COMPUTATIONAL APPROACHES TO THE STUDY OF GENOMIC ROLES OF REPEATED DNA SEQUENCES

by

LOUIE NATHAN VAN DE LAGEMAAT

B.A.Sc, The University of British Columbia, 1996

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

(Genetics Graduate Program)

THE UNIVERSITY OF BRITISH COLUMBIA

April 2006

Abstract

Repeated sequences make up nearly half of the bulk of mammalian genomes and vary widely in structure and function. This thesis describes computational approaches for assessment of interaction of repeats and their host genomes. Following public release of the human genome sequence, initial investigations focused on overall distributions of retroelements with respect to sequence composition and genic position. Exclusion of various retroelements from regions both within and surrounding protein-coding genes suggested selection against the presence of these elements. Directional biases of accepted retroelements in these regions further supported this notion. Directional biases are understood to reflect differential mutagenicity by sequences in one direction vs. the other. To examine the relationship between protein-coding genes and mobile elements further, mappings of genomic transposable elements in the human and mouse genomes were examined in relationship to positions of all exons of protein-coding gene mRNAs. I found that approximately one quarter of mRNAs of protein-coding genes harbor sequence contributed by transposable elements. The fact that transposable element sequence is most often found in untranslated regions (UTRs) suggests a highly significant role for these sequences in modulation of translation efficiency in addition to roles in transcription. Further investigations used directional biases of retroelements in transcribed regions in humans and mice to show that transposable elements transcribed by RNA polymerase II (pol II) exert varying effects upon insertion, depending on the sequence of the element. Finally, a bioinformatic study done on global insertion patterns of retroelements since human-chimpanzee divergence revealed that some transposable elements polymorphic for presence or absence in primate genomes actually represented

deletions rather than *de novo* insertions. These deletions were flanked by short tracts of identical sequence, suggesting deletion by recombinational mechanisms. The relative rarity of these events lends support to the assumed stability of transposable element insertions while illustrating the recombinational activity of even low-copy, nonadjacent, short repeated sequences, such as those found flanking transposable element insertions. In summary, these bioinformatic studies lend insight into the biological roles and genomic effects of mammalian genomic repeats, especially transposable elements.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ALV	Avian Leukosis Virus
BLAST	Basic Local Alignment Search Tool
BLAT	Blast-like Alignment Tool
DSB	double strand break
ERV	endogenous retrovirus(like)
ETn	Early Transposon
GO	Gene Ontology
HERV	human ERV
HIV	Human Immunodeficiency Virus
HR	homologous recombination
IAP	intracisternal A particle
indel	insertion/deletion
IPR	InterPro
IR	inverted repeat
KOG	euKaryotic clusters of Orthologous Groups
LINE	long interspersed nuclear element
LTR	long terminal repeat
MaLR	Mammalian apparent LTR Retrotransposon
MER4	MEdium-Reiterated repeat 4
MLT	Mammalian LTR Transposon
MLV	(Moloney) Murine Leukemia Virus
MRN	MRE11/RAD50/NBS1 (protein complex)
NCBI	National Center for Biotechnology Information
NHEJ	nonhomologous end joining
ORF	open reading frame
PCR	polymerase chain reaction
pol II	RNA polymerase II
RPA	replication protein A
SINE	short interspersed nuclear element
SIV	Simian Immumodeficiency Virus
SSA	single strand anealing
ssDNA	single stranded DNA
TE	transposable element
TPRT	target-primed reverse transcription
TSD	target site duplication
UCSC	University of California Santa Cruz
UTR	untranslated region

Acknowledgements

Reflecting over the work described here, I would like to thank many people for their input and give them credit for the role they have played in my work. First of all, many thanks go to my supervisor, Dr. Dixie Mager. Besides being a smart and insightful person, she has been a fun and interactive supervisor, and conversation with her has always been profitable. I really appreciated the measure of freedom she gave me to pursue particular aspects of the questions at hand, and she was always ready to discuss new data, especially when there were graphs to look at. Furthermore, it's been fun getting to know past and present members of the Mager Lab and the broader BC Cancer Research Centre community, and I thank them all for their friendship and input.

Further credit is due to many people who have been involved with my projects in various supportive roles. First in this regard have been my thesis committee members, who have contributed in various ways to my success. Drs. Ann Rose and Steven Jones have been a source of good questions about my data, and in more than one instance, have led to positive development of my projects. Dr. Holger Hoos's contributions in the realm of algorithm development have led to some of the software developed in the course of this work. In addition to my committee, I thank the National Science and Engineering Research Council of Canada and the Canadian Institutes of Health Research for generous funding during the course of my work.

