2005

Amplification dynamics of primate retrotransposons

Dale James Hedges
Louisiana State University and Agricultural and Mechanical College, dalehedges@hotmail.com

AMPLIFICATION DYNAMICS OF PRIMATE RETROTRANSPOSONS

A Dissertation

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

in

The Department of Biological Sciences

by
Dale James Hedges
B.A., Duke University, 1998
May 2005

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Mark Batzer, for ensuring that I maintained a sufficient amount of forward momentum in my research, despite my inclination to scribble things on sticky notes and gaze out of the window. I would also like to thank my committee members, Drs. Mohamed Noor, Michael Hellberg, William Hansel, and David Pollock for their advice and encouragement during the course of my research and throughout the process of preparing my dissertation. In addition, I'd like to express my gratitude to my family for their many forms of support, particularly the food and money forms.

I'd like to acknowledge the following collaborators for their contributions to various aspects of this research. For their contributions to chapters two, three, and four: Pauline Callinan and Jinchuan Xing. For their contribution to chapter three: Dr. Abdel-Halim Salem, Dr. Richard Cordaux, Dr. Scott Watkins, Dr. Michael Bamshad, Dr. Randall Garber, and Dr. Lynn Jorde. For their contribution to chapter four: Jerrilyn Walker, Dr. J. G. Shewale, Dr. S. K. Sunhat.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The rapidly increasing amount of sequence data has brought about a new appreciation for the tremendous influence mobile elements have had in shaping eukaryotic genomes. Despite their ubiquity, however, the factors governing the proliferation of mobile elements—or, in some cases, the lack of proliferation—across diverse taxa remain poorly understood. Analysis of *Alu* activity in humans and chimpanzees since their divergence indicates a two-fold increase in human *Alu* activity compared to that of the chimpanzee. This human retrotransposition increase is accompanied by a roughly two-fold higher level of chimpanzee *Alu* diversity. We prepose a model, wherein smaller effective population sizes in humans brought about a shift in host-element dynamic, ultimately leading to increased *Alu* activity in humans. We also survey *Alu*-associated diversity on the human sex chromosomes in order to examine whether *Alu* elements behave similarly to genetic marker systems. Our results suggest that, comparable to other genetic systems, *Alu* elements exhibit reduced diversity on the sex chromosomes. Our data provide no evidence for retrotransposon targeted biology influencing *Alu* insertion frequencies. We go on to synthesize several recent advances in the mobile element field and propose a novel hypothesis concerning how retrotransposon lineages manage to largely lie below the radar of population-level negative selection.

CHAPTER ONE

INTRODUCTION

The latter part of the twentieth century witnessed the field of biology slowly coming to terms with the notion that genomes are more than blueprints for the production of proteins. With the revelation that the chemical structure of DNA could be used to store and propagate information about protein sequences, it had seemed reasonable to envision the genome as a storehouse for protein coding instructions. There might be the occassional regulatory segment amidst these coding stretches but natural selection would surely have sculpted a compact, ruthlessly efficient vehicle to transmit its vital information. Then in the late 1960s researchers began to examine the size of genomes of a wide range of taxa, and the notion of the genome as a tidy repository of protein coding data began to rapidly unravel.

Problems first emerged in the context of what has come to be called the "C-value paradox" (Rosbash et al. 1974; Zuckerkandl 1976). The paradox centered around the observation that seemingly simple eukaryotic organisms were frequently endowed with gargantuan genomes—sometimes orders of magnitude larger than our own. For an example, at a mere 3.5 pg, the human genome is dwarved by that of the red-bellied newt (29.89 pg) (Becak et al. 1970). As the more sophisticated species, we found it relatively straightforward to infer that complexity and genomic size were not necessarily correlated. Although a number of purported solutions to the paradox were proposed over the years, there remained little consensus on the nature of this excess genomic material nor the factors that determined the amount of it present. The identification of introns in 1977 provided clear evidence of a "matrix of noncoding DNA" enveloping expressed coding sequence. McClintock's pioneering work on mobile elements provided yet another clue as to what manner of things might be lurking about the genomic landscape. Yet only with the advent of large scale sequencing did the vastness of the noncoding DNA component of genomes become apparent. And featured prominently within that vastness

were mobile elements. Currently, it is estimated that some 45-50% of the human genome is comprised of repetitive elements (Lander et al. 2001). And that's a conservative number; it's what we can recognize. Compare that to 2-3% coding sequence, and it clear that the human genome is anything but tidy and efficient. In humans, long-dead elements constitute the bulk of these repetitive sequences. These "molecular fossils" were inherited from our early primate and mammalian ancestors (Smit and Riggs 1996). A smaller percentage consists of actively proliferating elements. The situation varies considerably across eukaryotic taxa, particularly in the age distribution of mobile element sequence, but an abundance of repetitive sequence appears to be a common theme (Lander et al. 2001; Waterston et al. 2002). Among vertebrates, one rare exception is the pufferfish (Brenner et al. 1993; Crollius et al. 2000), which sports a compact genome with nominal mobile element activity. Why pufferfish are an exception isn't altogether clear.

With the knowledge that repetetive and other noncoding sequence could inflate genomes independently of their actual coding content, a substantial portion of the paradox surrounding genome size appeared to be resolved. Much of what determines the size of a genome is the level of repetitive element activity in its history. The question that inevitably follows is what determines the level of mobile element activity in a genome? Presenting itself neither as a "paradox," nor as a challenge to our supremacy among earthly organisms, this question proved to be far less captivating than the C-value paradox and was largely ignored for some time. The increasing availability of whole genome sequences is changing that situation somewhat. The sheer mass of mobile elements within most annotated genomes has brought renewed interest in repetetive sequences, and some of this attention has turned to the forces which constrain or promote mobile element activity and diversity (Brookfield 2005; Deceliere et al. 2005; Vieira

and Biemont 2004; Vieira et al. 1999). In addition, the utlitity of retrotransposon insertions as markers for evolutionary and population genetic studies has also brought increased awareness of their ubiquity (Shedlock and Okada 2000). The work which follows focuses primarily on retrotransposons in primates, and in particular on the *Alu* family, one of only three mobile element lineages known to be actively retrotransposing in humans. We examine retrotransposition activity in primates in an attempt to understand the dynamic relationship between mobile elements and their host genomes. In addition, in the spirit of current mobile element research, we explore the use of these elements as tools for other avenues of genetic investigation. While the sphere of these studies remains within the primate order, the implications of the processes involved, particularly those discussed in chapters two and five, may prove applicable across a wide array of taxa.

In chapter two, we make use of newly generated data from the chimpanzee genome sequencing project to address key questions conerning mobile element activity in primates. By comparing large segments of human and chimpanzee sequence, evolutionary recent *Alu* insertions within both species were identified. Using Gorilla as an outgroup for comparison, we were able to examine the relative amount of *Alu* retrotranspositional activity occurring within both species subsequent to their divergence. We assessed both the level of insertion polymorphism as well as the sequence architecture of lineage-specific insertions that were discovered. Using information gleaned from our analyses, we propose a model of the dynamic relationship that exists between retrotransposons and their host populations.

In chapter three, we survey *Alu* associated diversity associated with the human sex chromosomes. Mobile elements represent a novel class of genetic markers with which sex chromosome diversity can be characterized. Comparing these data to similar studies conducted

on autosomes, we ask whether the population-level forces acting on mobile elements on the sex chromosomes are similar to those acting upon other genetic markers such as microsatellites and SNPs. Any differences detected in the behavior of mobile elements compared with other genetic marker systems might suggest as-yet-unknown retrotransposon-targeted biological processes at work in the genome. At the same time, such discrepancies would also cast doubt on the utility of mobile elements for population genetic studies.

Chapter four represents a fortuitous byproduct of our sex chromosome survey work. In the course of our study, we were able to devise a novel method for discriminating between male and female human DNA samples based on the unique features and history of mobile elements. In this chapter, the methodology is described and validated using over 700 human samples.

In the concluding chapter, I review several recent developments in the field of primate retrotransposons. I attempt to integrate the contents of the preceding chapters, as well as other projects with which I have been involved, into our overall understanding of retrotransposon amplification dynamics. By considering the population framework in which mobile elements evolve, a new understanding of the underlying strategy of primate retrotransposon proliferation begins to emerge, one that may be broadly applicable to other classes of mobile elements across diverse taxa.

References

Becak, W., M.L. Becak, G. Schreiber, D. Lavalle, and F.O. Amorim. 1970. Interspecific variability of DNA content in Amphibia. *Experientia* **26:** 204-206.

Brenner, S., G. Elgar, R. Sandford, A. Macrae, B. Venkatesh, and S. Aparicio. 1993. Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366:** 265-268.

Brookfield, J.F. 2005. The ecology of the genome - mobile DNA elements and their hosts. *Nat Rev Genet*.

Crollius, H.R., O. Jaillon, C. Dasilva, C. Ozouf-Costaz, C. Fizames, C. Fischer, L. Bouneau, A. Billault, F. Quetier, W. Saurin et al. 2000. Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis. *Genome Res* **10:** 939-949.

Deceliere, G., S. Charles, and C. Biemont. 2005. The dynamics of transposable elements in structured populations. *Genetics* **169:** 467-474.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Rosbash, M., P.J. Ford, and J.O. Bishop. 1974. Analysis of the C-value paradox by molecular hybridization. *Proc Natl Acad Sci U S A* **71:** 3746-3750.

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

Smit, A.F. and A.D. Riggs. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* **93:** 1443-1448.

Vieira, C. and C. Biemont. 2004. Transposable element dynamics in two sibling species: Drosophila melanogaster and Drosophila simulans. *Genetica* **120:** 115-123.

Vieira, C., D. Lepetit, S. Dumont, and C. Biemont. 1999. Wake up of transposable elements following Drosophila simulans worldwide colonization. *Mol Biol Evol* **16:** 1251-1255.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Zuckerkandl, E. 1976. Gene control in eukaryotes and the c-value paradox "excess" DNA as an impediment to transcription of coding sequences. *J Mol Evol* **9:** 73-104.

CHAPTER TWO

DIFFERENTIAL ALU MOBILIZATION AND POLYMORPHISM AMONG THE HUMAN AND CHIMPANZEE LINEAGES*

Introduction

Alu elements are primate-specific members of the SINE (**S**hort **IN**terspersed **E**lement) family of retroposons. They have enjoyed enormous success over the course of primate evolution and, by conservative estimates, comprise some 10% of the human genome (Lander et al. 2001; Schmid 1996). Due in large part to the human genome project, a wealth of knowledge has been accumulated concerning the underlying biology, retroposition activity, and associated population genetics of *Alu* repeats (Batzer and Deininger 2002; Schmid 1998). The ubiquitous presence of *Alu* sequences within primate genomes has been the cumulative result of a "copy and paste" mechanism, in which an RNA polymerase III generated transcript is reverse-transcribed and integrated into the genome (Burke et al. 1999). In addition to being wholly dependent upon host cellular processes for their transmission through the germline, *Alu* elements also lack the ability to generate the endonuclease and reverse transcriptase necessary for their own retroposition. Instead, they must appropriate the necessary enzymatic machinery from L1, a member of the LINE (**L**ong **IN**terspersed **E**lement) retroposon family (Jurka 1997; Kajikawa and Okada 2002). As a result of this obligatory relationship with their genomic host and other transposable elements, the *Alu* family has been characterized as a "parasite's parasite" (Schmid 2003). Despite their various designations as "junk," "parasites," and "selfish DNA," researchers have been reluctant to dismiss them as entirely self-serving genomic entities. A number of authors have suggested a potential role for *Alu* elements within their host genomes, and recent implications of *Alu* element involvement in alternative splicing, segmental duplications, and DNA repair serve to further fuel these arguments (Bailey et al. 2003; Lev-Maor et al. 2003; Morrish et al. 2002; Salem et al. 2003a). Whether these observations constitute adaptations, exaptations (i.e. they have been commandeered for their current roles, despite not having evolved

for them) (Brosius 1999), or are simply coincidental by-products of their presence in the genome remains a subject of debate. Although a great deal of progress has been made in understanding the mechanism of *Alu* retroposition, much about the factors governing their evolutionary dynamics remain unresolved. To address these and other questions will require a better understanding of the manner in which *Alu* elements have propagated and adapted themselves within nonhuman primate lineages. As the fate of the *Alu* retroposon is necessarily linked to that of its genomic host, major events in primate evolutionary history will likely have left their mark within the *Alu* "fossil record" that is present in the genomes of all living primates.

Given the relatively recent divergence time (5-6 mya) of the human and chimpanzee lineages (Wildman et al. 2003), it would be reasonable to expect *Alu* transpositional activity and the underlying molecular biology associated with retrotransposition in the chimpanzee might closely parallel that of humans. However, initial examination of ~10.6 Mb of sequence from multiple primate genomes by Liu *et al.* revealed a significant deficit in chimpanzee *Alu* insertions as compared to humans and baboons (Liu et al. 2003). Their results suggest that substantial variation in transposition and/or fixation rates may exist among primate lineages. Whether these differences are attributable to underlying differences in biology, stochastic fluctuations in *Alu* proliferation, and/or broader population-genetic processes remains to be determined.

Here we present the first chromosomal-level comparison of *Alu* retroposition dynamics and associated polymorphism between chimpanzees and humans. We have surveyed common chimpanzee chromosome 22 and its human homologue, chromosome 21, for lineage-specific *Alu* sequences and determined the insertion polymorphism associated with each of these insertions. We also examined the nucleotide composition of the observed inserts to better understand

evolutionarily recent *Alu* activity. Finally, we propose a population-based model to account for fluctuations in *Alu* activity within and between primate lineages. In contrast to prior studies of *Alu* diversity, which have largely relied upon inferred "young" *Alu* sequence characteristics to identify loci for investigation, the present comparative approach allows for a more unfiltered appraisal of *Alu* retroposition activity since we last parted ways with our chimpanzee relatives.

Results

Alu Insertion Levels

For the purpose of our comparison, all available sequence from human chromosome 21 and chimpanzee chromosome 22 was first aligned using a local installation of BLAT (Kent 2002), resulting in approximately 32 Mb of aligned sequence that was subsequently screened for evidence of lineage specific *Alu* insertions (see *Methods*)**.** In order to reduce the likelihood of misidentifying deletion events in one lineage as insertions in the other, the identification of *Alu* insertions was restricted to loci exhibiting distinct, individually inserted *Alu* elements (see *Methods*). As a consequence, several questionable insertion/deletions from both the human and chimpanzee were excluded as probable lineage specific deletion events. Of the remaining putative insertions, the possibility of deletion events masquerading as *Alu* insertion events was further excluded by using the gorilla as an outgroup to determine the ancestral state of the locus. In all, 46 lineage-specific *Alu* insertions were identified in chimpanzee chromosome 22 while 101 lineage-specific elements were identified in human chromosome 21, demonstrating a 2.2X increase in the number of detectable human insertions (Table 1). These results are in excellent agreement with Liu *et al*, who found 11 chimpanzee and 23 human insertions (2.1X) in their ~10.6Mb human-chimp comparison (Liu et al. 2003); as their sequence data was derived from multiple genomic locations, this correspondence suggests that our data are reflective of the

Table 2.1 – Lineage-specific *Alu* insertions

Lineage-Specific *Alu* Insertions			
	Human	*Human/Chimp Ratio*	*Chimpanzee*
Observed Inserted Total	**101**	***2.20***	**46**
PCR Tested	**78**	**---**	**43**
Fixed Present	**63**	**---**	**26**
Observed Polymorphic	**16**	**---**	**18**
Observed Polymorphic Fraction	**0.21**	***.50***	**0.41**
Adjusted Polymorphic [a]	**31 -- 33**	**---**	**35 -- 37**
Adjusted Polymorphic Fraction	**0.33 -- 0.34**	***0.56 -- 0.60***	**0.57 -- 0.59**
Adjusted Inserted Total	**116 -- 118**	***1.84 -- 1.93***	**61 -- 63**

[a] Adjusted polymorphic fraction was calculated based upon simulation of the frequency of polymorphic *Alu* elements observed in a given genome by sampling alleles from a uniform frequency distribution (see Methods). Ranges indicated were generated based on 95% confidence intervals derived by simulation.

genome as a whole and not endemic to the particular chromosomes surveyed.

While the cross-species comparison allowed us to classify loci as putatively specific to either the human or chimpanzee lineage, there remained the possibility that *(a)* some of the insertions were shared polymorphisms in which only one lineage's sequenced individual possessed the insertion *(b)* there were "fixed present" insertions in one species that remained polymorphic in the other. Extensive surveys of hundreds of human *AluYa5, AluYb8 and AluYc1*

insertions in which representative common chimpanzee and bonobo (*Pan paniscus)* samples were analyzed in nonhuman primate controls have demonstrated that the sharing of *Alu* polymorphism between species for these young *Alu* subfamilies would be negligible (Carroll et al. 2001; Roy-Engel et al. 2002a; Roy-Engel et al. 2001). In addition, theoretical estimates of the rate of decay of shared polymorphism (Clark 1997), as well as empirical nucleotide data from human, chimpanzee, and gorilla sequences (Hacia et al. 1999), indicate that the number of shared polymorphisms expected given the number of loci involved in our study would be at most one, and therefore this effect would not appreciably alter our results. However, to address the possibility that some unknown property of *Alu* insertions might cause them to deviate substantially from these expectations, we evaluated all non-Ya5/Yb8/Yc1 human insertions (most likely to be shared) and 25 chimpanzee-specific insertions in population panels (80 humans and 12 common chimpanzees) from the opposite species and found no instances of shared *Alu* polymorphism. In addition, these results also give no indication that an appreciable number of elements fixed in human populations remain polymorphic in the chimpanzee. This is further evidenced by the fact that surveys of human *Alu* elements found that shared insertion in chimpanzee was extremely rare (Carroll et al. 2001; Roy-Engel et al. 2001). Were there a significant number of fixed human elements remaining polymorphic in the chimpanzee, insertion status of the chimpanzee reference samples in these large surveys would have occurred with higher frequency.

To aid in distinguishing whether the observed *Alu* insertion disparity represents a decrease in the chimpanzee *Alu* retroposition rate or an increase in the human retroposition rate within a local phylogenetic context (human, chimpanzee, gorilla, orangutan), we examined a 1.5Mb segment of homologous 7q31 sequence available in all three species for *Alu* insertions

specific to a given species. The results of this comparison indicate a gorilla *Alu* transposition/fixation level that is near that of *Pan troglodytes*, with four *Alu* inserts in *Gorilla gorilla* compared to three in *Pan troglodytes* and eight in humans. The small amount of gorilla sequence available for comparison resulted in too few *Alu* insertions to yield significant results ($p \sim .25$). However, the trend exhibited between humans and chimpanzees in this region (8:3) echoes that of our larger chromosome 21 survey, leading us to believe that the gorilla insertion numbers are also representative of its genome. Although more extensive sequence comparisons using gorillas and orangutans will be required before definitive conclusions can be drawn, our data favor a human-specific increase in *Alu* retroposition activity within the local phylogenetic context. Examination of the subfamily composition of human and chimpanzee elements (see below) lends further support to this interpretation.

Distribution of Insertions

Qualitatively, the evolutionarily recent *Alu* insertions were found distributed relatively evenly throughout the chimpanzee and human chromosomes, with expected lower densities near telomeric and centromeric regions primarily due to unsequenced heterochromatic regions. *Alu* density has previously been established to be strongly correlated with both GC-content and gene density (Lander et al. 2001; Schmid 1996). Chromosome 21 exhibits a 42% GC content, compared with 48% on chromosome 22 and 49% on chromosome 19, which contains both the highest GC content and highest gene density (Lander et al. 2001). Correspondingly, overall *Alu* density is highest on chromosome 19, followed by chromosome 22 (Chen et al. 2002). Chromosome 21 is relatively gene poor with an average density of approximately 7 genes per Mb compared to the 11.1 per Mb genomic average (Hattori et al. 2000). However, recent genomic surveys of young AluYb8 and AluYa5 subfamilies demonstrate no significant deficit of

young subfamily insertions on chromosome 21 ((Carter et al. 2004); unpublished data). This may partially be attributable to the fact that the *Alu* GC and genic distribution bias appears to be more pronounced for evolutionarily older insertions (Jurka et al. 2004; Lander et al. 2001). As a result of the relatively small numbers of recently inserted *Alu* elements in our survey, larger genome-wide comparisons of young *Alu* inserts will be necessary for adequately detecting any changes in distribution between species. However, we do note here that, in agreement with previous studies of total *Alu* content (Chen et al. 2002; Lander et al. 2001), human and chimpanzee specific insertions on chromosomes 21/22 had a tendency to insert in GC-rich genic regions, with over 20% of the insertions in our survey being located within the introns of known genes, and an even higher frequency (>50%) when predicted genes are considered. Based on estimates of known and predicted gene number and average chromosome 21 gene sizes, we estimate that these gene categories span approximately 20% and 8% of the sequenced region of the chromosome respectively. In addition, *DSCAM*, an alternatively spliced gene involved in neural development (Yamakawa et al. 1998), demonstrated a total of five human-specific insertions. This may not in-itself be remarkable, as *DSCAM* spans 840kb, making it a rather large target for insertion. However, all five inserts are in the antisense orientation relative to gene transcription, a feature that has been linked to alternative splicing (Lev-Maor et al. 2003). Given intronic *Alu* orientation frequencies of 0.47 (sense) and 0.53 (antisense) calculated from a survey of 179 AluYb8 and AluYa5 gene insertions, this configuration of antisense *Alu* elements deviates significantly from expectation ($p < .05$).

Anomalous Loci

In addition to the lineage-specific insertions found in our study, one element, designated *CS12*, was determined to be exclusive to gorilla and chimpanzee genomes and not present in

human, implying a relationship contrary to the orthodox phylogeny of ((HC),(G)). Such discrepancies have been reported elsewhere (Salem et al. 2003b) and most likely represent lineage sorting of an ancestral polymorphism present in the common ancestor of humans, chimpanzee, and gorilla. The existence of such sorting events serve to highlight the relatively short period of time, evolutionarily speaking, during which these three lineages emerged. For the purposes of this study, however, putative lineage sorting events were excluded from further analysis, as they could not be classified as lineage-specific for either humans or chimpanzee.

Another locus, *HS6*, exhibited phylogenetic inconsistencies that were less readily explained. PCR analysis of the locus showed insertions in orangutan, gorilla, and human to the exclusion of chimpanzee. The maintenance of a polymorphism over this period of time--approximately 6 myrs from the branching of orangutan to the divergence of humans and chimpanzees--would be unlikely, prompting us to consider the possibility of an *Alu* excision at the chimpanzee locus. For further examination, we sequenced the orthologous loci in *Gorilla gorilla*, *Pan paniscus*, and *Pongo pygmaeus* (Figure 1). The *HS6* insertions in human, gorilla, and orangutan contained direct repeats that were identical in both sequence and length, strongly indicating identical by descent insertions. Unexpectedly, the chimpanzee locus was a perfect pre-integration site, consisting of only one copy of the direct repeat (Figure 1). In the only previously reported instance where an *Alu* element appeared to be excised from a genome, remnants of the *Alu* insertion remained in the sequence (Edwards and Gibbs 1992). As the precise excision of an *Alu* insertion appeared to be a remote possibility, we began to explore other potential explanations for our observations. One such possibility is that a segmental duplication in a great ape common ancestor produced a pair of paralogous loci, only one of which received an *Alu* insertion. This paralogous loci, which would itself be polymorphic and

subject to lineage sorting, could have resolved itself into the observed phylogenetic situation. Our inability to detect evidence through PCR for more than one uninserted locus among the tested species indicates that this long-term maintenance of a duplication polymorphism is no more probable than that of a long-lived *Alu* insertion polymorphism. However, when considered together, these alternative pathways to the same observed state makes the observed insertion states somewhat more likely. On further examination of the HS6 locus, we discovered two immune-related genes, *CXADR* and *CHODL*, within 1Mb of HS6. It is conceivable that balancing selection acting at these nearby loci served to maintain the HS6 polymorphism, ultimately resulting in the unusual phylogenetic distribution of this *Alu* insertion. Additional investigation of the genes at this locus will be required to verify this hypothesis.

Figure 2.1

Figure 2.1 Reconstructed *Alu* HS6 insertion sites in human and nonhuman primates. Shaded area indicates direct repeat region. Chimpanzee site demonstrates no evidence for an extracted insertio

Subfamily Composition

Human *Alu* elements inserted on chromosome 21 were classified according to subfamily structure as previously reported (Batzer et al. 1996) (Figure 2). All human-specific insertions were members of the *Alu*Y subfamily or one of its derivatives. Of these, the *Alu*Ya5 and *Alu*Yb8 subfamily comprised the largest percentage, comprising 25% and 38% of the loci respectively. For those elements categorized as members of *Alu*Y, their sequences were screened against the human genome database to determine if they belonged to previously uncharacterized

subfamilies. Several of these elements appeared to be members of small (10-100 member) *Alu* subfamilies that had previously remained unidentified. Comparative analysis of additional chromosomes will likely reveal additional small subfamily structure that remained undetected by previous molecular and computational methods.

