2014-07-11

Examining Selective Pressures Shaping Retrotransposon Distribution Within Genes

Sara B. Linker
University of Miami, sara.linker@gmail.com

Follow this and additional works at: https://scholarlyrepository.miami.edu/oa_dissertations

Recommended Citation
https://scholarlyrepository.miami.edu/oa_dissertations/1256

This Open access is brought to you for free and open access by the Electronic Theses and Dissertations at Scholarly Repository. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of Scholarly Repository. For more information, please contact repository.library@miami.edu.
UNIVERSITY OF MIAMI

EXAMINING SELECTIVE PRESSURES SHAPING RETROTRANSPOSON DISTRIBUTION WITHIN GENES

By

Sara Linker

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida
August 2014
A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

EXAMINING SELECTIVE PRESSURES SHAPING RETROTRANSPOSON DISTRIBUTION WITHIN GENES

Sara Linker

Approved:

Derek Dykxhoorn, Ph.D.
Professor of Human Genetics

Dale Hedges, Ph.D.
Professor of Human Genetics
Ohio State University

Juan Young, Ph.D.
Professor of Human Genetics

Eden Martin, Ph.D.
Professor of Human Genetics

Alexandra Wilson, Ph.D.
Professor of Biology

M. Brian Blake, Ph.D.
Dean of the Graduate School
Retrotransposons (RTs) comprise approximately half of the human genome and contribute to chromatin structure, regulatory motifs, and protein-coding sequences. Since RT insertions can disrupt functional genetic elements as well as introduce new sequence motifs to a region, they have the potential to effect the function of genes that harbour insertions as well as those nearby. A detailed characterization of RT density in genes could help inform predictions of the functional consequence of de novo as well as polymorphic insertions. The research presented here examines the selective pressures that modify the distribution of RTs within genes. It further examines these pressures by determining the association of RTs with evidence of differential gene expression. The findings of this study indicate that RTs are under purifying selection within genes. These results further indicate that factors that are associated with the RTs (e.g.: orientation with respect to the the gene, presence of a polyadenylation motif), and factors associated with the gene (e.g: tissue-specificity, coding sequence density) are modifiers of this selection. Moreover these results show that the presence of a RT is associated with differential gene expression, indicating that altered gene expression may be a potential mediator for the phenotypic change that results in the selection of RTs within genes. Future studies can use the results presented here to create a model for predicting the degree of impact of new RT insertions.
## Contents

List of Figures vi

List of Tables viii

1 INTRODUCTION 1

1.1 Retrotransposons ............................................... 2
1.2 Subfamilies of Retrotransposons ........................................ 7
1.3 Process of Retrotransposition ........................................... 10
1.4 Selection on the Sex Chromosomes .................................... 11
1.5 Selection and the Distribution of Retrotransposons ....................... 12
1.6 Antisense Bias of Retrotransposons in Genes .......................... 16
1.7 The Dynamics of Regulating Gene Expression .......................... 16
1.8 Variation in Gene Expression Between Humans and Non-human Primates ... 17
1.9 Retrotransposons as Non-coding Sources of Regulating Gene Expression ... 18
1.10 Structural Rearrangements Associated with Retrotransposition ......... 19
1.11 Phenotypic Effects of Retrotransposons ............................... 22
1.12 A New Path for Retrotransposon Research ............................ 24
1.13 Motivation ..................................................... 25

2 ORIGINS OF ANTISENSE BIAS 27

2.1 Retrotransposon Antisense Bias and the Process of Integration ........ 28
2.1.1 Shared Retrotransposons are More Common than Population-specific Elements ........................................ 29
2.1.2 Young Antisense and Sense Alu are Present in Equal Proportions ... 30
2.1.3 Antisense Bias is Present in DNA Transposons ........................... 32
2.2 Antisense Bias Increases Over Evolutionary Time ....................... 33
2.2.1 Antisense Bias of Alu is Constant Across Reference Subfamilies ........................ 33
2.2.2 Antisense Bias of L1 Increases across Reference Subfamilies ........ 33
2.2.3 Antisense Bias of Alu Increases Quickly Over Evolutionary Time .......................... 34
2.3 Evidence of Selection Acting on Alu and L1 ........................... 36
2.3.1 Selection of L1 within Genes ...................................... 37
2.3.2 Selection of Alu within Genes ...................................... 39
2.3.3 Evidence of Purifying Selection ..................................... 40
2.3.4 Evidence of Purifying Selection on Sense and Antisense Alu .................. 43
2.3.5 Evidence of Purifying and Positive Selection on Sense and Antisense L1 ........ 43
2.4 Conclusions ..................................................... 44
3 RT DISTRIBUTION IN GENES
  3.1 The Current Distribution of Alu and L1 within Genes ........................................... 46
    3.1.1 L1 and Alu Antisense Bias Exhibits a Linear Decay Across Gene Deciles .......... 47
  3.2 Gene-specific Variables Associated with Retrotransposon Distribution ................. 50
    3.2.1 Linear Decay of AS Bias is Associated with Coding Sequence Density .......... 50
    3.2.2 Alternative Splicing ......................................................................................... 50
    3.2.3 Alu and L1 are Decreased Near 5’-UTR ......................................................... 51
    3.2.4 Coding Sequence Accounts for Variability Near 5’-UTR ................................ 56
    3.2.5 Tissue-Specificity of the Gene Effects Antisense Bias ................................. 59
  3.3 Retrotransposon-specific Variables are Associated with Retrotransposon Dis-
     tribution ................................................................................................................... 60
    3.3.1 Truncated L1 Escape Depletion Near the 5’-UTR ............................................. 62
    3.3.2 Protein Binding Motifs in Alu ............................................................................ 65
    3.3.3 Protein Binding Motifs in L1 ............................................................................ 67
    3.3.4 Conservation at the Basepair Level of Binding Motifs ..................................... 68
  3.4 Interaction between Gene-specific and Retrotransposon-specific Variables .......... 71
    3.4.1 Presence of RT-contained Binding Motifs Based on Gene Type ...................... 72
  3.5 Conclusions ............................................................................................................... 74

4 SEX-SPECIFIC EXPRESSION ......................................................................................... 76
  4.1 Mobile Element Insertions between Humans and Non-human Primates ............... 77
    4.1.1 Gene Expression Decreased in Humans with a Mobile Element Insertion ......... 78
    4.1.2 Male Dependent Association between MEI Status and Expression .................. 82
    4.1.3 Analysis with Removal of NHP-specific Retrotransposons .............................. 83
  4.2 Mobile Element Insertions that are Polymorphic Between Humans ...................... 84
    4.2.1 Genes containing MEI-Alu are Differentially Expressed Between Male
         and Female Humans ............................................................................................. 84
  4.3 Replication of Sex-specific Results in Humans ....................................................... 85
    4.3.1 MEIs Segregating in Humans Associated with Decreased Expression .............. 86
    4.3.2 MEI-seg Associated Decrease in Expression is Specific to Males .................... 86
    4.3.3 Genes Containing MEI are Functionally Enriched for Autism Suscep-
         tibility Genes ...................................................................................................... 88
  4.4 Conclusions ............................................................................................................... 89

5 METHODS ...................................................................................................................... 90
  5.1 Chapter 2 methods .................................................................................................... 90
    5.1.1 Data Sources ...................................................................................................... 90
    5.1.2 Determining the Antisense Bias of Alu Retrotransposons ............................... 90
    5.1.3 Determining Ranks for Retrotransposon Subfamilies ....................................... 91
  5.2 Chapter 3 methods .................................................................................................... 91
    5.2.1 Data Sources ...................................................................................................... 91
    5.2.2 Binning Reference Locations ............................................................................ 92
    5.2.3 Identifying Genes with Secondary UTRs ......................................................... 92
    5.2.4 Distribution Analysis ......................................................................................... 92
    5.2.5 Tissue Specific Gene Sets ................................................................................. 93
    5.2.6 Motif Analysis .................................................................................................. 93
  5.3 Chapter 4 methods .................................................................................................... 94
5.3.1 Identifying Differential Mobile Element Insertions Between Humans and NHPs ........................................... 94
5.3.2 Expression in NHPs Versus Humans .............................................. 94
5.3.3 Expression in Humans versus Humans ........................................... 95

6 Discussion .............................................. 96
   6.1 Motivation .............................................. 96
       6.1.1 The System .............................................. 98
   6.2 Lack of Antisense Bias in Newly Integrated Elements ......................... 98
   6.3 Purifying Selection of Retrotransposons within Genes .......................... 99
   6.4 Deviations in Selection Between Alu and L1 .................................... 100
   6.5 Decrease of Retrotransposons Near Exons ..................................... 102
   6.6 Binding Motifs .............................................. 103
   6.7 Tissue Specific Association with Antisense Bias .............................. 103
   6.8 Examining Signs of Selection for Variables other than Antisense Bias .... 104
   6.9 Decreased Gene Expression in the Presence of a Retrotransposon .......... 105
   6.10 Differences in Effect on Expression Between Retrotransposon Types ....... 105
   6.11 Interaction of Sex of the Individual with Changes in Gene Expression in the 
       Presence of a Retrotransposon .............................................. 106
   6.12 Alu and Autism Spectrum Disorder ............................................. 107

7 Appendix Tables ........................................... 109

References ............................................. 110
# List of Figures

1.1 Structure of Human Retrotransposons ........................................... 5  
1.2 Evolution of the Human Alu Sequence ........................................... 8  
1.3 Example Mechanisms of RT-mediated Deletions ................................. 21  

2.1 Similar AFs of shared Alu ............................................................. 30  
2.2 Change in pAS Across Reference Alu ............................................. 34  
2.3 Change in pAS Across Reference L1 ............................................. 35  
2.4 Age of L1 By Subfamily ............................................................. 38  
2.5 Age of Alu By Subfamily ........................................................... 40  
2.6 Clustering of Alu by Age ............................................................. 42  
2.7 Clustering of L1 by Age ............................................................. 43  

3.1 Normalizing Genes by Length ......................................................... 46  
3.2 AS bias of Alu decays linearly across gene boundaries ....................... 48  
3.3 AS bias of L1 decays linearly across gene boundaries ....................... 49  
3.4 Distribution of Coding Sequence Exons in RefSeq Genes ...................... 51  
3.5 Examples of Genes Containing 2° 5'-UTRs ..................................... 52  
3.6 Design for analyzing the impact of internal 5'-UTRs on RT density ........ 53  
3.7 Distribution of 5'-UTRs in genes with no 2° 5'-UTRs ........................ 54  
3.8 Distribution of 5'-UTRs in genes containing 2° 5'-UTRs ..................... 55  
3.9 Change in AS Alu density in the presence of a 2° 5'-UTR ................... 55  
3.10 Change in S Alu density in the presence of a secondary 5'-UTR ............ 56  
3.11 Change in AS L1 density in the presence of a secondary 5'-UTR .......... 57  
3.12 Change in S L1 density in the presence of a 2° 5'-UTR ..................... 57  
3.13 CDS density in 2° 5'-UTR gene groups ........................................ 58  
3.14 5'-UTR density and predictions based on CDS density ..................... 60  
3.15 Examples of Alu Consensus Sequences that Contain Different Binding Motifs 62  
3.16 Start position of short L1 (< 1kb) within the full-length L1 consensus sequence 63  
3.17 Effect of L1 Size on L1 density: Antisense .................................. 64  
3.18 Effect of L1 Size on L1 density: Sense ......................................... 64  
3.19 Effect of L1 Size on L1 density: Antisense, decile 1 ........................ 65  
3.20 Presence of Alu-contained protein-binding motifs within genes .......... 66  
3.21 Presence of L1-contained protein-binding motifs within genes ........... 67  
3.22 Basewise conservation of polyadenylation signals within S and AS L1 ... 69  
3.23 Basewise conservation of polyadenylation signals within S and AS Alu ... 70  
3.24 Basewise conservation of Pol-III A-box region in AS and S Alu .......... 71
3.25 Significance Alu-contained binding motifs in tissue-specific and housekeeping genes ............................................................ 72
3.26 L1-contained binding motifs in tissue-specific and housekeeping genes .................................................. 73
3.27 Significance of L1-contained binding motifs in tissue-specific and housekeeping genes ............................................................ 74

4.1 The effect of MEI status on gene expression between primates .......................................................... 80
4.2 Odds ratios based on sex of origin ........................................................................................................ 83
4.3 Differential expression between sexes in individuals containing the MEI ........................................... 87
4.4 Differential expression between sexes in individuals lacking the MEI ................................................. 87
4.5 Comparison within groups of the same sex, between MEI+ and MEI- individuals 88
List of Tables

2.1 Groups of Polymorphic Retrotransposons ........................................ 31
2.2 AS bias of young RT insertions ...................................................... 31
2.3 Antisense Bias of Reference Alu and L1 ......................................... 32
2.4 Aggregated ages of Alu and L1 by chromosome .............................. 41
2.5 Significant comparisons of age ..................................................... 42

3.1 Slope of linear decay for Alu and L1 within the internal deciles of genes . 49
3.2 CDS density in genes with 2° 5'-UTRs ............................................ 59
3.3 Correlation of CDS and Alu across gene categories .......................... 61
3.4 Size of L1 elements in the human reference genome ......................... 63
3.5 Context Dependent Proportions of the SRY Motif in L1 elements .......... 68

4.1 Gene counts per MEI-type and non-human primate (NHP) .............. 78
4.2 Contingency table for expression differences between human and non-human primate .......................................................... 79
4.3 Contingency table values and results of the two-tailed permutation tests for MEI-expression tests ..................................................... 81
4.4 Results of ANOVA comparing effects of sex, tissue, NHP, and MEI ...... 82
4.5 **ANOVA: Chimp and Gorilla Only** A single regression equation was calculated for significant odds ratios regressed on sex, tissue, primate, and mei-type. The model was then analyzed by ANOVA to determine the respective contributions of each variable. Df = degrees of freedom. 84
4.6 Contingency table for expression differences between males and females of the same species ......................................................... 85

7.1 Significance and direction of age differences across Alu groups .......... 109
7.2 Significance and direction of age differences across L1 groups .......... 109
...Seeing the common, everyday thing in this new light I was often transfixed. The moment one gives close attention to anything, even a blade of grass, it becomes a mysterious, awesome, indescribable magnified world in itself. Almost an unrecognizable world. The [scientist] waits in ambush for these unique moments. He pounces on his little grain of nothingness like a beast of prey. It is the moment of full awakening, of union and absorption.

- Henry Miller (Plexus, The Rosy Crucifixion 1963)

Retrotransposons (RTs), a class of mobile DNA elements, are present in almost all major groups of eukaryotes (Arkhipova & Meselson, 2000). They replicate through a prolific "copy and paste" mechanism, facilitating amplification of the element with each new integration. Despite their success at infiltrating many host organisms, RTs have the potential to confer selectable phenotypic changes to an organism, some of which may be deleterious. A primary example is that of RTs disrupting amino acid sequences. For example, in a case of sporadic colorectal cancer, an L1 RT integrated into the last exon of the tumor suppressor gene APC. This insertion altered the coding nature of the gene triggering the disease (Miki et al., 1992).

Future disease studies must be able to predict disruptions by RTs in a manner similar to the now common practice of filtering single nucleotide polymorphisms (SNPs). In compari-
son to SNP studies, the analysis of RTs is complicated by the multitude of consequences that insertions can have on a gene. For example, sequence motifs specific to RTs can induce functional changes by sequestering protein complexes such as chromatin remodellers, enhancers, and polymerases (Humphrey et al., 1996; Norris et al., 1995; Vansant & Reynolds, 1995; Chen et al., 2009). Furthermore, RT insertions can interrupt coding sequences, disrupt splice signals, and mediate chromosomal rearrangements (Belancio, Hedges, & Deininger, 2008; Belancio, Roy-Engel, & Deininger, 2008; Nevers & Saedler, 1977).

An often debated question is whether the functional attributes of RTs confer a selective advantage or disadvantage to the organism. I hold that these two options are not mutually exclusive. It is likely that the selectable effect of an insertion is dependent upon a dynamic action between factors associated with the RT and factors associated with the insertion location. Indeed studies of the distribution of RTs demonstrate that there is an unequal level of maintenance of RTs depending on location in the genome, indicating that the selection may vary based on genomic context.

Herein is presented a study of RTs and their effects on genes in the human genome. Evidence is described suggesting that selection shapes RT distribution within genes in a manner that is dependent on genomic context. Moreover the data presented indicate that RTs affect the selectable phenotype of gene expression in humans.

1.1 Retrotransposons

Mobile elements are composed of two major classes: Type I and Type II transposons. Type I transposons (i.e: Retrotransposons, RTs) integrate into the genome via an RNA intermediate (Slotkin & Martienssen, 2007). This is often referred to as a "copy and paste" mechanism due to "copying" of the parent DNA sequence into mRNA that is then "pasted" into a new location in the genome. The mechanism underlying this form of integration will be described in more detail in section 1.3. Type II transposons, conversely, insert through a "cut and paste" mechanism, where the parent element is excised from the genome and directly re-inserted into a new location.
Although both RTs and Type II transposons are capable of mobilizing, they have divergent insertion mechanisms as well as varied effects on the host genome. The important distinction of RTs is that with each new insertion, the number of mobile elements in the host genome increases by one, thereby amplifying the number of RT elements present in the genome. Through this amplification, RTs now make up almost 40% of the current human genome. This is striking considering the 1.5% of the genome made up of protein-coding sequence (33% including introns) (Lander et al., 2001).

This large class of RTs can be further subdivided into the subgroups of long-terminal repeat containing RTs (LTR) and non-LTR RTs. LTR-RTs contain long-terminal repeats that flank the internal protein coding sequences. These non-coding repeats facilitate expression as well as integration into the host genome (Havecker et al., 2004). LTR-RTs have a similar structure to retroviruses in that they contain the \textit{gag} and \textit{pol} genes, though they lack the envelope protein required for the ability to exist extracellularly. These elements are present in primate genomes, however most human LTR-RTs are thought to be inactive remnants of ancient viral insertions. There are a small number of human-specific LTR-RTs that may still be active in the human genome, though at much lower levels of integration than the non-LTR retrotransposons (Pi et al., 2004).

Non-LTR-RTs (hereafter referred to as RTs) are the most common form of transposons in the human genome, and will be the focus of the study presented here. Their origins can be traced back to the primitive eukaryote \textit{Giardi lamblia}, predating the origin of viruses and LTR-RTs (Burke et al., 2002). Although their origin can be traced to near the split of prokaryotes and eukaryotes, non-LTR RTs expanded to near present day levels during mammalian evolution (Belancio, Roy-Engel, & Deininger, 2008), and are inherited through mostly vertical transmission (Malik et al., 1999).

There are two major classes of non-LTR RTs, Long INterspersed Elements (LINEs) which are defined as being longer than 500 bp, and Short INterspersed Elements (SINEs) which are smaller. In primates there are two SINEs and one LINE that make up the majority of active elements. These two SINEs are Alu and SINE-VNTR-Alu (SVA), and the LINE is
LINE-1 (L1) (figure 1.1). Alu were originally identified in 1975 through early renaturation techniques that indicated there were numerous 300 bp duplexes apparent in the human genome (Schmid & Deininger, 1975), a majority of these duplexes were cleaved under AluI enzymatic digestion, and were therefore termed Alu elements (Rubin et al., 1980). LINE-1 were originally differentiated from Alu and microsatellites due to their large size ( > 1.2 kilobases (kb)) and digestion by HindIII or KpnI enzymes (Schmeckpeper et al., 1981; Manuelidis, 1982; Shafit-Zagardo et al., 1982). Due to their long length in comparison to Alu and other SINEs they were termed "long" interspersed elements.

A major distinction of L1 is their ability to encode the proteins required for retrotransposition. Depending on this ability to facilitate their own integration, RTs are termed either "autonomous" or "non-autonomous". The proteins encoded by autonomous RTs must be capable of endonuclease and reverse transcriptase activity (Mathias et al., 1991; Feng et al., 1996). Without autonomous elements, the non-autonomous elements would be immobile. The non-autonomous elements, therefore, are able to be inserted even though they do not encode the protein machinery themselves.

In the current mammalian genome, L1 is the most prevalent and active autonomous element (Feng et al., 1996; Mathias et al., 1991). The full-length L1 is approximately 6kb in humans and 7.5kb in mouse (figure 1.1) (Belancio, Roy-Engel, & Deininger, 2008). The mRNA is bicistronic, encoding two open reading frames (ORF1 and ORF2) promoted via an internal Polymerase-II (Pol-II) promoter (Minakami et al., 1992; Swergold, 1990). Transcription of L1 is initiated at the first bp, thereby enabling the promoter to be mobilized upon integration of the full RT, and allowing maintenance of transcriptional activity upon re-integration (Minakami et al., 1992). This Pol-II promoter is bidirectional and therefore has the ability to enhance transcription both of the L1 sequence, and of sequence upstream of the L1 promoter (Mätlik et al., 2006). Though the upstream sequence can be transcribed by the antisense promoter, and potentially act as a non-coding RNA (Cruickshanks et al., 2013), it would not likely be retrotransposed since it would not contain the hallmark sequences required for retrotransposition.
Figure 1.1 – Structure of Human Retrotransposons

The simplified structures of L1NE-1 (L1), Alu, and SINE-VNTR-Alu (SVA) are shown here. **L1** is broken up into three main regions: 5’- UTR, the protein coding region, and the 3’- region and is approximately 6kb long. The 5’-UTR regions contains a Pol-II promoter sequence with bidirectional activity (green arrows). This region also contains homology to the Alu sequence. Sense transcription of L1 copies the bicistronic mRNA encoding two open reading frames (ORF1 and ORF2; red boxes). There is also a 3’ region of the L1 transcript that contains a 3’-UTR (dark green) and an active polyadenylation sequence as well as a polyA tail (light green).

**Alu** is transcribed by the Pol-III via internal A and B box sequences (red boxes) and is approximately 280 - 300 bp long. The center region of Alu is rich in adenosines and often is of the sequence A<sub>6</sub>TACA<sub>6</sub>. The 3’ region of the Alu also contains a polyA tail (light green). **SVA** is a composite elements that is roughly 2kb long. The 5’-portion of SVA contains a CCCTCT<sub>n</sub> repeat sequence. This is linked to an Alu-like sequence, followed by a region containing a variable number of tandem repeats (VNTR). This is then linked to a SINE-like sequence, and finally there is a polyA sequence at the 3’end.

The ORFs encoded by L1 have very different roles in retrotransposition. **ORF1** protein (ORF1p) has nucleic acid binding and chaperone activity and is essential for L1, but not Alu, retrotransposition. The exact role of ORF1p is currently unknown, however mutations in conserved arginine residues abolish L1 retrotransposition. These same mutations reduce ORF1p’s affinity for single-strand RNA as well as DNA melting and strand-exchange ability (Martin et al., 2005). Structurally, ORF1p associates via a coiled-coil domain with itself to create a trimer that then binds to the L1 RNA (Martin et al., 2003). ORF1p also associates with ORF2p in
the ribonucleoprotein particle that is required for functional retrotransposition, and mutations in ORFp1 disrupt the formation of this particle (Doucet et al., 2010).

The ORF2 product of L1 (ORF2p) does most of the work in retrotransposition, encoding an endonuclease near the N-terminus, a reverse transcriptase, and a zinc-finger domain at the C-terminus (Moran et al., 1996; Ostertag & Kazazian, 2001). The ORF2p reverse transcriptase is highly processive, able to copy cDNA at least five times longer than the Moloney murine leukemia viral reverse transcriptase. It has slower processivity, however, through runs of polyG’s that coincide with hairpin-loop structures (Piskareva & Schmatchenko, 2006). The endonuclease encoded by ORF2 is in the internal portion of the protein. The binding specificity of this endonuclease is for 5’-TTAAAA motifs in the target DNA (Cost & Boeke, 1998; Szak et al., 2002), though there are other factors that modify the selection of integration sites. For example, DNA that is in chromatin state is less likely to be nicked by the L1 endonuclease than naked DNA in vitro (Cost et al., 2001). There are some DNA sequences, however, where nicking is enhanced in the presence of chromatin (Cost et al., 2001). This finding was not surprising since the L1 endonuclease shows specificity for super-coiled DNA (Feng et al., 1996). The final, 3’, portion of ORF2p is the cytosine-rich, zinc-finger-like C-terminus. Mutations in this zinc-finger domain significantly reduces retrotransposition (Moran et al., 1996). It is hypothesized that this region facilitates the autonomous nature of L1 through non-specific binding to non-autonomous RNA (Piskareva et al., 2013).