Very importantly, I feel very grateful to Patrik Medstrand for his involvement all my PhD long. I suppose our relationship goes beyond a supervisor-student relationship into the realm of personal friendship. I have learned a great deal from Patrik, and the exchange visits to his lab in Sweden have been productive, and more than that, lots of

fun. Thanks also goes to Patrik's wife Lilly, and my hosts in Lund, the Bruce family.

Last and definitely not least, I thank my parents, Wulf and Henrietta, for all their love and interest in my work. I have always enjoyed the discussions we have had.

Chapter 1: Introduction

1.1 Thesis overview

The goal of this thesis project was to use global computational genomic analyses of repeated sequences in mammalian genomes to understand interactions of these elements with their host genome. Repetitive sequences, including transposable elements and tandem and segmental duplications, make up nearly half of mammalian genomes. Repeated sequences influence the host genome in several main ways, from the obvious role in genome expansion to generation of distributed similar sequences susceptible to ectopic recombination, to provision of transcriptional and regulatory signals. As early as Barbara McClintock's experiments in maize in the 1950s, there was some appreciation of the cytogenetic and concomitant regulatory consequences of DNA that could move about in a genome. However, only more recently have molecular studies begun to unravel some of the mechanisms involved in amplification of repetitive sequence and the mechanisms mediating the regulatory roles of repetitive elements fixed in mammalian genomes. These advances in understanding, coupled with the availability of the sequenced genomes of humans and various model organisms, have enabled further genome-wide studies of repeated sequences and their role in organismal biology. The research reported in this thesis attempted to clarify the dual roles of transposable elements and other repetitive sequences and their potentials for both benefit and harm. Primarily, bioinformatic and mapping techniques were used to elucidate these roles with occasional additional validation using wet laboratory approaches. These approaches make it possible to address questions regarding global genomic processes and how repetitive DNA has shaped genome architecture.

1.2 Sequence motion: what, how, why, and where to?

1.2.1 Discovery of repetitive DNA

Mammalian genomes may best be described as a patchwork of many types of sequences, including genes, control regions, and, not mutually exclusively, repeated sequences. The first hints of the pervasive presence of repeated sequence in genomes date back to the early 1950s, when it was observed that the DNA content of cells, which was presumed to contain the genes, was poorly correlated with the overall level of complexity of the organism (the so-called C-value paradox), reviewed in Gregory (2001). One early explanation proposed that the extra DNA was 'junk', or non-coding pseudogenes, while later theories suggested it to be self-replicating 'selfish DNA' (Doolittle and Sapienza 1980; Orgel and Crick 1980). The discovery of mobile or transposable elements by Barbara McClintock a half century ago (McClintock 1950; McClintock 1956) and subsequent discovery of retrotransposable elements has better explained the nature of this 'selfishness'.

1.2.2 Mobile elements and their modes of transposition

Thorough study of the human genome has shown that recognizable repeats make up nearly half of its bulk (International Human Genome Sequencing Consortium 2001). Lesser coverage by repeats in other mammalian genomes, for example in rodents, has been attributed to the incompleteness of the library of known rodent repeats as well as faster substitution rates in the rodent lineage, and therefore the estimates of approximately 40% repetitiveness of rodent genomes are considered to be lower bounds (Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004).

Several broad categories of repetitive sequence exist. One of the most basic distinctions one can make is related to copy number. Low copy number repeats, including tandem and segmental duplications, are generated by recombinational mechanisms and are discussed in Section 1.4. High copy number repeats, also known as mobile or transposable elements (TEs), are considered to move about using proteins encoded either by themselves or *in trans* by another mobile element (Table 1.1). These elements have amplified over the course of evolution and attained copy numbers ranging from several tens to over a million in mammalian genomes. TEs may be characterized on the basis of several aspects, including whether or not the element is autonomous, presence or absence of terminal repeats, and the mechanisms by which the elements move.

Table 1.1 Human TE types, copy numbers, genomic coverage, and mutation occurrence

Type/family	*Copies per haploid human genome (X1000)*[a]	*Genomic coverage (%)*[a]	*Human mutations documented*[b]
SINEs	1558	13.14	
Alu	1090	10.60	22
MIR	468	2.54	
LINEs	868	20.42	
L1	516	16.89	13
L2	315	3.22	
L3	37	0.31	
LTR elements	443	8.29	
ERV class I	112	2.89	
ERV class II (ERV-K)	8	0.31	
ERV class III (ERV-L)	83	1.44	
MaLR	240	3.65	
SVA	3[c]	0.15	4
DNA elements	294	2.84	
Total		**44.84**	**39**

[a]Taken from International Human Genome Sequencing Consortium (2001), unless otherwise noted
[b]From Chen et al. (2005)
[c]From Wang et al. (2005)