At present, very little is known about the subfamily structure of *Alu* elements within the chimpanzee genome. Multiple alignments of all observed *Pan troglodytes* chromosome 22 lineage-specific inserts uncovered two candidates for active subfamilies. The first group, consisting of 27 elements, has a consensus sequence identical to that of *Alu*Yc1 in humans. Whether this subfamily is identical by descent or state to its human counterpart is unclear, as *Alu*Yc1 differs from the canonical *Alu*Y sequence by a single **G→A** nucleotide substitution. Human *Alu*Yc1 insertions exhibit a relatively young (1-3 myr) average age (Garber et al. *in press*). Our estimates of the chimpanzee *Alu*Yc1 family place it between 1.2-2.6 myrs old. While this is suggestive of an independent parallel mutation, the human *Alu*Yc1 elements may have remained relatively dormant in the human genome until some time subsequent to *Pan-Homo* split. To better localize the chimpanzee *Alu*Yc1 activity in time, we examined the insertion status of 18 *Pan troglodytes* specific *Alu*Yc1-like elements in a representative bonobo (*Pan paniscus*), estimated to have diverged from *Pan troglodytes* approximately 1.8 mya (Yu et al. 2003). Eleven elements were present in the *Pan troglodytes* population but absent from our *Pan paniscus* individual and 7 elements were present in both species, indicating that the chimpanzee *Alu*Yc1-like subfamily had began amplifying prior to the *Pan troglodytes-Pan paniscus* divergence. This places a lower bound on the chimpanzee *Alu*Yc1 family age of approximately two million years, not ruling out the possibility that these subfamilies are of common descent.

The second group of four elements (designated YCV1) were distinguished by five diagnostic mutations from the *Alu*Y consensus. Screening of the human genome database revealed several matches within humans, indicating that this subfamily was not restricted to the chimpanzee lineage and has been amplifying, albeit slowly, since before the human-chimpanzee split. Here, there is little possibility of a parallel forward mutation event, as YCV1 is distinguished by five mutations.

Alu Insertion Polymorphism

To assess the diversity of individual lineage-specific *Alu* insertions on human chromosome 21, 78 *Alu* elements that were amenable to PCR were amplified on a panel of 80 human individuals from four geographically diverse populations (African-American, Asian, German Caucasian, and South American). Among the four represented populations, 16 of 78 (20.51%) elements demonstrated polymorphism in our panel. Allele frequencies of all polymorphisms, as well as primers used in this study, are available at our website (http://batzerlab.lsu.edu). Forty-three chimpanzee-specific insertions were evaluated on our chimpanzee panel of twelve unrelated *Pan troglodytes*. Due to the small size of our *Pan troglodytes* sample, we assessed its adequacy in evaluating loci for polymorphism (see *Methods*). Assuming a uniform distribution of *Alu* allele frequencies, we estimated that our 12 individual (24 chromosome) sample would capture approximately 88-93% of the polymorphism present at the examined loci. In all, 18 of 43 (41.86%) elements exhibited polymorphism in our chimpanzee panel. The 2.0 ratio of human to chimpanzee polymorphism fraction is somewhat higher than the 1.5 ratio of a recent nucleotide heterozygosity study (Yu et al. 2003). If adjustments for unequal polymorphism levels are made, however, the values become closer (see *Discussion*).

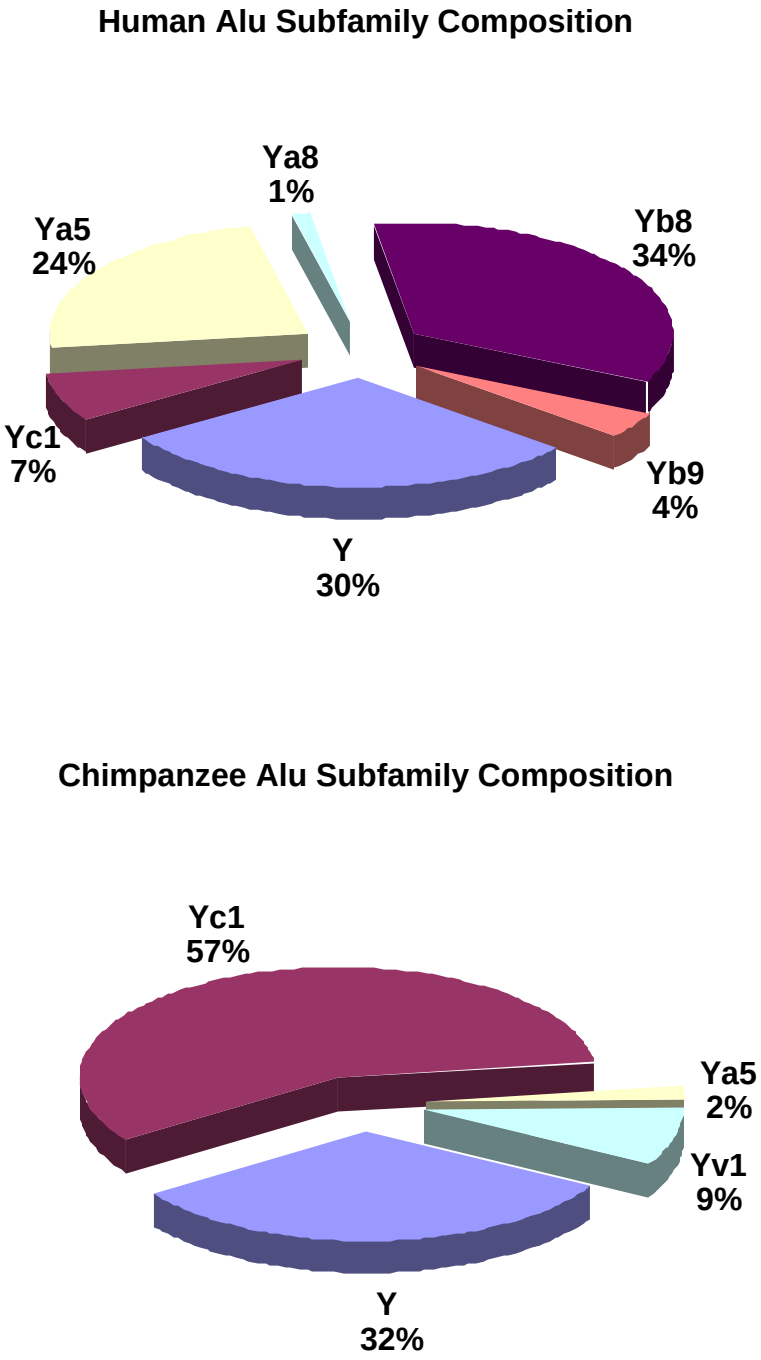


Figure 2.2 Subfamily composition of lineage-specific *Alu* insertions in humans and common chimpanzee.

Discussion

Alu Transposition Levels and Subfamily Structure

Our results suggest that an elevation in human *Alu* retroposition activity, largely mediated by two human *Alu* subfamilies (*Alu*Ya5 and *Alu*Yb8), occurred some time subsequent to the divergence of the human and chimpanzee lineages. The most current estimates for the ages of these subfamilies place them amplifying between 2.5-3.5 mya (Carroll et al. 2001). A survey of a 4Mb X-Y translocation event (Schwartz et al. 1998), which has previously been dated to approximately 3.5-4 mya (Sargent et al. 2001) suggests no appreciable retroposition activity of *Alu*Ya5 and *Alu*Yb8 families prior to that time period. This is indicated by the absence of *Alu*Yb8 and *Alu*Ya5 elements duplicated at the time of the translocation event. These observations place the onset of significant *Alu*Ya5 and *Alu*Yb8 mobilization subsequent to the divergence of the human and chimpanzee lineages, indicating that a contraction in population size during or immediately following speciation does not account for the chimpanzee-human *Alu* disparity.

The question arises as to whether or not the *Alu*Ya5 and *Alu*Yb8 subfamily expansions were simultaneous or distinct events. While current age estimates date them to roughly the same period, polymorphism levels of *Alu*Yb8 (20%) and *Alu*Ya5 (25%) suggest a somewhat younger overall age for the *Alu*Ya5 subfamily, as more of its members remain unfixed in the population (Carroll et al. 2001). However, the polymorphism fraction may only serve to indicate that the bulk of *Alu*Ya5 insertions are distributed closer to the present than that of *Alu*Yb8, and is not necessarily reflective of the initial appearance date of the subfamily.

An additional factor with the potential to influence the estimated ratio of *Alu* insertion numbers in species is the existence of unequal diversity levels within humans and chimpanzees

for *Alu* insertions. Using the observed *Alu* diversity in chimpanzee and human, we estimated the extent to which this effect may have skewed our results (see *Methods*). Our estimates suggest that in 95% of cases 42-58% of the polymorphic *Alu* insertion loci would be missed by sequencing a single representative human genome or chimpanzee genome. When we adjust insertion numbers within both lineages for these missed *Alu* loci, our estimate of the human-chimpanzee insertion ratio is 1.84 - 1.93 (Table 1).

The paucity of evolutionarily recent *Alu* insertions observed on the *Pan troglodytes* chromosome 22 restricts our ability to completely capture the chimpanzee *Alu* substructure. However, assuming that young *Alu* subfamily dispersal in humans is distributed proportional to chromosome size, the chance of missing a major young *Alu* family (>300 elements) in our chimpanzee chromosome 22 survey would be remote (less than 5%). Our data indicate that the major lineages that constitute the bulk of recent human activity, *Alu*Ya5 and *Alu*Yb8, are only present at negligible levels in *Pan troglodytes*. A solitary *Alu*Ya5 element was found on chimpanzee chromosome 22, and although Genbank database queries indicate that a small number of authentic *Alu*Yb8 chimpanzee insertions are present in the *Pan troglodytes* genome, quantitative PCR results suggest the their copy number is negligible compared to humans (Walker et al. 2003). The *Alu*Yc1-type subfamily appears to dominate the *Pan* lineage (Figure 2), but we can not conclusively say if it is identical by descent to the subfamily that is found in humans. If it is indeed the same family, it would be curious that, given their estimated ages (1-3myr), the source sequence would have remained relatively dormant in both lineages only to become active, independently, at a later time. Alternatively, the independent, parallel success of these source mutations may suggest a selective advantage for the **G→A** consensus substitution,

or it could simply be a base position where such change is tolerated in the *Alu* source or "master" genes.

While several of the *Alu* polymorphic loci in chimpanzee contained sequence characteristics that were present in only a single copy on chromosome 22, these insertions will serve as excellent starting points to search for further chimpanzee *Alu* family substructure, as they likely represent chromosome 22 representatives of smaller, active *Alu* subfamilies analogous to those recovered in the human sequence.

The presence of *Alu*Yb8 and *Alu*Ya5 members in small copy numbers within the chimpanzee and gorilla genomes (Leeflang et al. 1993) demonstrates that the sequence evolution of successful subfamilies begins well before their peak activity. These subfamilies appear to undergo a lengthy period during which low baseline mobilization occurs. A chance insertion within a suitable genomic context, however, could initiate a burst of activity from the locus within a given host lineage. In conjunction with L1 enzyme availability and population genetic factors (see below), such fortuitous insertions would initiate the expansion phase of the *Alu* subfamily.

Alu Insertion Polymorphism

Our *Alu* insertion diversity data demonstrate two times higher *Alu* polymorphism in chimpanzee compared to humans. If we adjust the estimates of polymorphic *Alu* loci by accounting for the insertion polymorphisms that were predicted to be missed in chimpanzee and human sequences (see *Methods*), our ratio of chimpanzee to human *Alu* polymorphism decreases to 1.67 - 1.78. A number of previous studies, making use of multiple genetic systems, have attempted to assess the level of genetic diversity of chimpanzees relative to that of humans. Mitochondrial and nuclear genome surveys have generated seemingly conflicting depictions of

chimpanzee diversity. Mitochondrial diversity has been estimated to be as much as 10 times higher among chimpanzees than humans (Rogers and Jorde 1995). Nuclear nucleotide diversity estimates, in contrast, have yielded chimpanzee heterozygosities that are lower than human levels for protein-coding loci (King and Wilson 1975; Satta 2001). Surveys of additional coding and noncoding loci have reported nucleotide heterozygosity estimates 3-4X higher in chimpanzee than humans (Deinard and Kidd 1999; Kaessmann et al. 1999). Our range of 1.67 - 1.78 times higher common chimpanzee diversity best corresponds to that of Yu *et al.*, who estimated nucleotide diversity in common chimpanzee at 1.5 times higher than that of human, with a lower value for bonobo (Yu et al. 2003).

The previously reported disparity of heterozygosity values exhibited by different genetic systems (mitochondrial, microsatellite, nuclear SNPs) can potentially be explained by a population bottleneck in humans which had a more severe effect on mitochondrial diversity due to its smaller (1/4 autosomal) effective population size (Yu et al. 2003). The existence of a bottleneck in human evolutionary history has been suggested by many studies (Chen and Li 2001; Harpending et al. 1998; Lonjou et al. 2003). While our chromosome 21/22 data are consistent with this scenario, we can not exclude other possibilities, such as selective sweeps reducing mitochondrial diversity.

If the correspondence between *Alu* insertion polymorphism ratios and the nucleotide diversity ratios between humans and chimpanzees is not simply coincidental, it would appear that the effective population size is the dominant influence determining the fraction of *Alu* insertion polymorphisms in these genomes. That is, despite markedly different subfamily composition and retroposition histories between the two lineages, *Alu* insertion polymorphism generally parallels nucleotide polymorphism in behavior. This is a somewhat surprising result,

given that fluctuations in *Alu* activity over time could result in one lineage having an excess or deficit of younger, polymorphic *Alu* insertions relative to the other lineage, largely independent of effective population size. However, this situation could conceivably be explained if the more dramatic changes in *Alu* insertion rates occurred in more distant evolutionary history and have had little influence on current polymorphism levels. In this scenario, relatively uniform insertion rates within individual lineages over recent evolution history has resulted in effective population size being the dominant determinant of polymorphism levels. Further resolution of the insertion dates of human and chimpanzee *Alu* elements will be necessary to clarify this issue.

A Population-based Model for Fluctuations in *Alu* Mobilization

Under standard neutral or "nearly neutral" population genetics theory, three scenarios could conceivably account for the relative increase in fixed *Alu* insertions within humans as compared to chimpanzees. First, a smaller long-term effective population size in the human lineage could have resulted in the fixation of otherwise slightly deleterious *Alu* insertions at a higher rate in humans. Under this scenario, the roughly two-fold increase in observed human insertions would need to be accounted for by deleterious elements. While this possibility can not presently be excluded, the fixation of hundreds of deleterious *Alu* insertion loci would no doubt represent a considerable burden to a population. An explanation that avoided such a genetic calamity would appear to be more parsimonious. A second scenario would be that the existing *Alu* polymorphism which was present at the time of human-chimpanzee speciation was funneled through a *Homo* lineage bottleneck, resulting in an increased fixation of *Alu* elements within humans. In this situation, the differences in *Alu* insertion number would be attributable to many more of these ancestral polymorphisms fixing in the human lineage than the chimpanzee. This scenario is unlikely as well, however, as the sequence structure of *Alu* insertions of humans,

comprised largely of two young subfamilies, differs considerably from that of chimpanzee (Figure 2). This suggests that they were not derived from a common pool of *Alu* insertions that were polymorphic at the time of speciation. In addition, the major retroposition activity within the *Alu*Ya5 and *Alu*Yb8 subfamilies can be reasonably dated by independent lines of evidence to a period subsequent to the human-chimpanzee speciation (see *Results*). The third possibility, which we favor, is an increase in the *Alu* retroposition rate itself. This would be analogous to an increased nucleotide mutation rate within a given lineage. However, in the case of retroposition, there is an added layer of complexity in the interaction between insertion rates, fixation rates, and population size that must be addressed.

The population dynamics of *Alu* elements within their hosts can account for much of the insertion variance observed within and between primate lineages. The basic components of our model are as follows. **1)** Variation in source *Alu*-producing loci exist in the population **2)** Stochastic sampling of these source variants either at speciation or during bottleneck events alters the population-level *Alu* transposition activity (*insertions per birth*) **3)** While the previous two conditions are sufficient to produce variation within and between lineages, smaller effective population sizes will both increase the sampling variance of *Alu* sources and reduce a given population's ability to select against deleterious source loci. This may result in a substantially increased population-level *Alu* activity (*insertions per birth*) brought about by environmental insults, speciation events, etc.

Aside from their observed GC-rich distribution bias, there has been no evidence indicating that *Alu* insertions behave appreciably different than nucleotide polymorphisms as genetic markers once inserted in the genome (Bamshad et al. 2003; Perna et al. 1992; Stoneking et al. 1997; Watkins et al. 2001; Watkins et al. 2003). As such, the behavior of *Alu* elements

should be consistent with other neutral or "nearly neutral" characters. The probability of a given *Alu* insertion reaching fixation in a population is therefore contingent upon its initial frequency in the population, $1/2N$, where N is the population size (Kimura 1983). In the context of *Alu* retrotransposition, however, not all of the further assumptions of neutral theory hold. While the number of novel nucleotide mutations arising each generation in a population is dictated by the size of the population (i.e. total number of mutable sites) and the frequency of mutations arising each generation, the number of novel *Alu* insertions has a more complex relationship with population size. As the majority of new *Alu* copies are known to arise from a select number of 'master' or source loci, these loci themselves will be subject to allelic variation in both transpositional competency and/or insertion status. Evidence for such allelic variation in retrotransposition capability has been observed in members of the L1 subfamily (Lutz et al. 2003) and within *Alu* may be attributable to variation at PolIII promoter efficiency, variation in target-primed reverse transcription, oligo dA tail instability (Roy-Engel et al. 2002b), and insertion status polymorphism for the source locus itself. Additional evidence from L1 sequence transduction events demonstrate that retroposon source sequences can produce "offspring" that proceed to fixation while the parent sequences are ultimately lost (Boissinot et al. 2001). As a consequence of this source allele variation, a reduction in overall human population size may occur while the number of novel *Alu* insertions *per individual birth* actually *increases* due to the stochastic effects of sampling the active source variants (Figure 3). In effect, unlike nucleotide substitution rates, the equivalent of the *Alu* substitution rate will itself fluctuate along with population size. The intensity of these fluctuations will increase as the population size becomes smaller. Simultaneously, a reduced effective population size is less capable of selecting against detrimental source variants as the population size grows smaller. This effect is exacerbated

because the *Alu* source is effectively "screened" by its indirect relationship to the deleterious insertion loci it generates. As a consequence, transposition may run rampant when the population size is no longer large enough to effectively select against *Alu* "hyperactivity." Within a window of selective pressure, deleterious insertions would still be effectively removed from the genome,

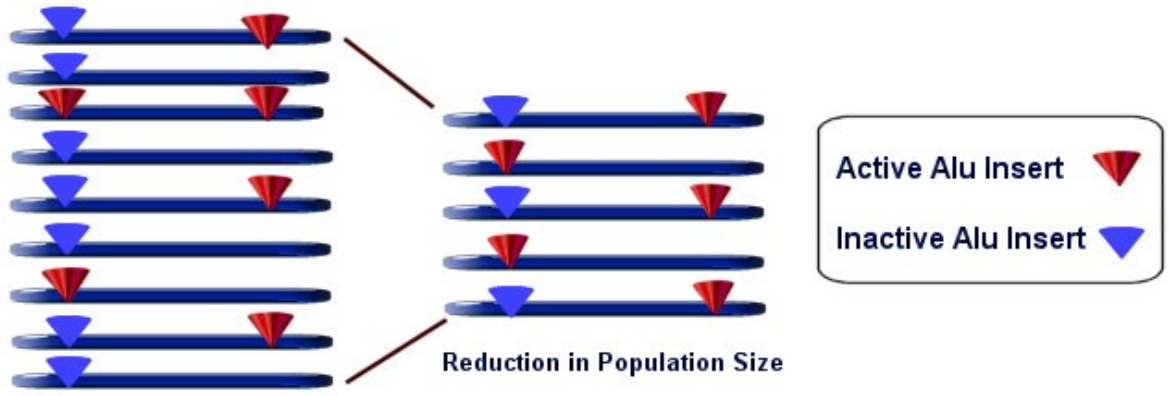


Figure 2.3 Variation in the insertion status and retroposition capability of *Alu* elements at two loci. Reduction in population size leads to variation in the number of active elements.

but the source or sources generating the deleterious insertions become(s) essentially neutral (i.e. having a selective coefficient << 1/2N).

An attractive feature of this explanation is that it does not necessitate the presence of a large number of fixed deleterious loci to account for differential lineage *Alu* insertion counts. Furthermore, it does not require the invocation of any novel biology to account for changes in the relative number of insertions per generation. One prediction of the model is that the onset of increased *Alu* transposition activity would tend to be coincident with population size decreases and, as a consequence, *Alu* transposition rates may change *rapidly* within and between lineages. By developing better analytical tools to estimate the ages of individual *Alu* insertions, it may be

possible to localize transposition events in time and estimate the rate at which *Alu* transposition activity fluctuates. A further prediction is that isolated, inbred populations would be at an increased risk for *Alu* "hyperactivity", as they would experience a decreased capacity to select against active source loci. Genomic display, ATLAS, and similar methodologies that have the potential to exhaustively examine retroposon insertions within individual genomes will allow testing in extant populations for evidence of this effect.

Materials and Methods

DNA Samples

Cell lines used to isolate DNA samples were as follows: A chimpanzee diversity panel of twelve *Pan troglodytes* of unknown geographic origin was obtained from the SouthWest foundation for Biomedical Research, gorilla (*Gorilla gorilla*), lowland gorilla Coriell AG05253A, owl monkey (*Aotus trivirgatus*), ATCCCRL1556, and pygmy chimpanzee (*Pan paniscus), Coriell AG05253A*. Human DNA from South American populations was purchased as part of the Human Variation Panel available from the Coriell Institute for Medical Research. DNA samples from the European, African American and Asian population groups were isolated from peripheral blood lymphocytes available from previous studies.

Human-Chimpanzee Comparison

DNA sequences for chromosome 22 (approximately 43Mb, including overlapping sequence) were obtained from The Chimpanzee Chromosome 22 Sequencing Consortium (http://chimp22pub.gsc.riken.go.jp). Sequence for human chromosome 21 was obtained from UCSC June 2003 assembly data. Human chromosome 21 and chimpanzee chromosome 22 alignments were generated using a local installation of BLAT (Blast-Like Alignment Search Tool) (Kent 2002), resulting in approximately 32 Mb of aligned sequence out of an estimated

33.8 Mb total chromosome 21 sequence (Hattori et al. 2000). BLAT results were subsequently screened using a Perl script for all insertions/deletions of sizes 100-1000bp. These sequences, along with 200bp of flanking sequence, were extracted for further examination. In addition, a separate manual BLAT screen of the human genome database (using UCSC web interface) using the chimpanzee chromosome 22 sequence was conducted to assess the accuracy of our script-generated results. Indel sequences were screened using a local installation of RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) to determine their repetitive element content. Subsequent sequence alignments were done with MEGALIGN program, part of the DNASTAR package. Redundant repeat insertions resulting from overlapping BLAT fragments were excluded by verifying unique flanking sequence. An additional ~1.5 Mb of human, chimpanzee, and gorilla homologous sequence from chromosome 7 was obtained from the NIH Intramural Sequencing Center (www.nisc.nih.gov). Sequences were aligned with BLAT and/or MEGALIGN to identify species-specific indels and RepeatMasker was used to determine their repetitive element content.

All putative *Alu* insertions were manually verified as authentic by determining if the insertions met established criteria for evolutionarily recent *Alu* insertions. Authentic *Alu* insertions were required to have only 5' truncations, as 3' truncations have not been observed to occur upon insertion. Any "partial" *Alu* indels in which a fragment of the *Alu* is already present at the locus prior to the indel event were excluded, as these are more characteristic of partial deletions of elements. *Alus* that were contained within larger insertion/deletion events were also excluded, as these did not represent authentic *Alu* transposition events. To further resolve ambiguities, all putative insertions were amplified from the gorilla genome to determine the ancestral state of the insertion.

Statistical Methods

Estimating the Number of Detected Polymorphic Alu Insertions

Estimations of the number of polymorphic insertions that would be detected in a single sequenced genome were conducted by generating 1000 samples of a genome (set detectable of alleles) from a uniform distribution of *Alu* insertion frequencies. This choice of distribution was based on observations of the allele frequencies of human *Alu* inserts (Carroll et al. 2001; Roy-Engel et al. 2001), and reasoning that the higher long-term effective population size of chimpanzee would result in an even more uniform (flat) distribution of *Alu* insertion frequencies due to the lack of recent bottlenecks and/or expansions (Harpending et al. 1998). In our simulation, the probability of discovering a given allele was proportional to its frequency in the population. The mean fraction of detections was 0.5, with a variance inversely proportional to the number of actual polymorphic loci. Our 1000 replicates using 100 loci yielded a standard deviation of 4%, which was used to calculate a 95% confidence interval for unsampled polymorphisms of 42% - 58%.