The non-autonomous Alu and SVA SINEs have successfully hijacked the retrotransposition capabilities of L1 to amplify in primate genomes. Similar SINEs exist in many species with some of the oldest members present since before the split of mammals and birds ~ 350 million years ago (mya) (Nishihara et al., 2006). The majority originated from 5S-RNA or tRNAs, however Alu, a primate specific RT, originated from the 7SL-RNA (7SL) (figure1.2). 7SL is an integral component of the signal recognition particle, and is responsible for trafficking secretory proteins through the ER (Gundelfinger et al., 1984). It is thought that an ancient 7SL duplicated into a pseudogene called the fossil Alu monomer (FAM) (Quentin, 1992b). FAM then underwent mutations and subsequent amplification as two separate monomeric repeats
(Quentin, 1992b). Since these repeats would become the left and right components of Alu, they were appropriately named Fossil Left Alu Monomer (FLAM) and Fossil Right Alu Monomer (FRAM). Eventually FLAM and FRAM were joined into a dimer by a central A-rich region which is the common ancestor of the modern Alu (Quentin, 1992a).

The present day Alu has succeeded in leveraging the retrotransposition protein machinery of L1 to expand to millions of copies in the human genome (Lander et al., 2001). One hypothesis for the success of Alu as a non-autonomous retrotransposon is the secondary structure imparted by the 7SL ancestor sequence. 7SL has high affinity for the SRP 9/14 proteins from the signal recognition particle. Dewannieux et al. hypothesized that the similar binding of Alu to SRP 9/14 facilitates shuttling Alu to the ER where the L1 ribonucleoprotein particles reside, thereby enhancing the mobility of Alu (Hsu et al., 1995; Dewannieux et al., 2003). They also posit that SRP 9/14 may perform some of the roles that ORF1p plays in L1 retrotransposition. The recent finding that SRP 14 was one of the factors associated with L1-ORF1p in a yeast two-hybrid screen (Goodier et al., 2013), indicates that the SRP proteins may in fact be involved in facilitating retrotransposition.

SVA, another primate-specific non-autonomous RT, is partially composed of Alu sequence. It is a composite element that contains regions with similarity to Alu and the human endogenous retrovirus (hERV), as well as a variable number tandem repeat region (Hancks & Kazazian, 2010). SVA are still active in the human genome, though at lower levels than either L1 or Alu (Stewart et al., 2011). Even younger than the primate-specific Alu, SVA appear to have evolved relatively recently. They are found only in primates and most commonly in humans, chimpanzee, and gorilla (Zhu et al., 1994). SVA contain all of the necessary requirements for non-autonomous retrotransposition, mainly an internal promoter, and a long polyadenylation stretch at the 3’-end (discussed in section 1.3).

1.2 Subfamilies of Retrotransposons

Each RT family (L1, Alu, and SVA), can be divided into subfamilies based on similar, diagnostic mutations in their sequence. The underlying assumption of grouping subfamilies by
Figure 1.2 – Evolution of the Human Alu Sequence

The middle green flow diagram depicts the proposed evolution of modern-day Alu sequences. A 7SL sequence directly evolved into Fossil Alu Monomer (FAM) by gene duplication. FAM then evolved into two distinct sequences, Fossil Left Alu Monomer (FLAM) and Fossil Right Alu Monomer (FRAM). These two sequences were joined by a central A-rich region, becoming the ancient version of Alu. Alu then subsequently amplified as well as mutated, thus giving rise to the subfamilies in modern primates (AluJ, AluS, and AluY). On the right in red is the approximate time in millions of years since the event occurred. On the left in orange are hallmarks in human evolution that occurred at similar time points.

diagnostic mutations, is that RTs with similar mutations (i.e: same subfamily) are likely to have mobilized around the same time in evolutionary history. This is a logical assumption since RTs are vertically transmitted, and successive similar mutations in divergent subfamilies of RTs is unlikely. There are highly active elements ("hot") that are responsible for most of the mobilization within subgroups (Cordaux & Batzer, 2009). These "hot" master elements may mobilize other active elements as well, which then produce more insertions, but they account for most of the mobilization. Therefore, the model of RT subfamilies is that a small number of parent elements are responsible for the retrotransposition of many elements within a discrete period of evolutionary time (Cordaux & Batzer, 2009). These parent elements gain mutations and are eventually rendered immobile. These accumulated mutations are therefore useful in determining the time-scale in which a RT element was amplified.
It is estimated that there are 80 - 100 master L1 elements actively mobile in humans, though this is likely to be an underestimate (Brouha et al., 2003). The subfamilies of L1 have been evolving for approximately 170 million years, since before the marsupial eutherian split (Smit, 1996), and in a direct lineage for over 40 million years (Khan et al., 2006). The most active L1 were present in the early primate genome and amplified between 40 and 12 mya (Khan et al., 2006). In humans, there are over 100 L1 subfamilies identified (Smit & Riggs, 1995; Smit, Afa, Hubley, R and Green, n.d.), with the major groups denoted as L1M, L1P, and L1Hs. These subfamilies were distinguished based on differences in the 3’-ends of the ORF2 sequence, therefore each subfamily is associated with a unique consensus sequence. Only a small portion of these mobilization competent L1 are thought to be "hot" elements (Brouha et al., 2003; Lutz et al., 2003). The "hot" L1 elements in the human genome are a part of the human-specific subfamily, L1-Ta (Brouha et al., 2003). Furthermore, L1-Ta is the only presently active subfamily in humans, and has a germline insertion rate of 1 in 270 births (Ewing & Kazazian, 2011).

The burst of Alu mobilization also occurred around 35 - 60 mya (Shen et al., 1991). Despite the recent formation of Alu, there are over 30 subfamilies with three major families, from oldest to youngest being: AluJ, AluS, and AluY. Although evidence from population studies suggests that the only active subfamilies of Alu are within the AluY lineage (Stewart et al., 2011), there is contested evidence to suggest that AluS may retain some ability for mobilization (Bennett et al., 2008). There are an estimated 143 master AluY presently active in the human genome (Bennett et al., 2008). Whereas for L1 all master elements came from the L1-Ta, there are multiple actively mobile subfamilies of AluY (e.g: AluYa5 and AluYb8). Together they have an combined insertion rate of 1 in 20 births (Cordaux et al., 2006).

SVA are the youngest of the three elements, originating around 25 mya (H. Wang et al., 2005). Due to their high homology to one another there are only six subfamilies of SVA. SVA, as a group, are largely promoted by external gene promoters located upstream of the SVA sequence (Damert et al., 2009). This process is termed 5’-transduction. The transduction of
these promoters can result in the mobilization of non-SVA DNA to new genomic locations, a concept that will be revisited in section 1.10.

1.3 Process of Retrotransposition

L1, Alu, and SVA are mobilized through the process of target-primed reverse transcription (TPRT) (Raiz et al., 2011; Dewannieux et al., 2003; Ostertag & Kazazian, 2001). TPRT is a "copy and paste" method of transposition, necessitating the involvement of an RNA-intermediate. In humans the enzymes that carry out TPRT are encoded by L1. The process begins by the transcription of the ORF1 and ORF2 components of full-length L1. ORF2p is required for integration of L1, however it is not necessary, but enhances, the integration of Alu (N. Wallace et al., 2008). It has been hypothesized that RNA transcribed by Pol-III is integrated more quickly than Pol-II (Kroutter et al., n.d.), offering a rationale for the abundance of Alu integrations. Indeed Alu, which are transcribed by Pol-III, integrate at higher rates than that of L1, which is transcribed by Pol-II. Upon transcription of the RT, the RNA is shuttled out of the nucleus where it forms an ribonucleoprotein particle (RNP) with the L1 ORF1p and ORF2p (Doucet et al., 2010; Hohjoh & Singer, 1996; Kulpa & Moran, 2005). A number of host factors are also present in this RNP that contribute to mobilization. For example, PABPC1, MOV10, UPF1, and PCNA have all been identified to associate within L1 RNPs (Taylor et al., 2013). A subset of these proteins repress retrotransposition, such as MOV10 (Arjan-Odedra et al., 2012), while some of them enhance it. For example, PABPC1 enhances retrotransposition through binding the polyadenosine (polyA) tail of mRNAs (del Campo, 2009). The knockdown of polyA binding proteins such as PABPC1 has been shown to reduce retrotransposition while retaining the transcription and translation capacity of L1 (Dai et al., 2012). Interestingly, it has also been shown that retrotransposons with longer A-tails are more efficient at retrotransposition (Comeaux et al., 2009). These results all indicate that the polyA tail present in L1, Alu, and SVA is an intrinsic modifier of the efficacy of retrotransposition.

Upon formation of the RNP, the complex is translocated into the nucleus. The process of insertion begins by the ORF2p cleaving a single stranded nick at the A-T phosphodiester bond
in an 5’-TTAAAA motif (Feng et al., 1996). Although there are known hotspots of integration (Wimmer et al., 2011), it is still hypothesized that the selection of the 5’-TTAAAA motif is random (Ovchinnikov et al., 2001). After the first strand is nicked, the stretch of adenosine nucleotides at the end of the mobilized RNA hybridizes to the ‘TTTT’ s of the nicked DNA. Reverse transcription begins from this primed region, copying the retrotransposon RNA into cDNA from the 3’-OH end of the ‘TTTT’. The reverse transcription proceeds from this point.

Since Alu are short, ~300bp elements, they are normally completely reverse transcribed. Full-length L1 however, are ~6kb in length, and the retrotransposition process does not always proceed until completion. The incomplete processing of L1 retrotransposons is termed 5’- truncation and many of L1 in the human genome exist in this form at approximately 1kb in length. The conclusion of TPRT results in a duplication of the flanking DNA sequence, termed target-site duplications (Kojima, 2010). These regions are often used to validate whether an insertion is real, or whether it is false positive. In rare events, the process of TPRT is accompanied by structural rearrangements in the region. The link between retrotransposition and structural rearrangements will be discussed more in section 1.10.

1.4 Selection on the Sex Chromosomes

Since this study will be, in part, examining forces of selection on RTs, it is important to note the differences in selection on the autosomes and the sex chromosomes. These differences will be used to develop hypotheses of the expected distributions of RTs on different chromosomes in the presence of selective forces.

The effective population size is a key determinant of the ability for selection to act in the context of genetic drift (Charlesworth, 2009). The autosomes and the sex chromosomes have different effect population sizes. This is partially due to the limited number of X and Y chromosomes in comparison to autosomes.

Another factor at play in the difference of selection between chrX and chrY is the effect termed "Muller’s ratchet". Muller posited that recombination is highly important for purifying selection. This theory states that there is a lack of an ability to purify deleterious mutations
from the chromosome due to reduced recombination. Each generation cannot "recombine out" deleterious mutations, and therefore has a higher genetic load of deleterious mutations to begin with. In the concept of Muller's ratchet, this mutational load is ever increasing. Effects such as Muller's ratchet that reduce the effective population size are hypothesized to have resulted in the degeneration of genes on chrY (Charlesworth & Charlesworth, 2000). Although studies have shown that deleterious mutations are accumulating on chrY, purifying selection is still possible (Wilson Sayres et al., 2014).

The advantage of studying selection on the sex chromosomes versus the autosomes is that each of these chromosome groups undergo varying selective forces. One example is that autosomes, chrX, and chrY have different effective population sizes, where the effective population size of autosomes is greater than that of chrX which is greater than chrY. Importantly, however, in the case of recessive mutations, chrX can eliminate these mutations more easily since it spends half of its time in the hemizygous state. Using this prior knowledge, predictions can be made regarding the expected relative frequencies of mutations between the chromosomes.

1.5 Selection and the Distribution of Retrotransposons

The distribution of RTs in the current human genome can be an indication of forces of selection or the effect that a RT is having on the organism. Evidence that selection is acting on RTs comes from the random distribution of young elements that becomes non-randomly distributed over long spans of evolutionary time (Pavlíček et al., 2001). Although this pattern could be explained in part by other forces such as genetic drift, imbalanced segregating allele frequencies indicate that selective forces are playing a role in these distributions (Stewart et al., 2011). RTs insert in the human genome at a rate of approximately 1 in 20 births for Alu, 1 in 270 births for L1, and 1 in 916 births for SVA (Cordaux et al., 2006; Ewing & Kazazian, 2010; Xing et al., 2009). Although the distribution immediately after insertion is unknown, polymorphic insertions are evenly distributed throughout the genome (Stewart et al., 2011), as are young L1 present in the human reference genome (Ovchinnikov et al., 2001).
Distribution of RTs: Change over Evolutionary Time

Intriguingly, localized hotspots of insertion are found in the human genome. NF1 is a hotspot for both Alu and L1 integration events (Wimmer et al., 2011). Furthermore, young RTs are often nested within older elements, which some studies have taken as indication of RTs creating hotspots for the integration of new RTs (Levy et al., 2010). Despite these limited findings of insertion hotspots, it is still accepted that RTs insert relatively randomly throughout the genome.

Contrary to the young elements, RTs with long residence times in a population show a non-random distribution across the genome. Aberrations in the otherwise random distribution of older elements can be taken as signs of local selection or genetic drift.

Interestingly, Alu and L1 do not show the same trends of altered densities. Alu tend to be present in higher numbers in in GC-rich regions and to cluster with one another, whereas L1 are increased in AT-rich regions (Pavlícek et al., 2001; Soriano et al., 1983). These altered distributions of Alu and L1 indicate that there may be differential selective pressures acting on RTs. It is therefore important to determine if there are selective pressures acting on RTs depending on the genomic context of integration. The presence of variable selective pressures would indicate that a) RTs exert a non-neutral effect on the genome, b) different regions of the genome may be more sensitive to RT insertions, and c) that polymorphic RTs may impart phenotypic variability to an organism. If polymorphic RTs impart phenotypic variability, then these elements have the potential to account for some of the missing phenotypic heritability observed in Genome-Wide Association Studies (Marian, 2012).

Distribution of RTs: Structural Rearrangements

One example of a genomic force leading to the altered distribution of Alu, is the process of non-allelic homologous recombination. Homologous recombination is a process of double-strand break repair that uses the availability of sister chromatids to repair the break. Non-allelic homologous recombination is similar in procedure to HR, however it uses two sequences that are similar, but not sister chromatids, to repair the broken region. Non-allelic homologous
recombination events often results in deletions, or rearrangements. Since Alu are highly interspersed and have strong homology to one another they are common substrates of non-allelic homologous recombination (Sen et al., 2006; Witherspoon et al., 2009). It is has even been shown that this increased propensity for non-allelic homologous recombination has resulted in Alu being responsible for a majority of the segmental duplications currently present in the human genome. Alu mediated non-allelic homologous recombination is also responsible for many of the rearrangements present between human and non-human primates (Bailey et al., 2003). When Alu are used as substrates for non-allelic homologous recombination, the result would be two Alu residing close to one another, rather than at farther, more random distances, and thus has been used as a hypothesis to describe the co-clustering of Alu in GC-rich regions (Jurka et al., 2004). Furthermore, the ubiquitous removal of newly inserted Alu, would reduce the load in AT-rich regions (Jurka et al., 2004).

**Distribution of RTs: Purifying Versus Positive Selection**

Despite large scale rearrangements of RT densities, there is evidence of selection at the level of the individual element. One study, indicating that RTs undergo purifying selection, examined the rate of heterozygosity of polymorphic RTs in various human populations. They found that Alu and L1 are present more often in the heterozygous state than expected (Stewart et al., 2011).

Although many groups suggest that RTs are undergoing purifying selection (Jurka et al., 2004; Stewart et al., 2011), there are also studies suggesting the potential of positive selection acting on RTs. The altered distribution of Alu in the draft reference genome with respect to GC-content initially caused speculation that RTs were under the influence of positive selection (Lander et al., 2001). A similar GC-bias is observed in B1 within mice. B1 are rodent-specific SINEs that are similar to Alu, but that are not derived from Alu. This finding of seemingly convergent evolutionary forces led the authors to claim that the distribution of Alu and B1 are present due to forces of positive selection (Tsirigos & Rigoutsos, 2009), though there is not much data to support the hypothesis of positive selection versus purifying selection in this
case. The population genetics community responded to the claim of positive selection in the Lander et al. 2001 paper with skepticism. Helen and Brookfield 2013 cited less complicated mechanisms as potential mediators of the altered RT distribution, such as recombination of Alu elements, hitchhiker effects, and the change in GC-content over time (Hellen & Brookfield, 2013).

One piece of evidence supporting positive selection is that specific bases within the Alu consensus sequences are mutated less often than expected at random (Britten, 1994). Some of these mutations, however, are necessary for Alu transcription and are therefore more likely to be maintained across subfamilies (Brookfield, 1994). For example, a large number of the mutations that were found were located within the Polymerase-III binding regions (Britten, 1994). These nucleotides would be required for retrotransposition and are therefore more likely to be present than other nucleotides. Since this original paper was published in 1994, almost a decade ahead of the draft genome sequence, it is possible that there are other nucleotides within Alu and other retrotransposon sequences are maintained that could not have been picked up with the small sample size they were working with (1500 elements) (Britten, 1994). It is important to revisit this question with the data currently available that would facilitate analysis of the conservation of each nucleotide.

**Distribution of RTs: High Copy Number**

Previous studies have taken advantage of the high copy number of RTs to study the relationships of mobile elements with the host genome, revealing the capacity of various genomic contexts to withstand retrotransposition. There are 1.1e06, 9.5e05, and 3.7e03 Alu, L1, and SVA respectively annotated in the human Hg19 reference genome (Kent et al., 2002). Furthermore, the 1000 genomes project has identified thousands of Alu, L1, and SVA respectively that are segregating within humans. These segregating, or polymorphic RTs allow us to study the effects of polymorphic elements on the genome, as well as to look at distributions of early integration.
1.6 Antisense Bias of Retrotransposons in Genes

The non-random distribution of RTs is likely a result of a combination of variables such as the selection coefficient of the mutation, the effective population size, and the rate of recombination in the region of the mutation. Interestingly, the distributions of RTs in genes differs depending on orientation of the RT with respect to the gene. Elements in the antisense orientation (AS) with respect to a gene are more common than elements that are oriented in the sense (S) direction, termed antisense bias (AS bias) (Van De Lagemaat et al., 2006). When in the same region, RTs in either orientation would be expected to be undergoing similar levels of genetic drift, effects of population size, and hitchhiker effects. The varying distribution is likely the result of either selection on the RT based on the orientation, or a bias due to TPRT. Antisense bias is observed in Alu, L1, and even the non-LTR human endogenous retrovirus. It is still unknown whether AS bias is due to the process of integration, or is due to a mechanism of selection over time. One of the focuses of the research presented here is to determine whether integration or selection is the driving force of AS bias of RTs.

1.7 The Dynamics of Regulating Gene Expression

The genome responds to spatio-temporal and environmental cues by regulating gene expression (Coulon et al., 2013; Spitz & Furlong, 2012), thereby coordinating tissue-type, stage of development, as well as responses to external stress. These concerted signals direct proper growth and maintenance by supplying proteins and non-coding RNA necessary for the stability of the cell. Regulating gene expression is a multi-facted process including chromatin folding, transcription complex formation, and post-transcriptional degradation. One of the core components of regulating gene expression is the interactive binding of multiple regulatory proteins to specific sequences within the DNA. These regulatory proteins, such as transcription factors and enhancers, are structured to fold and bind specifically to predetermined sequences or subtle variations thereof. The DNA bound protein can then recruit other enzymes to the region to modify gene expression or alter the local structure of the chromatin. For example, the RORA
hormone receptor binds to DNA with the sequence: 5’-[AT][TA]A[AT][CGT]TAGGTCA-3’ where brackets indicate positions which are recognized when any of the included nucleotides are present. The binding of RORA transcription factors to these DNA sequence motifs can result in increased expression of the downstream gene. Other similar proteins may induce decreased expression, alternative splicing, or coordinated regulation with other genes.

Aberrant regulation of gene expression is often associated with disease, and therefore many studies focus on the genetic determinants of altered gene expression. Variations associated with differential expression are termed expressed quantitative trait loci (eQTL). Although these variants may occur within the coding region (Delker & Quint, 2011), they more commonly occur near transcription start sites and within regulatory regions (Veyrieras et al., 2008). Despite there being many other forms of genetic variation, these eQTL studies generally examine small nucleotide variants, such as SNPs and small indels. Larger variations are often missed such as large structural rearrangements, copy number variations, and mobile element insertions. Importantly, it is these larger variations that track most with large scale changes in phenotype, such as the loss of conserved non-protein coding elements through deletions (Hiller et al., 2012). Moreover, chromosomes that have undergone large scale rearrangements have a 2 fold faster protein evolution rate between humans and chimps (Navarro & Barton, 2003)

### 1.8 Variation in Gene Expression Between Humans and Non-human Primates

The coding genome of humans and non-human primates (NHP) are highly similar, with humans and chimps differing by only 4% of the protein-coding regions (Varki & Altheide, 2005). Despite the conservation of coding sequence, many phenotypes are distinctive between the two species. For example, the structure of organs such as the salt-wasting kidneys, the immune response, and social organization are variable between humans and NHPs (Varki & Altheide, 2005). Attempts to determine the genetic differences underlying these and other phenotypic changes have revealed that not only are protein-coding sequences highly similar, but 3’-UTRs also exhibit baseline mutation rates (Y. Li & Su, 2006; H.-Y. Wang et al., 2007). Gene expression levels, however, are significantly different between NHPs and humans (Blekhman et
al., 2008; Gilad et al., 2006), indicating that the regulation of gene expression may play a underlying role in the development of larger phenotypic differences.

Since non-coding sequences, such as transcription factor binding sites, are important for coordinating regulation of gene expression, it is likely that they also play a large role in species-specific differential expression. Indeed, genes with increased expression in humans versus non-human primates are enriched for transcription factors, indicating that changes in transcriptional regulation may play an important role in the differences in gene variation between species (Gilad et al., 2006).

1.9 Retrotransposons as Non-coding Sources of Regulating Gene Expression

Retrotransposons have shaped the regulatory code of the human genome throughout primate evolution. They contribute to splice sites (Sorek et al., 2002; Belancio, Roy-Engel, & Deininger, 2008; Krull et al., 2005; Lev-Maor et al., 2008), A-to-I RNA editing (Athanasiadis et al., 2004), nucleosome binding signals (England et al., 1993; Englander & Howard, 1995; Tanaka et al., 2010), histone modification signals (Huda et al., 2010), DNA cytosine methylation (Liu & Schmid, 1993; Schmid, 1991), transcription factor binding sites (Yang et al., 1996; Vansant & Reynolds, 1995; Pandey et al., 2011), and primate-specific enhancers (Jacques et al., 2013). The attribute of RTs that makes them capable of having these effects is that RTs contain sequence motifs that can act as binding sites for transcription factors (Polak & Domany, 2006), polymerases (Fuhrman et al., 1981; Minakami et al., 1992; Swergold, 1990), polyadenylation enzymes (Kaer et al., 2011), and other regulatory proteins. Therefore, when RTs integrate into a new location they may impact regulatory mechanisms as well as the protein coding regions of the gene.

One way that RT insertions effect transcription is through altering splice sites. Alu, L1, and SVA all contain potential splice donor and acceptor sites, and upon inserting into gene boundaries these signals can effect the structure of the final mRNA. SVA is efficient at co-opting splicing enzymes often resulting in 5’transduction of non-repetitive sequence during SVA mobilization (Damert et al., 2009). SVA can also effect splicing of the endogenous transcript. For
example, an insertion of SVA into the 3’-UTR of the fukutin gene resulted in aberrant splicing and the development of Fukuyama muscular dystrophy in Japanese individuals (Taniguchi-Ikeda et al., 2011). Alu and L1 have also often provided splice signals and are responsible for a large proportion of alternative isoforms in humans (Lev-Maor et al., 2008).

Alternative splicing due to RTs can result in the inclusion of a RT into the final mRNA, termed exonization. Since Alu have expanded to millions of copies in the human genome, it is possible that the genome has developed a response to limit the number of Alu exonization events (Zarnack et al., 2013). Heterogeneous nuclear ribonucleoprotein C (hnRNP C) is a RNA-binding protein that competes with splicing factors to inhibit this inclusion of Alu into the final mRNA. Another link between Alu and alternative splicing is through the propensity of Alu to undergo Adenosine to Inosine (A-to-I) RNA editing. Alu are common targets of A-to-I editing in the genome because they often reside as inverted repeats in close proximity to one another (Athanasiadis et al., 2004). The secondary structure created by the binding of these elements is highly prone to editing by the enzyme ADAR1 (Iizasa & Nishikura, 2009). Due to fluctuations in the secondary structure, the number and placement of sites edited within an mRNA is variable, depending on the length and secondary structure of the molecule (Rieder et al., 2013; Blow et al., 2006). The effects of these editing events are still under study, however it is known that when Alu are highly edited, the Alu alters exon choice leading to alternative splicing (Athanasiadis et al., 2004; Lev-Maor et al., 2008).

