, 30 sec at the annealing temperature, and 1 min of extension at 72 °C, followed by a final extension step of 10 min at 72 °C. The PCR products were loaded on 1–2% agarose gels, depending on the product sizes, stained with ethidium bromide, and visualized using UV fluorescence (Bio-Rad, Hercules, CA). In cases where the expected size of the PCR product was greater than 1.2 kb, iTaq (Bio-Rad, Hercules, CA), Ex Taq polymerase (TaKaRa, Otsu, Shiga, Japan) or KOD Hifi DNA polymerase (Novagen, Madison, WI) were used following the manufacturer’s instructions.

If needed, individual PCR products were purified from the agarose gels using the Wizard gel purification kit (Promega, Madison, WI) and cloned into vectors using TOPO-TA Cloning kit (Invitrogen, Carlsbad, CA) according to the manufacturer’s instructions. For each sample, three colonies were randomly selected and subject to colony PCR. The sequencing of the colony PCR products was performed using dideoxy chain-termination sequencing on an Applied Biosystems ABI3130XL Genetic Analyzer (Applied Biosystems, Foster City, CA). Raw sequence data were analyzed using DNASTAR’s Seqman program in the Lasergene version 5.0 software package (http://www.dnastar.com).

**Identification of Ancestral State for RRMI**

To identify the ancestral (i.e., pre-inversion) state of each RRMI locus, we combined three methods: target-site duplication (TSD) analysis, BLAT search, and PCR assay. L1 and *Alu* elements are accompanied on both sides by short direct repeats termed TSDs, which range in size from 7 to 20 bp and are nearly identical to one another (Fanning and Singer, 1987). Each element tends to have unique TSDs and rarely share TSD sequences with other elements. Given this, we scrutinized the TSDs of the L1 and *Alu* elements that spanned each inversion breakpoint (Figure 1).
Figure 4.5. RRMI between human and chimpanzee lineages. The mechanism underlying RRMI is shown at the left. In the illustration of the ancestral state, the two retrotransposons have intact TSDs whose sequence is listed in the colored boxes. The shape “X” indicates recombination between the retrotransposons. In the illustration of the human-specific inversion, both retrotransposons are chimeric, and no longer have matching TSDs. For both illustrations, two arrows indicate the positions where each oligonucleotide primer anneals to for PCR amplification. Agarose-gel chromatographs of PCR products are shown on the right. The upper gel picture displays the ancestral state of the RRMI, while the lower gel picture displays the human-specific inversion. The DNA templates used in each PCR reaction are shown on top of the gel pictures.

4.5). If an RRMI event had occurred, the breakpoint-spanning elements would become chimeric, and the TSDs for these elements would no longer match one another. The determination of the ancestral state of each locus could therefore be made based upon the presence of matching TSDs.

Identification of the ancestral state using BLAT searches involved the use of orangutan and rhesus macaque as out groups. We used the human inverted sequences as queries for BLAT searches against four genome assemblies: the human (hg18), chimpanzee (panTro2), orangutan (ponAbe2), and rhesus macaque (rheMac2). Human-specific inversions were characterized by a
pattern in which all genomes except the human showed similar orientation patterns in the graphical results window provided by BLAT. In contrast, cases of chimpanzee-specific inversions produced patterns in which only the chimpanzee genome showed different graphical patterns from the others.

For those RRMI loci whose ancestral state was still ambiguous, despite both TSD and BLAT analyses, we experimentally confirmed the ancestral state using PCR assays. We designed one oligonucleotide primer from the flanking sequence of the inversion and the other from the internal sequence between two repeats. To decide the ancestral state of the RRMI, we then compared PCR products from human, chimpanzee, gorilla, and orangutan (Figure 4.5).

**Analysis of RRMI Franking Sequences**

To estimate the gene density of genomic regions neighboring the RRMI loci, we counted the number of genes within the 4 Mb of sequence flanking the 5’ and 3’ ends of each RRMI locus, using the NCBI Map Viewer utility, run on Build 36.3 of the *Homo sapiens* genome and Build 2.1 of the *Pan troglodytes* genome (http://www.ncbi.nlm.nih.gov/mapview). For GC content analysis, 10 kb of flanking sequence in either direction of each RRMI locus was collected. The GC content of the combined 20 kb of flanking sequences was then calculated using the Mobyle geecee utility (http://mobyle.pasteur.fr/cgi-bin/MobylePortal/portal.py?form=geecce). All DNA sequences generated during the course of this study were deposited in GenBank (http://www.ncbi.nlm.nih.gov/Genbank) under accession numbers FJ167604-FJ167607.

**References**


CHAPTER FIVE:

SUMMARY
The recent sequencing of a number of primate genomes has consistently shown that repetitive DNA segments comprise much more of the genome than coding DNA sequences. The majority of these repetitive DNA segments consist of mobile elements, that make up ~50% of primate genomes. Many of these elements are currently active, inserting new copies of themselves into new loci in their host genomes, resulting in an increase in genome size. By contrast, these elements also have ability to decrease the size of their host genome through insertion-mediated and recombination-mediated deletions. As such, mobile elements are a substantial and vigorous source genome size variation and chromosomal rearrangements in primate lineages. This dissertation aimed to examine the activities of L1 and Alu elements in creating genomic variation in human and chimpanzee lineages. L1 and Alu elements utilize various mechanisms to shape their host genomes, and among the mechanisms, we focus on the effects of de novo insertion, recombination-mediated deletion, and recombination-mediated inversion.