DNA transposons, no longer mobile in mammals, are exemplified by the P-

elements in Drosophila (Pinsker et al. 2001). Active elements of this type move by a cut-and-paste mechanism (Kazazian 2004). Autonomous elements encode their own transposase protein, which binds in a sequence-specific manner to the terminal inverted repeats flanking the element and cleaves the DNA, excising the element (Miskey et al. 2005). Non-coding DNA with similar flanking inverted repeats is also susceptible to cutting and pasting by the same proteins. In rice, multiple DNA transposons exist, including autonomous *Pong* and non-autonomous *mPing* and *Mutator*-like elements (Jiang et al. 2003; Jiang et al. 2004). Proliferation of *Mutator*-like non-autonomous elements harboring coding-competent genic sequence has also been documented (Jiang et al. 2004). In mammals, these elements are no longer functional. However, their recombinase functions persist in the co-opted RAG genes used in V(D)J recombination (Brandt and Roth 2004; Schatz 2004).

Retroelements, in contrast to DNA transposons, move by a copy-and-paste mechanism (Kazazian 2004). RNA intermediates are reverse transcribed into DNA using a reverse transcriptase protein. Similar to DNA transposons, autonomous retroelements code for their own reverse transcriptase, while non-autonomous elements depend on this protein *in trans*.

Several broad classes of retroelements exist, including long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), pseudogenes, and elements with long terminal repeats (LTRs) (Kazazian 2004). Approximately 100 families of LTR-containing elements have been found in humans, varying in length from several hundred base pairs in size for solitary LTRs to full length elements as long as 10 kb (Jurka 2000; International Human Genome Sequencing Consortium 2001). These include endogenous retroviruses (ERVs), which are presumed to have resulted from

germline infections by exogenous viruses, LTR retrotransposons, and repetitive elements with an LTR-like structure for which no corresponding full-length structure has been identified (Mager and Medstrand 2003; Medstrand et al. 2005). Approximately 85% of LTR-retroelement insertions exist in the human genome as solitary LTRs, a result of recombination between the terminal repeats (International Human Genome Sequencing Consortium 2001).

ERVs are so named due to their structural similarity to the integrated provirus form of exogenous retroviruses (Figure 1.1). The flanking LTRs contain necessary regulatory motifs for their transcription in three regions, termed U3, R, and U5, described further below. The internal sequence encodes the proteins necessary for their retrotransposition. Both autonomous ERVs and LTR retrotransposons increase their copy number through mRNA intermediates which are reverse transcribed and reinserted into the host genome, and their overall genomic organization includes several features related to their life cycle (Wilkinson et al. 1994). Just 3' of the upstream LTR are a tRNA primer binding site, which primes reverse strand synthesis, and a packaging signal which interacts with the nucleocapsid protein. The internal sequence of these elements includes *gag* and *pol* genes, which code for a nucleocapsid protein, a protease, RNAse H, reverse transcriptase, integrase, and other genes. Just 5' of the downstream LTR is a poly-purine tract. Finally, the LTR begins and ends with TG and CA dinucleotides, which are important for insertion. Some ERVs have an additional *env* gene, which codes for an envelope protein and allows ERVs to form infectious particles that can escape the host cell and reinfect adjacent cells. However, it should be noted that in humans the vast majority of these elements are defective, with mutations in some, usually all, of their internal sequences. No disease-causing ERV insertions have been found in humans,

although several loci are polymorphic for presence or absence of an ERV in humans (Turner et al. 2001; Bennett et al. 2004; Belshaw et al. 2005). As with all retroviruses, insertions of ERVs are flanked by short tracts of identical sequence, called a target site duplication (TSD), usually 4-6 bp in length.