1.10 Structural Rearrangements Associated with Retrotransposition

Other than the direct amplification of the RT, there are other structural effects, such as deletions, rearrangements, and transduction events that are associated with the process of retrotransposition. Deletions often result from the process of transposon insertion evidenced from non-LTR RTs in humans, HO cleavage of the mating type locus in *Saccharomyces cerevisiae* (Moore & Haber, 1996), as well as IS1-mediated deletions in *Escherichia coli* (Nevers & Saedler, 1977). Non-LTR RT-mediated deletions in primates are responsible for an estimated 7.5 megabases (Mb) of sequence lost. Furthermore, there is an estimated 18 kilobases (kb) between human
and chimp due to L1 integration, and 9kb due to Alu alone (K. Han et al., 2005; Callinan et al., 2005). Large (2 -14 kb) deletions are common in cell culture (Gilbert et al., 2005, 2002; Symer et al., 2002), and range from local to intrachromosomal deletions (Gilbert et al., 2005). Studies comparing primate genomes, however, reveal most deletions are very small ( < 100 bp) and follow a Poisson distribution with a maximum size of 5kb (K. Han et al., 2005). It is possible that the discrepancy between the size range of deletions in cell culture and primate comparison studies is due to purifying selection of large deletion events. Two mechanisms proposed for RT-mediated deletions are NAHR and L1-endonuclease independent retrotransposition (figure 1.3 (Callinan et al., 2005)). Alu are particularly good substrates for NAHR because they are highly interspersed in the genome, and have a high GC content, which can stabilize the repair structure. L1-endonuclease independent retrotransposition is the process by where the initial double strand break is created by a means other than the L1 endonuclease. This break is then "filled in" using available template from the RT RNA and the L1 reverse transcriptase. There is evidence that this form of RT plays a role in stabilizing telomeres (Morrish et al., 2007). These deletions create variability in the human population, but may also be a process associated with disease such as cancer and this connection will be elaborated on more in section 1.11(Raschke et al., 2005; Florl & Schulz, 2003).

Genomic duplications are another by-product of RT-mediated NAHR, with Alu thought to be responsible for a large proportion of the highly-homologous segmental duplications in humans (Bailey et al., 2003). Most of these duplications are larger than the observed deletions, averaging 1 Mb (Bailey et al., 2003). The majority were facilitated through the younger AluY and AluS subfamilies, and most commonly from AluS. This indicates that large scale genomic duplication events in the time range of AluS, after the divergence of new and old world monkeys, are due to Alu-Alu recombinations.

The process of post-insertional transposon-mediated recombination has been observed across many species including human, mouse, Saccharomyces cerevisiae, and of course the famous Zea Mays (Mieczkowski et al., 2006; Edelmann et al., 1989; W. Hu et al., 1998). A 120kb inversion on chr12 was identified using an L1 integration cassette, and is thought to have occurred as
Figure 1.3 – Example Mechanisms of RT-mediated Deletions

a) Non-Allelic Homologous Recombination (NAHR). The medium green and dark green boxes represent non-repetitive portions of the genomic DNA. The light green box represents a gene. The orange arrows are RTs where the arrow is pointed in the direction of transcription. The dotted line represents a site of a DNA double strand break (DSB), the L1 Endonuclease is not required for this DSB. The regions of homology between the two inverted RTs bind to each other and facilitate initiation of the homologous recombination pathway. When the DSB is resolved along the red line, a deletion will occur. b) L1-Endonuclease Independent Retrotransposition. The medium green and dark green boxes again represent the non-repetitive portion of the genomic DNA. The short orange boxes represent genomic regions with homology to the retrotransposon (long orange box). Upon creation of an L1-endonuclease independent DSB, end-repair removes nucleotides from the region near the break. The RT then binds to the regions of homology and is used as a template for reverse transcription (red line), thereby healing the DSB at the cost of a small deletion.

a result of the integration process (Symer et al., 2002). RT-mediated structural rearrangements are commonly thought to be a result of post-insertion processes, similar to those resulting in deletions and duplications. There are at least 252 RT-mediated inversions between humans and chimps that have altered the structure, while maintaining the mass, of the genome (Lee et al., 2008). Three of these loci include exons, indicating a potential for altering phenotypes between species (Lee et al., 2008).

Another mechanism of structural change due to RT mobilization is the transduction of non-repeat sequence. Sequence transduction occurs either to the 5’ or 3’ side of the RT, and has been observed both through in silico analysis as well as in vitro (Symer et al., 2002; Damert et al.,...
2009; Pickeral et al., 2000). 5’transduction occurs when the retrotransposon is promoted by an upstream gene promoter. This is often the case in SVA mobilization (Damert et al., 2009), where approximately 8% of all SVA elements have identified 5’- sequence transductions. One ancient transduction event in particular, of the MAST2 promoter, has enhanced the transcription activity and facilitated the insertion of an entire group of SVA elements (Damert et al., 2009). 3’transduction is more common than 5’- and occurs with both SVA and L1 retrotransposition (Damert et al., 2009; Pickeral et al., 2000). In this case, the polymerase continues to transcribe beyond the RT polyadenylation signal and into non-repeat sequence. On average the length of the transduced sequence ranges from 200 - 300 bp (Goodier et al., 2000; Pickeral et al., 2000). If this sequence has homology to the sequence at the insertion point, it can then be inserted upon reintegration. The 3’region is not always mobilized with the element, even if it is transcribed. For example, Alu have weak Pol-III termination sequences and therefore often terminate at a region downstream of the RT sequence. This downstream unique sequence is not reintegrated, since priming on the polyT sequence is facilitated through the polyA tail within the Alu. Despite the requirement for homology for reincorporation, approximately 15% of L1 show evidence of 3’- transduction events. It has been proposed that the mechanisms of 5’- and 3’- transduction are capable of "shuffling" exons in the human genome, thereby adding variation. An extension to this is the trans mobilization of retropseudogenes and non-processed RNA from the L1 reverse transcriptase but without the coincident transcription of the RT (Ejima & Yang, 2003; Esnault et al., 2000; Dewannieux & Heidmann, 2005).

1.11 Phenotypic Effects of Retrotransposons

RTs can lead to both disease-causing and advantageous phenotypic changes. RT insertions into coding exon sequences are often associated with diseases such as the insertion of Alu into factor IX leading to hemophilia (Vidaud et al., 1993), BRCA2 in breast cancer (Miki et al., 1996; Teugels et al., 2005), and NF1 leading to neurofibromatosis (M. R. Wallace et al., 1991). Similarly L1 insertions can also disrupt coding sequence by inserting into exons (Kazazian et al., 1988; T. H. Li & Schmid, 2001; Mukherjee et al., 2004). Interestingly RT insertions into
the introns can also lead to disease mainly through the creation of cryptic splice sites. Such was the case with the insertion of Alu into the intron of ALPS which resulted in skipping of exon 8 and led to autoimmune lymphoproliferative syndrome (Tighe et al., 2002). Alternatively, the insertion of L1 into the intron created a cryptic splice site that included the L1 in the mature mRNA (van den Hurk et al., 2003).

As mentioned in section 1.10, RTs are prolific templates for recombination, especially Alu which are associated with at least 50 human diseases through Alu-Alu recombination (Callinan et al., 2005). For example, familial instances of Duchenne’s muscular dystrophy (X. Y. Hu et al., 1991), Fabry-disease (Kornreich et al., 1990), Tay Sachs (Myerowitz & Hogikyan, 1987), Lesch-Nyhan disease (Brooks et al., 2001), and many others have been associated with Alu recombination events. Furthermore, there have been many independent discoveries of Alu-mediated recombination in the low-density liporeceptor gene associated with hypercholesterolemia (Chae et al., 1997; Lehrman et al., 1987, 1985; Rüdiger et al., 1991; Yamakawa et al., 1989; Goldmann et al., 2010). L1 recombination events have also been associated with diseases such as Alport syndrome-diffuse leiomyomatosis (Segal et al., 1999), though the discovery of disease causing L1-recombination events is less common than that of Alu. Interestingly there are also reports of diseases due to deletions from Alu-L1 recombination events, such as maple-syrup urine disease (Silao et al., 2004) and Duchenne muscular dystrophy (Suminaga et al., 2000), indicating the promiscuity of sequence allowed for recombination.

There are many other cases and mechanisms of RTs leading to disease, such as through the deletions associated with insertions as well as 3’transduction events. Although these mechanisms are well described and published in plenty, there may be a skew in the literature in association with disease-causing insertions, due to ascertainment bias. As increasing numbers of "unaffected" individuals are sequenced the RT component underlying normal variation should begin to be revealed, though there are cases available to date. For example, the case of Wilms’ tumor (WT1) gene that is under tight control during development. WT1 contains a full length Alu in the promoter region. When the Alu is deleted, the promoter loses the ability to appropriately silence the WT1 gene in HeLa cells (Hewitt et al., 1995).
Another case was within a primate conserved Alu/L1 composite sequence within the SLC7A2 gene. A mutation in this sequence resulted in infantile encephalopathy, indicating that the RT sequence had been co-opted by the host for a conserved phenotypic function. Upon further analysis it was revealed that this portion of the SLC7A2 gene was transcribed as a ncRNA in the brain, that when knocked down increased neural apoptosis. Furthermore the region containing the mutated RT created a hairpin loop structure, necessary for proper function of the ncRNA (Cartault et al., 2012).

Although there are many more examples of RTs which have evolved to create useful functions in humans, the last one I will share is that of a particular locus on chr19. This region is the source of at least 50 miRNAs, all of which are transcribed via promoters from Alu RTs. Interestingly, being promoted by Alu, these miRNAs are all Pol-III transcribed as opposed to the canonical Pol-II transcription of most miRNAs (Borchert et al., 2006).

1.12 A New Path for Retrotransposon Research

The study of the effects of RTs on genes has previously been largely hampered by inadequate methods to identify sufficient numbers RT polymorphisms. Recent advances both in sequencing technology and bioinformatics analyses have enabled researchers to begin to identify polymorphic RTs and therefore to study the effect that these elements have on phenotype. In 2010 three seminal papers published in Cell and Genome Research described feasible techniques for sequencing L1 retrotransposons (Ewing & Kazazian, 2010; Iskow et al., 2010; Beck et al., 2010). Combined with an array-based technique published the following year, these methods filled in the first missing piece of identifying the contribution of RTs to human variation (Baillie et al., 2011). That was identifying the variants. Combined with novel bioinformatics algorithms to map these variations to the genome (Ye et al., 2009; Chen et al., 2009) these techniques facilitated the completion of the RT portion of the 1000 genomes project, which annotated hundreds of insertions that are polymorphic between human populations. Now that the techniques are available to identify the presence or absence of RTs between individuals, it is necessary to examine the phenotypic consequences of these insertions.
1.13 Motivation

Herein I present an examination of the consequences of RT insertions within genes. The hypothesis examined is that, when inserted into genes, RTs affect the gene in a selectable manner. It is therefore postulated that the AS bias has arisen through selection as opposed to integration. Furthermore, it is hypothesized that the sequence of the RT as well as the genomic context can modify the resulting distribution of the RT. Finally it is hypothesized that the presence of a RT is associated with changes in the phenotype of gene expression. Evidence presented herein shows that RT insertions are under negative selective pressure within genes, genetic factors modify this selection of RTs, and that there is an association between RTs and differential expression of the genes they reside within. This research will aid future disease studies attempting to link genetic variability of RTs to phenotypic change by aiding the ability to annotate RT insertions as potentially deleterious based on variables such as genomic location and RT-type. These results were attained through the following specific aims:

**Aim 1: Determine whether RT antisense bias is due to integration or selection**

In order to identify the contexts that modify selection of RTs it must first be determined whether RTs are integrated uniformly within genes or if there is a bias upon integration. I therefore identified if RTs were integrated uniformly with respect to the orientation within genes (i.e.: AS or S). This was achieved by examining the AS bias of RTs of varying allele frequencies.

**Aim 2: Identify variables associated with RTs that modify RT distribution within genes**

This analysis examined which RT and gene contexts modified RT distribution within genes. The distribution of RTs was tested against a random distribution for a variety of RT variables (e.g.: regulatory sequences) and gene contexts (e.g.: tissue-specificity of gene expression).

**Aim 3: Determine association between RTs and the phenotype of gene expression**

To further explore the potential for RTs to effect genes in a selectable manner, it was examined
how RTs were associated with the phenotype of gene expression. In order to test this, the association between RT insertions and differential gene expression was analyzed between humans and well as between humans and non-human primates.
Chapter 2  ORIGINS OF ANTISENSE BIAS

The non-random distribution of RTs can be used to facilitate studies into the variable effects that RTs impart on genes— effects that may be accompanied by phenotypic changes. The distribution of RTs is only useful in these studies, however, if it is due to forces acting on phenotypes (i.e.: selection) as opposed to non-phenotype associated mechanisms such as integration preferences. Studies attempting to determine the mechanistic underpinnings leading to changes in RT distribution over time have identified that selection plays a role in the genome-wide distribution of RTs (Stewart et al., 2011), but it is unknown if the altered distribution observed within genes is also due to forces of selection.

Despite the knowledge that selection acts on RTs in general, there are two major questions that are yet to be resolved, i) is the altered distribution of RTs, specifically within genes, due to selection and ii) are the selective forces acting on RTs positive or negative (purifying). The first question (i) is vital to future studies attempting to discern the phenotypic effects of RTs on genes. There are altered distributions of RTs within genes that, if they are due to selection, can be used to predict the phenotypic impact of future insertions. This knowledge would therefore inform future disease studies investigating the association between RTs and phenotype. If instead the altered distribution of RTs in genes is due to the non phenotype-dependent force of integration then these predictions would be unfounded, and perhaps misleading. Therefore, before further analyses on phenotypic associations with RTs distributions are analyzed, it must be determined whether RT integration is the force driving altered distribution of RTs within genes, or if forces associated with phenotype (i.e.: selection) are the driving force.
The RT antisense bias (AS bias) is a powerful example of altered distributions of RTs within genes that can be used to delineate between effects of selection versus integration. AS bias is the tendency of RTs to be present in the antisense (AS) orientation within genes more often than the sense (S) orientation. An interesting feature of AS bias is that it peaks near transcription start sites (TSS), as do the total number of RTs (Medstrand et al., 2002).

This recurring altered distribution of AS bias towards transcription start sites may indicate a mechanism of selection, however it may also indicate a preference for integration locations. An argument for the enrichment of RTs near the TSS is that there is increased open chromatin near the TSS and therefore increased accessibility to the L1 ORF2p endonuclease. However, if open chromatin was the reason for the increase in RTs then this preference should be replicated in both the sense and antisense orientations. Moreover, since the L1 ORF2p endonuclease cleaves dsDNA (Feng et al., 1996), there should be no preference for integration due to single-strand accessibility. Also, it has been shown that L1 elements inserted in the S orientation were associated with decreased expression of a coincident gene, however the same effect is not true of L1 in the AS orientation (J. S. Han et al., 2004). For these reasons, I hypothesized that the AS bias of RTs in genes was due to a selective process.

To distinguish whether AS bias was truly a result of selection over time or if it was due to integration, the distribution of recent Alu insertions were examined, as well as the change in AS bias over evolutionary time.

2.1 Retrotransposon Antisense Bias and the Process of Integration

Under the hypothesis that selection is the driving force of AS bias in genes, the difference in AS and S elements should accumulate over time, and newly inserted RTs should have no observable AS bias. Relatively young Alu and L1 were therefore studied to examine the AS bias of newly inserted elements. Three groups of young Alu and L1 were used for this analysis. Groups a and b were filtered groups of Alu and L1 that were annotated as insertions in the 1000 genomes project (Stewart et al., 2011). Group a were elements that were populations-specific, and group b were elements that were shared between all three populations. Group c Alu and
L1 were annotated as germline insertions when identified in somatic retrotransposition studies (Baillie et al., 2011) (table 2.1. All three groups represented RTs that were polymorphic in the population (young). Here, polymorphic refers to the status of presence or absence of the entire element where any mutations within the element were irrelevant to the status.

### 2.1.1 Shared Retrotransposons are More Common than Population-specific Elements

Since the goal was to determine if either the immediate force of integration or the longitudinal force of selection was responsible for AS bias, elements were grouped based on residence time in the population. The allele frequency of each insertion was used as a proxy for residence time in the population. Alu in groups \(a\) and \(b\) were initially identified in the 1000 genomes project as being polymorphic between individuals (Stewart et al., 2011). The allele frequencies of these elements were calculated from three populations; Yoruban (YRI); European (CEU); and Chinese/Japanese (CHBJPT) (Stewart et al., 2011). For the current analysis, 1000 genomes annotated Alu were separated into the two groups mentioned above: \(a\) population-specific Alu (i.e.: Allele Frequency (AF) > 0 in only one population), and \(b\) Alu present among all three populations. The expectation was that elements present in only one population were more likely to have small residence times \(a\). Similarly elements present prior to the split of the three populations would be present in all three populations and would likely have longer residence times \(b\). The mean AFs of group \(a\) Alu were 0.091 \((±0.071)\), 0.067 \((±0.047)\), and 0.070 \((±0.067)\) in YRI, CEU, and CHBJPT respectively. To identify the mean AF of group \(b\) Alu, the maximum population-specific AF was identified per element. The average of these maximum AFs was significantly higher than the population-specific sets 0.279 \((±0.16)\) (Wilcoxon rank sum \(p_{YRI} < 1.3e-184; p_{CEU} < 2.6e-62; p_{CHBJPT} < 8.5e-49\)). The AFs of group \(c\) Alu were similar in all three populations (figure 2.1). Since AFs are correlated with residence times in the population (W. H. Li, 1975), these AFs indicate that group \(a\) Alu have shorter residence times (younger) than group \(b\).
2.1.2 Young Antisense and Sense Alu are Present in Equal Proportions

Although the allele frequencies of group \(c\) RTs were unknown, they were identified in Baillie et al. as being present in the germline insertions. Group \(a\) and \(c\) RTs all represent young polymorphic insertions that were significantly less frequent than group \(b\). To reduce type II error by increasing sample size the counts were combined across groups \(a\) and \(c\) ("young") (table 2.1 and table 2.2).

To determine if there was an antisense bias in newly integrated RTs, the probability of choosing an AS RT in the "young" group was tested versus an expected probability of 0.50 using the binomial sign test. The required sample size to have sufficient power (power = 0.95) to reject the null hypothesis, with an \(\alpha\) of 0.05, was 331 total elements, which our sample size exceeded (\(N = 1130\)). This value was computed using an effect size of 0.09, which was equivalent to effect required to observe the antisense bias of Alu and L1 in the reference genome (table 2.3). The hypothesis that young insertions had an AS bias equal to 0.50 could not be re-
Table 2.1 – Groups of Polymorphic Retrotransposons

<table>
<thead>
<tr>
<th>Group</th>
<th>Reference</th>
<th>Study Type</th>
<th>Allele Sharing</th>
<th>Allele Frequency</th>
<th>Combined</th>
<th>Nickname</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>Stewart 2011</td>
<td>population</td>
<td>population-specific</td>
<td>0.076</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td>b</td>
<td>Stewart 2011</td>
<td>population</td>
<td>shared</td>
<td>0.279</td>
<td>×</td>
</tr>
<tr>
<td>3</td>
<td>c</td>
<td>Baillie 2011</td>
<td>individual</td>
<td>NA</td>
<td>NA</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2.2 – AS bias of young RT insertions

<table>
<thead>
<tr>
<th>Group</th>
<th>Reference</th>
<th>Element</th>
<th>Population</th>
<th>AS</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>a</td>
<td>L1</td>
<td>YRI</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>a</td>
<td>L1</td>
<td>CEU</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>a</td>
<td>L1</td>
<td>CHB/JPT</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>a</td>
<td>Alu</td>
<td>YRI</td>
<td>150</td>
<td>154</td>
</tr>
<tr>
<td>5</td>
<td>a</td>
<td>Alu</td>
<td>CEU</td>
<td>29</td>
<td>20</td>
</tr>
<tr>
<td>6</td>
<td>a</td>
<td>Alu</td>
<td>CHB/JPT</td>
<td>21</td>
<td>18</td>
</tr>
<tr>
<td>7</td>
<td>c</td>
<td>L1</td>
<td>NBB</td>
<td>169</td>
<td>178</td>
</tr>
<tr>
<td>8</td>
<td>c</td>
<td>Alu</td>
<td>NBB</td>
<td>166</td>
<td>177</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>569</td>
<td>561</td>
</tr>
</tbody>
</table>

Since the test of newly inserted RT AS bias versus a null expectation was not significant, newly inserted RTs were compared versus the AS bias of the reference RTs. This analysis allowed for the ability to determine if the AS bias of young elements was significantly different from the bias that is often cited in the literature (Zhang et al., 2011). The proportion of AS
elements in the young group of Alu and L1 were compared versus the AS bias of reference Alu and L1 (Kent et al., 2002; Smit, Afa, Hubley, R, and Green, n.d.).

First the AS bias of reference Alu and L1 were identified. As has been previously documented, the proportions of AS and S Alu and L1 in the Hg19 reference were significantly different from the null hypothesis of equal distribution (table 2.3). A test of equal proportions was used to calculate the deviation from the null hypothesis that young AS RTs occur in equal proportion to that of reference RTs. It was identified that reference Alu and L1 RTs had a significantly different AS bias than newly inserted Alu (proportion AS_{young} = 0.50; proportion AS_{Ref} = 0.59; 95% CI = 0.12 - 0.057; p < 4.44e-09). This was evidence that the commonly cited AS bias of reference Alu was not present initially upon integration.

<table>
<thead>
<tr>
<th>Element</th>
<th>AS</th>
<th>S</th>
<th>p(AS)</th>
<th>CI, 95%</th>
<th>Binomial sign test p &lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Alu</td>
<td>281711</td>
<td>231331</td>
<td>0.55</td>
<td>0.548-0.550</td>
<td>2.47e-323</td>
</tr>
<tr>
<td>2 L1</td>
<td>208098</td>
<td>109206</td>
<td>0.66</td>
<td>0.654-0.657</td>
<td>1.98e-323</td>
</tr>
<tr>
<td>3 SVA</td>
<td>1585</td>
<td>468</td>
<td>0.77</td>
<td>0.753-0.790</td>
<td>2.36e-141</td>
</tr>
<tr>
<td>3 Alu and L1</td>
<td>489809</td>
<td>340537</td>
<td>0.59</td>
<td>0.589-0.590</td>
<td>2.96e-323</td>
</tr>
</tbody>
</table>

Table 2.3 – Antisense Bias of Reference Alu and L1
Reference Alu were identified using RepeatMasker annotations on the Hg19 build of the human genome. AS: antisense; S: sense; p(AS): probability of antisense; CI: confidence interval.

2.1.3 Antisense Bias is Present in DNA Transposons

The evidence presented in section 2.1.2 has shown that there is no discernible AS bias upon integration, indicating that the process of Alu integration (ie: TPRT) does not likely contribute to AS bias. Alu, L1, and SVA are all integrated through the process of TPRT, and all have significant differences in the proportion of AS and S elements (table 2.3). Therefore, it was pertinent to this study to determine if AS bias was restricted to elements which were integrated by TPRT. The null hypothesis that transposons inserted by non-TPRT mechanisms had equal proportions of S and AS elements was examined. Similar to the RTs Alu and L1, a pooled group of DNA transposons had a significant increase of AS elements over S elements (probability of AS = 0.540; 95% CI = 0.538 - 0.543; p < 1.64e−190), indicating that AS bias is likely due to a process that acts independently of the mechanism of integration.
2.2 Antisense Bias Increases Over Evolutionary Time

Since Alu present in the reference genome have a significantly larger proportion of AS elements than newly integrated elements, it is likely that this change in proportion was due to a progressive action over evolutionary time. It was therefore important to examine the change in AS bias over time. To examine the change of $p(\text{AS})$ over time, AS bias of Alu and L1 were individually examined after separation into their predefined subfamilies. The relative age of a subfamily is often measured as the average distance of all elements within a subfamily from the respective consensus sequence (millidiv). Since elements within a subfamily inserted at similar points in evolutionary time, elements within a subfamily are expected to have similar values of millidiv. The average millidiv across all elements within a subfamily was therefore used as a measurement of time where smaller values of millidiv corresponded to younger subfamilies with less residence time in the population, and larger millidiv corresponded to older subfamilies with longer residence times.

2.2.1 Antisense Bias of Alu is Constant Across Reference Subfamilies

Since Alu originated in the primate lineage, the scale of Alu age spans from early primate evolution, to modern human evolution. AS bias was examined across Alu subfamilies ranked by millidiv (figure 2.2). Although there was a slight increase in AS bias with increasing millidiv, this increase was not significant (slope: $4.1e-04$; F-test $p = 0.38$). This indicated that the effect leading to AS bias of Alu likely occurred soon after the time point of integration.