Detection of Polymorphism

The probability of detecting a *Alu* insertion polymorphism at a given locus is contingent upon its minor allele frequency **1 - [(1 - q)^N]**, where **q** is the minor allele frequency and **N** is the number of sampled chromosomes. Consequently, the number of detectable *Alu* variants will be subject to the distribution of allele frequencies in the population. If we assume this is roughly uniform, then summing over ***i*** minor allele frequencies $\sum$**[[1 - [(1 - q$_i$)^N]]** yields the fraction of polymorphic sequences detected**.** By simulating 1000 trial detections of uniformly distributed minor alleles, we estimate that 95% of the time our human panel of 80 individuals (160 chromosomes) would detect 97.3 - 99.7% and our chimpanzee panel of 12 individuals (24

chromosomes) would detect 89 - 93% of the polymorphism at PCR evaluated loci. Within the observed polymorphism, there should be a skew towards higher frequency alleles, as these are more likely to appear in a given sequenced genome. Since we restricted our analysis to polymorphic/fixed status this bias should not affect our conclusions.

PCR Analysis

Oligonucleotide primers for the PCR amplification of each *Alu* element were designed using the 700-1200 base pair flanking unique sequence fragments and Primer3 software (Whitehead Institute for Biomedical Research, Cambridge, MA, USA) (http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). The sequences of the oligonucleotide primers, annealing temperatures, PCR product sizes and chromosomal locations for all *Alu* elements in this study can be found on our website (http://batzerlab.lsu.edu). PCR amplification was performed in 25 µl reactions using 10-50ng of target DNA, 200nM of each oligonucleotide primer, 200µM dNTP's in 50 mM KCl, 1.5 mM $MgCl_2$, 10mM Tris-HCl (pH 8.4) and 1 unit *Taq* DNA polymerase. Each sample was subjected to an initial denaturation step of 94° C for 150 seconds, followed by 32 cycles of PCR at one minute of denaturation at 94° C, one minute at the annealing temperature, one minute of extension at 72° C, followed by a final extension step at 72° C for ten minutes. The resulting products were then evaluated for polymorphism on EtBr-stained 2% agarose gels and visualized with UV lighting.

DNA Sequencing

DNA sequencing was performed on gel purified PCR products that had been cloned using the TOPO TA cloning vector (Invitrogen) using chain termination sequencing on an Applied Biosystems 3100 automated DNA sequencer. All sequences generated in this study are available in the Genbank database (Accession #s AY569161--AY569170).

References

Bailey, J.A., G. Liu, and E.E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73:** 823-834.

Bamshad, M.J., S. Wooding, W.S. Watkins, C.T. Ostler, M.A. Batzer, and L.B. Jorde. 2003. Human population genetic structure and inference of group membership. *Am J Hum Genet* **72:** 578-589.

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Batzer, M.A., P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz, and E. Zuckerkandl. 1996. Standardized nomenclature for Alu repeats. *J Mol Evol* **42:** 3-6.

Boissinot, S., A. Entezam, and A.V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* **18:** 926-935.

Brosius, J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. *Gene* **238:** 115-134.

Burke, W.D., H.S. Malik, J.P. Jones, and T.H. Eickbush. 1999. The domain structure and retrotransposition mechanism of R2 elements are conserved throughout arthropods. *Mol Biol Evol* **16:** 502-511.

Carroll, M.L., A.M. Roy-Engel, S.V. Nguyen, A.H. Salem, E. Vogel, B. Vincent, J. Myers, Z. Ahmad, L. Nguyen, M. Sammarco et al. 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* **311:** 17-40.

Carter, A.B., A.-H. Salem, D.J. Hedges, C. Nguyen Keegan, B. Kimball, J.A. Walker, W.S. Watkins, L.B. Jorder, and M.A. Batzer. 2004. Genome wide analysis of the human Yb lineage. *Human Genomics* **1:** 167-168.

Chen, C., A.J. Gentles, J. Jurka, and S. Karlin. 2002. Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *Proc Natl Acad Sci U S A* **99:** 2930-2935.

Chen, F.C. and W.H. Li. 2001. Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am J Hum Genet* **68:** 444-456.

Clark, A.G. 1997. Neutral behavior of shared polymorphism. *Proc Natl Acad Sci U S A* **94:** 7730-7734.

Deinard, A. and K. Kidd. 1999. Evolution of a HOXB6 intergenic region within the great apes and humans. *J Hum Evol* **36:** 687-703.

Edwards, M.C. and R.A. Gibbs. 1992. A human dimorphism resulting from loss of an Alu. *Genomics* **14:** 590-597.

Garber, R.K., D.J. Hedges, S.W. Herke, N.W. Hazard, and M.A. Batzer. in press. The Alu Yc1 subfamily: sorting the wheat from the chaff. *Cytogenetics and Genome Research*.

Hacia, J.G., J.B. Fan, O. Ryder, L. Jin, K. Edgemon, G. Ghandour, R.A. Mayer, B. Sun, L. Hsie, C.M. Robbins et al. 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat Genet* **22:** 164-167.

Harpending, H.C., M.A. Batzer, M. Gurven, L.B. Jorde, A.R. Rogers, and S.T. Sherry. 1998. Genetic traces of ancient demography. *Proc Natl Acad Sci U S A* **95:** 1961-1967.

Hattori, M., A. Fujiyama, T.D. Taylor, H. Watanabe, T. Yada, H.S. Park, A. Toyoda, K. Ishii, Y. Totoki, D.K. Choi et al. 2000. The DNA sequence of human chromosome 21. *Nature* **405:** 311-319.

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

Jurka, J., O. Kohany, A. Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* **101:** 1268-1272.

Kaessmann, H., V. Wiebe, and S. Paabo. 1999. Extensive nuclear DNA sequence diversity among chimpanzees. *Science* **286:** 1159-1162.

Kajikawa, M. and N. Okada. 2002. LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* **111:** 433-444.

Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.

Kimura, M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

King, M.C. and A.C. Wilson. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188:** 107-116.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Leeflang, E.P., W.M. Liu, I.N. Chesnokov, and C.W. Schmid. 1993. Phylogenetic isolation of a human Alu founder gene: drift to new subfamily identity [corrected]. *J Mol Evol* **37:** 559-565.

Lev-Maor, G., R. Sorek, N. Shomron, and G. Ast. 2003. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* **300:** 1288-1291.

Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.

Lonjou, C., W. Zhang, A. Collins, W.J. Tapper, E. Elahi, N. Maniatis, and N.E. Morton. 2003. Linkage disequilibrium in human populations. *Proc Natl Acad Sci U S A* **100:** 6069-6074.

Lutz, S.M., B.J. Vincent, H.H. Kazazian, Jr., M.A. Batzer, and J.V. Moran. 2003. Allelic heterogeneity in LINE-1 retrotransposition activity. *Am J Hum Genet* **73:** 1431-1437.

Morrish, T.A., N. Gilbert, J.S. Myers, B.J. Vincent, T.D. Stamato, G.E. Taccioli, M.A. Batzer, and J.V. Moran. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31:** 159-165.

Perna, N.T., M.A. Batzer, P.L. Deininger, and M. Stoneking. 1992. Alu insertion polymorphism: a new type of marker for human population studies. *Hum Biol* **64:** 641-648.

Rogers, A.R. and L.B. Jorde. 1995. Genetic evidence on modern human origins. *Hum Biol* **67:** 1-36.

Roy-Engel, A.M., M.L. Carroll, M. El-Sawy, A.H. Salem, R.K. Garber, S.V. Nguyen, P.L. Deininger, and M.A. Batzer. 2002a. Non-traditional Alu evolution and primate genomic diversity. *J Mol Biol* **316:** 1033-1040.

Roy-Engel, A.M., M.L. Carroll, E. Vogel, R.K. Garber, S.V. Nguyen, A.H. Salem, M.A. Batzer, and P.L. Deininger. 2001. Alu insertion polymorphisms for the study of human genomic diversity. *Genetics* **159:** 279-290.

Roy-Engel, A.M., A.H. Salem, O.O. Oyeniran, L. Deininger, D.J. Hedges, G.E. Kilroy, M.A. Batzer, and P.L. Deininger. 2002b. Active Alu element "A-tails": size does matter. *Genome Res* **12:** 1333-1344.

Salem, A.H., G.E. Kilroy, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2003a. Recently integrated Alu elements and human genomic diversity. *Mol Biol Evol* **20:** 1349-1361.

Salem, A.H., D.A. Ray, J. Xing, P.A. Callinan, J.S. Myers, D.J. Hedges, R.K. Garber, D.J. Witherspoon, L.B. Jorde, and M.A. Batzer. 2003b. Alu elements and hominid phylogenetics. *Proc Natl Acad Sci U S A* **100:** 12787-12791.

Sargent, C.A., C.A. Boucher, P. Blanco, I.J. Chalmers, L. Highet, N. Hall, N. Ross, T. Crow, and N.A. Affara. 2001. Characterization of the human Xq21.3/Yp11 homology block and conservation of organization in primates. *Genomics* **73:** 77-85.

Satta, Y. 2001. Comparison of DNA and protein polymorphisms between humans and chimpanzees. *Genes Genet Syst* **76:** 159-168.

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

Schmid, C.W. 1998. Does SINE evolution preclude Alu function? *Nucleic Acids Res* **26:** 4541-4550.

Schmid, C.W. 2003. Alu: a parasite's parasite? *Nat Genet* **35:** 15-16.

Schwartz, A., D.C. Chan, L.G. Brown, R. Alagappan, D. Pettay, C. Disteche, B. McGillivray, A. de la Chapelle, and D.C. Page. 1998. Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum Mol Genet* **7:** 1-11.

Stoneking, M., J.J. Fontius, S.L. Clifford, H. Soodyall, S.S. Arcot, N. Saha, T. Jenkins, M.A. Tahir, P.L. Deininger, and M.A. Batzer. 1997. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res* **7:** 1061-1071.

Walker, J.A., D.A. Hughes, B.A. Anders, J. Shewale, S.K. Sinha, and M.A. Batzer. 2003. Quantitative intra-short interspersed element PCR for species-specific DNA identification. *Anal Biochem* **316:** 259-269.

Watkins, W.S., C.E. Ricker, M.J. Bamshad, M.L. Carroll, S.V. Nguyen, M.A. Batzer, H.C. Harpending, A.R. Rogers, and L.B. Jorde. 2001. Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. *Am J Hum Genet* **68:** 738-752.

Watkins, W.S., A.R. Rogers, C.T. Ostler, S. Wooding, M.J. Bamshad, A.M. Brassington, M.L. Carroll, S.V. Nguyen, J.A. Walker, B.V. Prasad et al. 2003. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res* **13:** 1607-1618.

Wildman, D.E., M. Uddin, G. Liu, L.I. Grossman, and M. Goodman. 2003. Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: enlarging genus Homo. *Proc Natl Acad Sci U S A* **100:** 7181-7188.

Yamakawa, K., Y.K. Huot, M.A. Haendelt, R. Hubert, X.N. Chen, G.E. Lyons, and J.R. Korenberg. 1998. DSCAM: a novel member of the immunoglobulin superfamily maps in a Down syndrome region and is involved in the development of the nervous system. *Hum Mol Genet* **7:** 227-237.

Yu, N., M.I. Jensen-Seaman, L. Chemnick, J.R. Kidd, A.S. Deinard, O. Ryder, K.K. Kidd, and W.H. Li. 2003. Low nucleotide diversity in chimpanzees and bonobos. *Genetics* **164:** 1511-1518.

CHAPTER THREE

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

Reprinted with permission from *Gene

Introduction

Recently Integrated Alu Insertions in the Human Genome

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

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

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

Repetitive Elements and Genetic Variation on the Sex Chromosomes

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

While no previous research specifically addresses repetitive element generated insertion polymorphisms on the sex chromosomes, studies using other classes of genetic markers have

shown reduced genetic variation on the X and Y chromosomes of humans and other organisms (Begun and Whitley 2000; Nachman 1997; Yu et al. 2001). This reduction of observed polymorphism has largely been attributed to reduced recombination and lower effective population sizes of these chromosomes (Begun and Whitley 2000; Nachman 1997). The current study affords the opportunity to assess human sex chromosome variability with a novel class of genetic markers.

Materials and Methods

Cell Lines and DNA Samples

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

Identification of Alu Elements

Alu elements from the recently integrated *Alu* subfamilies Ya5, Ya5a2, Ya8, Yb8, Yb9, Yc1, Yd3, and Yd6 were identified from the August 2001 release of the UC Santa Cruz draft sequence (http://genome.ucsc.edu/). *Alu* subfamily members were located by two complementary methods. A local installation of RepeatMasker (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) was used to screen sequences on chromosomes X and Y for the positions of recently integrated *Alu* elements.

Exceptions to this were the Yc1 and Yc2 subfamilies, which were not identified by the software at the time of the study. In addition, subfamily specific oligonucleotides (Table 3.1) were utilized in a local installation of the National Center for Biotechnology Information basic local alignment search tool (BLAST) software (Altschul et al. 1990) to identify exact complements within the draft human genomic sequence as previously described. Results from these analyses were pooled and cross-checked to remove duplicate elements. *Alu* elements were then extracted from their locations within the chromosome and aligned with MEGALIGN (DNASTAR V 3.1.7) for subfamily verification and further analysis. Lists of all the *Alu* elements identified in the database searches and full alignments of all the recovered *Alu* elements are available under the publications section of our website (http://batzerlab.lsu.edu).

Table 3.1 *Alu* subfamily specific oligonucleotides [a]

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

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

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

Primer Design and Amplification

Oligonucleotide primers for the polymerase chain reaction (PCR) amplification of each *Alu* element were designed using the Primer3 program (http://www-genome.wi.mit.edu/cgi-

bin/primer/primer3_www.cgi). Sequences flanking the *Alu* insertions were first masked with RepeatMasker to remove all repetitive elements. Primer3 was then utilized to design PCR primers within the remaining flanking unique DNA sequences. PCR amplification was accomplished in 25µl reactions using either 60ng of template DNA (human populations) or 15ng (non-human primates), 0.2µM of each oligonucleotide primer, 200 µM deoxynucleotide-triphosphates, 1.5mM $MgCl_2$, 10mM Tris-HCl (pH 8.4) and Taq® DNA polymerase (1 unit). Each sample was subjected to the same amplification cycle as follows: initial denaturation of 150 seconds at 94ºC, 32 cycles of one minute of denaturation at 94ºC, one minute at the specific annealing temperature (shown in appendix 1), one minute of extension at 72ºC, followed by a final extension at 72ºC for 10 minutes. For analysis, 20µl of the PCR products were fractionated on a 2% agarose gel which contained 0.25µg/ml of ethidium bromide. PCR products were visualized using ultra violet (UV) fluorescence. Twenty individuals from four populations (African-American, Asian, European and either Egyptian or South American) were screened to test each locus for insertion polymorphism. Additional male DNA samples from the following populations; French (8 individuals); Indian (15); African-American (15) were used to confirm polymorphism on the Y chromosome.

Results

Subfamily Copy Number and Distribution

Following a computational search of the human draft sequence, using both diagnostic oligonucleotide queries of the database and RepeatMasker screening, 345 *Alu* repeat elements from eight young *Alu* subfamilies (*Alu* Ya5; Alu Ya8; Alu Ya5a2; *Alu* Yb8; *Alu* Yb9; *Alu* Yc1; *Alu* Yd3; and *Alu* Yd6) were identified. Of these, 264 recently integrated *Alu* subfamily members were found on human chromosome X, while chromosome Y contained 80. The

expected distributions of young *Alu* subfamilies on the sex chromosomes were calculated based on the size of each *Alu* subfamily and the proportion of the human draft sequence represented by the respective chromosome (chromosome sizes and sequenced base pair totals taken from the August 2001 freeze UC Santa Cruz summary statistics) as reported previously for human chromosome 19 (Arcot et al. 1998). The results of the database screening and expected numbers are given in Table 3.2. While several subfamilies were represented at or near expected levels, some deviated substantially. In particular, the number of *Alu* Ya5 elements was double that expected on the Y chromosome, but nearly equal to that expected on the X. The number of Yb8 subfamily members was consistent with expected numbers on both sex chromosomes. The Yc1 subfamily had approximately twice the expected number of elements on both the X and Y chromosomes. However, the excess of Yc1 *Alu* elements probably reflects the erroneous detection of Y subfamily elements that have had a fortuitous single base pair mutation to the Yc1 consensus sequence (Roy-Engel et al. 2001).

Age of Alu Insertions on the Sex Chromosomes

The average ages of the recently integrated *Alu* insertions on the X and Y chromosomes were estimated and compared to previous subfamily age estimates to determine if the amplification dynamics of recently integrated *Alu* elements on the sex chromosomes is comparable to that of the rest of the nuclear genome. In order to estimate the average age for each *Alu* subfamily the number of substitutions at CpG and non-CpG sites was determined. The mutation density for each of these mutation classes is different as a result of the methylation and

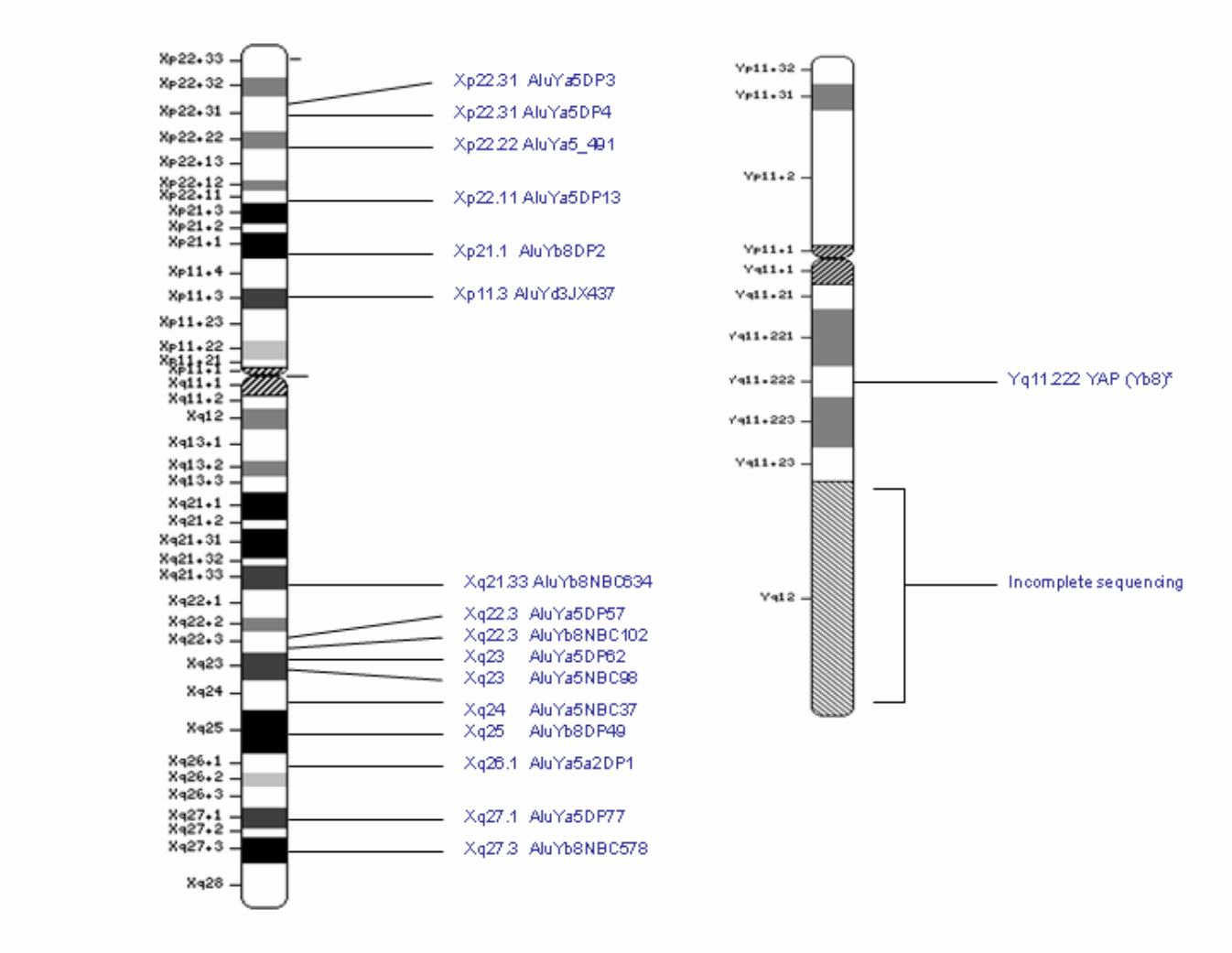


Figure 3.1

Fig. 3.1 - Idiogram of human sex chromosome specific *Alu* insertion polymorphisms.The physical location of each *Alu* insertion polymorphism was determined using the sequence map from each chromosome as a framework to localize the elements. The sequence from the q12 portion of the human Y chromosome has not yet been completed and therefore the Alu elements within this portion of the Y chromosome have not yet been analyzed. All of the *Alu* insertion polymorphisms from the recently integrated subfamilies of elements are shown in the figure. The * denotes the previously reported YAP *Alu* element (Hammer 1994)

Table 3.2 Expected and observed distribution of recently integrated *Alu* elements on the X and Y chromosomes.

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

[a] Copy numbers based on previous estimated size of the subfamilies (Batzer and Deininger 2002; Xing et al. Submitted).

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

Table 3.3 Estimated ages of sex-chromosome specific *Alu* subfamilies

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

Table 3.4 X chromosome *Alu* insertion polymorphism, genotypes and heterozygosity

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

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

2 Average heterozygosity is the average of the population heterozygosity across all four populations

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

subsequent spontaneous deamination of 5 methyl-cytosine bases (Bird 1980) and is approximately 10 fold higher in CpG than non-CpG base positions within *Alu* elements (Batzer et al. 1990; Labuda and Striker 1989). The average age for each *Alu* subfamily is then estimated by using the mutation density and a neutral rate of evolution of 0.15% per million years for non-CpG sequences (Miyamoto et al. 1987) and 1.5% per million years for CpG sequences as described previously. All deletions, insertions, simple sequence repeat expansions, and truncations were eliminated from the age calculations. All of the *Alu* elements that were identified in the draft sequence and were less than 100 bp in length were eliminated from the analysis. The estimated ages of Ya5, Yb8, and Yc1 are in line with the age estimates which were reported previously (Carroll et al. 2001; Roy-Engel et al. 2001; Xing et al. Submitted) of 2.1-4.2 Myr and are summarized in Table 3.3. Subfamilies with less than five representatives on the sex chromosomes were excluded as there was not enough sequence for accurate estimates to be made. It is important to note that the mutation rate for X and Y chromosome DNA sequences is different (Huang et al. 1997), and these differences may influence these age estimates. However, this difference should be minimal.

An evolutionary analysis of the time of origin of the *Alu* elements located on the human sex chromosomes was determined within the primate lineage was determined by PCR amplification of the individual loci using chimpanzee and gorilla DNA as templates. From the 225 recently integrated *Alu* elements analyzed in this study, three X chromosome loci (Yc1DP26, Yc1DP8 and Ya5DP38) and three Y chromosome loci (Yc1AD168, Yc1AD242, Yc1AD244) contained insertions within the chimpanzee and/or gorilla genomes, confirming that the overwhelming majority of the sex-chromosome specific *Alu* elements inserted in the human genome after the human and African ape divergence which is thought to have occurred within

the last 4-6 million years. It is interesting to note that most of the putative recently integrated *Alu* elements that were also found in non-human primate genomes were members of the Yc1 family. This is not surprising since a single base mutation differentiates this subfamily from the *Alu* Y subfamily as mentioned above (Roy-Engel et al. 2001).

Human Genomic Diversity

Individual *Alu* elements were screened for polymorphism by amplification of a panel of diverse human DNA samples, which included 20 African-Americans, 20 Europeans, 20 Asians, and either 20 Egyptians or S. Americans. A total of eighty individuals were screened comprising approximately 120 X chromosomes and 40 Y chromosomes (Table 3.4). 121 sex-chromosome specific *Alu* elements were not amplified by PCR, 109 of which were positioned within repeat-saturated regions of the genome, making the design of unique primers impossible. The remaining 12 elements either generated paralogous PCR products, or failed to amplify for unknown reasons that may include mutations within the sites where the oligonucleotide primers anneal, small deletions or even larger recombination events between adjacent sequences such as mobile elements.