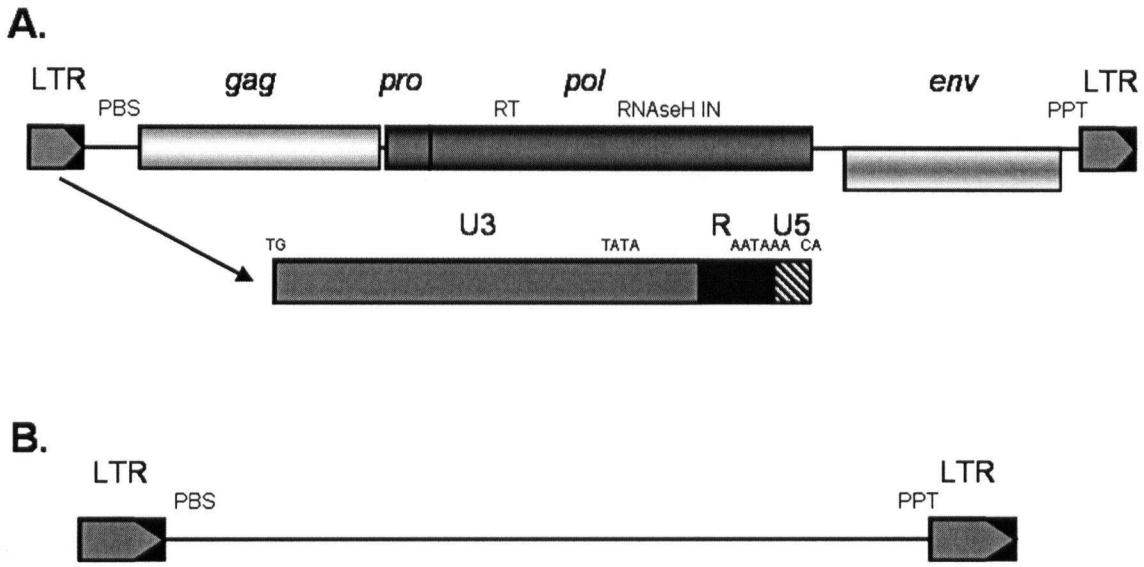

Figure 1.1 Full length endogenous retrovirus-like elements. A. Full-length HERV-H, as described by Jern et al. (2005). HERV-H has genomic organization and genes related to those of infectious retroviruses, as well as canonical pol-II transcriptional signals. This suggests that it has been an autonomous element that entered the primate germline as an infection by an exogenous virus. PBS and PPT are primer binding site and polypurine tract, respectively. B. HUERS-P1 element, first described by Harada et al. (1987). HUERS-P1 element lacks discernable open reading frames, although having presumptive pol-II transcriptional signals in its LTRs, and therefore is presumed to be non-autonomous.

Another class of LTR-containing elements exists whose internal sequence bears little or no resemblance to that of exogenous viruses. These elements include the Mammalian apparent LTR retrotransposons (MaLRs) and so-called medium-reiterated 4 elements (MER4s) which lack internal similarity to retroviral genes (Smit 1993). However, these elements do contain the obligatory LTRs with regulatory motifs as well as polypurine tracts, which leaves open the question of how these elements have been

mobilized.

Active full-length LINE elements are typified by the L1 family (Figure 1.2). These elements are transcribed from an internal RNA polymerase II (pol II) promoter (Swergold 1990; Tchenio et al. 2000; Yang et al. 2003; Athanikar et al. 2004) and have two open reading frames (ORFs), which encode a nucleic acid binding protein, a reverse transcriptase, and an endonuclease; these proteins are necessary for their own transposition (Kazazian 2004). These elements terminate with a polyadenylation signal and poly-A tract. The origin of LINE elements is unknown, however many classes of eukaryotes contain them, including mammals, fish, invertebrates, plants, and fungi (Furano 2000). L1s exhibit a marked *cis* preference, which means that proteins encoded most often act only on the mRNA that encoded them (Wei et al. 2001). In addition, however, L1s have also been shown to mobilize other RNA species, for example processed mRNAs and SINEs in human (Esnault et al. 2000; Dewannieux et al. 2003). Antisense promoter activity from the 5' UTR has also been reported (Nigumann et al. 2002). Species of mRNA mobilized by L1s usually have a TSD 7-20 bp in length.

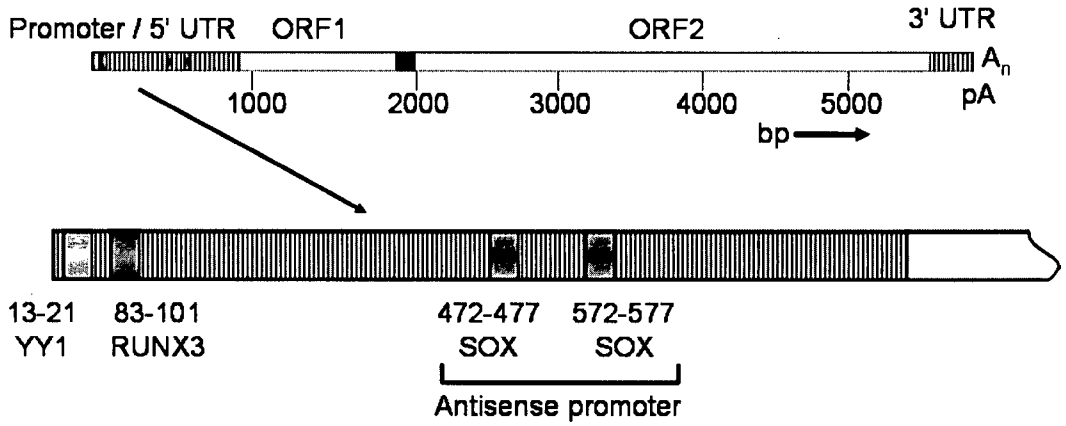

Figure 1.2 L1 structure. The overall genomic organization of the currently active L1 (Hs) consists of a 5' UTR/promoter, two open reading frames (ORF1 and ORF2), and a short 3' UTR, which terminates in a canonical polyadenylation signal followed immediately by a poly-A tract. The promoter region has recognized binding sites for YY1, RUNX3, and SOX-family transcription factors (Swergold 1990; Tchenio et al. 2000; Yang et al. 2003; Athanikar et al. 2004). An antisense promoter has been described (Nigumann et al. 2002).