2.2.2 Antisense Bias of L1 Increases across Reference Subfamilies

Although Alu and L1 both have a significant AS bias in the genome, it is possible that they are undergoing different degrees of selection. This is evidenced by the finding that the AS bias of L1 overall is significantly larger than that of Alu ($\text{prop } AS_{L1} = 0.656$; prop $AS_{Alu} = 0.549$ ; $p < 2.2e-16$).
L1 AS bias was also examined across subfamilies ordered by millidiv. Contrary to the findings in Alu, there was a significant trend towards an increase in AS bias over time (F-test $p < 2.66e-06$) (figure 2.3). These results indicated that, in L1, AS bias increased dynamically over long periods of evolutionary time.

### 2.2.3 Antisense Bias of Alu Increases Quickly Over Evolutionary Time

The findings so far have shown that AS and S RTs have no bias upon integration, however there is an AS bias of Alu in the reference genome. For Alu, this AS bias in reference Alu was independent of the age of the subfamily. It was therefore hypothesized that AS bias in Alu was achieved rapidly after integration. To examine the change of AS proportion on shorter time scales than the previous analysis, $p(\text{AS})$ of young Alu were compared to $p(\text{AS})$ of Alu that were shared among all three populations (section 2.1.1; table 2.1; group $b$). As mentioned previously, shared Alu were assumed to have longer residence times in the genome than young
Figure 2.3 – Change in pAS Across Reference L1
AS/S ratios were calculated for each L1 subfamily and plotted versus the mean deviation from the subfamily consensus sequence (millidiv). AS bias was regressed on millidiv to obtain the fitted line (blue; \( p < 2.66 \times 10^{-06} \)).

Alu (young = combined groups \( a \) and \( c \); table 2.1) because shared elements must have integrated prior to the split of YRI, CEU, and CHB|JPT, while young elements were assumed to have inserted after the split. The higher average allele frequencies of shared Alu compared to group population-specific Alu substantiated this assumption (section 2.1.2).

The AS bias was compared between shared and young Alu to determine if AS bias was present in this slightly older group of elements. Indeed the proportion of AS in the young subset of Alu was significantly smaller than polymorphic Alu that were shared among all three populations (\( b \)) (\( \text{propAS}_{\text{young}} = 0.50; \text{propAS}_{\text{shared}} = 0.54; 95\% \text{ CI} = -0.11 - -0.02; p < 0.01 \)). This indicated that the change to AS bias of Alu occurred relatively soon after integration, on an evolutionary time scale.

Overall, these results indicated that Alu and L1 were initially integrated with no AS bias. The proportion of AS and S Alu changed relatively quickly over evolutionary time scales. The
AS bias of L1 also changed over evolutionary time, and at much larger time scales than that of Alu.

2.3 Evidence of Selection Acting on Alu and L1

The results so far have shown that AS bias was not present upon integration but increased over evolutionary time, as would be expected under a model of selection. If selection within genes was playing a role in this change in AS bias, there are three hypotheses that can be tested: 1) Antisense RTs undergo positive selection within genes, 2) Sense RTs undergo negative (purifying) selection within genes, and 3) Both (1) and (2) are true. Previous reports have indicated that Alu undergo purifying selection on a genome-wide level (Stewart et al., 2011). However, two previous publications proposed that the increase of Alu in GC-rich regions and upstream of the transcription start site of genes was due to positive selection (Tsirigos & Rigoutsos, 2009; Lander et al., 2001) though one of these two articles gave no experimental evidence to support these conclusions. Neither of these studies have addressed the selection of RTs within gene boundaries, which is important to study due to the potential disruptions of RTs directly on transcription as well as regulation of both transcription and translation (Thornburg et al., 2006; X.-F. Wang et al., 2009).

In 1975 Li calculated that, on average, deleterious alleles have shorter residence times in a population. In other words, they will be younger than neutral or advantageous alleles (W. H. Li, 1975). The logic behind this finding lies in that, in populations of equal size, an allele is more likely to be passed on to the next generation if it is neutral or positive compared to an allele that is deleterious. The neutral or positive allele will therefore remain in the population longer, accumulating a longer residence time in the genome. Meanwhile, new alleles, both advantageous and deleterious will continually appear. While the advantageous alleles are retained, the deleterious alleles will continually be lost, with only young alleles that have replaced them remaining present. Therefore at any discrete point in time, the deleterious alleles will be younger on average than the advantageous or neutral alleles.
Using the rationale presented in Li, 1975; the relative force of selection can be compared between two, otherwise equal, retrotransposon groups by calculating their average residence times in a population. RTs within a subfamily are an optimal sample to study selective forces, since they have the advantage of being amplified at similar points in time. Therefore, elements of the same subfamily should have similar residence times in the population when calculated as millidiv. However, if there are deviations in the forces of selection between subsets of the subfamily (i.e: AS versus S, or intergenic versus within genes), then the residence times in the population should be unequal.

2.3.1 Selection of L1 within Genes

To study the potential of selection acting on RTs, it was important to determine if the age, in millidiv, of an L1 element was contingent on the placement within the genome (i.e.: intergenic, AS within genes, or S within genes). To examine the association of age with placement, millidiv was regressed on each possible location of the element (i.e.: intergenic, sense, or antisense) (equation 2.1).

Since each element was expected to have the respective subfamily contribute a larger effect to millidiv than the placement, the effect of subfamily needed to be taken into account in the model. For example, a currently active L1 would be expected to have a smaller millidiv value (i.e.: younger) than an L1 that was inserted before the split of primates (i.e.: older), independent of forces of selection. Each subfamily was therefore tested separately, and multiple testing correction was applied.

ANOVA was calculated for the effect of location for each subfamily and plotted as a function of the average millidiv in the entire subfamily within genes (figure 2.4). 79 of the 114 subfamilies had millidiv values that were significantly affected by genomic locations after Bonferroni correction for multiple testing (α = 4.3e-04). A subset of the subfamilies had relatively small sample sizes (blue gradient in figure 2.4). As expected due to a reduction in power, these subfamilies did not show an association between genomic location and millidiv. Overall, the significance of the association with location increased with increasing mean mil-
Figure 2.4 – Age of L1 By Subfamily

The age L1 were regressed on the placement of the L1 (i.e.: intergenic, sense, or antisense). Regression equations were calculated separately for each L1 subfamily. The significance of each test is plotted versus the average age of the entire subfamily in millidiv)

These results showed that with increasing age of an L1 subfamily, there was increased disparity in the ages of elements within different genomic regions. Since all L1 elements within a subfamily are expected to have similar millidiv, independent of the location within the genome, these results supported the hypothesis that a force of selection was acting on RTs in a manner that is dependent on genomic context.

\[
\text{millidiv} = \beta_0 + \beta_{\text{location}_{\text{subfamily},i}} + \epsilon
\]  

Next it was determined if the association between location and millidiv was driven more by the effect between genes and intergenic elements, or the effect between S and AS elements.
The majority of subfamilies with significant comparisons showed a significant difference between intergenic elements and elements within genes. 77 subfamilies had elements that were significantly different in age between intergenic and within gene regions (tests =114 ; Bonferroni corrected $\alpha = 4.4e-04$). 76 of these 77 subfamilies were older in intergenic regions than within genes. As mentioned previously, shorter residence times in a population are associated with a lower likelihood of being passed to the next generation. Therefore, these results indicated that L1 within genes were less likely to be passed on to the next generation than the same elements in intergenic regions. This substantiated the hypothesis that there were differential effects contributing to the selection of L1 within genes.

When only L1 within genes were analyzed, there was a smaller number of subfamilies with significant effects driven by genomic position. 10 subfamilies had significant differences in age between AS and S orientations, 7 of which were older in the antisense orientation. This indicated that S oriented L1 were less likely to be passed on to the next generation than AS oriented L1.

### 2.3.2 Selection of Alu within Genes

Alu were similarly analyzed for differences in age depending on position and orientation. Similar to L1, there was an increasingly significant effect of Alu placement on age (figure 2.5), and subfamilies with reduced sample size showed no observable association with location. 14 of the 14 subfamilies that were significantly different in age between intergenic and within genes had larger millidiv values intergenically (tests = 36 ; Bonferroni corrected $\alpha = 1.4e-03$). The trend of older Alu elements in intergenic regions indicated that Alu were commonly under a more negative selective force when within genes compared to the same subfamilies in intergenic regions. Similarly 20 of the 20 subfamilies that were significantly different in age between S and AS orientations were older in the AS orientation. This substantiates the hypothesis that S Alu were less likely to be passed down to the next generation than AS Alu of the same subfamilies.
2.3.3 Evidence of Purifying Selection

The above results indicated that both Alu and L1 were undergoing differential levels of selection within genes compared to intergenic elements. These results, however, did not show whether negative (purifying) or positive selection was the driving force. If intergenic elements were assumed to be neutral, then the conclusion would be that RTs within genes were undergoing purifying selection due to the younger ages of elements within genes.

To test this hypothesis of purifying selection on RTs in genes, the prior knowledge of the differing rates of selection on chrX, chrY, and the autosomes were used. It’s been shown previously that the ability for purifying selection to remove deleterious alleles is reduced on
chrY compared to chrX and the autosomes (section 1.4). Moreover, selection can be increased on chrX due to the time the mutation spends in the hemizygous state when in males. Therefore, if purifying selection is indeed acting on RTs within genes, then the age of RTs should be the lowest on chrX and the autosomes, while chrY would be expected to have accumulated older elements due to the lack of purifying selection.

To have a broad idea of the age of RTs across chromosomes, each genomic orientation group (i.e: intergenic, sense, antisense) was analyzed with all subfamilies aggregated together. Although aggregating all subfamilies into a single measure was expected to wash out much of the signal, there were significant differences between chromosomes when analyzed this way (table 2.4 and table 2.5).

<table>
<thead>
<tr>
<th>Element</th>
<th>Location</th>
<th>Chromosome</th>
<th>Millidiv, Mean</th>
<th>Millidiv, SD</th>
<th>N_elements</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alu</td>
<td>sense</td>
<td>x</td>
<td>114.70</td>
<td>38.70</td>
</tr>
<tr>
<td>2</td>
<td>Alu</td>
<td>sense</td>
<td>aut</td>
<td>120.62</td>
<td>39.34</td>
</tr>
<tr>
<td>3</td>
<td>Alu</td>
<td>sense</td>
<td>y</td>
<td>135.95</td>
<td>49.40</td>
</tr>
<tr>
<td>4</td>
<td>Alu</td>
<td>anti</td>
<td>x</td>
<td>116.53</td>
<td>37.92</td>
</tr>
<tr>
<td>5</td>
<td>Alu</td>
<td>anti</td>
<td>aut</td>
<td>120.78</td>
<td>38.99</td>
</tr>
<tr>
<td>6</td>
<td>Alu</td>
<td>anti</td>
<td>y</td>
<td>137.55</td>
<td>48.78</td>
</tr>
<tr>
<td>7</td>
<td>Alu</td>
<td>int</td>
<td>x</td>
<td>118.45</td>
<td>40.76</td>
</tr>
<tr>
<td>8</td>
<td>Alu</td>
<td>int</td>
<td>aut</td>
<td>121.31</td>
<td>40.43</td>
</tr>
<tr>
<td>9</td>
<td>Alu</td>
<td>int</td>
<td>y</td>
<td>149.29</td>
<td>55.50</td>
</tr>
<tr>
<td>10</td>
<td>L1</td>
<td>sense</td>
<td>x</td>
<td>186.74</td>
<td>80.64</td>
</tr>
<tr>
<td>11</td>
<td>L1</td>
<td>sense</td>
<td>aut</td>
<td>207.98</td>
<td>79.02</td>
</tr>
<tr>
<td>12</td>
<td>L1</td>
<td>sense</td>
<td>y</td>
<td>191.20</td>
<td>81.32</td>
</tr>
<tr>
<td>13</td>
<td>L1</td>
<td>anti</td>
<td>x</td>
<td>192.60</td>
<td>74.85</td>
</tr>
<tr>
<td>14</td>
<td>L1</td>
<td>anti</td>
<td>aut</td>
<td>210.52</td>
<td>74.06</td>
</tr>
<tr>
<td>15</td>
<td>L1</td>
<td>anti</td>
<td>y</td>
<td>213.14</td>
<td>76.01</td>
</tr>
<tr>
<td>16</td>
<td>L1</td>
<td>int</td>
<td>x</td>
<td>188.12</td>
<td>78.32</td>
</tr>
<tr>
<td>17</td>
<td>L1</td>
<td>int</td>
<td>aut</td>
<td>206.79</td>
<td>79.44</td>
</tr>
<tr>
<td>18</td>
<td>L1</td>
<td>int</td>
<td>y</td>
<td>201.73</td>
<td>82.99</td>
</tr>
</tbody>
</table>

Table 2.4 – Aggregated ages of Alu and L1 by chromosome
Location refers to the genomic location in which the RT resided where int = intergenic; sense = RT oriented in the sense direction with respect to the gene; anti = RT oriented in the antisense direction. Chromosomes were separated into chrX (x), chrY (y), and the autosomes (aut). Mean = mean of millidiv of all of the elements in the respective category. SD = standard deviation of millidiv of all of the elements in the respective category. N = number of elements in the respective category. The row denoting chrX had two values in the Significance column. The first value corresponds to the test of chrX versus autosomes, and the second value corresponds to the test of chrX versus ChrY. The row denoting autosomes values has one value in the Significance column. This corresponds to the test between autosomes and ChrY. The row denoting chrY values has no values in this column since the comparisons were presented in either the chrX column or the autosome column. In the Significance column 1 = Student’s t-test is significant after Bonferonni correction, α = 0.0028; 0 = not significant.
Table 2.5 – Significant comparisons of age
The age of elements within each group from table 2.4 were compared using a Student’s t-test. The significant comparisons are presented here. 1 = first chromosome in comparison; 2 = second chromosome in comparison.

<table>
<thead>
<tr>
<th>Element</th>
<th>Orientation</th>
<th>1 Direction</th>
<th>2 Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Alu</td>
<td>sense</td>
<td>x</td>
<td>&lt; aut</td>
</tr>
<tr>
<td>2 Alu</td>
<td>sense</td>
<td>x</td>
<td>&lt; y</td>
</tr>
<tr>
<td>3 Alu</td>
<td>sense</td>
<td>aut</td>
<td>&lt; y</td>
</tr>
<tr>
<td>4 Alu</td>
<td>anti</td>
<td>x</td>
<td>&lt; aut</td>
</tr>
<tr>
<td>5 Alu</td>
<td>anti</td>
<td>x</td>
<td>&lt; y</td>
</tr>
<tr>
<td>6 Alu</td>
<td>anti</td>
<td>aut</td>
<td>&lt; y</td>
</tr>
<tr>
<td>7 Alu</td>
<td>int</td>
<td>x</td>
<td>&lt; aut</td>
</tr>
<tr>
<td>8 Alu</td>
<td>int</td>
<td>x</td>
<td>&lt; y</td>
</tr>
<tr>
<td>9 Alu</td>
<td>int</td>
<td>aut</td>
<td>&lt; y</td>
</tr>
<tr>
<td>10 L1</td>
<td>sense</td>
<td>x</td>
<td>&lt; aut</td>
</tr>
<tr>
<td>11 L1</td>
<td>sense</td>
<td>aut</td>
<td>&gt; y</td>
</tr>
<tr>
<td>12 L1</td>
<td>anti</td>
<td>x</td>
<td>&lt; aut</td>
</tr>
<tr>
<td>13 L1</td>
<td>anti</td>
<td>x</td>
<td>&lt; y</td>
</tr>
<tr>
<td>14 L1</td>
<td>int</td>
<td>x</td>
<td>&lt; aut</td>
</tr>
<tr>
<td>15 L1</td>
<td>int</td>
<td>x</td>
<td>&lt; y</td>
</tr>
<tr>
<td>16 L1</td>
<td>int</td>
<td>aut</td>
<td>&gt; y</td>
</tr>
</tbody>
</table>

Figure 2.6 – Clustering of Alu by Age
Red: Sense; Green: Antisense; Black: Intergenic. A = Autosome, X = chrX; Y = chrY
2.3.4 Evidence of Purifying Selection on Sense and Antisense Alu

As expected, within sense, antisense, and intergenic Alu, elements on chrX had significantly lower millidiv than those on autosomes, which were significantly less than those on chrY. This result showed that chromosomes with heightened ability for purifying selection had elements that were significantly younger, supporting that purifying selection is indeed acting on Alu RTs. In fact, the effect of rate of selection as identified by type of chromosome had a larger effect on element age than the orientation, which could be observed when millidiv values were clustered together (figure 2.6). For example, sense Alu on chrY were significantly older than antisense Alu on autosomes.

2.3.5 Evidence of Purifying and Positive Selection on Sense and Antisense L1

Figure 2.7 – Clustering of L1 by Age
Red: Sense; Green: Antisense; Black: Intergenic. Intergenic. A = Autosome, X = chrX; Y = chrY
When L1 elements were clustered by age, there was a disparity in grouping of elements together by chromosome. L1 elements on autosomes were clustered in one major branch and L1 elements on chrX were clustered on a second major branch (figure 2.7). However, L1 on chrY were interspersed within chrX and autosomes. When compared to using Student’s t-tests S L1 on chrY were significantly younger than AS L1 on chrY, indicating a potential difference in selective force acting on S and AS L1 on chrY.

2.4 Conclusions

In summary the results within this chapter suggested that the AS bias of Alu and L1 RTs were due to forces of selection. It was found that the AS bias present in the current human genome was not present upon integration, and that this bias was produced quickly over evolutionary time in Alu, and over longer periods of time for L1. The ages associated with these RTs further indicated that there were varied levels of selective pressures exerted on Alu and L1 depending on their position and orientation within the genome. In general, Alu and L1 within genes were younger than the same subfamilies intergenically. This was evidence that Alu and L1 may be under purifying selection within genes. Moreover, in regions with restricted purifying selection (i.e.:chrY), there was a reduction of both AS bias and an increase in age of Alu RTs. These results substantiated the hypothesis that RTs were initially integrated without bias and became non-randomly distributed over time due to purifying selection. These findings warrant further study into the variables affecting the non-random distribution of RTs within genes to determine what variables associated with both RTs and with genes that effect the level of selection.
Retrotransposons (RTs) comprise approximately half of the human genome and contribute to chromatin structure, regulatory motifs, and protein-coding sequences (Lander et al., 2001; Krull et al., 2005; Englander & Howard, 1995). Upon insertion, RTs can disrupt functional genetic elements as well as introduce new sequence motifs to a region (Belancio, Hedges, & Deininger, 2008). They therefore have the potential to effect the function of genes in which they insert.

It is hypothesized that this phenotypic impact of RTs on genes is a contributing factor to the non-random distribution of RTs across the genome. An example of non-random distribution being the propensity for a decrease of RTs in the sense orientation within genes (i.e: antisense bias (AS bias)). Although some of the variables associated with non-random distributions of RTs are understood (section 1.5), many variables remain unstudied. Specifically, there are many variables that have the potential to impact gene expression that have not been analyzed in association with RT distribution.

In order to understand the impact of RTs within differing genomic locations, it is important to characterize the associations of different variables with RT distribution. Determining which variables are most associated with decreases in RT distribution could help inform predictions of the severity of phenotypic consequence of de novo as well as polymorphic RT insertions. RT distribution analyses would therefore inform future disease studies attempting to associate genetic variability with disease phenotypes.

There are many genomic variables that can be tested for association with changes in RT distribution. Variables that are associated with i.) the RT (e.g. length of element, orientation,
sequence) as well as variables that are associated with ii.) the gene (e.g. tissue expression, density of coding sequence) will be examined in this chapter.

3.1 The Current Distribution of Alu and L1 within Genes

Since the previous chapter indicated that the ratio of antisense to sense elements (AS bias) was due to a force of selection, the AS bias was used as the initial measure of distribution of RTs in relation to the gene. The dependency of AS bias on position within a gene was calculated for both Alu and L1, where position was defined as one of ten deciles of a gene. RefSeq genes were divided into these deciles using the boundaries of transcription start site (TSS) and transcription stop site, including introns (figure 3.1).

![Figure 3.1 – Normalizing Genes by Length](image)

**Figure 3.1 – Normalizing Genes by Length**
Example of a gene divided into deciles. Deciles are divided by black dashed lines and represented by the orange number above the divisions. Decile 1 contains the first 0-10% of the gene, and decile 10 contains the last 90-100% of the gene. Deciles were calculated from transcription start to transcription stop site from the RefSeq annotated genes. The long boxes at the start and stop represent untranslated regions, the smaller internal boxes represent coding exons, and the green line represents the transcribed gene including introns.

Previous research had used distance from the TSS to examine RT density. The alternative approach presented here of normalizing genes by length expanded the scope of the analysis to examine the effect of relative position on RT density. This enables the detection of regions of the gene that were more sensitive to RT insertions independent of length. A side-effect of normalizing by the TSS and transcription stop positions was that every decile 1 (first decile) and decile 10 (last decile) necessarily contained an exon. Since deciles 2 - 9 (internal deciles) were not defined by any discrete exon boundary, they were variable with respect to the presence of an exon. For this reason many analyses were restricted to the internal deciles. It must be
noted that although the first and last deciles inevitably contained an exon, it was not necessarily a coding exon (i.e.: 5’-UTR or 3’-UTR only).

3.1.1 L1 and Alu Antisense Bias Exhibits a Linear Decay Across Gene Deciles

When analyzed by raw distance in bp, RTs are increased in number near TSSs compared to transcription stop sites (Tsirigos & Rigoutsos, 2009). It was first important to determine if examining RT distribution, while normalizing genes by length, replicated these findings. AS and S RTs were counted within each decile across all RefSeq genes and normalized by the total number of RTs in all deciles (AS + S) for both Alu and L1.

**Alu**

AS and S Alu were counted in each decile and normalized by the total number of Alu within all deciles (figure 3.2). Replicating previous findings, there was an increase in elements near the TSS (decile 1, d1) compared to the end of the gene (decile 10, d10). This change was associated with a significant difference in the proportion of AS Alu between deciles 1 and 10 (Test of unequal proportions: prop\(_d1\) = 0.598 ; prop\(_d10\) = 0.479; p < 5.67e-262). Also as expected due to the known AS-bias, there were more AS Alu than S Alu overall. Alu in the S orientation were distributed relatively equally across the internal deciles with drops in density in deciles 1 and 10. Conversely, AS Alu exhibited a linear decay in distribution from decile 2 to decile 9. Similarly to Alu on the S strand, AS Alu also showed decreased in density in deciles 1 and 10. This decrease was outside what would be expected based on the linear trend from deciles 2 - 9. Therefore, although they were increased near the TSS more than the transcription stop site, they were decreased compared to decile 2.

**L1**

L1 had a similar distribution to Alu when within genes. Again, as expected due to the overall AS bias, there was an increase in AS L1 compared to S L1. Furthermore there was an increase of elements in d1 compared to d10. This increase was associated with a significant change in proportion of AS L1 (Test of unequal proportions: prop\(_d1\) = 0.709 ; prop\(_d10\) = 0.588; p <
Counts of Alu in the AS and S orientation per decile were normalized by the total Alu in all RefSeq genes. 
as = antisense, s = sense

1.32e-166). Throughout the internal deciles (2 - 9), S L1 remained at a relatively constant distribution, though it must be noted that there was more variability between these deciles, than observed with Alu. AS L1 exhibited a linear decay across the internal deciles (figure 3.3). The overall difference between AS L1 and S L1 within each decile, was larger than what was observed in Alu. This was emphasized by AS and S elements having reached a point of equal density in the last inner decile for Alu, while AS L1 remain more prevalent than S in every decile.

**Linear Decay of Alu and L1**

Both AS Alu and L1 began at a relatively high frequency in decile 2 and declined linearly until decile 9. The rate of AS decay across gene deciles was determined by analyzing these internal deciles only. This avoided the confounding effects from the process of normalization. Slopes were calculated with the counts of raw elements as opposed to the proportions to facilitate
Figure 3.3 – AS bias of L1 decays linearly across gene boundaries
Counts of L1 in the AS and S orientation per decile were normalized by the total L1 in all RefSeq genes. as = antisense, s = sense

interpretation (table 3.1). As expected, the decile number had a significant impact on AS density in both Alu and L1, with Alu reduced by 994 elements per decile and L1 by 1410 elements per decile. S elements were not significantly affected by the decile within which they resided (table 3.1).