The number of elements on the X chromosome which exhibited polymorphism within the human genomes that were surveyed consisted of nine Ya5's, five Yb8's, one Ya5a2, and one Yd3 element. All young subfamily members analyzed on the Y chromosome were found to be monomorphic, with the exception of one previously identified Yb8 *Alu* insertion, termed YAP (Y Alu polymorphism) (Hammer 1994), which is an intermediate frequency *Alu* insertion polymorphism. The remaining *Alu* insertion polymorphisms were classified as high, low or intermediate frequency as previously described and summarized in Table 4. Unbiased heterozygosity values for each of the polymorphisms were determined by allele counting. The

heterozygosity data suggests that the *Alu* insertion polymorphisms from the X chromosome will be useful as genetic markers for human population genetics. A schematic diagram showing the location of all the *Alu* insertion polymorphisms located on the human X and Y chromosomes is shown in Figure 3.1.

The levels of *Alu* insertion polymorphism on the X and Y chromosomes were compared to previous data on the detection of autosomal *Alu* insertion polymorphisms. The data in (Carroll et al. 2001) was adapted to exclude all elements on the sex chromosomes in order to make comparisons against autosomal loci only. Chromosome X showed 14.06% (9/64) polymorphism for the Ya5 subfamily, 100% (1/1) for Ya5a2, 20% (1/5) for the Yd3 subfamily and 8.77% (5/57) for the Yb8 subfamily. On the Y chromosome 6.66% (3/45) polymorphism was observed for the Ya5 subfamily, 10.53% (2/19) for the Yb8 subfamily, and 50% (1/2) for the Yb9 subfamily. Compared to previously reported levels of *Alu* insertion polymorphism throughout the genome of 25% (Ya5), 80% (Ya5a2), 20% (Yb8), and 25% (Yc1) (Batzer and Deininger 2002), our data indicate that there is a slight reduction in *Alu* insertion polymorphism on the human sex chromosomes.

Discussion

Distribution of Alu Elements

The expected chromosomal distribution of recently integrated *Alu* elements was calculated based on the estimated subfamily size and the relative percentage of the draft sequence constituted by each chromosome. The distribution bias in the observed numbers of *Alu* elements appears to be subfamily specific and is in good agreement with a recently published analysis sex chromosome mobile elements (Jurka et al. 2002). For example, the Ya5 subfamily has approximately twice the number of *Alu* elements expected on the Y chromosome but nearly

equal the number expected on the X chromosome. In contrast, the distribution of Yb8 subfamily members was consistent with estimated expectations on both chromosomes. Population genetics theory predicts that smaller effective populations should result in more frequent fixation of slightly deleterious insertions. Similarly, the virtual lack of recombination on the Y and reduced recombination on the X increases the extent of background selection and selective sweeps, further lowering the effective population size. Previous studies have reported a higher percentage of repetitive elements on the Y chromosome relative to autosomes and the X chromosome (Erlandsson et al. 2000). Boissinot and coworkers (Boissinot et al. 2001) previously reported an over-representation of full length and >500bp LINE elements, but no enrichment of SINEs on the sex chromosomes. In addition, the mobilization of *Alu* repeats has recently been suggested to be male germline specific (Jurka et al. 2002), suggesting yet another mechanism for the differential accumulation of *Alu* repeats within the human genome. Therefore, we conclude the distribution of different classes of mobile elements on the sex chromosomes in different species is the result of a number of complex processes such as mobilization mechanism and integration site preferences that are mobile element specific.

Age of Alu Subfamily Members

The ages of recently integrated *Alu* elements on the sex chromosomes was estimated based upon CpG and non-CpG mutation densities as reported previously. The estimated ages for the sex chromosome specific *Alu* elements are in good agreement with those reported previously (Carroll et al. 2001; Roy-Engel et al. 2001). It is possible that the higher mutation rate in the male germline (Huang et al. 1997) would result in increased divergence and therefore higher estimated ages for *Alu* subfamily members on the Y chromosome. This effect, however, may be more detectable in older *Alu* subfamilies which have had more time to acquire mutations than in

the recently integrated *Alu* subfamilies and certainly should not act selectively upon a single family of elements. This is in good agreement with a previous computational analysis of Y chromosome-specific mobile elements which demonstrated that the older *Alu* J and *Alu* S subfamilies showed significantly higher divergence on the Y chromosome, while the younger *Alu* Y subfamily divergence did not exhibit a significant difference (Erlandsson et al. 2000). Similarly, due to the increased male mutation rate, X-linked loci should theoretically exhibit a lower mutation rate than their autosomal counterparts since only one out of three X chromosomes is transmitted through the male germline each generation. However, this effect is likely minimal and is not reflected in the ages of the young Alu elements.

Population Dynamics

The recently integrated *Alu* subfamily members on the X and Y chromosomes exhibited reduced polymorphism as compared to their autosomal counterparts. Age estimates and data from orthologous inserts in non-human primates indicate that this reduction in polymorphism is not the result of increased age of *Alu* insertions found on the sex chromosomes. Rather, the results are consistent with neutral theory, given that lower effective population size should result in more rapid fixation of elements, lowering overall polymorphism levels on the sex chromosomes. Reduced recombination on the X and Y chromosomes may exacerbate this effect by increasing the extent of background selection and selective sweeps which further remove polymorphism (Charlesworth et al. 1994; Lander et al. 2001). The current findings are in agreement with several previously published studies in humans and other organisms that have found reduced polymorphism on the sex chromosomes (Hammer 1994; Jorde et al. 2000; Lander et al. 2001; Yu et al. 2001).

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

References

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J Mol Biol* **215:** 403-410.

Arcot, S.S., A.W. Adamson, G.W. Risch, J. LaFleur, M.B. Robichaux, J.E. Lamerdin, A.V. Carrano, and M.A. Batzer. 1998. High-resolution cartography of recently integrated human chromosome 19-specific Alu fossils. *J Mol Biol* **281:** 843-856.

Bamshad, M., T. Kivisild, W.S. Watkins, M.E. Dixon, C.E. Ricker, B.B. Rao, J.M. Naidu, B.V. Prasad, P.G. Reddy, A. Rasanayagam et al. 2001. Genetic evidence on the origins of Indian caste populations. *Genome Res* **11:** 994-1004.

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Batzer, M.A., P.L. Deininger, U. Hellmann-Blumberg, J. Jurka, D. Labuda, C.M. Rubin, C.W. Schmid, E. Zietkiewicz, and E. Zuckerkandl. 1996. Standardized nomenclature for Alu repeats. *J Mol Evol* **42:** 3-6.

Batzer, M.A., G.E. Kilroy, P.E. Richard, T.H. Shaikh, T.D. Desselle, C.L. Hoppens, and P.L. Deininger. 1990. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res* **18:** 6793-6798.

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

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

Boeke, J.D. 1997. LINEs and Alus--the polyA connection. *Nature Genetics* **16:** 6-7.

Boissinot, S., A. Entezam, and A.V. Furano. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol* **18:** 926-935.

Carroll, M.L., A.M. Roy-Engel, S.V. Nguyen, A.H. Salem, E. Vogel, B. Vincent, J. Myers, Z. Ahmad, L. Nguyen, M. Sammarco et al. 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* **311:** 17-40.

Charlesworth, B., P. Sniegowski, and W. Stephan. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371:** 215-220.

Deininger, P.L. and M.A. Batzer. 1993. Evolution of retroposons. *Evolutionary Biology* **27:** 157-196.

Deininger, P.L., M.A. Batzer, C.A. Hutchison, 3rd, and M.H. Edgell. 1992. Master genes in mammalian repetitive DNA amplification. *Trends Genet* **8:** 307-311.

Erlandsson, R., J.F. Wilson, and S. Paabo. 2000. Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. *Mol Biol Evol* **17:** 804-812.

Feng, Q., J.V. Moran, H.H. Kazazian, Jr., and J.D. Boeke. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87:** 905-916.

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

Huang, W., B.H. Chang, X. Gu, D. Hewett-Emmett, and W. Li. 1997. Sex differences in mutation rate in higher primates estimated from AMG intron sequences. *J Mol Evol* **44:** 463-465.

Jorde, L.B., W.S. Watkins, M.J. Bamshad, M.E. Dixon, C.E. Ricker, M.T. Seielstad, and M.A. Batzer. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* **66:** 979-988.

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

Jurka, J., M. Krnjaic, V. Kapitonov, J.E. Stenger, and O. Kokhanyy. 2002. Active Alu elements are passed primarily through paternal germ lines. *Theoretical Population Biology*.

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

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

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Luan, D.D., M.H. Korman, J.L. Jakubczak, and T.H. Eickbush. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72:** 595-605.

Miyamoto, M.M., J.L. Slightom, and M. Goodman. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the psi eta-globin region. *Science* **238:** 369-373.

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

Nikaido, M., F. Matsuno, H. Hamilton, R.L. Brownell, Jr., Y. Cao, W. Ding, Z. Zuoyan, A.M. Shedlock, R.E. Fordyce, M. Hasegawa et al. 2001. Retroposon analysis of major cetacean lineages: the monophyly of toothed whales and the paraphyly of river dolphins. *Proc Natl Acad Sci U S A* **98:** 7384-7389.

Nikaido, M., A.P. Rooney, and N. Okada. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: hippopotamuses are the closest extant relatives of whales. *Proc Natl Acad Sci U S A* **96:** 10261-10266.

Roy-Engel, A.M., M.L. Carroll, M. El-Sawy, A.H. Salem, R.K. Garber, S.V. Nguyen, P.L. Deininger, and M.A. Batzer. 2002. Non-traditional Alu evolution and primate genomic diversity. *J Mol Biol* **316:** 1033-1040.

Roy-Engel, A.M., M.L. Carroll, E. Vogel, R.K. Garber, S.V. Nguyen, A.H. Salem, M.A. Batzer, and P.L. Deininger. 2001. Alu insertion polymorphisms for the study of human genomic diversity. *Genetics* **159:** 279-290.

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

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

Sinnett, D., C. Richer, J.M. Deragon, and D. Labuda. 1992. Alu RNA transcripts in human embryonal carcinoma cells. Model of post-transcriptional selection of master sequences. *J Mol Biol* **226:** 689-706.

Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657-663.

Wichman, H.A., R.A. Van den Bussche, M.J. Hamilton, and R.J. Baker. 1992. Transposable elements and the evolution of genome organization in mammals. *Genetica* **86:** 287-293.

Xing, J., A.-H. Salem, D.J. Hedges, G.E. Kilroy, W.S. Watkins, J.E. Schienman, C.-B. Stewart, J. Jurka, L.B. Jorde, and M.A. Batzer. Submitted. Comprehensive analysis of two Alu Yd subfamilies. *Journal of Molecular Evolution*.

Yu, N., Z. Zhao, Y.X. Fu, N. Sambuughin, M. Ramsay, T. Jenkins, E. Leskinen, L. Patthy, L.B. Jorde, T. Kuromori et al. 2001. Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. *Mol Biol Evol* **18:** 214-222.

CHAPTER FOUR

A MOBILE ELEMENT BASED ASSAY FOR HUMAN GENDER DETERMINATION*

***Reprinted with permission of Analytical Biochemistry**

Determination of gender from human DNA samples is a common problem in forensics laboratories. While several PCR-based assays are currently available for human sex typing, each of the current approaches has limitations. Methods based on male-specific amplification, such as the amplification of the SRY locus (Sinclair et al. 1990), lack an internal positive control to discriminate between female DNA and male DNA which has failed to amplify for technical reasons. Restriction fragment length polymorphism (RFLP) assays based on sex-specific mutations at the ZFX/ZFY (Reynolds and Varlaro 1996) require a second enzyme digestion or hybridization step following the initial PCR amplification. A recent method proposed by Cali *et al*. based on a single adenine insertion within a tandem repeat array at the DXYS156 locus (Cali et al. 2002) requires access to allele detection equipment potentially unavailable to forensics labs with limited resources. The most widely used approach is based on the *Amelogenin* locus, which yields different sized polymerase chain reaction (PCR) amplicons for the X and Y chromosome versions of the *Amelogenin* gene (Sullivan et al. 1993). However, this method misidentifies males as females in some cases due to a deletion in the *AMEL Y* region (Santos et al. 1998; Steinlechner et al. 2002; Thangaraj et al. 2002). This deletion has previously been reported to be present at a frequency of 0.018% in Caucasian males, 1.85 % among Indians, and as high as 8% in Sri-Lankans (Santos et al. 1998; Steinlechner et al. 2002; Thangaraj et al. 2002) . While the frequency of the deletion is relatively low, the crucial nature of forensic test results in circumstances such as rape and prenatal gender determination, where there is risk for male-specific inherited disorders, makes any source of error a legitimate cause for concern. This has lead several researchers to recommend that *Amelogenin* not be relied upon as the sole determinant of gender (Brinkmann 2002; Santos et al. 1998; Steinlechner et al. 2002; Thangaraj

et al. 2002). Here, we present an alternative PCR method of human gender identification based on the presence/absence of *Alu* sequences.

Alu elements are transposable elements which have amplified throughout primate evolution and comprise roughly 10% of the human genome (Batzer and Deininger 2002). *Alu* insertions are generally considered to be homoplasy-free with respect to human population genetics, as the probability of two *Alu* elements independently inserting in the same genomic location is extremely small (Batzer and Deininger 2002). The insertion of an *Alu* element into a non-recombining X-Y homologous region creates a way of differentiating between inserted and non-inserted chromosomes based on PCR amplicon size. While some recently integrated *Alu* insertions remain polymorphic in the human population, many ultimately reach fixation for the presence of the *Alu* insertion (9). Fixed insertions on either the X or Y chromosome provide a way of identifying the respective chromosome, as the inserted chromosome yields a larger fragment when the homologous region is amplified with PCR (Figure 4.1). By screening X-Y homologous *Alu* insertions for levels of insertion polymorphism, we identified two monomorphic *Alu* insertions that meet the necessary criteria for a gender determination assay, one fixed on the X chromosome, *AluSTXa*, and one fixed on the Y chromosome, *AluSTYa*. Both of the *Alu* elements presumably inserted and reached fixation in the human lineage prior to the radiation of modern humans from Africa. Amplification of DNA samples from 778 diverse (African-American, European-American, and Hispanic-American) individuals of defined sex from paternity/identity cases for both the *AluSTYa* and *AluSTXa* loci showed 100% accuracy in gender identification. The DNA samples used in the study consisted of 389 females (278 African-American, 102 European-American, and 9 Hispanic-American) and 389 males (288 African-American, 90 European-American, and 11 Hispanic-American).

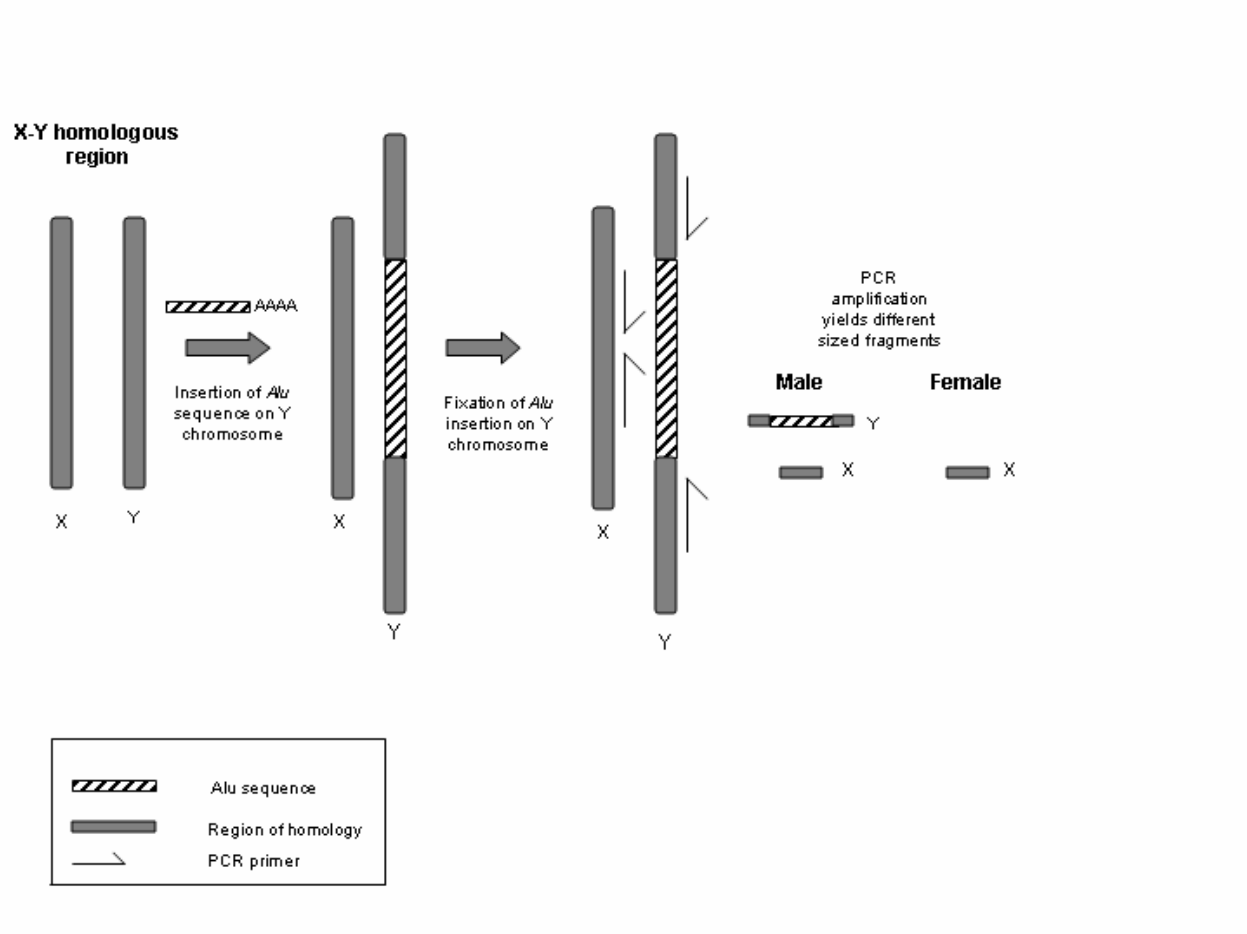


Figure 4.1

Figure 4.1 - Schematic diagram of mobile element based gender determination. In the diagram an *Alu* insertion has occurred on the Y chromosome within an X-Y homologous region. Once fixed in the population, the *Alu* insertion sequence results in a larger amplicon on the Y chromosome, allowing for the differentiation of the sex chromosomes via PCR amplification. X chromosome-specific insertions function in the same manner.Amplification of the loci was conducted via a PCR reaction and fragments were resolved on a 2% agarose gel (Figure 4.2). The primers used for the Y insertion, *AluSTYa*, were Forward 5'- CATGTATTTGATGGGGATAGAGG -3' and Reverse 5'- CCTTTTCATCCAACTACCACTGA -3', yielding an *Alu* filled site (Y chromosome) fragment of 528bp and an empty site (X chromosome) fragment of 199bp. Primers for the X insertion, *AluSTXa*, were Forward 5'- TGAAGAAATTCAGTTCATAGCTTGT -3' and Reverse 5'- CAGGAGATCCTGAGATTATGTGG -3', yielding an inserted (X chromosome) fragment of 878bp and an empty site (Y chromosome) fragment of 556bp. For both loci, males are distinguished as having two DNA fragments present, while females only have a single fragment (Figure 4.2).

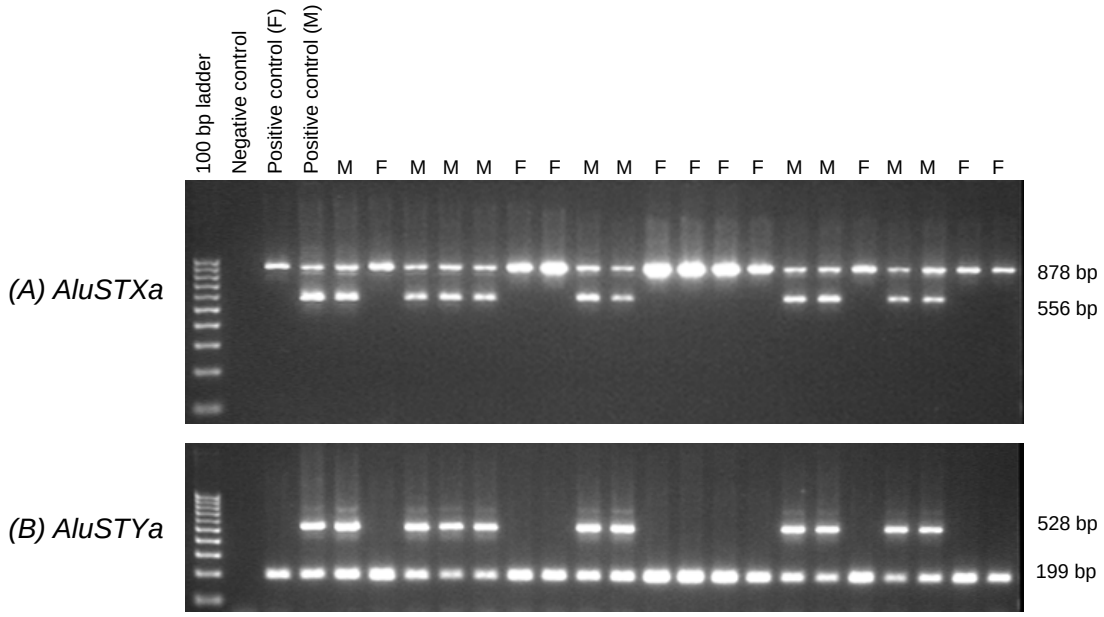


Figure 4.2

Figure 4.2 - Mobile element based gender determination. In the figure an agarose gel chromatograph from the analysis of twenty-four individuals using the genetic systems (a) *AluSTXa* and (b) *AluSTYa* is shown. Males are distinguished by the presence of two DNA fragments, while females have a single amplicon. F (female) and M (male) above each sample indicate the known gender. Individual PCR amplifications were performed in 25µl reactions using 25 ng of template DNA, 0.2µM of each oligonucleotide primer, 200 µM deoxynucleotide-triphosphates, 1.5mM $MgCl_2$, 10mM Tris-HCl (pH 8.4) and Taq® DNA polymerase (1 unit). Each sample was subjected to the same amplification cycle as follows: initial denaturation of 150 seconds at 94ºC, 32 cycles of one minute of denaturation at 94ºC, one minute at the specific annealing temperature (58ºC for *AluSTYa* and 60ºC for *AluSTXa*), one minute of extension at 72ºC, followed by a final extension at 72ºC for 10 minutes. For analysis, 20µl of the PCR products were fractionated on a 2% agarose gel which contained 0.25µg/ml of ethidium bromide. PCR products were visualized using ultra violet (UV) fluorescence.

Combining these loci together for human gender identification will provide increased accuracy for sex typing since local deletions or other types of mutations that eliminate PCR would have to occur in at least two independent genomic locations. The speed and ease of agarose based genotyping due to the ~300bp difference between filled and empty alleles will also enhance the utility of the assay in forensic laboratories. This approach should also be amenable to fluorescence-based amplicon detection, and quantitative PCR to resolve male and female contributions to sex-mixed samples. Furthermore, similar approaches based on repetitive element insertions located in homologous sex chromosome regions should be useful for gender determination in other taxa of heterogametic sex.

References

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Brinkmann, B. 2002. Is the amelogenin sex test valid? *Int J Legal Med* **116:** 63.

Cali, F., P. Forster, C. Kersting, M.G. Mirisola, R. D'Anna, G. De Leo, and V. Romano. 2002. DXYS156: a multi-purpose short tandem repeat locus for determination of sex, paternal and maternal geographic origins and DNA fingerprinting. *Int J Legal Med* **116:** 133-138.

Reynolds, R. and J. Varlaro. 1996. Gender determination of forensic samples using PCR amplification of ZFX/ZFY gene sequences. *J Forensic Sci* **41:** 279-286.

Santos, F.R., A. Pandya, and C. Tyler-Smith. 1998. Reliability of DNA-based sex tests. *Nat Genet* **18:** 103.

Sinclair, A.H., P. Berta, M.S. Palmer, J.R. Hawkins, B.L. Griffiths, M.J. Smith, J.W. Foster, A.M. Frischauf, R. Lovell-Badge, and P.N. Goodfellow. 1990. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature* **346:** 240-244.

Steinlechner, M., B. Berger, H. Niederstatter, and W. Parson. 2002. Rare failures in the amelogenin sex test. *Int J Legal Med* **116:** 117-120.

Sullivan, K.M., A. Mannucci, C.P. Kimpton, and P. Gill. 1993. A rapid and quantitative DNA sex test: fluorescence-based PCR analysis of X-Y homologous gene amelogenin. *Biotechniques* **15:** 636-638, 640-631.

Thangaraj, K., A.G. Reddy, and L. Singh. 2002. Is the amelogenin gene reliable for gender identification in forensic casework and prenatal diagnosis? *Int J Legal Med* **116:** 121-123.