Active SINE elements are typified in humans by the Alus and SVA elements. Alus are non-protein coding, RNA polymerase III-driven, 7SL RNA-derived dimeric elements (Ullu and Tschudi 1984) (Figure 1.3). Alus number in excess of one million copies in the haploid human genome (International Human Genome Sequencing Consortium 2001) and are still actively retrotransposing in the primate lineage. Their current amplification rate, estimated at one in 200 live births (Deininger and Batzer 1999), is 100-fold lower than at the peak of their activity (Shen et al. 1991). Elements polymorphic for presence or absence in the human population, which number approximately 1000 in the average human (Bennett et al. 2004), have also demonstrated usefulness in distinguishing relationships of human populations (Batzer and Deininger 2002). In addition to Alus, primate genomes contain a family of mammalian-wide tRNA-derived SINEs called MIR, which are older than Alus and are not known to be transcribed in humans (Smit and Riggs 1995). Other tRNA-derived SINEs are active in mice (Dewannieux and Heidmann 2005).

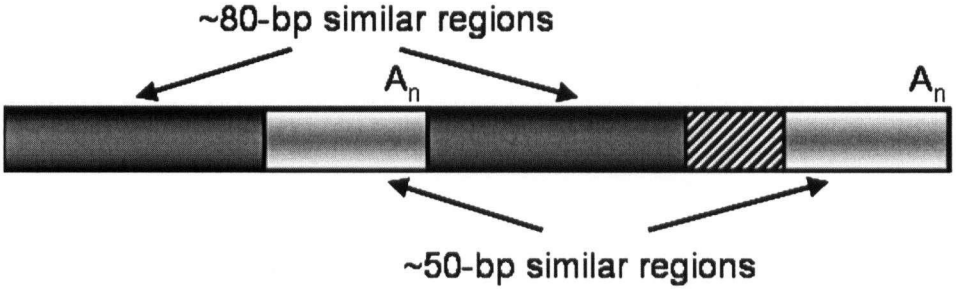

Figure 1.3 Typical Alu element of ~300 bp. Similar regions share ~70% identity.

In addition to Alu elements, SVAs comprise another relatively numerous superfamily of SINE elements. Numbering 2762 in the haploid human genome (Ono et al. 1987; Wang et al. 2005), these elements are believed to be confined to the hominoid

primates, indicating a relatively recent evolutionary origin (Kim et al. 1999; Wang et al. 2005). They consist of a hexamer repeat, homologies to inverted Alu elements, a variable number of tandem repeats, and a deleted partial copy of an ERV-K element including the terminal part of an *env* gene and a partially deleted LTR (Figure 1.4). SVAs end with a polyadenylation signal and poly-A tract (Ono et al. 1987; Zhu et al. 1992; Shen et al. 1994; Wang et al. 2005). The fact that SVA elements contain a major portion of an ERV-K LTR suggests that these active elements may have LTR-like effects. Furthermore, mRNA evidence exists supporting a role for these elements as alternative promoters, for example the hyaluronoglucosaminidase 1 and the G protein-coupled receptor MRGX3 genes (University of California Santa Cruz Genome Browser, http://genome.ucsc.edu).

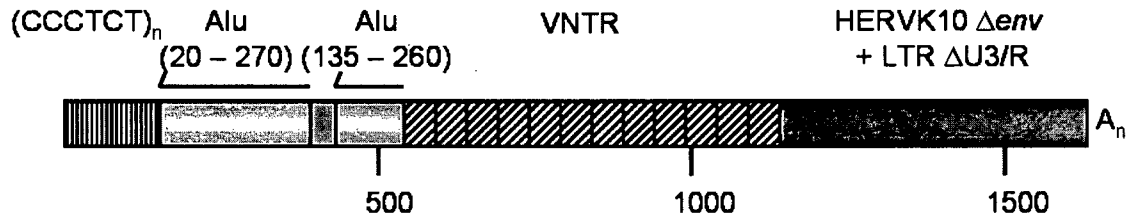

Figure 1.4 Structure of a typical SVA element, found at human chromosome 13q14.11. SVAs consist of a tract of CCCTCT hexamer repeats, two regions homologous to Alu elements, a region containing a variable number of tandem repeats (VNTR), and a partial HERVK10 *env* gene and partial LTR, ending with the polyadenylation signal and an immediate poly-A tract.