<table>
<thead>
<tr>
<th>Element</th>
<th>Orientation</th>
<th>Estimate</th>
<th>$p_{\text{intercept}}$</th>
<th>$p_{\text{decile}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Alu</td>
<td>AS</td>
<td>-1410.67</td>
<td>7.45e-12</td>
<td>5.53e-08</td>
</tr>
<tr>
<td>2 Alu</td>
<td>S</td>
<td>-27.17</td>
<td>2.33e-10</td>
<td>6.08e-01</td>
</tr>
<tr>
<td>3 L1</td>
<td>AS</td>
<td>-993.60</td>
<td>2.40e-08</td>
<td>2.00e-04</td>
</tr>
<tr>
<td>4 L1</td>
<td>S</td>
<td>32.88</td>
<td>3.09e-07</td>
<td>6.91e-01</td>
</tr>
</tbody>
</table>

Table 3.1 – Slope of linear decay for Alu and L1 within the internal deciles of genes
Decile number was regressed on element count in the antisense (AS) and sense (S) orientations for each element. $p_{\text{intercept}}$ = F-test p-value of the intercept of the regression line. $p_{\text{decile}}$ = F-test p-value of the effect of decile on the regression line.

Previous research suggested that RTs increased in a peak around the TSS. The results presented here using normalized gene lengths revealed that RTs are not only increased in density near the TSS, but are increased as a function of the normalized distance from the TSS. Moreover, the enrichment of elements is restricted to AS elements. If the increase near the TSS
was purely due to a structural change in accessibility, it would be expected that both AS and S elements would be increased near the TSS. Therefore, these results indicated that there may be a functional link between the orientation of elements and the propensity to which they are maintained within genes. Furthermore, there is a difference in this propensity near the TSS as opposed to near the 3'-end of a gene.

3.2 Gene-specific Variables Associated with Retrotransposon Distribution

The results so far have indicated that the AS bias of RTs within genes is dependent on the positional context of the element within the gene. Elements that integrate near the 5’-end of a gene were distributed differently with respect to orientation than the elements at the 3’-end of the gene. It is possible that variables associated with the RT as well as variables associated with the gene effect this distribution. First, variables associated with the gene will be explored.

3.2.1 Linear Decay of AS Bias is Associated with Coding Sequence Density

The linear decay of RTs across gene deciles may be contributed to by a number of factors. The most likely of these is the presence of coding sequence exons (CDS), since these are the most conserved units of a gene. Furthermore, Zhang et. al. showed that RTs were less often located in close proximity to an exon. Therefore the density of CDS within each decile was calculated. Indeed, CDS density increased at a linear rate across gene deciles (figure 3.4). The first decile of genes contained more CDS than the second decile. AS Alu were inversely correlated with this distribution of CDS (R² = -0.94; one-tailed p < 2.64e-05). S Alu, however, were not correlated with CDS (R² = -0.30; one-tailed p = 0.200). These results suggest that AS RTs are highly inversely correlated with CDS exon density.

3.2.2 Alternative Splicing

Alu and L1 both contain cryptic splice sequences that are often used for alternative splicing. It was therefore investigated whether genes with alternative isoforms have a different AS bias than constitutively expressed genes. Using the knownAlt annotation track from Hg19 (UCSC
Figure 3.4 – Distribution of Coding Sequence Exons in RefSeq Genes
The number of coding sequence exons (CDS) in the sense orientation were normalized by all sense oriented CDS in RefSeq genes.

Genome Browser), we separated genes into those with known alternative splicing events, and those that were constitutively expressed. We found no significant differences between the distribution of AS and S between the two groups (chi-square p > 0.25, p > 0.23 respectively). We further calculated the correlation of AS Alu density and CDS density within alternatively spliced genes. The negative correlation that is observed across all genes was replicated when examining only genes that are alternatively spliced (Pearson’s correlation coefficient: -0.78, p < 0.008). Therefore, there was no observable difference between alternatively spliced and constitutively spliced genes.

3.2.3  Alu and L1 are Decreased Near 5’-UTR

Although previous studies using raw distances showed that RTs were increased near TSSs, here it was noted that AS were increased over S at a ratio less than expected based on the linear trend from the internal deciles. Therefore, although there was an increase near the TSS, this increase was modified by another factor such as the exon density. As mentioned earlier, the first decile was defined by the TSS and was therefore enriched for the presence of 5’-UTR exons (figure 3.7), however the presence of a coding exon (CDS) was not obligatory by the normalization procedure (figure 3.4). Although the RT density was highly correlated with CDS density, it was important to study if AS RT density near the TSS was modified by the presence of the 5’-UTR.
This was especially important because it had been previously claimed that RTs were increased near the TSS due to positive selection (Tsirigos & Rigoutsos, 2009), a hypothesis that would be discounted if instead the increase near the TSS was due only to a reduction in CDS.

To determine the effect of the 5′-UTR on RT density, the effect of the 5′-UTR had to be separated from any latent effects correlated with the first decile. To separate out these effects, genes were identified that contained a second 5′-UTR (2° 5′-UTR) within deciles other than the first decile. These 2° 5′-UTRs were commonly due to either alternative start positions of transcription, or transcription of a separate gene that began within the boundaries of another gene. Examples of genes that contained 2° 5′-UTRs are depicted in figure 3.5.

Once genes with 2° 5′-UTRs were identified, they were further grouped into subsets based on the decile location of the 2° 5′-UTRs. For example, Group I genes contained all 2° 5′-UTRs in the first decile. Group II contained genes where the 2° 5′-UTRs were in the second decile, and so on for all ten deciles (figure 3.6). The distribution of 5′-UTRs within genes with no 2° 5′-UTR is presented in figure 3.7, and the distribution within genes these containing a 2° 5′-UTR is presented in figure 3.8. As expected, the number of 5′-UTRs were generally higher.
in the decile containing the 2° 5'-UTR (normalized decile 0). This indicated that the procedure of normalizing by 2° 5'-UTR enriched for 5'-UTR sequences in deciles other than the first decile.

In order to interpret the findings properly, it is important to note that each normalized decile did not contain equal numbers of measurements. This is due to the fact that the calculations of upstream and downstream values were contingent on the placement of the 0 decile. Therefore if, before normalization, the zero decile was originally decile 9, there would only be one decile downstream capable of being analyzed and the remaining downstream deciles would not have a value attributed to them. For this reason all deciles are shown in the analyses, however only the 0 decile was used to calculate changes in genes with a 2° 5'-UTR versus those that did not have an internal 5'-UTR.
Alu

The frequencies of all AS and S L1 and Alu were calculated within 2° 5′-UTR gene groups and compared to expectation based on genes lacking any 2° 5′-UTRs. Overall AS Alu in genes containing 2° 5′-UTRs were present in similar proportions to the AS Alu in genes lacking 2° UTRs (figure 3.9). The normalized 0 decile was significantly reduced compared to the same decile in genes without 2° UTRs (mean\textsubscript{with 2° 5′-UTR} = 0.049; mean\textsubscript{without 2° 5′-UTR} = 0.058; Student’s t-test p < 0.039). S Alu were also similar to expectation except for within the 0 decile (figure 3.10). The 0 decile was significantly decreased beyond expectation (mean\textsubscript{with 2° 5′-UTR} = 0.41; mean\textsubscript{without 2° 5′-UTR} = 0.46; Student’s t-test p < 5.2e-04) similar to the AS Alu.

L1

The distribution of AS L1 within genes containing a 2° 5′-UTR was somewhat similar to that observed with AS Alu (figure 3.11). The 0 decile of AS L1 was significantly reduced
Figure 3.8 – Distribution of 5′-UTRs in genes containing 2nd 5′-UTRs
Proportion of 2nd 5′-UTRs were calculated after normalization as described in figure 3.6. Decile 0 is the decile containing the 2nd 5′-UTR. All negative deciles are upstream of the 2nd 5′-UTR, and all positive deciles are downstream.

Figure 3.9 – Change in AS Alu density in the presence of a 2nd 5′-UTR
Proportion of 2nd AS Alu proportions were calculated as described in figure 3.6. All proportions were normalized by the expected proportion based on RefSeq genes that have no 2nd within the gene boundaries.
Figure 3.10 – Change in S Alu density in the presence of a secondary 5′-UTR
Calculated similarly to figure 3.9

over expectation (mean<sub>with 2° 5′-UTR</sub> 0.057 ; mean<sub>without 2° 5′-UTR</sub> 0.071 ; Student’s t-test p < 0.007), although the S L1 were not significantly reduced (p = 0.47) (figure 3.12).

Together the decrease of AS Alu and L1 in the decile containing the 2° 5′-UTR indicated that there was an effect of the 5′-UTR on RT density that was independent of the position within the gene. Interestingly, this finding is contrary to expectation. Since the RT density was increased near the TSS and therefore near the 5′-UTR it was hypothesized that there would be an increase of RTs near the 5′-UTR.

3.2.4 Coding Sequence Accounts for Variability Near 5′-UTR

The results so far suggested that, contrary to expectation, there was a decrease of RTs near the 2° 5′-UTR when the position within the gene was controlled for. Importantly, in section 3.2.1 it was shown that RTs were inversely correlated with CDS density. Since the first CDS exon of a gene is often, but not always, connected to the 5′-UTR, it was important to determine if the CDS density had an effect on the RT distribution observed in genes containing 2° 5′-UTRs.
CDS Density Differs Near 2° 5’-UTR

First, the CDS density within genes containing 2° 5’-UTRs was identified. Indeed the CDS density within genes that containing a 2° 5’-UTRs were altered compared to the CDS density in genes with no internal 2° 5’-UTRs (figure 3.13 and figure 3.4). The most striking feature
was the CDS density in the decile containing the $2^\circ$ 5'-UTR. These genes showed a significant increase in CDS density compared to expectation based on genes lacking internal 5’-UTRs (table 3.2). This further substantiated studying the effect of CDS density on RT density in genes containing $2^\circ$ 5'-UTRs.

**Effect of CDS Density on Alu and L1**

The CDS density and Alu density within non-$2^\circ$ 5'-UTR containing genes was used to predict the expected Alu density in $2^\circ$ 5’-UTR containing genes. To accomplish this, Alu density from non-$2^\circ$ 5’-UTR containing genes was regressed on CDS density from the same genes. The expected RT density was then predicted using the CDS density from genes containing $2^\circ$ 5’-UTRs. For Alu, the actual values were plotted next to the predicted values in figure 3.14. In the AS orientation, there was no difference in Alu or L1 density from what was expected based on CDS density. Conversely, in the S orientation, there were less Alu and L1 in regions containing a 5’-UTR than expected based on CDS density alone (Student’s t-test $p_{Alu} < 0.03$ ; Student’s t-test $p_{L1} < 0.001$ ).
Table 3.2 – CDS density in genes with 2° 5’-UTRs

<table>
<thead>
<tr>
<th>2° decile</th>
<th>CDS²</th>
<th>CDS all deciles²</th>
<th>CDS⁰no²</th>
<th>Prop²</th>
<th>Prop⁰no²</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>157</td>
<td>1774</td>
<td>12526</td>
<td>0.09</td>
<td>0.06</td>
<td>3.06e-07</td>
</tr>
<tr>
<td>3</td>
<td>118</td>
<td>1080</td>
<td>14264</td>
<td>0.11</td>
<td>0.07</td>
<td>7.55e-08</td>
</tr>
<tr>
<td>4</td>
<td>110</td>
<td>984</td>
<td>16055</td>
<td>0.11</td>
<td>0.08</td>
<td>3.24e-05</td>
</tr>
<tr>
<td>5</td>
<td>111</td>
<td>943</td>
<td>18119</td>
<td>0.12</td>
<td>0.09</td>
<td>6.08e-04</td>
</tr>
<tr>
<td>6</td>
<td>101</td>
<td>794</td>
<td>20213</td>
<td>0.13</td>
<td>0.10</td>
<td>3.00e-03</td>
</tr>
<tr>
<td>7</td>
<td>103</td>
<td>781</td>
<td>22518</td>
<td>0.13</td>
<td>0.11</td>
<td>2.63e-02</td>
</tr>
<tr>
<td>8</td>
<td>104</td>
<td>774</td>
<td>24746</td>
<td>0.13</td>
<td>0.12</td>
<td>1.50e-01</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>659</td>
<td>28504</td>
<td>0.15</td>
<td>0.14</td>
<td>2.31e-01</td>
</tr>
</tbody>
</table>

CDS were counted in genes containing 2° 5’-UTRs (column CDS²) and those with no internal 2° 5’-UTRs (column CDS⁰no²). The counts in each row correspond to the decile tabulated in the column denoted as 2° decile. For genes containing a 2° 5’-UTR, this was the decile containing the 2° 5’-UTR. The proportions of CDS were calculated by normalizing versus all of the CDS in all deciles. For genes containing a 2° 5’-UTR this was the value in column CDS all deciles². For genes with no 2° 5’-UTR, the total CDS in all deciles was 211134. The proportions were compared between genes with and without 2° 5’-UTR in the same deciles using a test of equal proportions.

In summary of the effect of CDS and 5’-UTRs on RT density, there was no evidence of an increase near the 5’-UTR when position within the gene was normalized. Instead, there was significant decrease in Alu density in the S orientation. Furthermore, the decrease of AS Alu was largely correlated with an increase in CDS sequence near 5’-UTRs. This suggested that the increase of AS RTs near the 5’-UTR was not an effect of the UTR, but an effect of the lack of CDS sequences in deciles that were near the 5’-end of a gene.

3.2.5 Tissue-Specificity of the Gene Effects Antisense Bias

Housekeeping genes have a predictably higher level of Alu frequency than other gene categories (Eller et al., 2007). It was therefore tested whether AS bias remained linear when subdivided by tissue-expression category. The effect of tissue-specificity was tested by both analyzing the slope of the density of AS as well as by correlating the distribution of Alu with the distribution of CDS within tissue-specific gene sets. The linear AS slope remained present in housekeeping genes, but it was not maintained in tissue-specific gene sets. Furthermore, the correlation between AS frequency and CDS distribution was reduced within tissue-specific gene sets (table 3.3). Although the sample size of each gene set trended with the R², sample size was not significantly correlated with R² (R² = -0.68; p = 0.062).
The density of Alu was calculated within each 2° 5' UTR gene group separately. These values are represented by the boxplots labelled "actual". The values predicted by the correlation of Alu with coding sequences in genes with no 2° 5' UTR are represented by the boxplots labelled "predicted".

To determine if differences in the levels of gene expression across tissue types was confounding this result, we tested the correlation between AS frequency and exon density as a function of relative expression in housekeeping and tissue specific genes. There was no association between expression level in the respective tissue and the correlation between AS and exon density ($R^2 = 0.60$; $p = 0.15$).

Overall these results indicated that the tissue-specificity of the gene in which a RT is located was associated with the AS bias. AS bias was more prevalent in housekeeping genes, which also have higher numbers of RTs, and less prevalent in genes that were expressed specifically in the pancreas, colon, and liver.

### 3.3 Retrotransposon-specific Variables are Associated with Retrotransposon Distribution

The consensus sequences for Alu and L1 contain many functional DNA binding motifs. Some of these binding motifs act as enhancers, or repressors of gene expression. Although the con-
sensus sequence of RTs contain many binding motifs, each individual RT in the genome is variable as to whether or not it contains these motifs. The loss or gain of a motif may occur by mutation at the nucleotide level (figure 3.15), through retrotransposition of a parent element that has already lost or gained the motif sequence, or through truncation of the element during integration.

Protein binding motifs can impact both RT expression as well as gene expression. For example, the 5’ portion of the polymerase III promoter (Pol-III) contains the A-box motif which is necessary for transcription factor binding and expression of Alu retrotransposons. The YY-1 binding motif contained within L1 is both a component of the core promoter necessary for L1 transcription as well as a general transcriptional repressor. The SRY motif is a binding domain for the testis-determining factor gene SRY which increases L1 transcription in 293 cells, and RUNX3 binding can act as both an activator and repressor of transcription and whose repression is linked with tumor progression.

These binding motifs can have an effect on gene expression, and some of these motifs are orientation specific (e.g.: polyadenylation sequences). The relative presence of these motifs within genes was compared to intergenic motifs to determine whether some motifs were excluded from genes, thereby indicating a functional impact on the organism.

<table>
<thead>
<tr>
<th>Category</th>
<th>N, genes</th>
<th>$R^2_{AS,Alu}$</th>
<th>one-tailed $p_{AS}$</th>
<th>$R^2_{S,Alu}$</th>
<th>one-tailed $p_{S}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Housekeeping</td>
<td>2237</td>
<td>-0.96</td>
<td>5.34e-06</td>
<td>-0.40</td>
<td>0.126</td>
</tr>
<tr>
<td>Neutrophil</td>
<td>231</td>
<td>-0.76</td>
<td>5.37e-03</td>
<td>-0.21</td>
<td>0.280</td>
</tr>
<tr>
<td>Muscle</td>
<td>111</td>
<td>-0.53</td>
<td>0.058</td>
<td>0.02</td>
<td>0.478</td>
</tr>
<tr>
<td>Cartilage</td>
<td>133</td>
<td>-0.53</td>
<td>0.058</td>
<td>0.23</td>
<td>0.258</td>
</tr>
<tr>
<td>Testis</td>
<td>252</td>
<td>-0.51</td>
<td>0.066</td>
<td>-0.38</td>
<td>0.139</td>
</tr>
<tr>
<td>Liver</td>
<td>305</td>
<td>-0.43</td>
<td>0.107</td>
<td>-0.22</td>
<td>0.270</td>
</tr>
<tr>
<td>Colon</td>
<td>122</td>
<td>-0.16</td>
<td>0.329</td>
<td>0.03</td>
<td>0.467</td>
</tr>
<tr>
<td>Pancreas</td>
<td>125</td>
<td>-0.08</td>
<td>0.413</td>
<td>-0.47</td>
<td>0.085</td>
</tr>
</tbody>
</table>

Table 3.3 – Correlation of CDS and Alu across gene categories
Correlation of the proportion of Alu in each decile versus the proportion of CDS in each decile.
Figure 3.15 – Examples Alu Consensus Sequences that Contain Different Binding Motifs

AluJo and AluSc represent two Alu subfamilies. The consensus sequences for these two elements are presented respectively. As expected, both consensus sequences contain the RNA Polymerase III A-box that is required for Alu expression, and therefore Alu retrotransposition. One further protein binding region is highlighted in each sequence, though these binding regions are exclusive to the subfamily of origin. AluJo contains a polyadenylation sequence within the central A-rich region. This polyadenylation signal has been lost in the AluSc consensus sequence (AAATAAA | AAATAGA). Similarly AluSc contains a hepatocyte nuclear factor 4 binding site that is not present in AluJo (GAGGTCAAGAGATCG | GAGCCCAGGAGTTCG).

3.3.1 Truncated L1 Escape Depletion Near the 5’-UTR

The full-length human L1 consensus sequence is approximately 6kb, but varies by subfamily. More commonly, L1 are approximately 900 bp on average. The shorter average sequence compared to full-length is due to malfunctions during retrotransposition where L1 are often truncated. Reverse transcription of the RNA template initiates near the 3’end of RNA and often the L1 retrotransposase is unable to complete retrotransposition of the entire element into the insertion site, resulting in truncation of the 5’end.

This 5’-truncation of L1 results in variable sizes of L1 in the human genome, with a majority being less than 1 kb (table 3.4). The 5’-region of L1, that is often lost during truncation, contains the Polymerase II promoter which can effect expression of neighboring genes. It is likely that this promoter region of L1 effects gene expression. To determine if L1 with promot-
Table 3.4 – Size of L1 elements in the human reference genome
Counts were tallied of all L1 in the human reference genome.

<table>
<thead>
<tr>
<th>size, kb</th>
<th>L1 count</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 1</td>
<td>833043</td>
</tr>
<tr>
<td>1 - 2</td>
<td>76631</td>
</tr>
<tr>
<td>2 - 3</td>
<td>20349</td>
</tr>
<tr>
<td>3 - 4</td>
<td>8174</td>
</tr>
<tr>
<td>4 - 5</td>
<td>3787</td>
</tr>
<tr>
<td>5 +</td>
<td>10181</td>
</tr>
</tbody>
</table>

ers have a different distribution within genes than L1 without promoters, the distributions of L1 of varying sizes were calculated.

Since size was used as a proxy for the presence of the L1 promoter, it was important to determine the presence of the L1 promoter within short L1 (< 1kb). The L1 promoter region is located within the first 2kb of the L1 consensus sequence. To identify whether L1 that were <1kb lacked the promoter region, the start position of each short L1 (<1kb) within the respective consensus. A histogram of the start positions within the consensus sequence are presented in figure 3.16). The enrichment near 6kb indicated that the 5’-most portion of short L1 was located many kb downstream of the L1 promoter sequence (N, \( L1 > 2kb \) = 654963 ; 87%). Therefore, the majority of short L1 did not contain the canonical promoter sequence. It
was therefore hypothesized that short L1 would not be under the same purifying selection as full-length L1, and would be present at higher levels than full-length L1.

**Figure 3.17 – Effect of L1 Size on L1 density: Antisense**
L1 density was calculated within all deciles. L1 were first grouped into sizes (< 1 kb, 1-2, 2-3, 3-4, 4-5, and 5+).

**Figure 3.18 – Effect of L1 Size on L1 density: Sense**
L1 density was calculated within all deciles. L1 were first grouped into sizes (< 1 kb, 1-2, 2-3, 3-4, 4-5, and 5+).

To test whether the linear decay of antisense oriented L1 was related to the size of the element, the proportions of L1 <1kb, 1-2kb, 3-4kb, 4-5kb, and 5kb+ were calculated within all genes. All groups of AS L1 had a similar distribution to all AS L1 (figure 3.17). When all size ranges were compared within the first decile it was noted that the proportion of AS L1 was higher with decreased length of the element (figure 3.19). Importantly, L1 < 1kb in length were present at a significantly higher level than full-length L1 (prop <1kb = 0.0574 ; prop 5+ kb = 0.0437 ; Test for equalities of proportion p < 1.01e-04).

S L1 grouped by length were also similarly distributed to all S, however full-length L1 were significantly reduced in the first decile compared to short L1 (prop <1kb = 0.0237 ; prop 5+ kb = 0.0163 ; Test for equalities of proportion p < 0.0012).
Figure 3.19 – Effect of L1 Size on L1 density: Antisense, decile 1
L1 density was calculated within the first decile. L1 were first grouped into sizes (< 1 kb, 1-2, 2-3, 3-4, 4-5, and 5+). The lines on either side of the point represent the 95% confidence interval. Prop: Proportion of antisense L1 per all L1 of the same size within genes. kb: maximum length of the group of L1 in kilobases (6000 represent L1 that are 5kb or greater). The minimum of each group is 1kb less than the maximum except for L1 5kb and greater.

3.3.2 Protein Binding Motifs in Alu

Protein binding motifs other than the Pol-II promoter are highly concentrated in both Alu and L1 consensus sequences. Binding motifs within the Alu consensus sequence were analyzed for changes versus expectation. As noted previously, most motifs were only present in a subset of the consensus sequences. Therefore, binding motifs were first counted in a subset of Alu consensus sequences to determine the relative expectation of identifying each motifs. Some motifs were present multiple times within an individual consensus sequence (e.g.: CpG, OTX2), while some motifs were not present in any of the consensus sequences analyzed (e.g: BRN2, DR2). Other than the Pol-III (Pol3) promoter sequence, OTX2, PITX2, and LYF1 were present in all of the consensus sequences in the table.

The proportions of motifs within intergenic Alu were used to normalize the proportions of binding motifs identified within genes (figure 3.20 and . To correct for differences in the total
number of Alu within different regions of the genome, the count of motifs were first normalized by the number of elements within each region analyzed. After this normalization procedure, the proportion of any given motif was effected more by the placement within a gene, than the sequence of the motif. For example, all of the motifs located within AS Alu had values above 1.2, independent of the type of motif. Similarly, all of the motifs within S Alu had values below 0.8.

Figure 3.20 – Presence of Alu-contained protein-binding motifs within genes
Protein binding motifs were examined for their presence within Alu both in genes and intergenically. The y-axis represents the ratio of the presence within genes to the presence intergenically. Alu within genes were separated into categories based on the linear decline of AS-bias. AS_5prime: Antisense Alu in the 10-20% decile; S_5prime: Sense Alu in the 10-20% decile; AS_3prime: Antisense Alu in the 80-90% decile; S_3prime: Sense Alu in the 80-90% decile. The RNA Polymerase III A-box promoter (Pol3) is highlighted by an orange dashed box.

ANOVA was performed to determine if the effect of placement (i.e: sense, antisense, in genes, 5'-, 3'-) was significantly more than the effect of the sequence of the motif (equations
3.1 and 3.2). Placement did indeed have a more significant effect on the proportion of motif than the sequence of motif (p_{placement} < 2.2e-16; p_{sequence} = 0.99).

\[ y = \beta_o + \beta_{motif_i} + ... + \beta_{motif_n} + \varepsilon \]  
\[ (3.1) \]

\[ y = \beta_o + \beta_{position_i} + ... + \beta_{position_n} + \varepsilon \]  
\[ (3.2) \]

### 3.3.3 Protein Binding Motifs in L1

![Figure 3.21 – Presence of L1-contained protein-binding motifs within genes](image)

Protein binding motifs were examined similarly to those within Alu (figure 3.20).