CHAPTER FIVE

CONCLUSION

Introduction

While it is widely recognized that the majority of the human genome is not directly involved in the production of proteins, our understanding of the noncoding regions spanning between genes remains far from complete. There has been the temptation, particularly early on, to dismiss these geneless stretches as barren wastelands of no particular interest or significance. Yet even a casual survey of current genome annotation reveals these regions are populated by a diverse group of characters, including pseudogenes, retropseudogenes, DNA transposons, retrotransposons, and endogenous retroviruses, among others. In addition, comparative genomics has revealed a number of sequence motifs that have been highly conserved since placental mammals and monotremes last shared a common ancestor (Dermitzakis et al. 2005; Dermitzakis et al. 2003). Far from being the vast expanses of random sequence that were initially imagined, it is becoming increasingly clear that organized forms crowd the majority of this genetic terrain.

In this review we focus one group of inhabitants, mobile elements, and their role in primate evolution. Since Dawkins popularized the concept of the selfish gene in the 1970s, mobile elements have, whether justifiably or not, served to epitomize his idea, preoccupying themselves with their own replicative ambitions-sometimes to the detriment of their host genomes. It is estimated that approximately 50% of the human genome is composed of such repetitive sequences (Lander et al. 2001b). This is likely a conservative estimate as many other repeat-generated regions have degenerated beyond recognition. The majority of the elements comprising this statistic are "deceased." They either never possessed or have long since lost the ability to perform their most notable-arguably their *only*-activity, to move and/or generate new copies of

themselves. These "molecular fossils" are all but certainly fated to continue to decay until their existence is no longer detectable. Across diverse taxa, the relative number of young and active *vs.* fossil transposable elements inhabiting a given genome is remarkably varied (Lander et al. 2001b). In addition to dfferences in the age composition of mobile elements in genomes, the varieties of elements contained within these taxa also differ considerably. In some taxa, such as humans, we find relatively high mobilization levels arising from a small number of active families (Batzer and Deininger 2002). In other taxa, such as the pufferfish Tetraodon, lower activity is observed that is distributed across a greater diversity of families (Neafsey et al. 2004). One of the questions currently looming in the mobile element field concerns what set of factors govern the diversity and transposition activity levels of TEs across lineages. While there are hints that host genomic defense mechanisms (Neafsey et al. 2004) along with demographic factors (Hedges et al. 2004) underlie some of this variation, a considerable amount work of remains ahead of us.

With the sequencing of the human and chimpanzee genomes now effectively complete, we have an unprecedented opportunity to assess the impact of mobile element activity on primate evolution. Although the current data surveyed here are unavoidably chimpanzee and human-centered, we can nevertheless begin to deduce a picture of primate mobile element expansion and its associated repercussions. A number of excellent reviews exist in the literature which discuss the molecular genetics and diversity of transposable elements (Batzer and Deininger 2002; Kidwell and Lisch 2001; Ostertag and Kazazian 2001). Here, we focus on recent advances in our understanding of the evolutionary dynamic existing between transposable elements and

their primate hosts, and how this ongoing struggle for coexistence has shaped the genomic architecture of extant primates.

Origin and Structure of Primate Retrotransposons

The SINE family, Alu

The birth of the *Alu* lineage appears to have occurred shortly after the dawn of the primate lineage. As a result, *Alu* elements are found exclusively in primates. Ubiquitous in all simian and prosimian genomes examined to date, the *Alu* family is thought to have initially arisen from *7SLRNA*, an RNA gene involved in the protein signal recognition complex (Ullu and Tschudi 1984). This makes it somewhat unusual among SINEs (Short INterspersed Elements), the majority of which are derived from tRNA genes (Okada 1991).

At the early stages of its evolution, the *Alu* element structure was remarkably spartan, consisting of a RNA pol-III promoter, a short stretch of intervening sequence, and a poly-A tail (Figure 1). At under 200 basepairs, the ancestral monomeric *Alu* sequence is conspicuously lacking protein coding regions for the enzymatic machinery that makes transposition possible. How then can we account for their expansion? This apparent paradox was ultimately resolved when it was demonstrated that *Alu* is able to commandeer the requisite mobilization machinery from *L1*, another class of mammalian retrotransposon (Dewannieux et al. 2003; Kajikawa and Okada 2002). Similar "parasitic" relationships between SINEs and LINEs have been observed within other taxa (Kajikawa and Okada 2002; Okada and Hamada 1997). While fossil remnants of the ancestral *Alu* state still linger in extant primate genomes (and active lineage may well be found still persisting in unexamined genomes) early on in primate evolution two *Alu* monomer elements merged to form the modern, dimeric *Alu* structure (Figure 1)

(Zietkiewicz et al. 1998). Curiously, experimental evidence suggests that the dimeric structure is transpositionally *less* competent than its ancestral monomeric counterpart due to transcript instability (Li and Schmid 2004). While such an innovation would appear counterproductive to successful proliferation, it is nevertheless the case that this dimerization event occurred *prior* to the major expansion of *Alu* subfamilies 30-40 mya. This massive mobilization was largely carried out by dimeric *Alu*S subfamilies. As we discuss below, the evolutionary logic of dimerization and further seemingly "backwards" innovations may be more sensible than it at first appears.

The LINE family, L1

While it appears evident that primate *L1* sequences arose from ancestral mammalian LINEs, the origin of those earliest LINE (Long INterspersed Element) ancestors is something of an enigma (Malik and Eickbush 2001). What is clear is the extreme antiquity of the non-LTR retrotransposon lineage to which *L1* belongs. At roughly 6000 bp, the primate *L1* family is considerably bulkier than *Alu.* It consists of an RNA pol-II promoter along with two open reading frames (ORFs), a 3' UTR, and a poly-A tail (Ostertag and Kazazian 2001) (Figure 1). The better characterized second ORF encodes a protein possessing both endonuclease and reverse transcriptase activity (Jurka 1997; Mathias et al. 1991). The first ORF encodes a protein of an as-yet unknown function that has nevertheless been demonstrated to be necessary for the *L1* transposition process (Moran et al. 1996). While experimental evidence suggests a cis-preference for *L1* encoded proteins, (Wei et al. 2001) distantly related mouse *L1* protein machinery is able to mobilize human *Alu* elements in cell culture (Hagan et al. 2003). Thus, while *L1* transcripts may preferentially be retrotransposed by their own proteins, the *Alu* retrotransposition process appears more promiscuous. Although a number of

full length *L1*s exist in the human genome, the majority of *L1* inserts appear to have been 5' truncated upon insertion, rendering them “Dead On Arrival" (DOA) (Myers et al. 2002).

Endogenous Retroviruses, SVA Elements, and Further Mobile Element Diversity

While *L1* and *Alu* families constitute the bulk of primate-specific mobile element activity, particularly in recent evolutionary history, a number of additional lineages have also left their mark on primate genomes (Smit and Riggs 1996). These include DNA transposons, SINE-R, LTR retrotransposons, and endogenous retroviruses. Although active 80-90 million years ago in an early primate ancestor, "cut and paste" DNA transposons have apparently had more success in the rodent order. During its tenure in primate evolution, the DNA transposon Tigger gave rise to numerous smaller MITE (Miniature Inverted Repeat Element) sequences in the genome of an ancestral primate (Smit and Riggs 1996). With only two great ape genomes sequenced thus far, the extent to which these DNA transposon lineages may have survived in an active form in extant primates remains unclear, though all indications point to their having died out in the human and chimpanzee lineages (Medstrand and Mager 1998).

In addition to DNA transposons, endogenous retroviruses have also impacted the genetic landscape of primates. These sequences, largely consisting of remnants of ancient germline retroviral infections, are believed to comprise nearly 1% of the human genome (Sverdlov 2000). Subsequent to integration into germline DNA, endogenous retroviruses can be inherited as mendelian genes and, in some instances, will continue to generate new genomic copies by retrotransposition. Endogenous retroviral insertions have been demonstrated to alter expression in nearby genes and have been implicated

in conveying host resistance. The role of endongenous retroviruses in primate evolution is addressed extensively in (Sverdlov 2000).

The SVA (SINE, VNTR, ALU) family has a chimeric structure, consisting of an LTR component, an LTR repetitive region, an *Alu* component, and a poly-A tail (Figure 1) (Ostertag et al. 2003). Evidence indicates that it existed in its present form at least as far back as the human-chimpanzee common ancestor. As with *Alu*, these elements are pol-III transcribed and require *L1* to provide the proteins required for transposition. In terms of size, however, they are intermediate between *Alu* and *L1*, and this characteristic likely shapes their particular niche in the ecology of the genome. As part of their structure consists of an *Alu*-derived component (Figure 1), they must have arisen subsequent to the *Alu* lineage. Still active in human and chimpanzee, SVA contributes to both human disease and genetic diversity (Ostertag et al. 2003).

Assessing the Impact of Transposition

Human Disease

With the availability of full genomic sequences, and an ever-growing arsenal of molecular and computational tools at our disposal, we are only now beginning to fully appreciate the full scope of mobile element activity and influence in primates. Perhaps their most conspicuous effect is their role in the etiology of numerous genetic disorders, including neurofibromatosis type 1, hemophilia types A and B, and familial hypercholesterolemia (Chae et al. 1997; Ganguly et al. 2003; Vidaud et al. 1993; Wallace et al. 1991). Literature and database estimates indicate that .3-.5% of human genetic disorders result either directly from mobile element insertion or from nonhomologous recombinations between existing mobile elements. (Deininger and

Batzer 1999) However, technical constraints surrounding current disease mutation detection methods likely result in this figure being an underestimate (Li et al. 2001).

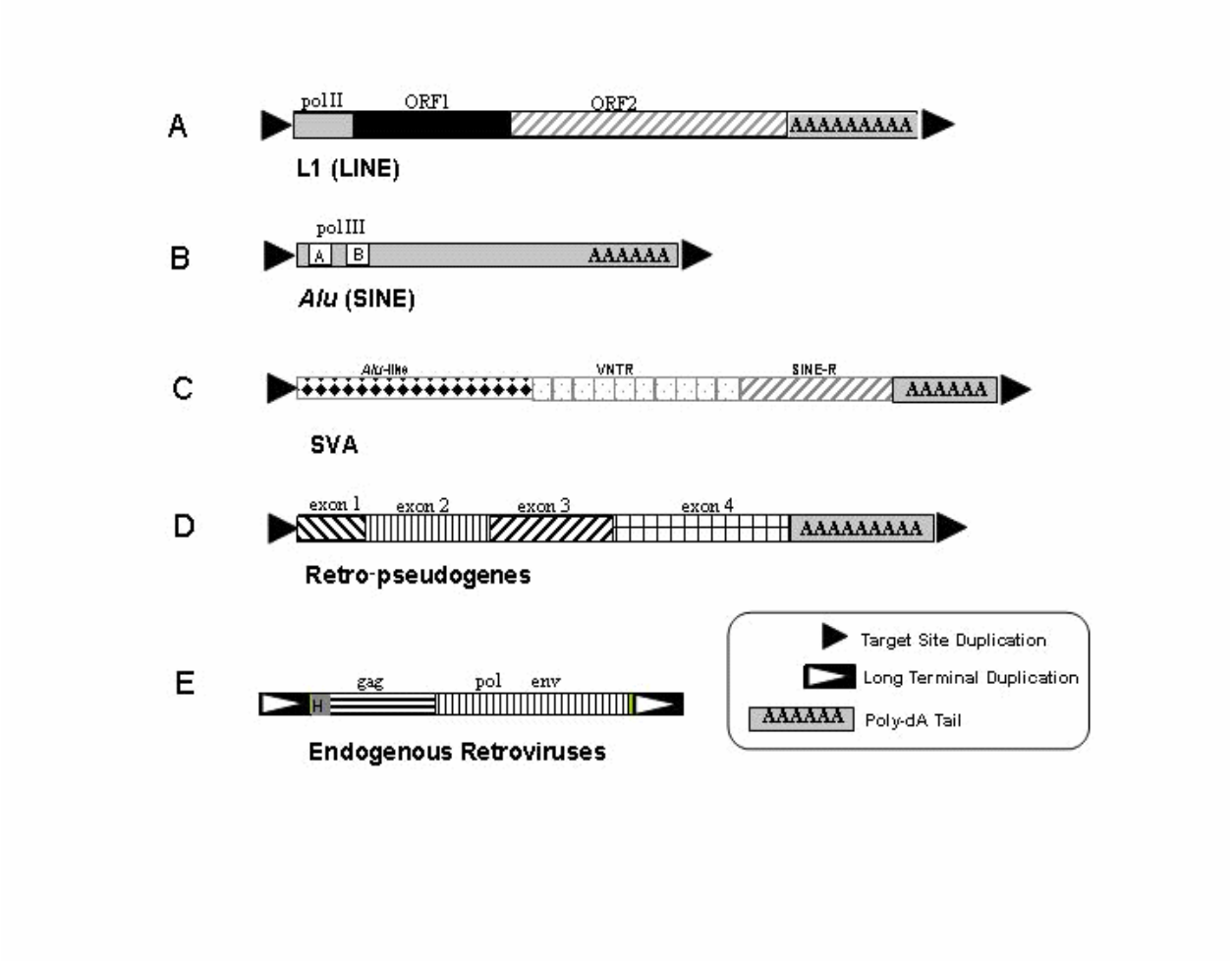


Figure 5.1

Figure 5.1 - S**tructure of Primate Mobile Elements**
Structure of major primate mobile elements. *A)* L1 *B)* Alu *C)* SVA *D)* Retropseudogene *E)* Endogenous Retrovirus

In addition to insertion and recombination-mediated gene disruptions, the ability of insertions to alter epigenetic regulation, seed microsatellite formation within introns, as in the case of Friedreich's ataxia (Justice et al. 2001), induce potentially maladaptive alternative splicing, (Lev-Maor et al. 2003) or premature truncation of transcripts (Han et

al. 2004; Perepelitsa-Belancio and Deininger 2003) may also contribute to disease states.

Genomic Variation and Size

Mobile elements also make a significant contribution to the genetic diversity existing currently among human populations. In humans, there are hundreds of mobile element insertions that exist as (primarily) neutral polymorphisms (Carroll et al. 2001). Population studies indicate that most of these insertion events occurred prior to the radiation of modern humans from Africa (Bamshad et al. 2003; Jorde et al. 2000; Watkins et al. 2001). In addition to these insertion-related polymorphisms, an abundance of polymorphic duplications and deletions generated from nonhomologous recombinations between mobile elements exist. (Gilbert et al. 2002; Pauline A. Callinan 2005; Salem et al. 2003a) Recent studies also indicate that *Alu* transposition may play an important role in the generation of segmental duplications that constitute roughly 5% of the human genome (Bailey et al. 2003). Due to the high CpG content of *Alu* elements and associated increase in nucleotide mutation rate (see below), *Alu* elements contain a substantial portion of the single nucleotide polymorphisms in the human genome. As mentioned above, the poly-A tails of *Alu* elements can also serve to seed microsatellite formation and expansion, (Arcot et al. 1995) which can in turn alter gene activity when in introns. We fully expect that many more incidents of gene alteration resulting from the regional influence, epigenetic or otherwise, of polymorphic mobile element insertions will be discovered as our knowledge of the genome and the etiology of genetic diseases expands.

In terms of genome size, comparative studies suggest that the activity of mobile elements has led to a roughly 10% expansion in the size of the human genome with

respect to chimpanzee (Liu et al. 2003). Across the various primate lineages, differential mobile element activity has likely resulted in similar genomic size fluctuations. If we take a more long-term evolutionary perspective, it is clear that the majority of the primate genome is repeat-laden, and mobile elements and their remnants compose the bulk of the substrate in which primate genes reside and evolve. Repeat driven genomic expansion may have, in addition to providing raw genetic material for evolution, also provided the necessary spatial context for evolutionary experimentation with regulatory schemes.

Exon Shuffling and Protein Evolution

The ability of *L1* to transduce considerable lengths of sequence beyond its 3' end has led to the speculation that *L1* elements might be able to move exons about the genome, facilitating protein evolution. The capacity of *L1* elements to transduce exons in this manner has been demonstrated *in vivo* (Moran et al. 1999). In addition to directly transducing sequences themselves, the protein machinery they produce also facilitate protein evolution *in trans*, as has been observed in the human Leptin receptor (Damert et al. 2004). While SVA lineage has also been shown to possess transduction capability, (Ostertag et al. 2003) there has been indication thus far that naturally occurring *Alu* sequences can transduce sequence. In addition to *L1* transduction events, inter and intrachromosal nonhomologous recombination, mediated by mobile element copy homology, can also lead to exon duplication and shuffling (van Rijk and Bloemendal 2003).

Genome GC content

Due to CpG methylation, many mammalian genomes, including primates, experience a unidirectional increase in C->T mutation rate at CpG loci, resulting in an

overall GC deficit (Waterston et al. 2002). The continued proliferation of GC-rich *Alu* sequences has served to replenish GC content within otherwise GC-poor primate genomes. While it has been proposed that *Alu* elements have been positively selected in GC isochores, (Lander et al. 2001a) there exists some evidence to the contrary, (Belle and Eyre-Walker 2002) and the time-scale over which this positive selection is purported to occur is not reconcilable with the existence of available *Alu* insertion/deletion polymorphism for natural selection to act upon (Brookfield 2001). For example, the expected coalescence time for a locus in a species with an effective population size of 10,000 individuals is approximately $4N_e$ or 1 myrs. Larger population sizes of ancestral primates would extend the expected persistance time of polymorphisms, but the concentration of *Alu* elements in GC regions only becomes evident with older (>5 yrs) *Alu* elements. This suggests that the processes underlying the Alu GC bias are occurring over a timescale far longer than the expected lifetime of Alu insertion polymorphisms. As the initial distribution of young Alu elements is slightly biased towards AT-rich regions, only the removal of *already fixed* Alu elements could account for the observed long-term distribution. Indeed, it has been proposed that purifying selection acting on such removal/deletion events (primarily occurring in the paternal germline) from regions of low GC content has resulted in the current Alu distribution (Jurka et al. 2004). The process of paternal deletion would putatively introduce new variation for selection to act upon. This explanation also presents something of a conundrum, however. As it is likely that most Alu elements would have reached fixation in population *prior* to the action of the force(s) that shape their distribution to GC regions (presumably these are deletion-based), these elements must have had either neutral or nearly neutral selection coefficients at the time of their

insertion and subsequent fixation. Why, then,would their selection coefficients subsequently change such that the Alu-containing allele becomes selected against? One might imagine a few such reversals occurring, but the idea that such selective flip-flops have occurred frequently enough to shape Alu distribution in primate genomes seems unlikely. Rather, while we suspect there may indeed be paternally based and other Alu-involved deletion events occurring in AT-rich regions, but we would argue that neutral drift, rather than selection, is what drives fixation of the “Alu-removed” alleles. The combination of this removal of *Alu* sequences through deletion in AT-rich regions, coupled with a tendency of gene-rich, GC-rich regions to not tolerate instability associated with such deletions, has likely resulted in the observed distribution of Alu insertions that we observe.

Gene Conversion

Although the underlying mechanisms are unclear, *Alu*-mediated gene conversion events have been well documented in the literature (Batzer and Deininger 2002). These events, where sequence is unidirectionally transferred from a donor to a target location, may have a considerable impact on the overall nucleotide diversity of the genome and, in particular, the evolution of mobile element families themselves. One such gene conversion event has been implicated in the deactivation of the CMP-N-acetylneuraminic acid hydroxylase gene, possibly a crucial step in the evolution of the modern human brain (Chou et al. 2002).

Gene Expression and Alternative Splicing

Perhaps the most significant events in which mobile elements have impacted primate evolutionary history remain to be discovered. Recent evidence indicates that *Alu* elements, when inserted in an inverse orientation to a gene transcript, can provide

alternative intron splicing sites, and numerous examples of *Alu*-incorporated ESTs have been detected. (Dagan et al. 2004; Sorek et al. 2002) In addition, it has been observed that Pol-II and Pol-III transcription factor binding sites can be carried by mobile elements, which may further serve to modulate gene expression. (Shankar et al. 2004) Significant epigenetic influences of mobile elements on surrounding chromatin is suggested by their exclusion from imprinted regions of the genome (Greally 2002). In addition, research has shown that L1 elements can alter gene expression when inserted within introns due to the reduced ability of the pol-II polymerase to read through L1 sequences (Han et al. 2004). While the full impact of these modifications on the genome has yet to be determined, they greatly expand the genetic repertoire with which mobile elements may influence primate evolution.

A Functional Role for Mobile Elements?

The interaction between mobile elements and their primate hosts can not adequately be addressed without tackling the question of whether or not these elements serve some necessary functional role. If the answer is yes, then the relationship between host and element must be addressed from within a symbiotic rather than a parasitic paradigm. Numerous functions have been proposed in the literature, including origins of replication, meiotic recombination, DNA repair, regulation of gene expression and others (reviewed in Ref 26), but none of these has been widely accepted. (Schmid 1998) It is important to distinguish between two fundamentally different kinds of beneficial "roles" that might be assumed by mobile elements. On the one hand, individual elements at specific chromosomal loci may occasionally provide a selective advantage to the host, either by altering the expression of a gene or, in rarer instances, being incorporated directly into the gene product itself and generating a novel protein.

That fact that such beneficial events occur is not itself in question, and numerous examples can be found in the literature. (Sarkar et al. 2003) Rather, the "question of function," as we will refer to it here, centers instead on whether mobile elements play a necessary and persistent role in their host organisms' survival. While an enormous amount of speculation has surrounded this issue, little conclusive evidence is presently available. The general tendency within popular scientific literature to classify mobile elements as "selfish" or "parasitic" clearly indicates where the broader biological community's sentiments lie. In support of this view is the demonstrably deleterious effect of some mobile element insertions, most notably in human diseases. The case against function can further be made from the infectious manner in which transposable elements colonize virgin genomes of sexually reproducing offspring, as, for example, in the case of Drosophila P-elements. Likewise, the conspicuous scarcity of retrotransposons within asexually reproducing lineages suggests they are not sustainable where sexual reproduction can not counter the fitness losses they impose (Arkhipova and Meselson 2000).

The case for function can also be compelling, however (Brosius and Gould 1992; Schmid 1998). Cellular stresses such as viral infections or heat shock, have been observed to result in *Alu*-specific transcription responses that down-regulate translational activity (Liu et al. 1995). From the closely related rodent order, there is evidence that a group of retrotransposons known as LTR class III plays a significant role in regulating gene expression in mouse Oocytes and preimplantation embryos (Peaston et al. 2004). In this case, promoter sequences from the terminal repeat region of the element initiate transcription and provide alternate 5’ exons for a number of genes. Such examples in rodents of TE recruitment in regulating critical developmental

processes increase the likelihood that similar TE functionality might also occur in primates.

There also remains the curious fact that *Alu* and *L1*, like SINE and LINEs in many other taxa, appear to have remained active among all extant primate lineages. This may simply signify the inability of genomes to eradicate these lineages. Theory indicates that as long as fitness costs incurred fall below two-fold, mobile elements can proliferate in sexual organisms (Bestor 1999). Yet theoretical approaches have difficulty accommodating the influence of repression mechanisms implemented by the host to control mobile element proliferation. If the cumulative burden of transposition on the host genome is high, any novel mutations that resulted in the repression of mobile element activity would be expected to rapidly sweep through the host population. With less than 300 bp of genomic sequence and no protein coding capability, the sparsely featured *Alu* family, for example, would appear as though it would have very limited avenues available with which to counter host suppression schemes. Is their continued persistence across so many primate lineages evidence of some conferred advantage? The various arguments for and against function are addressed in (Schmid 1998).

Despite all the uncertainty surrounding the issue of function, *Alu* has taken on a unmistakable role in recent human history. Owing largely to the pioneering efforts of Okada and colleagues working on nonprimate taxa, (Murata et al. 1993; Shimamura et al. 1997) mobile elements have proven to be powerful genomic tools for tackling several questions in primate phylogeny, notably in resolving the human/chimp/gorilla trichotomy, as well as resolving a number of branches of the prosimian (Roos et al. 2004; Schmitz et al. 2002; Schmitz et al. 2001) and old and new world monkey phylogenies (David A. Ray 2005; Jinchuan Xing 2005; Salem et al. 2003b). Since the ancestral state of an *Alu*

insertion is known to be the absence of the element, and they suffer essentially no homoplasy at the population level, polymorphic *Alu* insertions have also proven powerful tools for addressing questions about the history of human populations (Watkins et al. 2003). In addition to evolutionary studies, primate mobile element sequences are currently being capitalized upon in numerous forensic applications, including DNA quantitation, sex typing, inferring group membership of unknown samples (Bamshad et al. 2003). So despite their rather dubious role in primate evolutionary history, these "selfish" DNA elements have found a welcome home in the modern laboratory.