L1-driven retrotransposition is targeted to TT/AAAA sites, which are widely dispersed through mammalian genomes (Jurka 1997). The process is known as target primed reverse transcription (TPRT), reviewed in Kazazian (2004), and begins with opposite-strand nicking at the insertion site, uncovering a short poly-T tract complementary to the mRNA to be reverse-transcribed. The mRNA poly-A tail anneals

to the poly-T tract, which then serves as a primer for reverse transcription (Figure 1.5). After insertion, new elements segregate as Mendelian genes in the population. The likelihood of any given insertion reaching fixation in a population is very small.

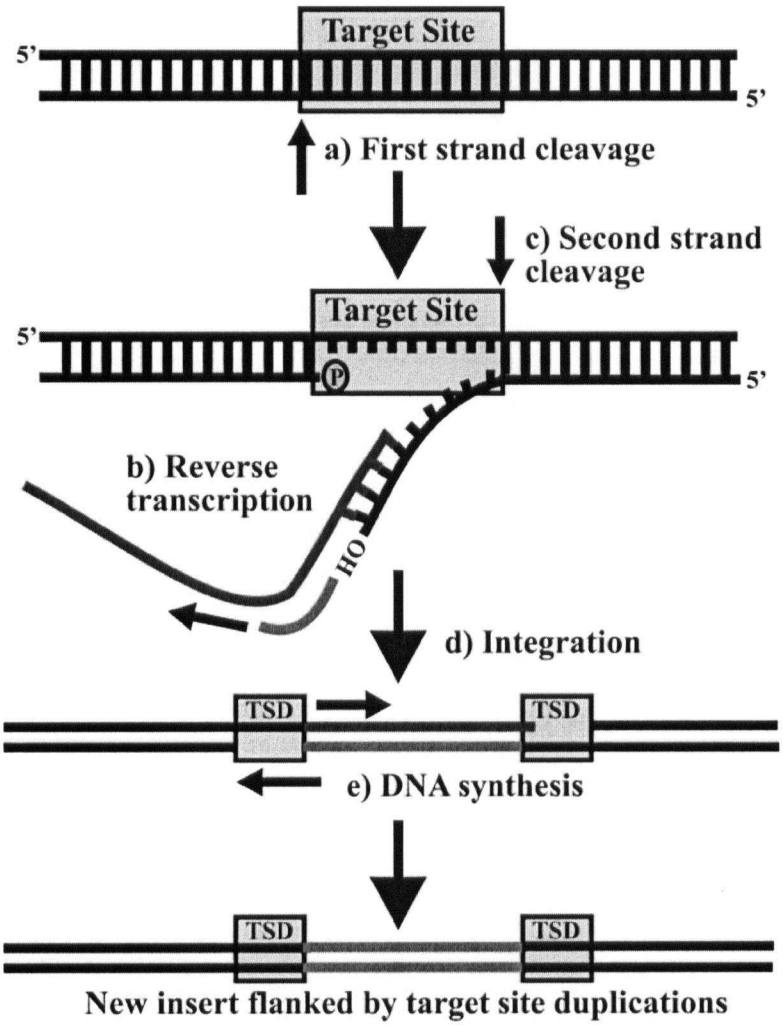

Figure 1.5 Target-primed reverse transcription (TPRT). Taken from Ostertag and Kazazian (2001).

Due to the copy-and-paste strategy employed by retroelements in their amplification, new retroelement insertions are ideally identical to their ancestor element. Independent mutations occurring in an individual element either during or after

retrotransposition introduce variation which is then faithfully inherited by derivative copies of the element. This process results in an increasing genomic population of elements marked by diagnostic mutations. These mutations may then be exploited to infer family relationships of accumulated retroelements (International Human Genome Sequencing Consortium 2001). Furthermore, the likely sequence of the ancestral elements can be reconstructed from sequence alignments of distributed copies of the element. Divergence from this consensus element can then be computed and, assuming a molecular clock, the approximate age of each element may be computed. This type of analysis has been described elsewhere (Smit 1993). The molecular clock is usually calibrated using fossil evidence whose age is calculated from radioactive dating of fossils based on the apparent age of the rocks where the fossils are found (Goodman et al. 1998).

1.2.3 Mutagenic roles of transposable elements

As early as Barbara McClintock's experiments with DNA transposons in corn, an appreciation of the 'gene-controlling' effect of these 'jumping genes' began to take root (McClintock 1950; McClintock 1956). In those experiments, it was noted even in physical examination of chromosomes that new insertions could drastically alter local chromosomal structure and these could account for alterations in phenotypic characteristics such as kernel color.