The proportions of motifs in intergenic L1 were compared to the proportions within genes. As before, the values were normalized by the total number of elements in the respective region.
prior to comparison versus intergenic L1 (figure 3.21). Contrary to the pattern observed with Alu, L1 showed no trend of enrichment based on the placement within the gene. Instead the sequence of the motif had a significant effect on the proportion of motifs \( (p_{\text{placement}} = 0.78; p_{\text{sequence}} < 3.56e-13) \). Furthermore, unlike Alu, all motifs within L1 were reduced in genes compared to intergenically. SRY was the motif with the largest depletion within genes compared to L1 in intergenic regions (table 3.5).

<table>
<thead>
<tr>
<th>Genomic Region</th>
<th>( N_{L1} ) with SRY</th>
<th>( N_{L1} )</th>
<th>Proportion</th>
<th>Prop. test vs. int. ( p &lt; )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Intergenic</td>
<td>15770</td>
<td>598467</td>
<td>2.64e-02</td>
<td>NA</td>
</tr>
<tr>
<td>2 Genes</td>
<td>11840</td>
<td>669036</td>
<td>1.77e-02</td>
<td>2.53e-243</td>
</tr>
<tr>
<td>3 Sense</td>
<td>3306</td>
<td>227066</td>
<td>1.46e-02</td>
<td>2.31e-222</td>
</tr>
<tr>
<td>4 Antisense</td>
<td>8534</td>
<td>441970</td>
<td>1.93e-02</td>
<td>4.10e-122</td>
</tr>
</tbody>
</table>

Table 3.5 – Context Dependent Proportions of the SRY Motif in L1 elements The number of L1 were counted in each region \( (N_{L1}) \), as well as the number of L1 containing an SRY motif \( (N_{L1 \text{withSRY}}) \). The proportion of SRY-containing L1 was compared within each gene region versus intergenic using a test of equal proportions \( (\text{Prop. test vs. int. } p <) \).

These results indicated that the sequence of both Alu and L1 were differentially maintained in genes versus intergenically. Binding motifs contained within Alu were reduced in a pattern similar to the distribution of the whole elements. Conversely, binding motifs within L1 were reduced depending on the sequence of the motif.

### 3.3.4 Conservation at the Basepair Level of Binding Motifs

The previous analysis examined the distribution of binding motifs within human genes and indicate a general decrease of RT-contained motifs within genes. This lack of motifs in genes could have occurred due to base-wise mutations of the motifs over time, or to a selection against elements as a whole. It was therefore important to analyze the base-wise conservation status of binding motifs which were differentially represented based on genomic context.

Conservation values used in this analysis were derived from the phyloP basewise conservation scores, where positive values indicated evidence of positive selection, and negative values indicated evidence of purifying selection. Importantly, conservation was only calculated within regions that matched the binding motif within the human reference sequence. Therefore, as opposed to the previous analysis which modeled the loss of binding motifs in the human ref-
Figure 3.22 – Base-wise conservation of polyadenylation signals within S and AS L1

Base-wise conservation scores were taken from the phyloP conservation scores, where positive values indicate potential positive selection, and negative values indicate possible negative selection. The scores of each bp were calculated for intergenic L1, AS L1, and S L1. AS (pink dots) and S (blue dots) L1 were then normalized versus the intergenic values.

This calculation analyzed the individual base pairs within motifs that were present in humans but potentially lost in non-human primates.

Polyadenylation sequences were the first motif analyzed since the 5’-AAATAAAA motif was present within both Alu and L1 consensus sequences. This motif was reduced within genes compared to intergenic regions in both Alu and L1. Base-wise conservation scores were averaged for each base within the polyadenylation sequence of either L1 (figure 3.22) or Alu (figure 3.23). L1 and Alu elements were first separated into AS and S elements before conservation scores were calculated. Interestingly, both Alu and L1 showed a decrease in conservation at the central T base. Moreover, polyadenylation signals of L1 in the sense orientation had a lower conservation score on average than those of AS oriented L1.

These results indicated that each position within a RT-contained binding motif had a different expectation of conservation. Interestingly, the polyadenylation sequence, had a similar
Figure 3.23 – Basewise conservation of polyadenylation signals within S and AS Alu
Basewise conservation scores were taken from the phyloP conservation scores, where positive values indicate potential positive selection, and negative values indicate possible negative selection. The scores of each bp were calculated for intergenic Alu, AS Alu, and S Alu. AS (pink dots) and S (blue dots) Alu were then normalized versus the intergenic values.

dynamic of conservation independent of the element within which it resided. Moreover, the S L1 had a complete reduction of the polyadenylation signal, whereas the AS L1, S Alu, and AS Alu all had reductions only in the central thymine.

Conservation of the Alu-specific Pol-III A-box (figure 3.24) was also examined in both the AS and S orientations. The Pol-III A-box is required for transcription and retrotransposition of Alu. When normalized versus intergenic values, there was no major difference between AS and S conservation signals, similar to what was observed for the polyadenylation signal within Alu.

There are four CpGs within the Pol-III A-box consensus sequence. The cytosine within a CpG context is more likely to undergo mutation than other nucleotides due to the common deamination of 5-methylcytosine. Surprisingly, only one of the four CpGs present within the queried sequence (position 20) was less conserved than the same position intergenically. Two
Figure 3.24 – Basewise conservation of Pol-III A-box region in AS and S Alu
Conservation scores were derived from phyloP conservation scores where a positive value indicates positive selection and a negative value indicates possible negative selection. AS and S values were normalized by intergenic values.

bp in the Pol-III sequence were more highly conserved within genes than intergenically. These were the cytosine at position 3 and guanine at position 15.

3.4 Interaction between Gene-specific and Retrotransposon-specific Variables

These results have shown that both gene-specific and RT-specific variables were associated with RT density. It was therefore important to analyze the interaction of these variables. The density of RTs was therefore analyzed based on both the tissue-specific expression of the gene as well as the binding motifs contained within the element.
3.4.1 Presence of RT-contained Binding Motifs Based on Gene Type

To reduce the number of statistical tests, analysis was restricted to the internal deciles that contained the extremes of Alu and L1 frequency. These were deciles 2 and 9 (figure 3.1), otherwise defined as the first and last bins within the linear range of Alu/L1 frequency (section 3.1.1). To correct for differences in overall abundance of RTs depending on genomic location, motif frequencies were normalized by the count of AS or S elements within the respective bin. This was repeated separately for tissue-specific (TS) genes and house-keeping (HK) genes. All values are tabulated in .

Alu

![Graph showing Alu-contained binding motifs in TS versus HK Genes](image)

**Figure 3.25 – Alu-contained binding motifs in tissue-specific and housekeeping genes**
The proportion of binding motifs within tissue-specific genes were tested versus those in housekeeping genes using a fisher’s exact test. The Bonferroni cut-off alpha is represented by a horizontal line.
The proportion of motifs in each placement within TS genes were compared to the counterpart in HK genes (figure 3.25). In general, proportions of motifs in TS genes were larger than HK genes (87 of 140; 62%; Z-test p < 5.2e-04). 29 of the 140 comparisons between HK and TS genes were significant after Bonferroni correction (α = 3.6e-04). Many of these comparisons were from polyadenylation sequences (N = 5 of 7 possible), and LUN1 (N = 4 of 7 possible). The majority of the significant comparisons were derived from whole gene comparisons (10 of 29 tests).

$L1$

Figure 3.26 – L1-contained binding motifs in tissue-specific and housekeeping genes
The proportions of motifs within all sense, all antisense, sense near the 5′-UTR, antisense near the 5′-UTR, sense near the 3′-UTR, and antisense near the 3′-UTR were plotted for L1 within tissue-specific (blue) and housekeeping (red) genes.

For L1-contained motifs, all proportions of motifs were greater in TS genes than the respective placement in HK genes (56 of 56) (figure 3.26). 35 of the 56 comparisons were significant
Figure 3.27 – L1-contained binding motifs in tissue-specific and housekeeping genes
The proportion of binding motifs within tissue-specific genes were tested versus those in housekeeping genes using a fisher’s exact test. The Bonferonni cut-off alpha is represented by a horizontal line.

after Bonferonni correction ($\alpha = 8.9e-04$) (figure 3.27). The majority of the significant comparisons were within polyadenylation sequences ($N = 7$ of 7 possible), and the second variation of RUNX3 ($N_b = 6$ of 7 possible). These results are further evidence that the binding motifs contained within RTs are under different selective pressures depending on the tissue-specificity of the gene.

3.5 Conclusions

Overall the results presented in this chapter suggested that variables associated with the gene (e.g.: tissue-specificity and coding sequence density) and variables associated with the RT (e.g.: orientation with respect to the gene and binding-motifs) altered the selective pressure on the RT. Interestingly, some of the selective forces were contrary to what had originally be proposed in the literature. Most specifically was the effect of the 5’-UTR, which was found here to have
a negligible or, at most, negative effect on RT density. This was contrary to the literature which had presumed that RTs were serving an advantageous function near the 5’-UTR.

It was also identified that binding motifs within RTs were under-represented within genes when compared to intergenic regions. Furthermore, binding motif distribution was modified based on the tissue-specificity of the gene. The differential representation of binding motifs was an important finding since the effects of de novo RT insertions, other than interrupting coding sequence, have largely been over-looked. Importantly, from this research, future studies can use the measures of distribution identified here to create a model for predicting the degree of impact that a new RT insertion has on an individual.
Transcriptional regulation is a highly coordinated process necessary for the proper function of an organism. Despite conserved mechanisms regulating expression, many genes are highly variable between species (Brawand et al., 2011; Carninci et al., 2005; Levine & Tjian, 2003; Su et al., 2002), between humans (Cheung et al., 2003; Storey et al., 2007; E. T. Wang et al., 2008), and between sexes (Rinn & Snyder, 2005; Trabzuni et al., 2013). Variations in expression are often correlated with measurable phenotypic changes such as dopamine receptor expression and locomotor activity in mice (Chesler et al., 2005). Transcriptome differences associated with disease status further indicate that gene expression is a key regulator of downstream phenotypes (Su et al., 2002; Voineagu et al., 2011). It is important to determine the underlying factors involved in gene expression variability to understand the plasticity of expression in unaffected individuals, as well as the limits of regulation, that when breached lead to disease.

Large-scale genome-wide analyses such as those published by the ENCODE project have described many of the organizing components of genes underlying expression variance (Birney et al., 2007; Dong et al., 2012; J. Wang et al., 2012). DNA methylation, histone modifications, and chromatin folding are all examples of coordinating components of transcriptional regulation (Bird et al., 2002; Jenuwein & Allis, 2001). Although many genetic components such as these have been tested for correlation with transcription, much of the repetitive content of the genome remains unexamined. Specifically, retrotransposons (RT) have yet to be characterized for their association with gene expression, even though they account for approximately 40% of the mass of the human genome (Lander et al., 2001).
Despite indications that RTs can have an effect on gene function, they are largely understudied due to inherent complications in identifying locations of polymorphic insertions. Recent advances in both sequencing technology and alignment algorithms have increased the reliability of identifying specific locations of RT insertions (Baillie et al., 2011; Beck et al., 2010; Ewing & Kazazian, 2010; Iskow et al., 2010; Stewart et al., 2011; Ye et al., 2009). As of yet, there have only been a few genome-wide studies comparing the presence of RTs with associated gene expression levels. Most importantly, Arora et al., identified that genes that are differentially expressed between humans and chimps are enriched for genes that contain RT-containing indels (Arora et al., 2011).

It is vitally important for disease studies to identify the effects of RTs on gene expression, as they are a major cause of genotypic variation inserting every 1:20 or 1:270 births (Cordaux et al., 2006; Ewing & Kazazian, 2011). If RTs indeed contribute to changes in gene expression, then the variability of RTs between humans may be influencing the phenotypic variability in the population as well as in disease.

In this chapter, the association between the presence of RTs and gene expression was examined. RTs were identified as either polymorphic between humans and a subset of non-human primates, or as polymorphic between human individuals. Expression data was then analyzed in these genes to determine if there was an association of RT polymorphisms with differential gene expression.

4.1 Mobile Element Insertions between Humans and Non-human Primates

To determine if retrotransposons (RTs) effect the phenotype of gene expression, insertions were tested for association with gene expression. The first class of data in which this was tested was in genes that contained a RT in humans but that were absent in a reference genome of non-human primates. Since there are many other variables that can be associated with changes in gene expression between humans and non-human primates, other than RTs, the tests of association used genes without insertions to control for inter-species differences.
Genes containing dMEI were defined as genes that contained a RT in humans and lacked the same RT in the non-human primate (NHP). The expression of genes containing a dMEI was compared to the expression of genes lacking a dMEI. There were three situations in which a gene could be labelled as lacking a dMEI. The first was where the gene contained the same RTs in both species, the second was where the gene lacked RTs in both species, and the third was where the NHP contained a RT while the human ortholog did not contain the same RT (NHP-specific RTs). If RTs indeed had a detectable effect on gene expression, then this third group of NHP-specific RTs had the potential to alter the interpretations of the results by altering gene expression in NHPs in relation to humans. It was therefore also important to specify which genes contained a RT in the NHP but not in the human and to remove these genes from analysis.

Out of the four NHPs studied only two (Chimpanzee and Gorilla) had sufficient annotation available to determine the presence of NHP-specific RTs. Therefore the analysis for this section was performed two ways. The first analysis used all NHPs available and allowed for the inclusion of the NHP-specific RTs. The second analysis used only Chimpanzee and Gorilla and removed the genes that contained NHP-specific RTs. There was strong agreement between these two approaches.

### 4.1.1 Gene Expression Decreased in Humans with a Mobile Element Insertion

<table>
<thead>
<tr>
<th>Group</th>
<th>Gene Set</th>
<th>RheMac3</th>
<th>PonAbe2</th>
<th>GorGor3</th>
<th>PanTro4</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>Expression Data</td>
<td>5963</td>
<td>6333</td>
<td>5982</td>
<td>6652</td>
</tr>
<tr>
<td>b</td>
<td>Filtered MEI-Alu</td>
<td>4721</td>
<td>4663</td>
<td>3889</td>
<td>2392</td>
</tr>
<tr>
<td>c</td>
<td>Filtered MEI-L1</td>
<td>226</td>
<td>181</td>
<td>152</td>
<td>69</td>
</tr>
<tr>
<td>d</td>
<td>Filtered MEI-SVA</td>
<td>271</td>
<td>564</td>
<td>420</td>
<td>331</td>
</tr>
<tr>
<td>e</td>
<td>a ∩ b</td>
<td>2645</td>
<td>2903</td>
<td>2471</td>
<td>1545</td>
</tr>
<tr>
<td>f</td>
<td>a ∩ c</td>
<td>144</td>
<td>124</td>
<td>105</td>
<td>46</td>
</tr>
<tr>
<td>g</td>
<td>a ∩ d</td>
<td>166</td>
<td>342</td>
<td>252</td>
<td>199</td>
</tr>
</tbody>
</table>

Table 4.1 – Gene counts per MEI-type and non-human primate (NHP)
Presented here are the number of genes analyzed between humans and the respective non-human primate. RheMac3: Rhesus Macaque; PonAbe2: Orangutan; GorGor3: Gorilla; PanTro4: Chimpanzee. Expression data: The genes with orthologs between humans and the non-human primate for which there was valid expression data. Filtered MEI: Genes containing mobile element insertions in humans but not the respective NHP, after filtering based on phylogenetic order.
RefSeq genes were identified in the human reference (hg19) that contained mobile elements of L1, Alu, or SVA. The presence of each human mobile element was then examined within each non-human primate (NHP) reference genome. If the mobile element was absent in the NHP, and present in the human then it was labelled a mobile element insertion (MEI). Genes were classified as an MEI-containing gene if the MEI was within the gene transcription start to transcription stop sites, including introns. The sample sizes of MEI-containing genes are presented in table 4.1.

These genes were then filtered to retain only genes with orthologs between hg19 and chimp (panTro4), gorilla (GorGor3), orangutan (ponAbe2), and rhesus macaque (rheMac3). Furthermore, in order to compare between the sexes, only genes on the autosomes were considered in this analysis. Using data from Brawand et al., expression values were obtained for all orthologous genes and for each species. This dataset contained expression for multiple tissues, including brain, heart, kidney, liver, and testis. The final counts of genes that had expression values in Brawand et al. are presented in table 4.1.

The differences in expression across all species were controlled for by inter-sample normalization procedures (Brawand et al., 2011). Comparing between genes with MEIs versus genes without MEIs further controlled for differences in expression present across all genes. To determine if the presence of an MEI was associated with a general shift in gene expression, odds ratios (OR) were calculated. The calculations represented the odds of having more genes with increased expression in humans that also contained an MEI (table 4.2). These ORs were calculated between human and each NHP for each tissue, MEI sub-type, and males and females separately (table 4.3).

<table>
<thead>
<tr>
<th></th>
<th>MEI</th>
<th>no-MEI</th>
<th>Total by MEI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher Expression</td>
<td>a</td>
<td>c</td>
<td>a+c</td>
</tr>
<tr>
<td>Human</td>
<td>b</td>
<td>d</td>
<td>b+d</td>
</tr>
<tr>
<td>NHP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total by Expression</td>
<td>a+b</td>
<td>c+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

Table 4.2 – Contingency table for expression differences between human and non-human primate

RefSeq genes were identified in the human reference (hg19) that contained mobile elements of L1, Alu, or SVA. The presence of each human mobile element was then examined within each non-human primate (NHP) reference genome. If the mobile element was absent in the NHP, and present in the human then it was labelled a mobile element insertion (MEI). Genes were classified as an MEI-containing gene if the MEI was within the gene transcription start to transcription stop sites, including introns. The sample sizes of MEI-containing genes are presented in table 4.1.

These genes were then filtered to retain only genes with orthologs between hg19 and chimp (panTro4), gorilla (GorGor3), orangutan (ponAbe2), and rhesus macaque (rheMac3). Furthermore, in order to compare between the sexes, only genes on the autosomes were considered in this analysis. Using data from Brawand et al., expression values were obtained for all orthologous genes and for each species. This dataset contained expression for multiple tissues, including brain, heart, kidney, liver, and testis. The final counts of genes that had expression values in Brawand et al. are presented in table 4.1.

The differences in expression across all species were controlled for by inter-sample normalization procedures (Brawand et al., 2011). Comparing between genes with MEIs versus genes without MEIs further controlled for differences in expression present across all genes. To determine if the presence of an MEI was associated with a general shift in gene expression, odds ratios (OR) were calculated. The calculations represented the odds of having more genes with increased expression in humans that also contained an MEI (table 4.2). These ORs were calculated between human and each NHP for each tissue, MEI sub-type, and males and females separately (table 4.3).
Female humans were missing expression values for testis and liver; these tissues were therefore excluded from analysis in females. To minimize type I error, p-values were calculated directly by permuting MEI-status and a strict Bonferroni cutoff was calculated based on the total number of tests across all MEI types (Number of OR tests = 96; Bonferroni cutoff $\alpha = 5.21\times 10^{-4}$) (table 4.3). Many ORs associated with Alu remained significant after Bonferroni correction (Alu: 16 of 32 ORs) (figure 4.1), indicating a tendency for gene expression to shift between species associated with MEI status. L1 and SVA ORs also showed significant results (L1: 1 of 32 ORs; SVA: 10 of 32 ORs).

![Figure 4.1](image)

**Figure 4.1 – The effect of MEI status on gene expression between primates**

Odds ratios for each tissue (brain, heart, liver, kidney, testis) and each primate (human, chimp, gorilla, orangutan, rhesus macaque) were plotted by MEI-type (Alu, L1, or SVA). Dots in black represented tests that did not pass Bonferonni significance ($\alpha = 1.25\times 10^{-3}$). Dots in red represent tests that did pass Bonferonni significance.

Interestingly, 76% of the significant ORs were < 1 (Alu: 11 of 16 sig. tests (69%); L1: 1 of 1 sig. tests (100%); SVA: 8 of 11 sig. tests (72%)), indicating an association of decreased gene expression in humans with the presence of an MEI.
<table>
<thead>
<tr>
<th>Test</th>
<th>Primate</th>
<th>Sex</th>
<th>Tissue</th>
<th>Element</th>
<th>OddsRatio</th>
<th>conf1</th>
<th>conf2</th>
<th>Raw_pvalue</th>
<th>MEI- Hs</th>
<th>MEI- NHP</th>
<th>MEI+ Hs</th>
<th>MEI+ NHP</th>
<th>pvalue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RheMac</td>
<td>male</td>
<td>brain</td>
<td>Alu</td>
<td>0.87</td>
<td>0.79</td>
<td>0.96</td>
<td>0.01</td>
<td>2708.00</td>
<td>7008.00</td>
<td>652.00</td>
<td>1940.00</td>
<td>4.00e-04</td>
</tr>
<tr>
<td>2</td>
<td>RheMac</td>
<td>male</td>
<td>kidney</td>
<td>Alu</td>
<td>0.83</td>
<td>0.75</td>
<td>0.91</td>
<td>0.00</td>
<td>6702.00</td>
<td>2896.00</td>
<td>1688.00</td>
<td>882.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>5</td>
<td>RheMac</td>
<td>male</td>
<td>testis</td>
<td>Alu</td>
<td>0.83</td>
<td>0.76</td>
<td>0.91</td>
<td>0.00</td>
<td>4411.00</td>
<td>5658.00</td>
<td>1039.00</td>
<td>1603.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>11</td>
<td>Orangutan</td>
<td>male</td>
<td>brain</td>
<td>Alu</td>
<td>0.77</td>
<td>0.69</td>
<td>0.86</td>
<td>0.00</td>
<td>2063.00</td>
<td>7375.00</td>
<td>509.00</td>
<td>2358.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>12</td>
<td>Orangutan</td>
<td>male</td>
<td>kidney</td>
<td>Alu</td>
<td>1.13</td>
<td>1.04</td>
<td>1.24</td>
<td>0.01</td>
<td>5827.00</td>
<td>3473.00</td>
<td>1863.00</td>
<td>981.00</td>
<td>8.00e-04</td>
</tr>
<tr>
<td>16</td>
<td>Orangutan</td>
<td>female</td>
<td>brain</td>
<td>Alu</td>
<td>1.66</td>
<td>1.48</td>
<td>1.86</td>
<td>0.00</td>
<td>7000.00</td>
<td>2142.00</td>
<td>2391.00</td>
<td>441.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>22</td>
<td>GorGor</td>
<td>male</td>
<td>kidney</td>
<td>Alu</td>
<td>0.73</td>
<td>0.66</td>
<td>0.81</td>
<td>0.00</td>
<td>3321.00</td>
<td>6416.00</td>
<td>669.00</td>
<td>1762.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>23</td>
<td>GorGor</td>
<td>male</td>
<td>liver</td>
<td>Alu</td>
<td>0.85</td>
<td>0.77</td>
<td>0.94</td>
<td>0.00</td>
<td>2824.00</td>
<td>6859.00</td>
<td>630.00</td>
<td>1794.00</td>
<td>2.00e-04</td>
</tr>
<tr>
<td>24</td>
<td>GorGor</td>
<td>male</td>
<td>heart</td>
<td>Alu</td>
<td>1.33</td>
<td>1.16</td>
<td>1.52</td>
<td>0.00</td>
<td>8053.00</td>
<td>1519.00</td>
<td>2093.00</td>
<td>297.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>25</td>
<td>GorGor</td>
<td>male</td>
<td>testis</td>
<td>Alu</td>
<td>0.69</td>
<td>0.63</td>
<td>0.76</td>
<td>0.00</td>
<td>4824.00</td>
<td>5405.00</td>
<td>941.00</td>
<td>1526.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>29</td>
<td>GorGor</td>
<td>female</td>
<td>heart</td>
<td>Alu</td>
<td>1.19</td>
<td>1.08</td>
<td>1.32</td>
<td>0.00</td>
<td>6278.00</td>
<td>2883.00</td>
<td>1695.00</td>
<td>654.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>32</td>
<td>Chimp</td>
<td>male</td>
<td>kidney</td>
<td>Alu</td>
<td>0.65</td>
<td>0.55</td>
<td>0.77</td>
<td>0.00</td>
<td>1792.00</td>
<td>8910.00</td>
<td>176.00</td>
<td>1345.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>34</td>
<td>Chimp</td>
<td>male</td>
<td>heart</td>
<td>Alu</td>
<td>0.70</td>
<td>0.63</td>
<td>0.78</td>
<td>0.00</td>
<td>5442.00</td>
<td>4966.00</td>
<td>651.00</td>
<td>847.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>35</td>
<td>Chimp</td>
<td>male</td>
<td>testis</td>
<td>Alu</td>
<td>0.77</td>
<td>0.69</td>
<td>0.85</td>
<td>0.00</td>
<td>5652.00</td>
<td>5507.00</td>
<td>678.00</td>
<td>864.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>36</td>
<td>Chimp</td>
<td>female</td>
<td>brain</td>
<td>Alu</td>
<td>1.41</td>
<td>1.23</td>
<td>1.62</td>
<td>0.00</td>
<td>7943.00</td>
<td>2620.00</td>
<td>1233.00</td>
<td>288.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>39</td>
<td>Chimp</td>
<td>female</td>
<td>heart</td>
<td>Alu</td>
<td>0.74</td>
<td>0.66</td>
<td>0.84</td>
<td>0.00</td>
<td>3922.00</td>
<td>6341.00</td>
<td>472.00</td>
<td>1024.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>42</td>
<td>RheMac</td>
<td>male</td>
<td>kidney</td>
<td>L1</td>
<td>0.49</td>
<td>0.35</td>
<td>0.71</td>
<td>0.00</td>
<td>8319.00</td>
<td>3714.00</td>
<td>71.00</td>
<td>64.00</td>
<td>6.00e-04</td>
</tr>
<tr>
<td>81</td>
<td>RheMac</td>
<td>male</td>
<td>brain</td>
<td>SVA</td>
<td>0.54</td>
<td>0.34</td>
<td>0.81</td>
<td>0.00</td>
<td>3332.00</td>
<td>8810.00</td>
<td>28.00</td>
<td>138.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>85</td>
<td>RheMac</td>
<td>male</td>
<td>testis</td>
<td>SVA</td>
<td>0.43</td>
<td>0.30</td>
<td>0.62</td>
<td>0.00</td>
<td>5409.00</td>
<td>7136.00</td>
<td>41.00</td>
<td>125.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>87</td>
<td>RheMac</td>
<td>female</td>
<td>kidney</td>
<td>SVA</td>
<td>1.93</td>
<td>1.22</td>
<td>3.17</td>
<td>0.00</td>
<td>9106.00</td>
<td>2697.00</td>
<td>143.00</td>
<td>22.00</td>
<td>8.00e-04</td>
</tr>
<tr>
<td>96</td>
<td>Orangutan</td>
<td>female</td>
<td>brain</td>
<td>SVA</td>
<td>2.21</td>
<td>1.57</td>
<td>3.19</td>
<td>0.00</td>
<td>9091.00</td>
<td>2545.00</td>
<td>300.00</td>
<td>38.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>102</td>
<td>GorGor</td>
<td>male</td>
<td>kidney</td>
<td>SVA</td>
<td>0.45</td>
<td>0.32</td>
<td>0.63</td>
<td>0.00</td>
<td>3945.00</td>
<td>7976.00</td>
<td>45.00</td>
<td>202.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>105</td>
<td>GorGor</td>
<td>male</td>
<td>testis</td>
<td>SVA</td>
<td>0.49</td>
<td>0.37</td>
<td>0.65</td>
<td>0.00</td>
<td>5691.00</td>
<td>6753.00</td>
<td>74.00</td>
<td>178.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>112</td>
<td>Chimp</td>
<td>male</td>
<td>kidney</td>
<td>SVA</td>
<td>0.46</td>
<td>0.26</td>
<td>0.77</td>
<td>0.00</td>
<td>1952.00</td>
<td>10076.00</td>
<td>16.00</td>
<td>179.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>114</td>
<td>Chimp</td>
<td>male</td>
<td>heart</td>
<td>SVA</td>
<td>0.47</td>
<td>0.34</td>
<td>0.64</td>
<td>0.00</td>
<td>6029.00</td>
<td>5684.00</td>
<td>64.00</td>
<td>129.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>115</td>
<td>Chimp</td>
<td>male</td>
<td>testis</td>
<td>SVA</td>
<td>0.49</td>
<td>0.36</td>
<td>0.67</td>
<td>0.00</td>
<td>6264.00</td>
<td>6238.00</td>
<td>66.00</td>
<td>133.00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>119</td>
<td>Chimp</td>
<td>female</td>
<td>heart</td>
<td>SVA</td>
<td>0.47</td>
<td>0.33</td>
<td>0.67</td>
<td>0.00</td>
<td>4351.00</td>
<td>7215.00</td>
<td>43.00</td>
<td>150.00</td>
<td>0.00e+00</td>
</tr>
</tbody>
</table>