Marching Across The Genetic Landscape

Mediating the overall impact of mobile elements is their ability to persist and proliferate within their respective host genomes. While it is clear that self-regulation and the efficiency of host repression mechanisms factor heavily in this equation, additional factors no doubt remain to be uncovered. Fortunately for the researcher, the topology of primate genomes is riddled with historical evidence of what can at best be described as "an uneasy coexistence."

Germline Specificity and Host Repression Mechanisms

There is increasing evidence that *Alu* and *L1* transposition in primates is largely restricted to the germline, with a possible bias toward the male germline (Jurka et al. 2002). From a "selfish" evolutionary perspective germline mobilization is very sensible, as there is little benefit for the retrotransposon in inserting itself within somatic chromosomes. The resulting copies would not be inherited and, more importantly, could greatly reduce the fitness of the host organism (and consequently the transposon itself). The ability of the "copy and paste" retrotransposon in particular to restrict its

activity to the germline is therefore critical in reducing its overall fitness burden on the host genome and paving the way for further propagation. Germline transposition specificity in primates, however, may itself have not been a mobile element adaptation so much as a consequence of the germ cell development process itself. The principle means by which primates are believed to regulate mobile element proliferation is DNA methylation (Liu et al. 1994; Yoder et al. 1997). During germline development, germline cells undergo a period of demethylation, allowing a window of opportunity for otherwise silent retrotransposons to mobilize.

Although methylation is considered the main regulatory mechanism in primates, other genomic defense systems may also exist. RNAi has been demonstrated to effectively quell mobile element activity in *C. elegans*, (Sijen and Plasterk 2003) and related mechanisms could conceivably be employed by primates. Despite claims of targeted mobile element excision mechanisms in primates (Jurka et al. 2004), we feel the evidence presented thus far is unconvincing. Were such removal mechanisms prevalent in mammals, the use of SINE elements as phylogenetic markers would have proven far more problematic than has been experienced to date. If, on the other hand, one contends that removal mechanisms act so rapidly and efficiently that they do not cause phylogenetic inconsistencies, then one would be hard-pressed to explain the genome's seemingly capricious decisions concerning when and where to excise elements. Why, for example, are disease-causing mobile element insertions not efficiently plucked out of the genome? If such mechanisms exist, it must be the case that when they invoked at a locus, they act with such ruthless efficiency that they generate no phylogenetic inconsistencies, and yet, when they would be most handy (rescuing disease insertion alleles, for instance), they are appear to be frequently not

invoked at all. For these and other reasons, active genomic removal mechanisms of retrotransposons in primates appear implausible to the authors at the present time. The distribution and diversity evidence that has been used to support the notion of retrotransposon removal in primates can, we believe, be accomodated by a combination of passive, nonspecific deletions and negative (purifying) selection. We intend to address these issues in detail in subsequent work.

Finally, the weeding out of deleterious insertions and their sources by natural selection due to reduced fitness of individual hosts can itself be conceived of as a type regulatory mechanism protecting against overly ambitious mobile elements. As we elaborate upon below, what is perhaps less evident is that the overall success of this form of regulation will be contingent on the population demographics of the host.

Amplification Strategies

Attempts to account for sequence diversity exhibited by primate retrotransposons have resulted in a number of tranposition models (Cordaux et al. 2004). Most notably, the "master gene" (MG) model posits a main driver or source sequence which generates large number of inert DOA copies (Shen et al. 1991). Further refinement of the model allows for the coexistence of multiple masters or sources. The MG model accounts for observed constraints in copy number expansion and sequence diversity as well as the nature of sequence substructure (*i.e.* the sharing of common diagnostic base motifs among hierarchical element families). Presumably, since the generated copies themselves are replicas of the original sequence, they remain inert because they lack additional factors present in the sequence surrounding the "master" sequence or sequences. Under the MG model, the probability of an existing master sequence generating a novel master sequence is contingent on the number of source-conducive

landing spots that are available in the host genome. Until recently, it was believed that this probability was vanishingly small due to a scarcity of suitable genomic locations. However, network-based analyses now suggest that *Alu* elements frequently spawn copies that are themselves retrotranspositionally competent "secondary sources" (Figure 2) (Cordaux et al. 2004). These secondary sources undermine the ability of the MG model alone to explain the constraint on retrotransposon numbers and diversity in primates.

Population Dynamics and "Stealth" Drivers

To fully appreciate the complexity of mobile element evolution, it is necessary to approach the issue from both a molecular and population genetics perspective. Despite considerable advances in understanding of the biology of mobile elements and a growing body of theoretical work, the integration of host population dynamics into the mobile element evolutionary framework remains incomplete. The consequence is the promulgation of hypotheses which, while biologically attractive, prove much less palatable when their population-level implications are considered. An increased effort, particular in the primate arena, must be made to re-examine mobile element evolution with both molecular and population considerations in mind. For example, while it is tempting to envision a fairly uniform insertion rate of mobile elements in genomes, source elements themselves can fluctuate in copy number, greatly affecting the overall number of element insertions occurring in the host genome population (Figure 3) (Hedges et al. 2004). Similarly, allelic variations of source elements may also fluctuate in the population, influencing the overall rate of transposition (Brouha et al. 2003; Lutz et al. 2003). In a relatively small primate population, a newly inserted element that is highly

active could alter in frequency (and hence the populations transposition rate) significantly over only a few generations (Figure 3).

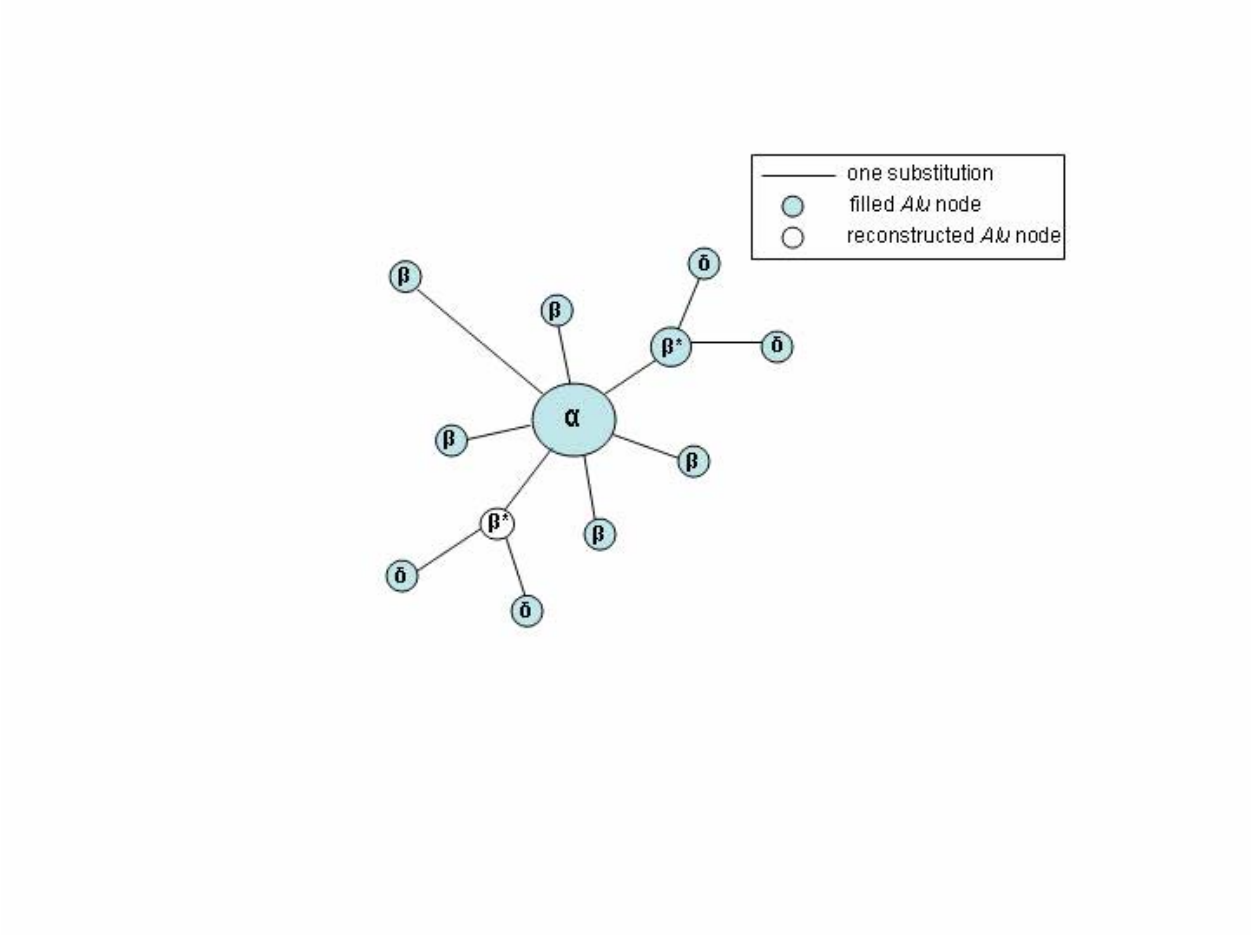


Figure 5.2

Figure 5.2 - *Alu* Network Phylogeny.
Example of a network phylogeny for a young *Alu* subfamily. Size of node indicates element copy number. Central alpha node (α) represents family consensus. Starred beta nodes (β*) depict Alu secondary source elements capable of producing "offspring.

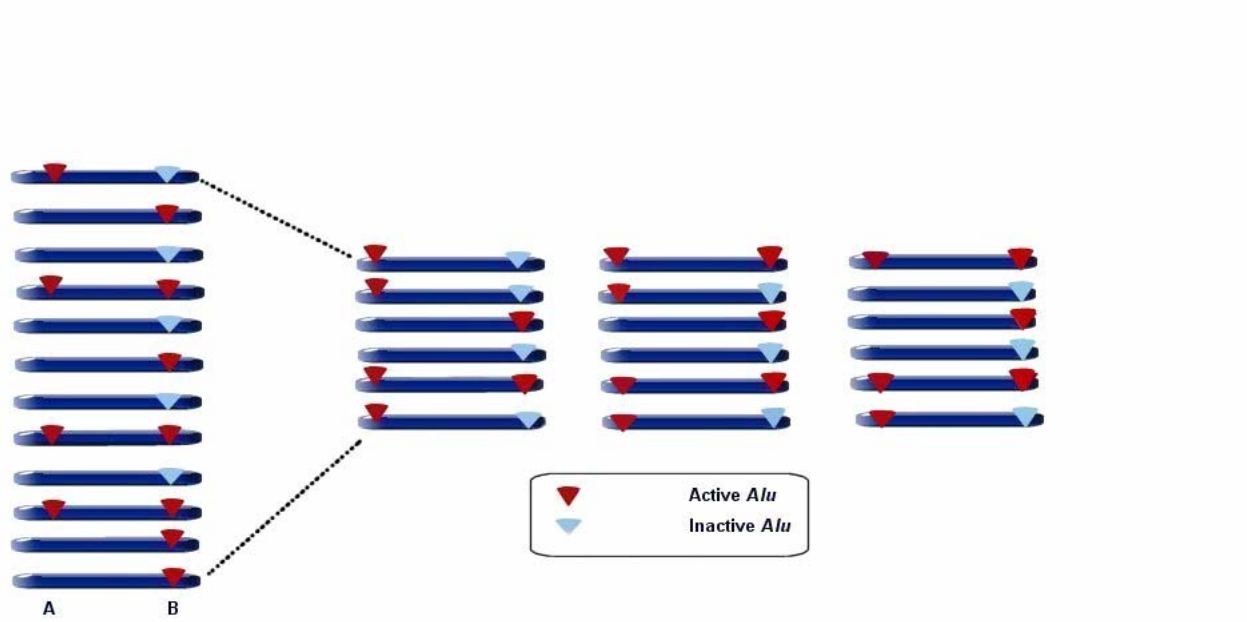


Figure 5.3

Figure 5.3 - Effect of Genetic Drift on Retrotransposition Level. Fluctuation of source element frequency in population. Two loci are, A and B, are depicted. At locus A, there exists a polymophism for the insertion/absence of an active *Alu* source element. At locus B, the inserted Alu is fixed in the population, but there is allelic variation for *Alu* activity. At each locus, the frequency of active *Alu* loci changes after passing through an population bottleneck event.

The recent evidence for appreciable numbers of *Alu* secondary sources further emphasizes that these population-level processes must be accommodated in our understanding of transposition dynamics.

So what becomes of newly generated secondary source elements? Even under neutral or nearly-neutral conditions, the vast majority will be lost rapidly to drift. These ephemeral source elements will likely have little influence on the overall structure of the genome, having had little time to produce new copies. A small fraction (roughly $1/2N_e$), however, will survive this initial stochastic barrier. If they are too transpositionally active, they will reduce host fitness and be subject to negative selection. However, it is

important to recognize that the deleterious alleles created by these active sources will, in all likelihood, not be physically linked to the chromosomal location of the source. They are, in effect, partially screened from negative selection. For example, if a "master" or source generates a copy which knocks out a gene resulting in a recessively inherited disorder, the newly formed disease allele will be selected against in subsequent generations far more intensely than the source locus that produced it.

Yet some disease alleles will be dominant in nature, and these-particularly dominant lethals-will lead to rapid removal of both disease and source loci together. Assuming an appreciable portion of mutants are dominant, exceedingly active sources should be efficiently purged through selection. What, then, is the Goldielocks level at which a source element should emit new progeny? It is clear that if the transposition level is too low, not enough offspring will establish themselves in the population to propagate the lineage. Neutral substitutions and deletions will accumulate in existing members and the lineage will be lost. On the other hand, if the transposition level is too high, selection will weed out the source before it can reach appreciable frequency in the population. As it turns out, the emission level that constitutes "just-right" for a mobile element is a moving target. The efficiency with which negative selection acts is contingent upon the selection coefficient of a loci and the effective population. Loci with selective coefficients sufficiently below $1/2N_e$ will drift as though neutral. Assuming a source can maintain a low enough emission level to stay below this threshold, it can fix in the genome. But the threshold will necessarily move up and down with the population size of the host. Hence, when population size drops, higher emission values are "tolerated" and overall transposition frequency in the population (i.e. number of insertions per birth) can increase. This may have been what resulted in an increase in

human *Alu* transposition compared to chimpanzee and gorilla (Hedges et al. 2004). Likewise, a larger population size may effectively squash mobile element duplication activity. Computational and analytical modelling of the above processes will ultimately be required to rigorously assess the impact of these forces on mobile element evolution.

As mentioned above, it can be expected that selective pressure against active elements will result in self-regulation. As a consequence, an effective retrotransposon survival strategy, which we have termed "stealth driver," can be envisioned. In this scenario, successful mobile element lineages will remain largely inactive over extended periods of evolutionary time due to a quiescent source. Occasionally, perhaps due to optimal population conditions, the source produces a highly active secondary source that rapidly expands the copy number of the lineage. Although selection ultimately culls this overactive element, the original "stealth driver" persists in genome, living to proliferate another day. In the interim, many element copies have been produced, one or more of which may become a "stealth driver" itself. Data from the two largest human *Alu* subfamilies, Ya5, and, more recently, Yb8, lend support this hypothesis (Kyudong Han 2005; Leeflang et al. 1993). These *Alu* families demonstrate extended quiescent periods followed by bursts of activity. While quiescence is key to longevity, punctuated bursts of secondary source activity may occasionally be required to ensure propagation of the lineage.

How do these "stealth drivers" maintain their low emission levels? The sequence context in which these elements reside is likely one component. Additionally, as mentioned above, in *Alu* elements the creation of a dimeric structure early in its evolutionary history actually resulted in *decreased* transposition activity. Likewise, it has been shown that key mutations in recent, successful *Alu* families also limit activity

(Aleman et al. 2000). In a similar manner, *L1* elements have been shown to contain numerous cryptic polyadenylation sites that serve to limit both the amount of transposition machinery they produce, as well as the number of full length transcripts (Perepelitsa-Belancio and Deininger 2003). In sum, there are now several lines of evidence that substantiating the notion that primate mobile elements are self-regulating. These regulation strategies may, however, only serve to allow elements to retain a low profile until more favorable expansion conditions exist. When such conditions arise, well-positioned progeny may significantly increase lineage numbers and, consequently, the overall burden of the elements on the host.

Conclusion

When a more complete understanding of genomics finally emerges, it is likely that the occupants of the genomic "wastelands" will prove every bit as interesting-and relevant to organismal biology-as the genes that accompany them. Mobile elements have played a large role in shaping the molecular evolution of extant primates. Understanding the dynamics of their proliferation will require the integration of numerous disciplines, including molecular biology, population genetics, and computational biology. Our failure to adequately draw upon of any one of these areas could result in our missing much of the rich tapestry of interactions underlying mobile element proliferation, and, consequently, major forces shaping genome evolution.

References

Aleman, C., A.M. Roy-Engel, T.H. Shaikh, and P.L. Deininger. 2000. Cis-acting influences on Alu RNA levels. *Nucleic Acids Res* **28:** 4755-4761.

Arcot, S.S., Z. Wang, J.L. Weber, P.L. Deininger, and M.A. Batzer. 1995. Alu repeats: a source for the genesis of primate microsatellites. *Genomics* **29:** 136-144.

Arkhipova, I. and M. Meselson. 2000. Transposable elements in sexual and ancient asexual taxa. *Proc Natl Acad Sci U S A* **97:** 14473-14477.

Bailey, J.A., G. Liu, and E.E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73:** 823-834.

Bamshad, M.J., S. Wooding, W.S. Watkins, C.T. Ostler, M.A. Batzer, and L.B. Jorde. 2003. Human population genetic structure and inference of group membership. *Am J Hum Genet* **72:** 578-589.

Batzer, M.A. and P.L. Deininger. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Belle, E.M. and A. Eyre-Walker. 2002. A test of whether selection maintains isochores using sites polymorphic for Alu and L1 element insertions. *Genetics* **160:** 815-817.

Bestor, T.H. 1999. Sex brings transposons and genomes into conflict. *Genetica* **107:** 289-295.

Brookfield, J.F. 2001. Selection on Alu sequences? *Curr Biol* **11:** R900-901.

Brosius, J. and S.J. Gould. 1992. On "genomenclature": a comprehensive (and respectful) taxonomy for pseudogenes and other "junk DNA". *Proc Natl Acad Sci U S A* **89:** 10706-10710.

Brouha, B., J. Schustak, R.M. Badge, S. Lutz-Prigge, A.H. Farley, J.V. Moran, and H.H. Kazazian, Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100:** 5280-5285.

Carroll, M.L., A.M. Roy-Engel, S.V. Nguyen, A.H. Salem, E. Vogel, B. Vincent, J. Myers, Z. Ahmad, L. Nguyen, M. Sammarco et al. 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* **311:** 17-40.

Chae, J.J., Y.B. Park, S.H. Kim, S.S. Hong, G.J. Song, K.H. Han, Y. Namkoong, H.S. Kim, and C.C. Lee. 1997. Two partial deletion mutations involving the same Alu sequence within intron 8 of the LDL receptor gene in Korean patients with familial hypercholesterolemia. *Hum Genet* **99:** 155-163.

Chou, H.H., T. Hayakawa, S. Diaz, M. Krings, E. Indriati, M. Leakey, S. Paabo, Y. Satta, N. Takahata, and A. Varki. 2002. Inactivation of CMP-N-acetylneuraminic acid hydroxylase occurred prior to brain expansion during human evolution. *Proc Natl Acad Sci U S A* **99:** 11736-11741.

Cordaux, R., D.J. Hedges, and M.A. Batzer. 2004. Retrotransposition of Alu elements: how many sources? *Trends Genet* **20:** 464-467.

Dagan, T., R. Sorek, E. Sharon, G. Ast, and D. Graur. 2004. AluGene: a database of Alu elements incorporated within protein-coding genes. *Nucleic Acids Res* **32 Database issue:** D489-492.

Damert, A., J. Lower, and R. Lower. 2004. Leptin receptor isoform 219.1: an example of protein evolution by LINE-1-mediated human-specific retrotransposition of a coding SVA element. *Mol Biol Evol* **21:** 647-651.

David A. Ray, J.X., Dale J. Heddges, Michael A. Hall, Meredith E. Laborde, Bridget A. Anders, Brittany R. White, Nadica Stoilova, Justin D. Fowlkes, Kate E. Landry, Leona G. Chemnick, Oliver A. Ryder, Mark A. Batzer. 2005. Alu insertion loci and platyrrhine primate phylogeny. *Molecular Phylogenetics and Evolution*.

Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol Genet Metab* **67:** 183-193.

Dermitzakis, E.T., A. Reymond, and S.E. Antonarakis. 2005. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet* **6:** 151-157.

Dermitzakis, E.T., A. Reymond, N. Scamuffa, C. Ucla, E. Kirkness, C. Rossier, and S.E. Antonarakis. 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science* **302:** 1033-1035.

Dewannieux, M., C. Esnault, and T. Heidmann. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35:** 41-48.

Ganguly, A., T. Dunbar, P. Chen, L. Godmilow, and T. Ganguly. 2003. Exon skipping caused by an intronic insertion of a young Alu Yb9 element leads to severe hemophilia A. *Hum Genet* **113:** 348-352.

Gilbert, N., S. Lutz-Prigge, and J.V. Moran. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.

Greally, J.M. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci U S A* **99:** 327-332.

Hagan, C.R., R.F. Sheffield, and C.M. Rudin. 2003. Human Alu element retrotransposition induced by genotoxic stress. *Nat Genet* **35:** 219-220.

Han, J.S., S.T. Szak, and J.D. Boeke. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429:** 268-274.

Hedges, D.J., P.A. Callinan, R. Cordaux, J. Xing, E. Barnes, and M.A. Batzer. 2004. Differential alu mobilization and polymorphism among the human and chimpanzee lineages. *Genome Res* **14:** 1068-1075.

Jinchuan Xing, H.W., David A. Ray, Cheney Huang, Oliver A. Ryder, Mark A. Batzer. 2005. Mobile element based phylogeny for Old World Monkeys. *Genetics*.

Jorde, L.B., W.S. Watkins, M.J. Bamshad, M.E. Dixon, C.E. Ricker, M.T. Seielstad, and M.A. Batzer. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* **66:** 979-988.

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

Jurka, J., O. Kohany, A. Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* **101:** 1268-1272.

Jurka, J., M. Krnjajic, V.V. Kapitonov, J.E. Stenger, and O. Kokhanyy. 2002. Active Alu elements are passed primarily through paternal germlines. *Theor Popul Biol* **61:** 519-530.

Justice, C.M., Z. Den, S.V. Nguyen, M. Stoneking, P.L. Deininger, M.A. Batzer, and B.J. Keats. 2001. Phylogenetic analysis of the Friedreich ataxia GAA trinucleotide repeat. *J Mol Evol* **52:** 232-238.

Kajikawa, M. and N. Okada. 2002. LINEs mobilize SINEs in the eel through a shared 3' sequence. *Cell* **111:** 433-444.

Kidwell, M.G. and D.R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int J Org Evolution* **55:** 1-24.

Kyudong Han, J.X., Hui Wang, Dale J. Hedges, Randall K. Garber, Richard Cordaux, Mark A. Batzer. 2005. Extended retrotranspositonal quiescence supports a "Back Seat Driver" model of Alu amplification. *Genome Res*.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001a. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001b. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Leeflang, E.P., W.M. Liu, I.N. Chesnokov, and C.W. Schmid. 1993. Phylogenetic isolation of a human Alu founder gene: drift to new subfamily identity [corrected]. *J Mol Evol* **37:** 559-565.

Lev-Maor, G., R. Sorek, N. Shomron, and G. Ast. 2003. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* **300:** 1288-1291.

Li, T.H. and C.W. Schmid. 2004. Alu's dimeric consensus sequence destabilizes its transcripts. *Gene* **324:** 191-200.

Li, X., W.A. Scaringe, K.A. Hill, S. Roberts, A. Mengos, D. Careri, M.T. Pinto, C.K. Kasper, and S.S. Sommer. 2001. Frequency of recent retrotransposition events in the human factor IX gene. *Hum Mutat* **17:** 511-519.

Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.

Liu, W.M., W.M. Chu, P.V. Choudary, and C.W. Schmid. 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* **23:** 1758-1765.

Liu, W.M., R.J. Maraia, C.M. Rubin, and C.W. Schmid. 1994. Alu transcripts: cytoplasmic localisation and regulation by DNA methylation. *Nucleic Acids Res* **22:** 1087-1095.

Lutz, S.M., B.J. Vincent, H.H. Kazazian, Jr., M.A. Batzer, and J.V. Moran. 2003. Allelic heterogeneity in LINE-1 retrotransposition activity. *Am J Hum Genet* **73:** 1431-1437.

Malik, H.S. and T.H. Eickbush. 2001. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. *Genome Res* **11:** 1187-1197.

Mathias, S.L., A.F. Scott, H.H. Kazazian, Jr., J.D. Boeke, and A. Gabriel. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254:** 1808-1810.