Mutagenic roles of TEs may be grouped into several types. Most basic of all, insertional mutagenesis involves disruption of a conserved and functionally important region of DNA. Perhaps because they are more readily analyzed, most well-studied cases of insertional mutagenesis involve disruption of a coding exon of a transcribed gene which then introduces a stop codon or frame-shift resulting in a prematurely terminated or non-functional transcript. Examples of this type of mutagenesis reported in the

literature include inactivation of the human cholinesterase gene by an Alu insertion (Muratani et al. 1991) and disruption of the Factor VIII gene by an L1 element causing hemophilia A (Kazazian et al. 1988). More examples are described by Deininger and Batzer (1999) and reviewed by Chen et al. (2005). As a variation on this theme, in one report a sequence believed to have been 3'-transduced by an SVA element was found to have disrupted exon 5 of the alpha spectrin gene, resulting in exon skipping (Hassoun et al. 1994; Ostertag et al. 2003).

Several more subtle regulatory roles exist for TEs, usually related to the structure of the consensus element. For example, although the phenomenon is apparently fairly rare, mutations in intronic antisense Alu 3' ends can lead to generation of an efficient splice acceptor site which may result in disease (Sorek et al. 2002; Lev-Maor et al. 2003; Sorek et al. 2004).

L1s, on the other hand, have recently been shown to act fairly universally as transcriptional rheostats, reducing transcription of genes they are found in (Han et al. 2004). This activity is believed to be due to the A-rich consensus element, which is believed to cause the pol-II holoenzyme to fall off its template with high frequency. Reduction in this A-richness while conserving the protein sequence of the element resulted in highly efficient transcription (Han and Boeke 2004). Furthermore, presence of A-rich L1 sequence in introns of genes was correlated with overall reduction in transcription efficiency. On the other hand, L1s integrating into introns in a direction antisense to the gene's direction of transcription have a demonstrated polyadenylation activity (Perepelitsa-Belancio and Deininger 2003; Han et al. 2004), resulting in reduction of transcription in either orientation by intronic L1s. Intronic L1s have also been linked to disease (Kimberland et al. 1999).

Another type of mutagenesis related to the element's sequence is that posed by ERVs and their LTRs. As described above, active elements of this type code for several genes related to their life cycle. In order to produce correct amounts of each protein, these elements often splice out part of their coding sequence, requiring the presence of functional splice donor and acceptor sites in the full-length element (Rabson and Graves 1997). Mutations due to these splice sites have been demonstrated in mice (Maksakova et al. 2006).

More importantly than donation of splicing motifs by full-length ERVs, LTRs harbor transcriptional control elements including fully functional promoters and polyadenylation signals. The general structure of LTRs has been studied extensively (Temin 1982; Rabson and Graves 1997). A classical LTR consists of three regions, termed U3, R, and U5. The U3 region extends from the 5' end of the LTR or proviral copy through the promoter to the start of transcription (Figure 1.1). Together with the R region, the U3 region does double duty as the 3' untranslated region (UTR) of retroviral transcripts. The R region extends from the start of transcription to the polyadenylation site of viral transcripts and contains the polyadenylation signal. Lastly, the U5 region, with the R region, forms the 5' UTR of the viral transcript.

The most fruitful mammalian examples of LTR mutagenesis come from inbred strains of mice, in which 10 percent of characterized new mutations are due to insertional mutagenesis as a result of ERV or LTR insertions (Maksakova et al. 2006). This is often associated with intronic localization (Baust et al. 2002) in the same transcriptional orientation as the gene, which often leads to premature polyadenylation (Maksakova et al. 2006).

The mutagenic nature of LTR polyadenylation motifs has been used in gene

trapping experiments, in which constructs with a selectable marker and LTR polyadenylation signal were randomly inserted into mouse embryonic stem cells (Friedrich and Soriano 1991; von Melchner et al. 1992; Boeke and Stoye 1997). A total of 28 clones with expression of the selectable marker were used to create transgenic lines of mice and then bred to homozygosity. Eleven of the 28, or approximately 40% of genic insertions proved to be embryonic lethal, highlighting a high-frequency outcome of mutagenesis by polyadenylation: death of the involved cell.

Potentially more insidious is the role of ERV insertions as ectopic promoters. LTR promoters consist of a core promoter and regulatory elements composed of arrays of protein binding sites that act as enhancers and repressors to control expression of viral genes (Temin 1982). For example, transcription of the Moloney Murine Leukemia Virus (MLV) is controlled by its LTR, shown in Figure 1.6. In general, the core MLV promoter consists of a TATA box and *cis*-acting CAAT enhancer (bound by C/EBP). The more distal enhancer region contains a tandemly duplicated group of transcription factor binding sites. While methylation likely silences many TE insertions (Lavie et al. 2005; Meunier et al. 2005), MLV is silenced by repressor binding sites found upstream of its LTR enhancer. A more complete discussion of transcriptional control by the LTRs of MLV and other exogenous viruses is found in Rabson and Graves (1997).