Table 4.3 – Contingency table values and results of the two-tailed permutation tests for MEI-expression tests

Conf1: lower confidence limit, Conf2: higher confidence limit, MEI- : Genes with no identifiable MEI within the gene boundaries; MEI+ : Genes with MEIs within the gene boundaries; Hs: Genes expressed higher in humans; NHP: Genes expressed higher in the respective non-human primate. pvalue: proportion of permutations that were more extreme than the actual odds ratio out of 5000 permutations.
4.1.2 Male Dependent Association between MEI Status and Expression

There were four variables that were uncontrolled for in the calculation of odds ratio: sex of the individual, non-human primate species, tissue of origin, and MEI-type. It was therefore important to determine if any of these variables had a significant effect on the outcome of odds ratio. To compare the effect of each variable type, ORs were regressed on sex, NHP, tissue, and MEI-type (equation 4.1). OR for the tissues testis and liver were removed due to the lack of data in female samples.

\[
\text{OddsRatio} = \beta_o + \beta_{\text{sex}} + \beta_{\text{NHP}} + \beta_{\text{tissue}} + \beta_{\text{MEI}} + \epsilon
\]  

These models were compared using a four-way ANOVA. Using ANOVA, all variables were examined to determine which had the largest effect on OR (sex, primate comparison, tissue, or MEI-type) (table 4.4). Sex of the individual (p < 5.2e-04) showed the most significant effect on OR.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Sex</td>
<td>1</td>
<td>13.02</td>
<td>5.2e-04</td>
</tr>
<tr>
<td>2 Tissue</td>
<td>4</td>
<td>2.98</td>
<td>0.02</td>
</tr>
<tr>
<td>3 Primate</td>
<td>3</td>
<td>2.81</td>
<td>0.04</td>
</tr>
<tr>
<td>4 MEI-type</td>
<td>2</td>
<td>1.59</td>
<td>0.21</td>
</tr>
</tbody>
</table>

Table 4.4 – Results of ANOVA comparing effects of sex, tissue, NHP, and MEI

A single regression equation was calculated for Odds ratios regressed on sex, tissue, primate, and mei-type. The model was then analyzed by ANOVA to determine the respective contributions of each variable. Df = degrees of freedom.

Indeed when grouping significant ORs based on sex, a majority of the significant ORs were derived from male samples (20/27 = 74%) as opposed to female samples (7/27 = 26%). The direction of effect (i.e.: increased versus decreased expression in humans versus NHP) was also associated with the sex of the individual in genes containing MEIs. ORs were more often < 1 (decreased expression) in male samples (18/20 = 90%) compared to female samples (2/7 = 28%). Overall the odds ratios between males were significantly smaller than those between females (figure 4.2).
Figure 4.2 – Odds ratios based on sex of origin
Odds ratios were calculated for each primate (human, chimp, gorilla, orangutan, and rhesus macaque) and each tissue (brain, heart, liver, kidney, testis). Odds ratios were separated by sex of origin. Black dots indicate odds ratios that were not significant after Bonferroni correction of the two-tailed permuted p-value. Red dots indicate odds ratios that were significant ($\alpha = 1.25\text{e}-03$).

4.1.3 Analysis with Removal of NHP-specific Retrotransposons

As mentioned earlier, the second analysis was restricted to Chimpanzee and Gorilla. All of the tests from Chimpanzee that were significant in the first analysis were similarly significant after removing genes containing NHP-specific RTs. All except for two tests from Gorilla that were initially significant were also significant in this second analysis. The two tests that were no longer significant were the tests of expression in the male and female heart with the presence of an Alu.

Similar to the initial analysis, there was a trend for odds ratios to be less than 1. Indeed, 13 of the 14 significant tests had an odds ratio of less than 1 ($z$-test $p < 0.001$) in this analysis. ANOVA was used to discriminate the effects of element type, sex, primate type, and tissue on the significant odds ratios. All variables except for the primate type had a significant effect.
(table 4.5). The sex-specific effects remained one of the most significant effects, after MEI-type.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Df</th>
<th>F-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Sex</td>
<td>1</td>
<td>1678</td>
<td>3.20e-05</td>
</tr>
<tr>
<td>2 Tissue</td>
<td>2</td>
<td>1909</td>
<td>2.20e-05</td>
</tr>
<tr>
<td>3 Primate</td>
<td>1</td>
<td>2.80</td>
<td>1.90e-01</td>
</tr>
<tr>
<td>4 MEI-type</td>
<td>1</td>
<td>1018</td>
<td>6.78e-05</td>
</tr>
</tbody>
</table>

Table 4.5 – ANOVA: Chimp and Gorilla Only

A single regression equation was calculated for significant odds ratios regressed on sex, tissue, primate, and mei-type. The model was then analyzed by ANOVA to determine the respective contributions of each variable. Df = degrees of freedom.

The agreement between these two approaches strengthens the finding that genes containing dMEI were more commonly genes that were decreased in expression in humans.

4.2 Mobile Element Insertions that are Polymorphic Between Humans

The bias in OR due to the sex of the individual indicated that there was a bias in gene expression between males and females that was restricted to genes containing MEIs. It was next determined if the expression differences between humans and NHPs were also present as differences in expression between male and female humans.

4.2.1 Genes containing MEI-Alu are Differentially Expressed Between Male and Female Humans

To test the differences between humans based on sex, expression was calculated within genes containing MEIs between chimps and humans. Expression was compared between females males using the contingency table presented in table 4.6.

To reduce the number of comparisons only one group of MEI was chosen to compare expression between the human sexes. This was MEI derived from Alu (MEI-Alu) that were present in human but absent in all other NHPs (equivalent to MEI-Alu-Chimp). To further restrict the number of tests expression was only examined in one tissue, the brain. Indeed, gene expression was more likely to be decreased in human males compared to human females when
Table 4.6 – Contingency table for expression differences between males and females of the same species

<table>
<thead>
<tr>
<th>MEI-Status</th>
<th>Males</th>
<th>Females</th>
<th>Total by Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher Expression</td>
<td>MEI</td>
<td>no-MEI</td>
<td>Total by MEI</td>
</tr>
<tr>
<td>Males</td>
<td>a</td>
<td>c</td>
<td>a+c</td>
</tr>
<tr>
<td>Females</td>
<td>b</td>
<td>d</td>
<td>b+d</td>
</tr>
<tr>
<td>Total by Expression</td>
<td>a+b</td>
<td>c+d</td>
<td>a+b+c+d</td>
</tr>
</tbody>
</table>

a,b,c, and represent counts of genes. Odds ratio = ad / bc

an MEI-Alu-Chimp was present within the gene (OR = 0.42; CI 95% = 0.33 – 0.52; Fisher’s p < 4.25 e-18).

4.3 Replication of Sex-specific Results in Humans

Although these results indicated that there was a sex-dependent difference between humans in genes containing an MEI; it was possible that the analysis detected an unrelated bias in the dataset due to the small sample size (N male = 3, N female = 2). Therefore, a larger replication set was examined. This larger dataset was used to test the claim that the MEI gene set was significantly associated with genes that were differentially expressed between males and females.

Trabzuni et al. characterized gene expression from human male (N=101) and female (N=36) brain and spinal cord tissues (Trabzuni et al., 2013). The top differentially expressed and alternatively spliced genes between males and females that were reported in this dataset were used for this analysis. To facilitate the comparison, MEI-Alu-Chimp were converted from Ensembl annotation to HUGO gene annotation (N MEI-Alu-Chimp converted = 2200). Of the 442 genes differentially expressed or spliced between the sexes in any brain region, 117 contained an MEI-Alu-Chimp. This was significantly more genes containing MEI-Alu-Chimp than expected by chance (OR= 2.8; CI 95% = 2.29 – 3.57; Fisher’s exact p < 2.2e-16). Total gene count was set at 19116 to reflect the number of genes uniquely annotated in HUGO (Flicek et al., 2013; Kent et al., 2002).
4.3.1 MEIs Segregating in Humans Associated with Decreased Expression

Alu, L1, and SVA are retrotranspositionally competent in humans resulting in mobile element variation within the human population (Ewing & Kazazian, 2010; Stewart et al., 2011). Stewart et al. annotated a subset of MEIs that are polymorphic between humans. It was next important to check if MEI-Alu locations that were segregating (MEI-seg) within the human population had a similar association with a change in gene expression as the MEIs previously identified between humans and primates.

MEI-segs were identified from the CEU individuals typed in the 1000 genomes project. Expression data for the same CEU individuals was available from the Montgomery et al. dataset. Gene expression from Montgomery et al. was filtered to include only genes that contained an MEI-seg between the transcription start and transcription stop sites. 547 genes matched this criterion. Gene expression was compared between males and females containing MEI-seg (MEI-seg(+)). 338 genes (62% of total, Z-test p < 3.8e-08) had a decrease in mean expression in males versus females (figure 4.3). Conversely when the same genes were compared between males and females without an MEI-seg (MEI-seg(-)), only 266 genes were decreased in males versus females (49% of total, Z-test p < 0.64) (figure 4.4). A Student’s t-test showed that this decrease in expression in MEI-seg(+) individuals was significantly larger than in MEI-seg(-) individuals (mean difexp log MEI-seg(+) = 0.009; mean difexp log MEI-seg(-) = -0.002; Student’s t-test p < 1.87 e-25).

4.3.2 MEI-seg Associated Decrease in Expression is Specific to Males

Expression values were compared within each sex group to determine if the effect was biased towards a decrease in males, or an increase of expression in females. The odds of decreased expression in the presence of an MEI-seg were significantly increased when the individual was male compared to female (OR = 2.36; CI 95% = 1.84 – 3.05; Fisher’s exact p < 5.90e-12). Furthermore, the differential expression between MEI-seg(+) and MEI-seg (-) males was significantly greater than that between females (Student’s t-test p < 7.93e-26) (figure 4.5). These
Figure 4.3 – Differential expression between sexes in individuals containing the MEI
Gene expression of females containing an MEI were compared to the same genes in males who also contained the same MEI. The t-test p-values of these comparisons were plotted versus the difference between the means of the two groups.

Figure 4.4 – Differential expression between sexes in individuals lacking the MEI
Gene expression of females lacking an MEI were compared to the same genes in males who also lacked the MEI. The t-test p-values of these comparisons were plotted versus the difference between the means of the two groups.
results indicated that the bias of decreased expression was largely driven by MEI-associated expression in males.

![Figure 4.5](image)

**Figure 4.5 – Comparison within groups of the same sex, between MEI+ and MEI- individuals**

Differences in expression were calculated between females containing the MEI and lacking the MEI (Fem) or between males containing the MEI and lacking the MEI (Male). A red line is drawn at zero, where no change would be expected.

### 4.3.3 Genes Containing MEI are Functionally Enriched for Autism Susceptibility Genes

The genes contained an MEI-Alu (Chimp) were tested for functional pathway enrichment. Splice variants were significantly enriched in this group \( (N = 1199, \text{Benjamini corrected } p < 4.1 \times 10^{-71}) \), as well as neuron projection development \( (N = 54, \text{Benjamini corrected } p < 7.7 \times 10^{-4}) \), brain expression \( (N = 1137, \text{Benjamini corrected } p < 9.2 \times 10^{-36}) \), and phosphatidylinositol signaling \( (N = 19, \text{Benjamini corrected } p < 9.4 \times 10^{-2}) \). The complete list is presented in supplementary table 2.

Baillie et al. 2011 previously showed that there is somatic retrotransposition in the human brain (Baillie et al., 2011). Within the list of genes that were deemed somatic inserts in Baillie et al. 2011, there was a significant overlap with genes associated with Autism Spectrum Dis-
order (ASD) (Hussman et al., 2011) (OR = 7.29; CI 95% = 5.01 – 10.73 ; Fisher’s exact p < 6.59e-27) (Supplementary table 2). Therefore an additional functional enrichment test was performed to determine if genes containing MEI were enriched for genes related to ASD. MEI-Alu (Chimp) were enriched for ASD genes. ASD genes were annotated using the Autism Database (AUTdb) (Basu et al., 2009). ASD genes were significantly enriched for genes containing MEI-Alu (Chimp) (OR = 1.83; CI 95% = 1.12 – 2.90; Fisher’s exact p < 0.01). More strikingly, ASD genes were significantly enriched within genes containing the human polymorphic MEI-segs (OR = 5.10; CI 95% = 3.48 – 7.41; Fisher’s exact p < 6.09e-16).

Similar to all MEI-seg genes, the ASD genes containing an MEI-seg showed sex-specific differences in expression. Differences in expression between MEI-seg (+) and MEI-seg (-) males were significantly greater than between MEI-seg (+) versus MEI-seg (-) females (mean dif logExp male = -0.008; mean dif logExp female = -0.002 ; Student’s p < 0.04) (figure 5b).

4.4 Conclusions

These results suggest that the presence of a RT within a gene is indeed associated with decreased expression of the same gene. More importantly, this effect of RTs on gene expression was sex-specific, and largely driven by effects in males. Lastly, the genes found to contain variations in RTs are enriched for Autism susceptibility genes, which are also enriched for differences in expression between the sexes. These results provide evidence for RTs in mediating sex-specific effects of human gene expression.
Chapter 5 METHODS

5.1 Chapter 2 methods

5.1.1 Data Sources

Insertion and deletion locations of retrotransposons (RT) that were polymorphic between populations, as well as those that were shared between populations, were downloaded from the 1000 genomes supplemental information (Stewart et al., 2011). This data was filtered for elements that were classified as insertions with respect to the Hg18 reference genome. Deletions in comparison to the reference genome were not used in this analysis. All coordinates were converted to Hg19 coordinates using the Genome LiftOver tool provided from UCSC genome browser (Kent et al., 2002). Somatic insertions of Alu in the brain were downloaded from Baillie et al. (Baillie et al., 2011).

5.1.2 Determining the Antisense Bias of Alu Retrotransposons

Allele frequencies (AF) identified in Stewart et al. were used for this analysis. Population specific subgroups were defined as Alu with an AF greater than 0 in either Yoruban (YRI), European (CEU), or the Chinese/Japanese (CHBJPT) populations, and an AF of 0 in the other two populations. Shared Alu were defined as those where the AF was greater than 0 in all three populations. The number of antisense and sense Alu were counted in each population-specific group and combined with the antisense and sense Alu from somatic insertion datasets (Bundo et al., 2013; Baillie et al., 2011). Antisense and sense counts were tested for differences in proportion using the R binomial sign test function; binom.test(). Comparisons in proportions
of antisense Alu between two groups were calculated using the R function prop.test() which is a test of equal proportions.

5.1.3 Determining Ranks for Retrotransposon Subfamilies

L1 were ranked first by subfamily group by the average divergence from their respective consensus sequences, as tabulated in the repeat masker track of the UCSC genome browser, in the column millidiv. The applied rankings very closely followed the subfamily naming convention. This approach follows conventional methods for dating retrotransposon families in many respects. It does, however, have its limitations, particularly over shorter time spans (i.e., levels of nucleotide divergence). We consider this approach more conservative as the timescale grows larger because we make no assumptions about mutation rates over time or across element sequences. The only assumption made is that mutations are gradually accumulated over time, and this should hold true in the majority of instances, barring gene conversion and rare substitutions that revert nucleotides back to the corresponding consensus base. As the purpose of our ranking is to show no systematic increase or decrease in AS/S ratio over longer evolutionary periods, this relative ranking method should be suitable for our purposes. Over shorter timescales, however, it fails to capture the variance in element age estimates, as well as the fact that subfamilies can propagate over extended periods of time and can often do so in parallel with other subfamilies. In this sense, considering rank positions as distinct time points is a gross oversimplification. As our purpose in this particular analysis is to examine large scale trends in the magnitude of antisense bias, however, we do not expect these simplifications to impact our results or interpretation.

5.2 Chapter 3 methods

5.2.1 Data Sources

Retrotransposon (RepeatMasker) coordinates were downloaded from the UCSC genome browser (Human, hg19). Genes were defined using the RefSeq database unless otherwise stated. Tran-
scription factor binding motifs were mined from sequence data from the Hg19 reference sequence.

### 5.2.2 Binning Reference Locations

Gene (RefSeq flat file track), and Retrotransposon (RepeatMasker) coordinates were downloaded from the UCSC genome browser (Human, hg19) (Kent et al., 2002). Each gene was subdivided into ten bins of equal size where decile 1 began at the transcription start and decile 10 ended at the transcription stop. This method of binning resulted in a concentration of exons in the first and last bins. Further calculations took this bias into account. Strandedness was taken into account so that the transcription start site was always in decile 1, irrespective of the direction of transcription.

### 5.2.3 Identifying Genes with Secondary UTRs

5’-UTR coordinates were downloaded from the UCSC genome browser RefSeq Gene track (Kent et al., 2002). RefSeq genes were divided into deciles as described above. The 5’-UTR coordinates were intersected with each gene deciles to determine which decile they were present in. The genes containing these internal 5’- UTRs were thus identified and grouped based on the decile within which the 5’-UTR was contained.

### 5.2.4 Distribution Analysis

Counts of elements within each decile were tabulated using an in-house script that intersected gene locations with RepeatMasker annotations to tabulate strand orientation counts per decile. Frequencies of elements within deciles were defined as the count of elements in a decile (for all genes) per total elements (across all genes). For S-only or AS-only plots, subfamily counts were normalized by the intergenic frequencies of the respective subfamilies in order to account for variable amplification rates. Trend lines were obtained via linear regression. For the analysis of L1 and Alu maintenance in genes as a function of rank by milliDiv; milliDiv was calculated as the average base mismatches in parts per thousand, as tabulated in Repeat-
Masker (UCSC Genome Browser, Hg19). A minimum sample size for comparison between subfamilies was calculated using G*Power3, using the Chi-Square test: Goodness-of-fit tests: Contingency tables test with an effect size of 0.1. This resulted in a minimum sample size of 1545, which we used as a minimum element count threshold for inclusion in subsequent linear regression analysis. To calculate the slopes of all Alu, L1, and exon counts across gene deciles, linear regression was performed on the internal deciles (i.e.: 2-9). Leading and trailing deciles were excluded to reduce the introduction of bias attributed to an Òexon anchoring effect.Ó Briefly, by using the transcription start and stop positions as gene boundaries, the first and last deciles will always contain a minimum of one exon each due to standard gene biology. Therefore, counts in these deciles are systematically skewed and less representative of the general pattern across the gene.

5.2.5 Tissue Specific Gene Sets

Gene categories were grouped based on the calculations in (Chang et al., 2011). Power calculations resulted in the restriction of analysis to tissues with > 100 genes. Decile and frequencies were performed as above, but with the restricted gene sets.

5.2.6 Motif Analysis

To reduce multiple testing we restricted motif analysis to sequences that have been previously identified in Alu or L1. Exact matches, unless otherwise stated, were searched in the necessary region (e.g.: in genes, intergenic) using an in-house program. Counts of each motif were tabulated in each region and, where necessary, were normalized by the relative element density. Truncated L1 were identified by comparing every known L1 (RepeatMasker, Hg19) versus known L1 consensus sequences using a local BLAST alignment. The start position of each L1 was denoted as the start position within its respective consensus sequence.
5.3 Chapter 4 methods

5.3.1 Identifying Differential Mobile Element Insertions Between Humans and NHPs

A full list of retrotransposons present in humans and absent in non-human primate is presented in supplementary table x. Genomic coordinates for each retrotransposon in each species were annotated using the RepeatMasker track in the UCSC Genome Browser (Kent et al., 2002; Smit, Afa, Hubley.R and Green, n.d.). Reference genomes used were Hg19 (human), PanTro4 (Chimpanzee), GorGor3 (Gorilla), PonAbe3 (Orangutan), and RheMac3 (Rhesus Macaque). Genomic coordinates of MEIs present in humans were examined for their presence in the respective NHP using a local version of LiftOver (Kent et al., 2002). Coordinates that were partially deleted were removed from analysis, resulting in a list of coordinates that were present in humans and fully deleted in the respective non-human primate. Importantly, in the comparison between humans and NHPs, a region may contain more than one differential MEI location. To increase the likelihood of valid insertions we filtered the differential insertions further based on known phylogenetic order. MEIs identified as present in humans but absent in chimps were required to be absent in all other NHPs. Those present in humans but absent in gorilla, were required to be absent in orangutan and rhesus macaque. Those present in humans but absent in orangutan were required to be absent in rhesus macaque, and those present in humans but absent in rhesus macaque were required to be present in gorilla. The genes overlapping these retrotransposon locations were identified using bedtools and the RefSeq annotation from transcription start to transcription stop site. Each gene was only analyzed once per species, even if more than one differential retrotransposon was identified within the gene boundaries.