Medstrand, P. and D.L. Mager. 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. *J Virol* **72:** 9782-9787.

Moran, J.V., R.J. DeBerardinis, and H.H. Kazazian, Jr. 1999. Exon shuffling by L1 retrotransposition. *Science* **283:** 1530-1534.

Moran, J.V., S.E. Holmes, T.P. Naas, R.J. DeBerardinis, J.D. Boeke, and H.H. Kazazian, Jr. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87:** 917-927.

Murata, S., N. Takasaki, M. Saitoh, and N. Okada. 1993. Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. *Proc Natl Acad Sci U S A* **90:** 6995-6999.

Myers, J.S., B.J. Vincent, H. Udall, W.S. Watkins, T.A. Morrish, G.E. Kilroy, G.D. Swergold, J. Henke, L. Henke, J.V. Moran et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71:** 312-326.

Neafsey, D.E., J.P. Blumenstiel, and D.L. Hartl. 2004. Different regulatory mechanisms underlie similar transposable element profiles in pufferfish and fruitflies. *Mol Biol Evol* **21:** 2310-2318.

Okada, N. 1991. SINEs. *Curr Opin Genet Dev* **1:** 498-504.

Okada, N. and M. Hamada. 1997. The 3' ends of tRNA-derived SINEs originated from the 3' ends of LINEs: a new example from the bovine genome. *J Mol Evol* **44 Suppl 1:** S52-56.

Ostertag, E.M., J.L. Goodier, Y. Zhang, and H.H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73:** 1444-1451.

Ostertag, E.M. and H.H. Kazazian, Jr. 2001. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35:** 501-538.

Pauline A. Callinan, P.L., Jianxing Wang, Scott Herke, Randy Garber, Mark Batzer. 2005. Alu Retrotransposition mediated deletion: A novel mechanism creating genetic instability in primate genomes.

Peaston, A.E., A.V. Evsikov, J.H. Graber, W.N. de Vries, A.E. Holbrook, D. Solter, and B.B. Knowles. 2004. Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. *Dev Cell* **7:** 597-606.

Perepelitsa-Belancio, V. and P. Deininger. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet* **35:** 363-366.

Roos, C., J. Schmitz, and H. Zischler. 2004. Primate jumping genes elucidate strepsirrhine phylogeny. *Proc Natl Acad Sci U S A* **101:** 10650-10654.

Salem, A.H., G.E. Kilroy, W.S. Watkins, L.B. Jorde, and M.A. Batzer. 2003a. Recently integrated Alu elements and human genomic diversity. *Mol Biol Evol* **20:** 1349-1361.

Salem, A.H., D.A. Ray, J. Xing, P.A. Callinan, J.S. Myers, D.J. Hedges, R.K. Garber, D.J. Witherspoon, L.B. Jorde, and M.A. Batzer. 2003b. Alu elements and hominid phylogenetics. *Proc Natl Acad Sci U S A* **100:** 12787-12791.

Sarkar, A., C. Sim, Y.S. Hong, J.R. Hogan, M.J. Fraser, H.M. Robertson, and F.H. Collins. 2003. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. *Mol Genet Genomics* **270:** 173-180.

Schmid, C.W. 1998. Does SINE evolution preclude Alu function? *Nucleic Acids Res* **26:** 4541-4550.

Schmitz, J., M. Ohme, B. Suryobroto, and H. Zischler. 2002. The colugo (Cynocephalus variegatus, Dermoptera): the primates' gliding sister? *Mol Biol Evol* **19:** 2308-2312.

Schmitz, J., M. Ohme, and H. Zischler. 2001. SINE insertions in cladistic analyses and the phylogenetic affiliations of Tarsius bancanus to other primates. *Genetics* **157:** 777-784.

Shankar, R., D. Grover, S.K. Brahmachari, and M. Mukerji. 2004. Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. *BMC Evol Biol* **4:** 37.

Shen, M.R., M.A. Batzer, and P.L. Deininger. 1991. Evolution of the master Alu gene(s). *J Mol Evol* **33:** 311-320.

Shimamura, M., H. Yasue, K. Ohshima, H. Abe, H. Kato, T. Kishiro, M. Goto, I. Munechika, and N. Okada. 1997. Molecular evidence from retroposons that whales form a clade within even-toed ungulates. *Nature* **388:** 666-670.

Sijen, T. and R.H. Plasterk. 2003. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature* **426:** 310-314.

Smit, A.F. and A.D. Riggs. 1996. Tiggers and DNA transposon fossils in the human genome. *Proc Natl Acad Sci U S A* **93:** 1443-1448.

Sorek, R., G. Ast, and D. Graur. 2002. Alu-containing exons are alternatively spliced. *Genome Res* **12:** 1060-1067.

Sverdlov, E.D. 2000. Retroviruses and primate evolution. *Bioessays* **22:** 161-171.

Ullu, E. and C. Tschudi. 1984. Alu sequences are processed 7SL RNA genes. *Nature* **312:** 171-172.

van Rijk, A. and H. Bloemendal. 2003. Molecular mechanisms of exon shuffling: illegitimate recombination. *Genetica* **118:** 245-249.

Vidaud, D., M. Vidaud, B.R. Bahnak, V. Siguret, S. Gispert Sanchez, Y. Laurian, D. Meyer, M. Goossens, and J.M. Lavergne. 1993. Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. *Eur J Hum Genet* **1:** 30-36.

Wallace, M.R., L.B. Andersen, A.M. Saulino, P.E. Gregory, T.W. Glover, and F.S. Collins. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353:** 864-866.

Waterston, R.H. K. Lindblad-Toh E. Birney J. Rogers J.F. Abril P. Agarwal R. Agarwala R. Ainscough M. Alexandersson P. An et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

Watkins, W.S., C.E. Ricker, M.J. Bamshad, M.L. Carroll, S.V. Nguyen, M.A. Batzer, H.C. Harpending, A.R. Rogers, and L.B. Jorde. 2001. Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. *Am J Hum Genet* **68:** 738-752.

Watkins, W.S., A.R. Rogers, C.T. Ostler, S. Wooding, M.J. Bamshad, A.M. Brassington, M.L. Carroll, S.V. Nguyen, J.A. Walker, B.V. Prasad et al. 2003. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res* **13:** 1607-1618.

Wei, W., N. Gilbert, S.L. Ooi, J.F. Lawler, E.M. Ostertag, H.H. Kazazian, J.D. Boeke, and J.V. Moran. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21:** 1429-1439.

Yoder, J.A., C.P. Walsh, and T.H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13:** 335-340.

Zietkiewicz, E., C. Richer, D. Sinnett, and D. Labuda. 1998. Monophyletic origin of Alu elements in primates. *J Mol Evol* **47:** 172-182.

APPENDIX A: SUPPLEMENTARY DATA

Sex chromosome *Alu* elements, Genbank numbers, PCR primers and conditions, human diversity and amplicon sizes.

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

		138008600						
Ya5DP65	AL512286	chrX:119941032-120032906	CAGGCTGACCACACAATCAT	GCTACAAGGGAAAACTGGCA	55	FP	456	159
Ya5DP15	AL451103	chrX:34868434-35043817	CAGGCTTGCACAAATATCCA	TTATATGAAGCACATTGAAGAAATG	55	FP	445	139
Ya5NBC326	AL133500	chrX:70223216-70424625	CCAAGAGACCACTTCCTATTTCA	AATGGGGGAGAGGACAGTCT	60	FP	539	216
Ya5 489	Z81364	chrX:130766117-130842210	CCATTATGACCAGTTGTGTGTTG	CCGGCCAAAAGCATTGTA	55	FP	433	115
Ya5 467	Z92844	chrX:42519788-42671585	CCCCTCCTCAGTTTTTGGAT	GGCTTAATAGCCAAGAGAGTGC	60	FP	400	85
Ya5 417	AF067122	chrX:155561893-155628434	CCTTCCCATAAACCCACTGA	CCAAAATTTGCTCCATGTTG	55	FP	441	121
Ya5NBC344	AL109853	chrX:132277087-132383551	CGTGAGAAAGCATAGGCAAC	TCCTTTCCTTATGCCTGCAA	60	FP	472	158
Ya5DP13	AC004470	chrX:21230949-21438905	CTATAGAGCCAAGCCTGATACTCTG	GTATGGGGAATGTGACAAGGAG	60	HF	430	141
Ya5DP18	AF241732	chrX:38416627-38459556	CTCAGTGTTCCCTCCTCTGG	ATGCGCTATGTCTTTTTGGG	55	FP	879	554
Ya5NBC80	AL590410	chrX:54568403-54757014	CTCTCCTGTGTCCATACTTCTT	CTGGCATGGAGATTTCTTAC	60	FP	368	47
Ya5DP88	AC005731	chrX:151553784-151697727	CTGAACCAAACTGGAAGGGA	GATTCACGTTGCACTTTTACCA	55	FP	490	175
Ya5DP5	AC019219	chrX:6134097-6314114	TATATGGGTAAAGATCCAAAGCAAGG	AGAATAATGCCTTAGCATTCAGCAG	60	FP	293	115
Ya5DP62	AL049591	chrX:114555491-114677890	GAATGAATGCAATGCCTAAGGT	AACCTATCTAGGGAGACCAGCAG	60	HF	410	115
Ya5DP77	AL356785	chrX:140674109-140839680	GAAGGATGATCTCTCCTTAC	TGCAAGGAGAGTTGGCATAA	55	HF	620	298
Ya5DP86	AL109654	chrX:148555591-148737740	GAGTAGTGTACATGAGGGGTTAT	AGGGCTGAGACAGTGTCTTC	55	FP	657	327
Ya5DP76	AL353788	chrX:138017665-138180403	GCAAATGTTCATTAAGAAAGCTGA	ATGGATTTTTGCTCTGCCC	55	FP	485	163
Ya5 455	AC002368	chrX:151258956-151583771	GCAACTTTCCCATGTTTTCC	TGGATGCAAGGTCTAAATTCG	55	FP	416	114
Ya5NBC170	Z94722	chrX:92120551-92227389	GCAAGACCTGTGTGTATGCTTAAAT	GAGAGTACACGAAAATACAGGCTTT	60	FP	521	195
Ya5 425	AL022166	chrX:54807015-54936240	GCACAGACAAGCTGCTCAAG	GAAGCCTGGCATGGAGATT	60	FP	431	110
Ya5DP53	AL359641	chrX:98554165-98729296	GCCAGGAACAGACAAGGTGT	TTGCCTTTTGGTGTTGTTCA	55	FP	490	177
Ya5DP40	AL031116	chrX:86290983-86441140	GCCTCATCCTGTACCATACTCC	TCCCACACTATTCTGATTTCTTCTT	55	FP	482	161
Ya5DP52	AL390027	chrX:98223595-98423785	GCCTGAGATGTGGGAGTAAAC	CAGCCTTCAAACTTGCACCT	55	FP	423	293
Ya5NBC37	AC002476	chrX:120184952-120332053	GCTTGAGGTTTTCATACTACTCTTATCTTT	ACTGTATAAGCATTTTCCTCTTTATCTTTC	60	IF	497	184
Ya5DP61	AL121878	chrX:114065698-114188586	GCTTTCTGCAGCAAAACTCA	CAGATGGCAAGAGCCTGAA	55	FP	684	370

Ya5NBC98	AL049591	chrX:114555491-114677890	TATAGCTAGTAAATGGTAGAGCCAGGA	CTGTCTAAGATAGTGATTGGACCTACTATG	55	HF	504	209
Ya5DP84	AL445258	chrX:147855595-148031077	GGAGCTGCAGGAGTTGTCTT	CCAGGAGCAGGAGAGAACAA	55	FP	496	173
Ya5 477	Z92844	chrX:42519788-42671585	GGCTTAATAGCCAAGAGAGTGC	AACCCCTCCTCAGTTTTTGG	55	FP	400	87
Ya5DP70	AL023799	chrX:130812222-130905926	GGGGAATGAGAGGGAAATGT	AAGACAGCCAAAATTCAGTTAAAAA	55	FP	1190	868
Ya5DP12	AC017058	chrX:19068390-19241039	GGGTTGATTTAGTGGCCCTT	TCCTTTCAGATTTTCGTGGG	55	FP	374	59
Ya5DP97	AC011142	chrX:12380392-12557081	TACTATATCCCCCATGCCCA	ACTTGGTCCTCTCTCCAGCA	55	FP	1075	749
Ya5DP59	AL360224	chrX:109503420-109660581	TAGAGAATGAGGGTGGCTGG	TCGTGACCTTAGCACATGGA	55	FP	472	158
Ya5NBC99	AL031312	chrX:146122637-146208640	TATACACACACACACAGAGAATGACTG	CCTGACTCGAAAGTACTGTTTTCTAAG	55	FP	515	198
Ya5DP22	AL590223	chrX:47743014-47959685	TCTAAACCTGCCCTAGCTAGATACC	TCCTTTCTCAAAACTGCTTTCC	60	FP	516	190
Ya5DP56	Z70051	chrX:104660637-104705312	TGAAGATGTTTCTCTCCCCAG	AGTGGAAGAGAAAGGGTGGG	55	FP	487	374
Ya5DP68	AL391002	chrX:126496085-126581721	TGATTTCACTATGAAACCCACTC	TGAAGGACTCAAAATTTTCCAC	55	FP	405	89
Ya5DP66	AC002377	chrX:120825392-120967170	TGGACTGCTATCTCACGCTG	TTGGTTTTCTGGCAAGTTCC	55	FP	938	624
Ya5DP41	AL137015	chrX:86883045-86982571	TGGAGACATGAATACATTTTAGACA	CCAACAGATTTCACTTTTTGCTT	60	X/Y	464	149
Ya5DP83	AL445258	chrX:147855595-148031077	TGGATTAAATACAGGCAGAAAGC	TGCAGCAAAGATCTTCCAGA	55	FP	478	164
Ya5DP6	AC073533	chrX:6458416-6640471	TGGGTGTTTGCATCAAGAAA	GCAGGCAGAGAGGACAGGTA	55	FP	731	412
Ya5DP44	AC004072	chrX:90436734-90607391	TGTCATCTTTATCTGCCTTGGA	ACGGAGATTCTGCTTCAACAA	55	X/Y	398	89
Ya5 466	AC002377	chrX:120960081-121101859	TGTCTTACAACTCCCCACTCAA	CCTGGCTCTTCCAAGTTAGG	60	FP	426	94
Ya5DP34	AL359885	chrX:79179019-79255815	TTAGGTCACCTCTCCCTTGC	CAAGTGTCTGCAAAAAGGCA	55	FP	1131	800
Ya5DP82	AL512285	chrX:146753057-146823003	TTTAAAAACATAACCCAGTTGAAAA	CACCCATTAATTCACTACCCAA	55	FP	1084	785
Ya5DP54	AL355593	chrX:98735910-98903974	TTTAAAGAAAGCCTGTGATGGA	AAATGAATTGGCCCACCTTT	55	FP	493	178
Ya5DP57	Z83850	chrX:105136491-105269471	TTACCTCAACAGTGACATAACAGCA	ATAGTGAAGCAGAGAACTGTTGGTT	60	HF	652	349
Ya8BGK21	AC016678	chrY:18083142-18225923	AATATCCACCAAGAACAGAAGCTTTAG	AATCTTTGACTAGGCCCTGTAAGTT				
Yb8DP1	AC079824	chrX:29704853-29824238	TCACCAATTATCCTCCTCCA	CGAGATGAATAAACACTGCACA	60	FP	442	235
Yb8DP2	AL049643	chrX:32572391-32691085	TCCTTTTATAAATTGGACAGAAAGC	TTCAAATGTCCAGCCAATTG	60	IF	400	48
Yb8DP3	AC022212	chrX:38096933-38284245	TTGTATTCCAGGGATCAGGC	GGGAGCCTGGGATTTTAGAG	60	FP	465	111

Yb8DP4	AC091810	chrX:39109332-39209804	TGGACTCCCACTGAGATGTG	ACTCACCCGCTAATTGTGCT	60	FP	499	145
Yb8DP5	AL023875	chrX:41894031-42016355	CCTTAATTTTGTTTCCCGCA	TTCACAGCTGGATCAGTTCAA	60	FP	451	102
Yb8DP7	AL034370	chrX:43613478-43733422	AAATGGTGGAAAAGATGCCA	CCCATCACAACTGTACCCAA	60	FP	485	119
Yb8DP8	AF196779	chrX:49459890-49643885	GAACTTAGAGAGAGCTAGTC	GTGCATCTTAGTATGAACTC	62	FP	673	358
Yb8DP9	AC078991	chrX:3366309-3536127	GAGACAGAGGCTACATGTGA	AACAGCAAATGAAATCGCCT	60	FP	1039	692
Yb8DP12	Z82211	chrX:56385934-56518162	ATGGACATCTCTGGTACGGC	CTAATTCCCCTGGCTGCATA	55	FP	489	151
Yb8DP13	AL158016	chrX:65925564-65996226	TAGGTTCATGAAGGCAAGGG	TGTCAATTAGAAGGCCTGGG	55	FP	479	258
Yb8DP18	Z98255	chrX:74382876-74552873	CAGTCTGTCTTCAGACCAGA	AGAAATGAATTAACGTGGC	62	FP	1026	626
Yb8DP22	AL358796	chrX:71193981-71539035	CTGGGGAAACAGACATAGTC	ACTTAGTGGACCTTCGTGGA	59	FP	727	485
Yb8DP25	AL591431	chrX:78222054-78373070	TGATGGGCATCACTGAAATC	CATTCTTAATGGGCCAATTTCT	60	FP	482	137
Yb8DP27	AL590031	chrX:78671333-78816485	TCATGCTGGAAAGGGCTATT	GCTTCCCACCTGAGCTAACA	60	FP	433	79
Yb8DP36	AL590043	chrX:94963848-95106616	AGTCAGTGACACCCACATGC	TGATGGAAGGATTTAAGCCAA	55	FP	500	142
Yb8DP38	AC003048	chrX:8164628-8205708	TACTGAGGCCATCGAGGAAC	CTCTCCTCACATCCCCGTAT	58	FP	491	145
Yb8DP39	AC002349	chrX:9399852-9559714	TGCAGATCTTATCAGCACATTG	ATTCATCCACCATCAGGGAA	55	FP	454	89
Yb8DP42	AC002449	chrX:113337879-113511645	GAAACCCAGTTTCACCATTTG	CAATGCATCTGTACCATGCTA	55	FP	670	318
Yb8DP43	AC005000	chrX:114817798-114925111	CCAAGGCAATCAATTTAGCC	TTCAAGATGCAGTCACTCGG	55	FP	897	544
Yb8DP44	AL357562	chrX:121846492-121975456	TTCATGTGGGCTTTTTGTGA	CAGCAAATTGTTCACAGTCCA	55	FP	471	123
Yb8DP45	AC002981	chrX:10814208-10967775	CCATCAATACATCGCTGGAA	TGTTCACCACCTTTCAACCA	62	FP	478	135
Yb8DP49	Ac002422	chrX:129115374-129275464	GACTAGGGGTTTGTGCCAGA	TCCCCCATTTCTGTTGTTGT	57	HF	459	138
Yb8DP51	AL138745	chrX:129973729-130197972	GCTTGCAACCTTACTGCCTC	GACAAAGCCTGAAGCCACTT	60	FP	414	68
Yb8DP52	AL022162	chrX:130258976-130259910	TGGGGGCACTTTACTAGGAT	CCACAGCTGGAGAACACTGA	60	FP	399	51
Yb8DP55	AL034400	chrX:133947057-134088818	GTGCTGCTGTAGCATTGCAT	GAAAGAACAGAGAACAGCCCA	60	FP	488	134
Yb8DP56	Z97196	chrX:134723484-134812365	AGACACCATCTGTGGGAAGG	ATTAAGGGCACTGTGCAACC	60	FP	461	120
Yb8DP58	AL390879	chrX:137999986-138172265	GTGGATGCCATTTTGGCTAC	TCCTTCATAGCCCGTAATGC	60	FP	494	161
Yb8DP59	AL022576	chrX:138442263-138579373	CTTGTGGGGACAACACTCCT	CTTCCTTCCACAGCCATTGT	60	FP	829	469

Yb8DP61	AL356785	chrX:140831401-140996972	GAGTAGCTACGTAAATACCC	TCCACACTTCATTCAAAGCC	59	FP	523	176
Yb8DP63	AL109653	chrX:147856085-148017425	CCCCTTCCTCTCACATAGCA	TTTATTCCCCCATTCCACAA	60	FP	1180	830
Yb8DP64	AC079383	chrX:12531947-12683222	CGTTTTCTATTTCCCACCACA	CCAACATTTTTCCTCCAAGG	55	FP	318	74
Yb8DP65	AC002524	chrX:13210194-13412733	CAGCTAGGCCTTGGAGATCA	TGCAAGCCAAATGAAAGAAA	55	FP	472	127
Yb8DP68	AF030876	chrX:157681205-157793960	CAAAGTCCTGTTGCGTACCTC	GCTGATGGCTACAACCCTGT	55	FP	953	630
Yb8DP70	AC078993	chrX:15756369-15970369	TTTGAATCAATATGTATATGGTGGA	CAGTTCCCATGACTTGGCTT	55	FP	437	71
Yb8DP76	AL592043.	chrX:33755847-33940359	GAGGCTAATATCAGCAAGCCA	TGTTTCAGCCAAAGAATGGA	60	FP	477	146
Yb8DP79	AL035088-AC016681	chrX:107155092-107301449, chrY:5852850-5921375	AGATTTCCAGAGGGAGCCAT	TTTCAACAGTCTTCTTTCGCA	60	X/Y	428	96
Yb8DP80	AL137065	chrX:107787396-107906706	CCATGATCATTTCCCTGACC	CCTGTCTGTTCTGCTTCTTTGG	57	FP	458	126
Yb8DP81	AC008162	chrX:120517329-120638169	CAGTTTCCTGGGTCCTGTGT	CAAGGCTTCCAGCTTAGGAA	57	FP	460	128
Yb8NBC8	Z98950	chrX:143336947-143460502	AAGAAAACTGATGGGGAAAG	CCAACTAGAGAAACGGAGAA	60	FP	599	198
Yb8NBC30	Z95124	chrX:84348492-84423053	TTGCCTTGGATGGCATATCT	AAATGGCCGGAGTAAGTCCT	55	IF	497	194
Yb8NBC38	AC002367	chrX:27624355-27772954	CGAGAGAAAGGGGTAGAAAGC	AATGCCTTCCAAGGACATCTT	60	FP	480	311
Yb8NBC62	AL031368	chrX:28485260-28629149	TGCCACACATTGTTCTAGGC	TGCCAACTATTGGAGGAGATG	45	FP	548	307
Yb8NBC75	Z68328	chrX:104956504-105000946	CCCACTGTGTTTATTGTTCC	GCTAAAGTACCCAGACCAAG	60	FP	519	200
Yb8NBC102	AL049591	chrX:114555491-114677890	TATAGCTAGTAAATGGTAGAGCCAGGA	CTGTCTAAGATAGTGATTGGACCTACTATG	60	HF	504	209
Yb8NBC133	Z84470	chrX:74641008-74790533	GCCATTGATCCCACAGAAAT	GCTGTGAATTCGTTGGTCCT	55	FP	536	232
Yb8NBC170	AL109653	chrX:147856085-148017425	TCCCCAAAGAAGGAGAGACA	TTCCCCCATTCCACAATTTA	60	FP	599	275
Yb8NBC221	AL034370	chrX:43613478-43733422	AATTCAAGCCAATGAACCAC	TCAGTGCTCTGAAGAAGCTCA	60	FP	431	97
Yb8NBC239	AF031078	chrX:157681205-157793960	TTGCTGACAGATCAGGGATG	TCCCCCTTCAAACCTATTCC	55	FP	730	419
Yb8NBC242	AC002349	chrX:9399852-9559714	ATCCACCATCAGGGAATCAA	TGCAGATCTTATCAGCACATTG	60	FP	450	117
Yb8NBC246	AC002981	chrX:10814208-10967775	CACCACCTTTCAACCAGGAA	ATCGCTGGAATGTGGTTCTC	60	FP	464	149
Yb8NBC247	AC002366	chrX:10014142-10273343	GCAGCACAAAGTAGTGGTTGG	TGCACCCACTTGATATGCTT	60	FP	551	259
Yb8NBC256	Z73986	chrX:100506131-100636835	CCCACAATTTCCACTTCAGG	GCATTGCTTCCCTTCTATTTC	55	FP	503	24