In chapter two, 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. Although human L1 subfamilies belong to the same lineage, we identified two lineages of recently integrated L1 subfamilies in the chimpanzee genome. These 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 could be 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. We believe that the chimpanzee genome constitutes an excellent model in which to analyze the evolutionary dynamics of L1 retrotransposons.

In chapter three, we compared the chimpanzee and human genomes to determine the extent of Alu recombination-mediated deletion in the chimpanzee genome since the divergence of the chimpanzee and human lineages. Combining computational data analysis and experimental verification, we have identified 663 chimpanzee lineage-specific deletions attributable to this process involving a total loss of ~771 kb of genomic sequence, essentially counteracting the genomic expansion caused by chimpanzee-specific Alu inserts. The RefSeq databases indicate that 13 exons in six genes are annotated as either demonstrably or putatively functional in the human genome, and 299 intronic regions have been deleted through ARMDs in the chimpanzee lineage. Therefore, our data suggest that this process may contribute to the genomic and phenotypic diversity between chimpanzees and humans. In addition, we found four independent ARMD events at orthologous loci in the gorilla or orangutan genomes. This suggests that human orthologs of loci at which ARMD events have already occurred in other nonhuman primate genomes may be “at-risk” motifs for future deletions, which may subsequently contribute to human lineage-specific genetic rearrangements and disorders.

In chapter four, we identified 49 retrotransposon recombination-mediated inversion loci between the human and chimpanzee genomes. Among them, six RRMI loci are responsible for inversions of exonic regions in known or predicted genes, which could result in phenotypic differences between the two lineages. In addition, we experimentally confirmed four loci found
to be polymorphic within a species. We suggest that inversions due to recombination between L1 and Alu elements could be a factor leading to genomic variation within human populations. We estimated that these elements are responsible 44% of the inversions found between the human and chimpanzee genomes. These findings suggest that L1 and Alu elements play a significant role in creating inversions between the two genomes.

As has been shown above, mobile elements have had a variety of impacts on their host genomes throughout primate evolution. Here the underlying distribution of de novo L1 insertions in the human and chimpanzee lineages has been examined. This study has shed new insight into the fundamental sequence diversity of L1s between the two lineages. Through the comparison of human and chimpanzee reference sequences, the genomic variation associated with de novo L1 insertions, Alu recombination-mediated deletions, and retrotransposon recombination-mediated inversions has been determined. These genomic variations are species-specific genomic rearrangements that have accumulated between the two lineages since their divergence. Some of these sequence variations presumably generate functional phenotypic variation, possibly having contributed to the isolation mechanisms present during the divergence of human and chimpanzee lineages. Therefore, this study shows the dynamic impact of mobile elements on the genomic variation in the human and chimpanzee lineages.

In this dissertation, genomic variations caused by L1 and Alu elements in the human and chimpanzee lineages have been analyzed. The results of this research support that these mobile elements are major drivers causing and accelerating the genomic differentiation not only between the human and chimpanzee genomes but also within a species. Advances in DNA sequencing technology have recently increased the number of available genome sequences for human individuals. Comparative genomics between these individuals should facilitate further
assessments of the impacts of mobile elements on human genome evolution. Using comparative genomics techniques and findings from this dissertation research, we will be able to elucidate mechanisms associated with genomic variations in the human lineage and quantify the levels of genomic variation between human individuals caused by mobile elements. These fascinating components of our genomes speak volumes about our shared evolutionary history, and our investigation of them will continue to enrich our understanding of human genomic architecture, disease, and diversity, both past and present.
This is a License Agreement between Jungnam Lee ("You") and Elsevier ("Elsevier"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

**Supplier**
Elsevier Limited, The Boulevard, Langford Lane
Kidlington, Oxford, OX5 1GB, UK

**Registered Company Number**
1982084

**Customer name**
Jungnam Lee

**Customer address**
7250 Perkins Rd. #217, Baton Rouge, LA 70808

**License Number**
2024361363421

**License date**
Sep 08, 2008

**Licensed content publisher**
Elsevier

**Licensed content publication**
Gene

**Licensed content title**
Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons

**Licensed content author**
Jungnam Lee, Richard Cordaux, Kyudong Han, Jianxin Wang, Dale J. Hedges, Ping Liang and Mark A. Batzer

**Licensed content date**
1 April 2007

**Volume number**
390

**Issue number**
1-2

**Pages**
10

**Type of Use**
Thesis / Dissertation

**Portion**
Full article

**Format**
Both print and electronic

**You are an author of the Elsevier article**
Yes

**Are you translating?**
No

**Purchase order number**

**Expected publication date**
Nov 2008

**Elsevier VAT number**
GB 494 6272 12

**Permissions price**
0.00 USD

**Value added tax 0.0%**
0.00 USD

**Total**
0.00 USD

**Terms and Conditions**
Dear Jungnam,

Thanks for your request of August 29, 2008. Under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.5/), which applies to all works we publish, you are free to copy, distribute, transmit, and adapt the work below, so long as the original authors and source are cited:


Best wishes,

Andy Collings
Publications Manager, PLoS Genetics
Website: http://www.plosgenetics.org/
Email Alerts: http://register.plos.org/
UK Tel: +44 (0)1223 463335
VITA

Jungnam Lee is the daughter of Eunsik Lee and Woosim Kim and has a husband, Kyudong Han, and a son, Andrew Han. She graduated with Bachelor of Science and Master of Science degrees in microbiology from Dankook University in Cheonan, Korea, in February of 1999 and February of 2003, respectively. She began her doctoral research in the spring of 2005 in the Department of Biological Sciences at Louisiana State University in Baton Rouge, Louisiana, under the direction of Professor Mark A. Batzer. Mrs. Lee will graduate with the degree of Doctor of Philosophy in December, 2008.