5.3.2 Expression in NHPs Versus Humans

Gene expression values in this study were taken from Brawand et al. supplementary data 2, which were expression values after normalization for between species differences. We subsequently averaged gene expression between groups, where groups were defined by the species of origin, sex, and tissue-type. A gene was denoted as having higher expression in humans if
expression in humans was greater than expression in the non-human primate. No fold-change
cutoff was applied. Due to a potential for interdependency between observations within sam-
ples p-values for each odds ratio test were calculated by permuting MEI-status. This was done
by first identifying the total number of genes that were expressed higher in humans (a) and
higher in the non-human primate (b). A simulated dataset was created with the same number
of genes expressed higher in humans (a) and higher in non-human primates (b). A sample size
equivalent to that which was tested in the original dataset was then randomly sampled from
this simulated dataset to place labels of "contained MEI" and "did not contain MEI" at random.
The odds ratio were then calculated as previously noted in the text and compared versus the
actual odds ratio. The simulated odds ratio was counted if it was greater than the actual odds
ratio or less than the reciprocal of the actual odds ratio. 5000 re-samplings were performed
for each analysis. The total number of counted simulated odds ratios was divided by 5000 to
obtain the two-tailed p-value.

5.3.3 Expression in Humans versus Humans

Expression values for CEU individuals within the 1000 genomes project were determined from
(Montgomery et al., 2010). Female and male status was identified using annotations from
Stewart et al.. Gene expression was calculated separately for individuals that were annotated
as having the alternate allele, those lacking the alternate allele, and subdivided by males and
females. Alternate alleles were identified in Stewart et. al 2011 through two analyses, one
using deletions and one using insertions. The alternate allele in both cases is the presence of
the retrotransposon. Student’s t-tests were produced in R. Genes were removed from analysis
if they contained less than 3 individuals in a given group. Due to this exclusion criteria there
are different genes present in different analyses, though the same gene set was used at the start
of the analysis.
6.1 Motivation

Genomic studies have been successful in characterizing a portion of the heritable component of complex human diseases. For example, through genomics, the human leukocyte antigen has been associated with the etiology of Multiple Sclerosis, and Complement Factor H with Age-Related Macular Degeneration. Despite these and other successes there remains a large portion of the heritability of complex diseases which is yet to be explained. This is a common problem that has arisen in the study of complex human diseases, and is often termed "missing heritability". There are many reasons why the heritable portion of complex disease etiology may remain elusive such as the limited power to detect small effects, gene-gene interactions, and epigenetic effects. Perhaps the most pertinent to this study are the technical limitations that result in the exclusion of retrotransposons (RTs) from analysis. It was the intent of this study to analyze the potential of RTs to account for phenotypic variation between individuals, and therefore the potential to be mediators of disease heritability.

RTs are common sources of heritable variation, and are capable of substantially disrupting gene function. For these reasons they are prime candidates for genetic variations that contribute to human disease. Current studies, however, focus on other forms of genetic variation such as single nucleotide polymorphisms, insertions/deletions, and large copy number variations. RTs therefore remain an under-examined source of phenotypic variation. The lack of research into the phenotypic associations with RTs is largely due to two major factors. These are 1.)
limitations in the methodology used to identify RT variants and 2.) difficulty in interpreting the effects of RT variants on the genome.

The lack of appropriate methodology stems from the highly repetitive and highly homologous sequence of RTs. Alu, for example, exist in millions of copies throughout the human genome. These Alu elements have high homology to the many other copies of Alu, as well as repetition within the individual elements themselves. These repetitive features of Alu and other RTs make them difficult to map to the reference genome when using modern short-read sequencing. Recent advances both in long read sequencers and clever bioinformatics mapping strategies have begun to advance the ability to map RTs to the reference genome. However, there remains a lack of methods with low false positive rates. Due to the advances in technology of annotating RTs, the problem of limited methodology was not the focused on in this study, though it remains a limitation. The second problem, and the focus of this study, is the difficulty of interpreting the downstream effects of the identified RT variants.

Local effects of RTs have been elucidated in the decades since the discovery of RTs. In vitro studies have established a host of effects that RTs can have on genes such as reduced expression or altered splice signals. Furthermore, in silico studies have identified a subset of genome-wide trends that indicate potential functional implications of RT insertions. For example, Zhang et al. showed that RTs are reduced within a few hundred basepairs of exons, inferring that RTs are more hazardous in these regions.

Despite these advances, it remains difficult to interpret the effects of RT variants on a genome-wide scale. This difficulty largely stems from the multitude of effects that a RT can have on a genomic region. RTs contain many structural features ranging purely from the size of the insertion, to the AT-content, to specific protein binding motifs contained within the sequence. All of these features will disrupt the genome in varying ways depending largely on the genomic context in which they insert. It was this ambiguity of the effect of RTs, that this study attempted to resolve by analyzing RTs in varying genomic contexts.
6.1.1 The System

Since I hypothesized that the difficulty in predicting effects of RTs stemmed from the interaction between the RT element and the genomic context, it was important to examine the roles of both of these factors on the effect of RTs within genes. *In silico* techniques were chosen over *in vitro* techniques to accomplish this. The advantages of an *in vitro* study, such as incorporating a tagged RT an observed chromatin chances, would be that the effects of RTs would be studied in a relatively controlled environment. The direct effects of RTs could therefore be determined. Indeed many experiments such as these have proven the effects of L1 insertions on reduced transcriptional processivity of RNA Polymerase-II. The major disadvantage of *in vitro* studies, however, would be that only singular genomic contexts could be queried at any given time. Since, the intention of this project is to determine common trends of the effects of RTs on genes, it is important to study the effects of RTs on many genes and in many contexts. The genome-wide data necessary for this type of analysis would be infeasible using current *in vitro* techniques. Therefore *in silico* techniques were used to examine genome-wide trends from information gathered from publically available databases.

In order to study genome-wide trends of the effects of RTs on genes, two measures of "effect" were selected. The first measure i.) was the distribution of RTs (a proxy for selection) and the second measure ii.) was gene expression. These two types of measurements resulted in an ability to distinguish the strength of an effect of a RT over long periods of evolutionary time, as well as the current biochemical change of gene expression. Together these studies both resulted in findings suggesting that RTs are not only capable of altering genes in a phenotypically selectable manner, many RTs that are presently polymorphic in the human population are associated with changes in gene expression.

6.2 Lack of Antisense Bias in Newly Integrated Elements

The initial finding presented in this dissertation was that newly integrated RTs are present in equal proportions in the sense and antisense orientations. This was contradictory to the trend
of older RTs that have a propensity to be observed in the antisense orientation when within a gene. This finding of equal proportions for newly inserted elements is important because it establishes that antisense bias is not due to integration, but is due to a process that occurs over some length of time.

The equal proportions of sense and antisense elements is further important for the currently expanding field of RT research. Through the advances in sequencing and bioinformatics technologies, the RT field is publishing an increasing number of papers on new RT variants. A caveat of interpreting these new findings is that the procedures for validating RT variants have been slower to advance than the ability to coarsely identify the insertions. Therefore, it is likely that many of the new RT variants being published are false positives. Although validation procedures will surely catch up to the annotation procedures with time, there is currently a difficulty in interpreting these findings. My research showing that RTs insert with equal proportions in the sense and antisense orientation can therefore be used as a quality control measure for these studies. Studies attempting to identify new RT variants can measure the antisense bias of the sample and compare this measure to both the antisense bias of newly inserted elements, and that of older elements. If the antisense bias of a study is more closely aligned to that of older elements, then this would warrant a more stringent attempt at validating the new insertions.

6.3 Purifying Selection of Retrotransposons within Genes

The research presented in Chapter 2 identified that, when within genes, RTs often undergo purifying selection. This was the first time that evidence had been provided for pervasive purifying selection of RTs within genes, though it had been suggested in a genome-wide study. Previously, researchers had alluded to advantageous roles of RTs as the reason for their accumulation in GC-rich regions and near transcription start sites of genes (Lander et al., 2001; Tsirigos & Rigoutsos, 2009). This concept of RTs as common mediators of an unknown regulatory mechanism is seductive in that it infers that there is a common advantageous result of an event that by all other thoughts should be mutagenic.
Conversely, here I provide support for the idea that RTs are more likely to be deleterious within genes, than advantageous. Importantly, these results were identified by analyzing the general trends of RTs. Therefore, the evidence of purifying selection does not exclude the possibility that individual RTs have been co-opted to serve an advantageous function. However, these results do suggest that the increased distribution of RTs near the transcription start site is not likely a result of co-option, and is instead a result of a lack of purifying selection compared to the rest of the gene. Although this is a less entertaining scenario it is a far more plausible hypothesis, and better supported by existing data.

It must be noted that the effects of genetic drift were not examined with respect to AS bias. This was because I had no reason to believe genetic drift would impact sense or antisense elements differentially.

6.4 Deviations in Selection Between Alu and L1

The two RTs studied here, Alu and L1, showed very different trends of purifying selection of sense elements. Initially upon integration, both Alu and L1 demonstrated no signs of antisense bias. For both elements, this then changed to an increase of antisense elements over sense elements. For reference Alu only, the antisense bias remained relatively constant across subfamilies independent of the length of time since they had been inserted. Conversely, the antisense bias of L1 subfamilies continued to increase as a function of longer residence times in the population.

There are a few possible explanations for this observed difference in trends between Alu and L1. One possibility is that the shorter time scale of Alu compared to L1 limits what can be observed for Alu elements. L1 elements have been present since the evolution of early mammals and therefore small changes would have had more time to accumulate to a large enough effect when compared to Alu that have only been present since the split of primates. The slight, but not significant, upward trend observed in Alu may be an indication that small changes are occurring over time. Future studies could attempt to distinguish the effect of time on observed changes in antisense bias by first calculating the time of divergence for each L1 and
Alu subfamily. These time scales could then be overlaid on the same axis, enabling the measure of differences in the slope of the antisense bias between the two major families. Although millidiv could be used as a proxy for time in this type of a study, if there are differences in mutation rates of Alu and L1 then this could lead to misinterpretation of the results.

Another compatible possibility for the differences in Alu and L1 is linked to the strength of the effect that each element has on a gene. This leads to a hypothesis of a multimodal distribution of selection coefficients of RTs within genes. Under this hypothesis both Alu and L1 are highly deleterious if inserted into a coding sequence, and only L1 are deleterious if inserted into less conserved regions, such as introns. More specifically, perhaps Alu and L1 insertions can be thought of as having one of three categories of selection coefficients. From largest to smallest these will be denoted here as \( a \), \( b \), or \( c \). This would mean that there are a subset of RT insertions that have large selection coefficients (\( a \)). For example, a full-length L1 inserted in the sense orientation within a highly conserved coding exon. It would be expected that this type of RT would have a large selection coefficient. Another set of inserts would have very small selection coefficients (\( c \)), such as Alu in the antisense orientation in large intronic regions. Lastly, another group of insertions could lie somewhere between the elements with large and small selection coefficients (\( b \)). Importantly this third set would only consist of L1. If Alu inserts can fall into either category \( a \) or \( b \), then this would explain the fast selection of the highly deleterious elements (\( a \)), and the consistent level of antisense bias over time (\( b \)). Similarly if L1 could either be of type \( a \), \( b \), or \( c \) then this would explain the quick selection of highly deleterious elements (\( a \)) as well as a change in antisense bias over time (\( b \)).

Furthermore, it is more likely that L1 would have elements of type \( b \) as opposed to Alu, since L1 are transcribed by RNA polymerase II and have protein binding sites that can disrupt local transcription.

Although the differences in selection between Alu and L1 may seem like a small matter, it has large implications for the study of RT variants. First, if L1 are indeed capable of having a larger regulatory effects on genes, then they are more likely than Alu to be mediators of human disease when inserted into introns. Secondly if Alu are more likely to have smaller initial
effects on the gene, then they are also more likely to be passed to the next generation. This gives Alu a higher chance of being co-opted by the system to serve a function since the Alu elements would have longer residence times in the population.

6.5 Decrease of Retrotransposons Near Exons

As mentioned in section 6.3 it was previously thought that RTs were increased near transcription start sites because they were conferring some sort of selective advantage to the gene. The research presented here shows the contrary. Our results indicated that, as had been seen in previous studies (Zhang & Mager, 2012), there was an increase of retrotransposons near the transcription start site. Interestingly, though, when genes were normalized by length, the RTs in the decile containing the transcription start site were reduced compared to what would be expected based on the trend across the rest of the deciles. Since the first decile was enriched for exons compared to the internal deciles, this dip in RT density indicated that RT density may be influenced by exon density, even near the 5’-UTR. Indeed, when this concept was followed up on, it was identified that not only are RTs reduced near coding exons, as has been previously shown, but when in the sense orientation they were also reduced near the 5’-UTR. This was a very unexpected finding since, as mentioned earlier, RTs are increased near the transcription start site, and therefore the 5’-UTR.

Despite incongruous interpretations, the previous findings and the current findings are compatible with one another. The finding that there is a lower density of coding exons near the transcription start site facilitates the interpretation of both models. This decrease in coding density results in a reciprocal increase in RT density that looks as though it is related to the 5’-UTR, as previous studies have identified. However, when the effect of 5’-UTRs was studied in regions of genes with higher levels of coding sequence, there were respective decreases in RTs. This supports the finding that in the absence of coding sequences, there is an increase of RTs. Therefore the increase of RTs near the 5’-UTR of genes is likely a red-herring, and instead is a function of coding sequence density.
6.6 Binding Motifs

Transcription factor binding sites may not have as large of an effect on a gene as an insertion that disrupts function by inserting directly into the coding sequence. For this reason it may be more difficult to pick up signs of the effects of RT-contained binding motifs on an organism. However, these smaller effects may have a larger potential of being co-opted by the system over time, due to longer residence times in the population.

Interestingly the conservation analysis of polyadenylation sequences showed that the central thymine (T) of the motif was less conserved than the external adenosines (A). Although it is possible that this is due to a fluctuation in the mutation rates between the T and the A’s, previous studies have shown that the internal T limits polyadenylation activity (Zarkower et al., 1986). Therefore, perhaps the lack of conservation of the internal T is reflective of the function of the polyadenylation signal. If so, perhaps the polyadenylation signals that are fixed in humans but mutated in non-human primates have either been co-opted by the system to serve as alternative termination signals, or are in regions that are less deleterious in humans. Either hypothesis warrants exploration into the differences of RT-contained polyadenylation signals between primates.

It must be noted that the subfamilies of the elements were not included as a confounding variable in this analysis. Since each subfamily encodes different binding motifs, it is important for future studies to account for variation in subfamily frequency within genes and in intergenic regions.

6.7 Tissue Specific Association with Antisense Bias

The finding that AS bias was dependent on the tissue-expression profile indicated that there may be different functional impacts of RTs depending on the transcriptional environment. Moreover, AS bias was reduced in tissue-specific gene sets as opposed to within housekeeping genes that contain a high number of RTs. Perhaps housekeeping genes are more permissive to RT integrations and malleability of gene regulation, while tissue-specific genes require a more
conserved model of regulation that would be otherwise obstructed by a new insertion. If this is true, it would underscore the importance of studying RTs in relation to human disease.

It must be noted that, although it didn’t reach significance, there was a subtle association of AS bias with the sample size of genes in the tissue-specific category. Since there is a reduction in power to detect significant results with smaller sample sizes it is important that the association of tissue-specific expression and RT AS bias is replicated in population studies once the data becomes available.

6.8 Examining Signs of Selection for Variables other than Antisense Bias

It is important to note that I did not examine the evidence of selection for variables other than the AS bias. It is possible that AS bias is due to selection while the other biases (ie: tissue-specific patterns of distribution and binding motif preference) are due to integration rather than selection. With regard to binding motifs, although an integration bias is possible, it is highly unlikely since the molecular biology of retrotransposition is not known to directly involve the internal binding motifs in any way.

There is a possibility of integration bias effecting RT distribution within tissue-specific genes, although it is unlikely. Since retrotransposition is influenced by chromatin dynamics, the structure of chromatin in the germ cell, specifically in the male germ cells, may play a role in bias of the number of insertions in different regions. If integration bias was the cause of differences in antisense bias between tissue-specific and housekeeping genes, it would be expected that the presence binding motifs would be similar between the two groups. That is unless the binding motifs aid in the transcription of the retrotransposon in certain tissues, in which case the binding sites would be expected to be differentially represented. Importantly, the major difference measured between housekeeping and tissue-specific genes was the change in antisense bias. Even if there was a bias for increased integration in certain genes depending on tissue of expression, this would not translate into deviations of antisense bias. Furthermore, studies analyzing mutations in tissue-specific genes have identified that selection rates vary de-
pending on the expression pattern of the gene (Duret & Mouchiroud, 2000). This supports the hypothesis that RTs are differentially selected depending on the tissue specificity of the gene.

6.9 Decreased Gene Expression in the Presence of a Retrotransposon

Results from the selection analyses indicated that RTs had a phenotypic effect on organisms that was selected on throughout primate evolution. It was therefore important to examine if any effects could be observed within genes that contained RTs in present-day humans. The main results of these analyses showed that the presence of a RT, and more specifically an Alu, was associated with a decrease in gene expression. Although there was a significant association with the presence of a RT and decrease in expression, not all genes showed a response to the presence of a RT. This indicated that there were latent variables effecting the statistical interaction of RTs and gene expression.

Again, this harkens back to the results from chapters 2 and 3 that showed that the distribution of RTs was largely contingent upon the context in which they inserted. For example, the orientation with respect to the gene, the proximity to a 5’-UTR, the proximity to a coding sequence, and the gene type in which they inserted are all modifiers of the distribution of RTs within genes. It is therefore likely that these variables are also modifiers of gene expression. Unfortunately there was not a large enough sample size of elements that also had expression values to calculate the effect of these variables on gene expression. In the future, when more data on RT polymorphisms becomes available, associations between these subtle variations and gene expression should be further explored.

6.10 Differences in Effect on Expression Between Retrotransposon Types

Most of the significant associations between RT and gene expression were associated with Alu and not L1 or SVA. It is possible that Alu have larger effects on gene expression than either L1 or SVA. However, it is also possible that Alu, L1, and SVA all have effects on gene expression, but that selective pressures and sample size issues reduce the ability to observe changes in the
latter two elements (req. effect size, Alu Chimp: 0.08; req. effect size, L1 Chimp: 0.48; req. effect size, SVA Chimp: 0.23).

Evidence in the section on selection indicated that L1 undergoes stronger forces of selection than Alu. If L1 have a larger effect on gene expression than Alu, then perhaps they do not survive in the population, and therefore they were not observed in this analysis. Conversely, and consistent with the hypothesis presented earlier, perhaps Alu have a smaller, but present, impact on gene expression. Therefore they are passed on through the population due to a low selective coefficient and the effect can therefore be observed. Regarding SVA, it is possible that these elements have an effect on genes that was not detected since SVA occur at much lower levels in the genome than the other elements. This reduction in sample size reduces the power to detect any significant results with small effect sizes.

6.11 Interaction of Sex of the Individual with Changes in Gene Expression in the Presence of a Retrotransposon

Perhaps the most intriguing finding regarding gene expression was that the decrease in expression was associated with RTs in a sex-specific manner. Importantly, the reduced gene expression in the presence of a polymorphic Alu was contingent on the individual being male. The finding that there was a difference in the effect of the RT on gene expression based on sex indicated the possibility of an interaction between RTs and sex-specific factors. Some sex-specific factors that could be facilitating this interaction are hormones such as estrogen and androgen.

RTs are known to contain hormone response elements within their consensus sequences. For example most of the estrogen-responsive DR2 elements in the human genome are present within Alu sequences. Estrogen, however, is a female specific hormone, and therefore if there was an interaction with DR2 elements, the response would be expected to be observed in females as opposed to males. RORA-1 is another hormone response element contained within Alu. More importantly, RORA-1 has sex-specific responses both to estrogen as well as an-
drogens. Future studies should investigate the role of RORA-1 and other hormone response elements in the effect of Alu on gene expression.

Even though the change in gene expression was male specific, it is possible that an unobserved effect is present within females as well. Since it has been shown that long stretches of As reduce transcription, it is likely that an insertion of Alu would decrease expression in both males and females. Perhaps in females estrogen binding elements play a role in opening up the region for higher expression, thereby negating the initial effect of the insertion. Although this hypothesis is more complex than the previous mentioned hypothesis, it should be examined as well in order to determine the roles of Alu in mediating sex-specific gene expression.

6.12 Alu and Autism Spectrum Disorder

In the analyses of altered gene expression in association with retrotransposon polymorphisms, there was a significant overlap between genes that were differentially expressed and genes that serve a brain specific function. Even more interesting was the finding that these genes were significantly enriched for genes that have been previously associated with Autism Spectrum Disorder (ASD). The enrichment of Alu in brain specific genes could be due to brain genes having larger introns on average than most genes. Therefore genes involved in brain function are more likely to have a RT insert within their boundaries than other genes. Even if there is a size-advantage of brain-related genes to be inserted by Alu than non-brain-related genes, this does not preclude the possibility that once Alu are inserted within these genes, they are capable of being co-opted by the genome to serve a function. Interestingly the RORA receptor, which is contained within Alu, has been implicated in Autism Spectrum Disorder independent of retrotransposon insertions. The association of RORA in Autism Spectrum disorder necessitates further research into the interaction between Alu elements and changes in gene expression at a molecular level.

In conclusions these results validate the hypothesis that RTs have effects on the genome that are dependent on the genomic context in which they integrate. It was identified that RTs are purified not only from within genes, but that this removal is increased when the RT is in
the sense orientation. Moreover, there seem to be sex-specific effects mediated by RTs on gene expression. These findings necessitate further study into the biochemical mechanisms underlying RT-mediation of gene expression in humans.
### Table 7.1 – Significance and direction of age differences across Alu groups

<table>
<thead>
<tr>
<th>ALU</th>
<th>Sense</th>
<th>chrX</th>
<th>Aut</th>
<th>chrY</th>
<th>Anti</th>
<th>chrX</th>
<th>Aut</th>
<th>chrY</th>
<th>Inter</th>
<th>chrX</th>
<th>Aut</th>
<th>chrY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>.</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>2</td>
<td>Sense</td>
<td>Aut</td>
<td>.</td>
<td>.</td>
<td>L</td>
<td>G</td>
<td>=</td>
<td>L</td>
<td>G</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>G</td>
<td>G</td>
<td>=</td>
<td>G</td>
<td>G</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>5</td>
<td>Antisense</td>
<td>Aut</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>L</td>
<td>G</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>G</td>
<td>G</td>
<td>L</td>
<td>G</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>8</td>
<td>Intergenic</td>
<td>Aut</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

red: Student’s t-test $p < 2.2e-16$; blue: Student’s t-test $p < 0.0014$; teal: Student’s t-test $p < 0.0018$ L: millidiv of row group is less than column group; G: millidiv of row group is greater than column group; =: millidiv of row group is equal to column group. Anti: Antisense; Inter: Intergenic

### Table 7.2 – Significance and direction of age differences across L1 groups

<table>
<thead>
<tr>
<th>L1</th>
<th>Sense</th>
<th>chrX</th>
<th>Aut</th>
<th>chrY</th>
<th>Anti</th>
<th>chrX</th>
<th>Aut</th>
<th>chrY</th>
<th>Inter</th>
<th>chrX</th>
<th>Aut</th>
<th>chrY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>.</td>
<td>L</td>
<td>=</td>
<td>L</td>
<td>L</td>
<td>=</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>2</td>
<td>Sense</td>
<td>Aut</td>
<td>.</td>
<td>G</td>
<td>G</td>
<td>L</td>
<td>=</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>3</td>
<td>Y</td>
<td>.</td>
<td>.</td>
<td>=</td>
<td>L</td>
<td>L</td>
<td>=</td>
<td>L</td>
<td>=</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>L</td>
<td>L</td>
<td>G</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>5</td>
<td>Antisense</td>
<td>Aut</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>=</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>6</td>
<td>Y</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>G</td>
<td>=</td>
<td>G</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>8</td>
<td>Intergenic</td>
<td>Aut</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
<td>G</td>
</tr>
<tr>
<td>9</td>
<td>Y</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

red: Student’s t-test $p < 2.2e-16$; blue: Student’s t-test $p < 0.0014$; L: millidiv of row group is less than column group; G: millidiv of row group is greater than column group; =: millidiv of row group is equal to column group. Anti: Antisense; Inter: Intergenic


Manuelidis, L. (1982). Nucleotide sequence definition of a major human repeated DNA, the Hind III 1.9 kb family. *Nucleic acids research*, 10(10), 3211–3219.