Yb8NBC269	AC091810-AF241734	chrX:39109332-39209804, chrX:38989413-39109331	CACGCTTAACCTCTACCACCA	TGGACTCCCACTGAGATGTG	60	FP	587	261
Yb8NBC483	AC012078	chrX:88400703-88612982, chrY:3556128-3732799	GGCCAAGAGCATTCCAAAAT	GCCAATTGGTCAGGGTACAA	58	X/Y	744	422
Yb8NBC578	AL159988	chrX:146926640-147038418	TTTTTGCAGATGCTTCCCTA	CCCTTGATCCAGATGTGATG	55	IF	380	72
Yb8NBC594	AC087225	chrX:158577659-158680667	AGCAGGTGGTTAGGTCTTGG	CAGGGGGAGGGAACATTAAC	60	FP	428	103
Yb8NBC613	AL158201	chrX:66488109-66630063	GTCGCTTCACCTTGCACTTT	CAATCTGTGAAGGCTGAGGA	55	IF	459	124
Yb8NBC634	AL390840	chrX:92693201-92890811	AACAGAAAGGCATCATTTGC	GGGGGCATTTATTACTGCTT	55	IF	420	95
Yb9DP1	AL050305	chrX:32824774-32964031	TGACGACAAAGCACAAGGAC	TGGGGAGAATTTTACAAAACTAGG	60	FP	499	165
Yb9DP10	AC002477	chrX:119582373-119706467	CCAATTCCACAAAGGCAAAT	TTAGCTGCCTGACACGTCC	62	FP	1144	825
Yb9DP13	AF277315	chrX:158097366-158244593	ATGGAAACTGCACAGAGAGG	CTCTCTGGGCAGACCACG	62	FP	620	531
Yb9NBC251	AC002477	chrX:119582373-119706467	CGGCCCTGATATGTCTTTGA	TCCACAAAGGCAAATGGATA	60	FP	838	500
Yc1DP2	AL353136	chrX:64692940-64885444	GGCCTATATTGCTATCACGCA	TTTTCTCTCAGGTTCTCTGTAAACT	60	FP	1050	721
Yc1DP4	AL357752	chrX:68485370-68664236	AAACATGGGAGGGAGGAAAG	GCTCAGAAACTCCCAACCAG	60	FP	486	318
Yc1DP5	AL121601	chrX:123991202-124124592	CAACCAGAGATCTTAAAATGTGA	TCAGCGTGAGAGCCCATATT	60	FP	452	330
Yc1DP7	AL031054	chrX:144887772-145086787	GACCCCAAAGGTTCAAGTCA	GCATGCCCACTAGCAGTGTA	60	FP	1072	731
Yc1DP8	AJ239323	chrX:50201890-50304742	CAATTTCCTGGCATTTGGAG	TTCAAGATGCAGTCACTCGG	60	FP	345	62
Yc1DP10	AJ239320	chrX:69939181-70231674	CACTTTTTCTTATTTGGCCCAG	ATGGGCAATTCAATGTTTCC	60	FP	428	65
Yc1DP11	Z75741	chrX:128441659-128443464	AACCTCACATTTTCCAAAGGTA	TCTTGCTTCCTGAGTCGGTT	60	FP	691	380
Yc1DP13	AL137013	chrX:73134712-73280970	AGGCCTCAAAGTTTAGGGGA	ATCAAAGGGGAATACTGGGG	60	FP	424	338
Yc1DP14	AL049643	chrX:32572391-32691085	CCACTGCAGGCAGGATTATT	GCATGCCTGATTCCACACTA	60	FP	480	314
Yc1DP16	Z86061	chrX:95243418-95361328	AGCATGCAAGGAAAGGGATA	TTCTCAGTTTCCAATCTTAGGGA	60	FP	486	134
Yc1DP18	Z98046	chrX:53964263-54042043	CAAGGTTTGGGTTCTGCTGT	CATGGACACAGTGGTGAAGG	60	FP	412	81
Yc1DP21	AL589872	chrX:53255422-53447504	CTTGAAGCTGCTCAGTAAGG	TAGCCATATCCACACA	60	FP	567	240
Yc1DP22	AL049562	chrX:128109596-128200796	GCAAAACTTTGCGCTAATCC	ATGGGAAGCTTTCCCTGACT	60	FP	746	415
Yc1DP24	AL158819	chrX:54387419-54562331	GGGGAAATGGGCCTAGTAAA	AATCACCTTAACGCCACAGC	60	FP	470	142
Yc1DP26	AL096861	chrX:150067750-150197435	TGCAATAAAGAGTGTTCCTCTCC	CCCAAACTTGGTAGGTGAAAA	60	FP	482	147

Yc1DP27	Z83823	chrX:125012174-125121452	TCACGTCTCTCCTTTGCTCA	CTCTGGAAGCCTGCTATTGG	60	FP	1072	775
Yc1DP30	AL591431	chrX:78222054-78373070	TGCCTTACCCAATACACATTT	AAGGCAAAAGTCCATAAAGCA	60	FP	498	172
Yc1DP32	AL365179	chrX:61340404-61521254	CCAAAGGAGGTGGCTACTCA	GCACCCTGGTGAGAAATTGT	60	FP	422	73
Yc1DP34	AL356317	chrX:62409559-62514092	TGGATCTGCTATCAGAATGGAC	TTTGTGCAAAATAGGACCCTT	60	FP	499	194
Yc1DP35	AL031319	chrX:109958712-110057481	GCCTTGGGCTGCTATCATAA	GGGCAGAATAACGCAAGATT	60	FP	500	185
Yc1DP38	AL359854	chrX:61176831-61340403	CCAAAGGAGGTGGCTACTCA	GCACCCTGGTGAGAAATTGT	60	FP	423	113
Yc1DP39	AC073614	chrX:25176210-25306010	CCAACAGACAGCTTTCCACA	CAAGTCGAGGTTCTCCCTCA	60	FP	498	200
Yd3JX170	AC005000	chrX:114817798-114925111	GTGATTGCTACTGCTTTTTGCTT	ACCTGATGAACATTTTAGGAACC	60	FP	570	255
Yd3JX757	AL139396	chrX:52597320-52775770	CATTAGAAATCAGAATGGCTTCG	CTTGGTTTATTCCTTTGCTATGC	60	FP	549	250
Yd3JX437	AL034412	chrX:46070143-46177191	TGGTGTACCTTAGTCCAAAGACC	TTTGCATCTCAGAACTTTTTCCT	60	IF	547	235
Yd3JX545	U73479	chrX:20177044-20213072	AGGTTATGAAAGGGTCTGCTTTT	GATATTTGGACACACACACCTAAA	60	FP	680	355
Yd3JXD75	AJ239320	chrX:69939181-70231674	TGTACTTGCCCCATCTTCTGTAT	TATTCTGAAAATCTTGGGGGTGT	60	FP	546	226
Yd6JX284	AL591591	chrX:32998640-33102756	TTTCCTGATGGAAGCAGTGTATT	TGTTAGCATAATTGATCCCAAAAT	60	FP	517	200
Yd6JX56	AC079173	chrX:3673291-3838308	ATACTTACCATTGCCTCGTCCTT	ATGTCATGATCGGCTAGTTCTTG	60	FP	530	216
Ya5a2AD3	AC006371	chrY:15065526-15267679	TGGGGAAATCGATGATTTAAGA	AAGACAACGCACAATACCTTTGA	55	X/Y	421	117
Ya5AD585	AC006983-AC024067	chrY:24548626-24728770,chrY:27613358-27721503	TAAAATATTGCAAGGGGATGA	CCAGGTCTGTGTCTTATTTTCTTT	56	FP	867	536
Ya5AD586	AC006983-AC006338-AC010088-AC025735	chrY:24548626-24728770,chrY:26134406-26321043,chrY:24321000-24428486,chrY:25944031-26029302	ACGCAGAACCTGAAATTGTGATT	ACCATGCATAAATAGTGCCAACT	60	FP	524	181
Ya5AD588	AC026061	chrY:22174780-22194121	TGAGCGTCTAATGTGTTAATGAAA	CAAATACTTCAGCCTTGTCAAGAA	60	FP	500	193
Ya5AD589	AC010086	chrY:22595725-22766459	TGCACATACTGCTATTGATG	TGGCTATGCTTTCTTCATCT	55	FP	549	232
Ya5AD591	AC073893	chrY:25211889-25276138	TTGTATTAAAGCCCGTAAAATGG	AAGAATTATCTAGGACAGCTTTGG	55	FP	544	223
Ya5AD592	AC007965-AC007359-AC016752-AC008175	chrY:24895138-25061373,chrY:23324934-23425360,chrY:24895138-25061373,chrY:23742819-23947855	CATCGTGATGGTCTAGATTTCTTT	TTAAGGCATCGGATTCTTTCT	55	X/Y	685	268
Ya5AD593	AC024067	chrY:27613358-27721503	AATTAAAAGCACCCCCAAGA	CTCACCTTCTCTGCTTAACAAAA	60	FP	543	227
Ya5AD594	AC010153-AC016728	chrY:25840084-25944030,chrY:26321044-26472895	TGTTTCAGAGAGGACAGAAA	AGTGATTGCCTTGACATAGT	55	X/Y	459	148

	AC006983-AC006338-AC010088-	chrY:24548626-24728770, chrY:26134406-26321043, chrY:24321000-24428486,						
Ya5AD595	AC025735	chrY:25944031-26029302	ACGCAGAACCTGAAATTGTGATT	AACCATGCATAAATAGTGCCAAC	60	X/Y	524	182
Ya5AD597	AC023274-AC006328	chrY:25351695-25489176, chrY:26636925-26814493	GTTTGCTCAAGCCCATTAAA	TAAATGTATCCTGGCACCAT	55	X/Y	434	115
Ya5AD598	AC023274-AC007562	chrY:25351695-25489176,chrY:26814494-26951370	AACGCCAAACACAATGACAA	TTTGGCTGCATGAATGTGTT	55	X/Y	592	277
Ya5AD600	AC010094-AC002509	chrY:3732800-3851035, chrX:88482028-88624126	AAAACAGCACAACGTTTTAT	TCTCAAAGCTCTAGGTTAGTTGA	60	FP	396	293
Ya5AD601	AC009491	chrY:8539647-8680380	AGTGGAAGCCATAAAACAAA	ACATAATCCAAGCATGATCC	60	FP	398	299
Ya5AD602	AC006040	chrY:2500001-2686304	CCCAAACCAAAACTGTTACT	TTTGTTCCTGCAGTCAATCT	60	FP	492	291
Ya5AD603	AC006376	chrY:14752949-14924755	TGAGGGAAGAACATTAAGGCATA	AGGTAAGCCAGATCCAGTTTTTA	60	FP	508	189
Ya5AD604	AC010723	chrY:15580278-15754497	AGCTGAAAGAGGACATCAAT	TGATATTCACCAGGGATTCT	55	FP	489	159
Ya5AD606	AC019060	chrY:4618247-4734841	TCTAAGGCAAACATGAGCTT	GAACATCTTAGAGCCTTCAAA	55	X/Y	1038	374
Ya5AD607	AC010977	chrY:5716765-5852849	AACATCAATTTGAAAACCTAGA	TGAGGAACAAAGGTTTTGAC	55	X/Y	472	141
Ya5AD608	AC009491 AL121881 Z95703	chrY:8539647-8680380 chrX:142771104-142956654 chrX:143720946-143847097	ATGAAAACTGTTCAGGGAGATATT	TGGTTAATATCCTGAAGGCAAAA	55	X/Y	629	314
Ya5AD609	AC015978	chrY:18788855-18967434	TTGGAAAGTACACCATAACCACA	GCCCTACTTGTCCATTTTTCAAT	60	FP	505	184
Ya5AD610	AC068541-AC007379	chrY:19834150-19871515, chrY:20027673-20201554	GATGCATGGATGATACAATTT	TGCTCAAGCCCTTTATTATT	55	FP	549	303
Ya5AD611	AC010133	chrY:20609301-20761174	ATACCTGGAGCTTTTTGTCA	CACGCATAGTCACAAGTTTT	55	FP	551	228
Ya5AD612	AC010889	chrY:20958342-21138265	ACGATTTTCAGAGTTGAAGC	AACTCTTATTTGGAGGGACA	55	FP	542	231
Ya5AD613	AC006998	chrY:16704663-16848722	GGAAACTTAAAGGAAAGGCACAT	CAAATCTTAAGAAAGCCAGTGGA	55	FP	710	400
Ya5AD614	AC016678	chrY:18083142-18225923	TCAGAGAAAATCAAGAAATGC	GAGTGAAAAGGGTGAAAATG	55	FP	549	204
Ya5AD615	AC006999	chrY:18504136-18616813	TTGCACATTTCTTGTTTTCCA	AAATGTGGGGAAATTGGTTT	57	FP	879	549
Ya5AD617	AC007967	chrY:8680381-8867727	ACATGTATACACATAAGTACATGTG	AATGCCAATTATCCTGACTT	55	FP	472	169
Ya5NBC9	AC006382 AC005704	chrY:16848723-17011332 chrX:5295540-5394572	CTTCCCTAGGATTTAAGTCACCATAAAGAC	TTTTCAACTTGTAACTGTAGAGGACAGGAC	60	X/Y	415	102
Ya5NBC153	AC005820	chrY:14465010-14615919	CCAATCTGGGAATTATGACAAGTAG	CTTCAGACTTCTGCTTGATTTCTTC	60	FP	496	186
Ya5NBC155	AC006565	chrY:14420131-14465009	TGTCAATATCAGACAGATCCATGAG	ACTTCCAACTATGTGGTCAGTTTTG	60	X/Y	505	182

Ya5NBC156	AC002531	chrY:14120145-14316044	TGTGGTAAGTGTAGTTTCAAAAGAGTTT	TAATCTCTGGACTGGAAACATAAAA	55	FP	480	148
Ya5NBC172	AC006371	chrY:15065526-15267679	CCAAACGTAAGATTGAGTGG	AGTGGTGTTCTCGGTATTTC	55	FP	473	155
Ya5NBC174	AC006462	chrY:17011333-17151126	TCACTCTTTGTCTTGCTGACTACAG	GCTATAGCTTCTATTTACGGGGAAT	55	FP	526	206
Ya5NBC218	AC006989	chrY:16294804-16452269	AGCCCAACATCTGGTTTTGT	TCCAGTCTCGTGTAAAATAGCTTG	55	FP	445	109
Ya5NBC219	AC006989	chrY:16294804-16452270	CCTGGCAACCACCATTCTAC	AAACCTGGAGGGCATTCTTT	58	FP	445	129
Ya5NBC325	AC009479	chrY:3222117-3377215	CTTCTCTCTCTGAAATGCCAAT	CAGTTGAAAGGTTTGACAATACACC	60	FP	501	184
Ya5NBC413	AC006040	chrY:2500001-2686304	GGGCATTTTCAATCTCTCCA	ATGAAGTTGGAGGGGCAGAG	60	FP	435	119
Ya5NBC503	AC019099	chrY:27901323-28009655	GCTGAAAAGCTGACTGACACC	CAGAAAGGTTTCCCAGTTCG	55	FP	456	156
Ya5NBC508	AC010723	chrY:15580278-15754497	GGTAAAATCCCTCCTTTGAG	GAACTAATTGGGAGAGAGCA	55	FP	405	96
Ya5NBC509	AC010135	chrY:17664290-17841040	TGCTTGTATCAGCAGTCCTCA	CCCTCCATCCATCGAAAAAT	60	FP	390	76
Yb8AD687	AC007320-AC023342	chrY:23555125-23742818, chrY:23425361-23494514	CCAGGAGCTAGGTAATCAACATTT	TGGAAGGGGCAAATAAGAAA	58	FP	622	322
Yb8AD689	AC010723	chrY:15580278-15754497	AAGAATTTGCCAACACAGGTT	TTGTGCACAGGATGATTTGA	60	FP	834	516
Yb8AD690	AC010726	chrY:15782642-15958965	TTAACTAACATGGGCACCAA	AAAAATAGATTGCTCTCCTTCA	55	FP	465	166
Yb8AD693	AC010972	chrY:16532607-16647043	ATGAAATGTCAGCCTGATTC	CTCCCATGAAATGACAAGAT	60	FP	471	122
Yb8AD720	AC025227	chrY:23494515-23555124	TCCTTCTTTGATGGACTTTC	AAGCTATGGTATCAGGGTGA	55	FP	626	314
Yb8AD721	AC012067	chrY:5187228-5351534	TTCTGCCATAGATGAAGGAT	GTATGTGCATGCATCTGTGT	55	FP	533	201
Yb8NBC108	AC010089-AC053490	chrY:26029303-26132458, chrY:24428487-24531718,	TGTCACTTGATTGTCCGCATA	TCAATGGCATCCTGAAAACA	60	FP	550	194
Yb8NBC109	AC006371	chrY:15065526-15267679	GTGCAACTTCAGTTTCTGCTAAGAT	CATGGTTATCTGCAAAGACTATGAC	55	FP	532	212
Yb8NBC110	AC006383	chrY:14960516-15065525	AATAGGCTGAATGCCCCAAT	CTAGCATTGCAATCCCTGCTTT	60	X/Y	507	186
Yb8NBC111	AC007320	chrY:23555125-23742818	CCAGTGTCATCATCCAGACTTATTC	TACACACACACACATGCATTCTAAG	60	FP	531	192
Yb8NBC112	AC006999	chrY:18504136-18616813	GCATCTTAACCTAAATACCTGATGC	CAGGGACATAGGGTGTGAGTTACTA	60	FP	503	192
Yb8NBC114	AC004617	chrY:13889626-14035646	GGGTGAGATAGCTTAAGGAAAGAGA	AGATCTTCCCAAGAAGCCTTTC	60	FP	510	164
Yb8NBC160	AC007284	chrY:7139521-7310769	CCACACATGGGTACCAGTCC	TTGCTTACCCACAGTCACCTC	60	FP	404	72
Yb8NBC268	AC016681 AL590492	chrY:5852850-5921375 chrX:91254000-91383356	TGGGGATAGAGGAAGAAGACAA	CCTTTTCATCCAACTACCACTG	60	X/Y	517	188
Yb8NBC496	AC010977	chrY:5716765-5852849	CTGGGATAAAACAAGAGATAACAGG	GGTGTGCAGATTTTTGAGTCAT	60	FP	407	68
Yb8NBC507	AC021107	chrY:22887518-23048118	GGCCACGTTCTGTTCTTGTT	TACCGCCTGAACTCCACTTT	53	FP	805	484

	AC012667	chrY:5351535-5426338						
Yb8NBC535	AL133274	chrX:90732320-90828381	CTGAATAGAATCAGGGCAACA	CCATCTGGGAATAGTGTGGTG	60	X/Y	482	150
Yb9AD60	AC007678	chrY:21877693-21986665	GGAAAACTGAAAGAATCCACACA	TCAGATGCAGGCTTTCTAACTTT	55	FP	439	114
	AC024703	chrY:4241197-4272897						
Yb9NBC416	AL162723	chrX:88944751-89173981	GCCTTTTGAAGCTTCTGTCG	TGTTCCTTTGGTTAGGCAGA	59	X/Y	506	187
Yc1AD246	AC010154	chrY:6291830-6346980	TGGGTGGGGCCAAATAAAGAA	TGGGGTTTATTCCTTCAGATGTT	60	FP	589	269
Yc1AD250	AC011751	chrY:17903627-18083141	GGTATGCAAAAAGAAGTGCT	TTCAGATATGTGACCTGCTT	60	FP	472	167
Yc1AD254	AC010877	chrY:14615920-14752948	TGAGCAGAACAGAAAACACA	TGTGTGGCTAGCAAGTTATT	60	FP	445	139
Yc1AD255	AC011302	chrY:13382453-13560389	AGCCGTAGTTCACAATGTTT	CACAGGGTGCATATTTTCTT	60	FP	481	154
	AC017019	chrY:9394276-9556454						
Yc1NBC28	AC010154	chrY:6291830-6346980	TGGTGAGTTCCTGGTCTTGCTG	TGCTCACTCTTTGGGTCCACAC	60	FP	414	99
Yc1RG 243	AC006998	chrY:16704663-16848722	GGTCTGCTTACCAAATGACTGAG	ACATTCCTGATTCACAGAAGCTC	60	FP	424	136
Yc1RG242	AC007043	chrY:18396934-18504135	GCAGGACACACTTCCTGTTTCT	GTCCAGCACAGAAGAGGAATAAA	60	FP	416	96
Yc1RG244	AC017020	chrY:17266120-17432322	CCTAGAGGATTAGAGTCTGCCCTA	TATCCCCTAAAACTCATGTGTGG	60	FP	459	131
Yd6AD16	AC007247	chrY:7310770-7427357	TGACCCTAAATATACCTTCCA	AGCAACCTTGAGAAGAGTTTT	60	FP	436	127
Yd6AD17	AC007247	chrY:7310770-7427357	TGGATTCTTCCTCTTTTTGG	TTGGCTTCCCTGAGAAAATA	55	FP	575	265

[1]. Annealing temperature.

[2]. Allele frequency was classified as: high frequency polymorphism (HF), intermediate frequency polymorphism (IF), low frequency polymorphism (LF) and fixed present (FP) as previously defined by Carroll et al., 2001. X/Y indicates a homologous region on the X and Y chromosomes.

Some of the reported *Alu* elements were detected in multiple sequencing contigs suggesting that they are either paralogous elements or the result of sequence assembly artifacts

APPENDIX B: LETTERS OF PERMISSION

19 January 2005 Our Ref: HG/HDN/JAN05/J058

Dale Hedges
Louisiana State University
37315 HWY 75
Plaquemine, 70764
USA

Dear Mr Hedges

GENE, Vol 317, 2003, pp 103 – 110, P Callinan et al, "Comprehensive…"

As per your letter dated 6th January 2005, we hereby grant you permission to reprint the aforementioned material at no charge **in your thesis** subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

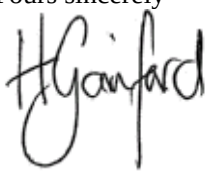
   "Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier".

3. Reproduction of this material is confined to the purpose for which permission is hereby given.

4. This permission is granted for non-exclusive world **English** rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form. Should you have a specific electronic project in mind please reapply for permission.

5. This includes permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

Yours sincerely



Helen Gainford
Rights Manager

ELSEVIER

19 January 2005 Our Ref: HG/HDN/JAN05/J062

Dale Hedges
Louisiana State University
37315 HWY 75
Plaquemine, 70764
USA

Dear Mr Hedges

ANALYTICAL BIOCHEMISTRY, Vol 312, No 1, 2003, pp 77 – 79, D Hedges et al, "A Mobile Element…"

As per your letter dated 6th January 2005, we hereby grant you permission to reprint the aforementioned material at no charge **in your thesis** subject to the following conditions:

1. If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies.

2. Suitable acknowledgment to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

   "Reprinted from Publication title, Vol number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier".

3. Reproduction of this material is confined to the purpose for which permission is hereby given.

4. This permission is granted for non-exclusive world **English** rights only. For other languages please reapply separately for each one required. Permission excludes use in an electronic form. Should you have a specific electronic project in mind please reapply for permission.

5. This includes permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

Yours sincerely

Helen Gainford Rights Manager

Sussman, Hillary

Subject: FW: republication question

Dear Mr. Hedges,

As per your request dated 27 January, we hereby grant you permission to reprint the article entitled "Differential Alu Mobilization and Polymorphism among the Human and Chimpanzee Lineages" (14:1068-1075) at no additional charge in your thesis, provided that the source article is suitably acknowledged as a footnote or reference list as follows: "Reprinted from Publication Title, Vol number, Author(s), Title of Article, Pages No., Copyright(Year), with permission from Cold Spring Harbor Laboratory Press.

Sincerely,

Hillary Sussman, Ph.D.
Executive Editor, Genome Research
500 Sunnyside Blvd.
Woodbury, NY 11797
Tel: (516)422-4014
Fax:(516)422-4092
hsussman@cshl.edu

Permission granted by the copyright owner, c[illegible]gent upon the consent of the original au[illegible]or, provided complete credit is given to the original source and copyright date.

By _____ 1/27/05
Date

COLD SPRING HARBOR LABORATORY PRESS

Dear Dr. Sussman,

I'm writing to inquire about how I would go about obtaining permission to reprint a complete, reformatted Genome Research article (14:1068-1075 2004) as part of my doctoral dissertation. The republication request forms on the cshl press website seemed more geared toward nonacademic situations. I was wondering if you could suggest who I might write/call to obtain a signed permission letter.

Thanks for you time,

Dale Hedges
dhedge1@lsu.edu
Batzer Laboratory
Louisiana State University
Department of Biological Sciences

VITA

Dale James Hedges is the son of Ernest and Linda Hedges. In the spring of 1994 he graduated from The Louisiana School for Math, Science, and the Arts in Nachitoches, Louisiana. He began his undergraduate studies at Duke university in Durham, North Carolina in the fall of 1994 and graduated with a Bachelor of Arts degree in biology and philosophy in 1998. After working for two years as a laboratory technician and analyst at the Duke University Center for Human Genetics, he began his doctoral research in the summer of 2001 in the department of Biological Sciences at Louisiana State University in Baton Rouge, Louisiana, under the direction of Professor Mark A. Batzer. Mr. Hedges will graduate with the degree of Doctor of Philosophy in May, 2005.