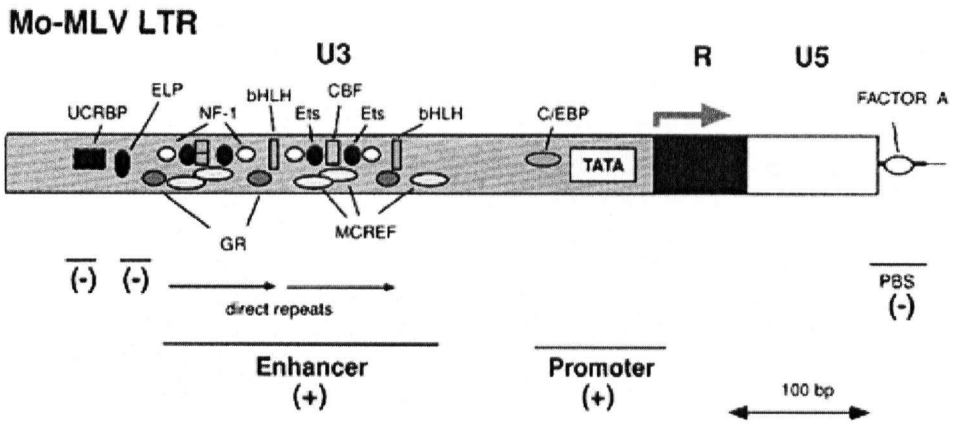

Figure 1.6 Transcription control elements of the MLV LTR. Taken from Rabson and Graves (1997).

Long experience with mutational mechanisms in mice, particularly due to spontaneous ERV insertions, has resulted in a large number of these mutations being characterized. The binding sites provided by *de novo* ERV insertions have frequently been found to cause oncogene activation in mice, functioning either as enhancers or promoters. Relevant mechanisms and frequency have been reviewed by Rosenberg and Jolicoeur (1997). Similarly for humans, an enhancer effect by exogenous lentiviral LTRs has been implicated in two cases of secondary leukemia after gene therapy treatment for X-SCID in 11 individuals (Hacein-Bey-Abina et al. 2003). However, this small number of trials leaves the overall expected frequency of adverse events in a gene therapy context in doubt. In another trial, the hematopoietic systems of 42 immunoablated rhesus monkeys were repopulated with cells having therapeutic insertions of MLV and simian immunodeficiency virus (SIV) vectors (Kiem et al. 2004; Dunbar 2005). Stable, polyclonal, virally marked hematopoiesis was observed, with no secondary leukemias over 6 months to 6 years.

In addition to their sense-oriented promoter, enhancer, and polyadenylation

effects, there is some evidence that LTRs can be damaging in the antisense direction, upstream of genes or within genes. In one documented case, loss of epigenetic silencing of an antisense intracisternal A particle (IAP) ERV upstream of the agouti gene in mice resulted in ectopic expression of the agouti gene from a cryptic ERV promoter (Morgan et al. 1999). Within introns, aberrant splicing of antisense ERV sequences from cryptic splice signals may be a prominent mutagenic mechanism of ERVs, as also highlighted for IAPs in mice in a recent review (Maksakova et al. 2006). The observation that LTRs are less likely to be found in genes in the antisense orientation than expected by random chance is in agreement with this view (see Chapter 4).

A couple of *de novo* mutagenic roles have been elucidated even for very old TEs. As discussed more fully in terms of mechanism in section 1.4.3, Alus have been shown to engage in Alu-Alu recombination (Deininger and Batzer 1999). The most recent literature describes many examples of this type of event, including mediation of MLL-CBP gene fusion in leukemia (Zhang et al. 2004), deletion in BRCA1 and BRCA2 genes in cancers (Tournier et al. 2004; Ward et al. 2005), and Factor VIII deletions in hemophilia (Nakaya et al. 2004).

Finally, for some mutagenic insertions such as those of the hominid SVA retroelement family, the precise mechanism of mutagenesis has not been determined. To date, four cases of *de novo* mutations due to SVA retroelements have been described. Interestingly, all were within the borders of known genes and are in the same transcriptional orientation as the enclosing gene (Chen et al. 2005). One hint as to a possible mode of mutagenesis comes from the fact that SVAs retain the HERV-K10 endogenous retroviral-derived hormone response element and the core enhancer sequences and a polyadenylation signal derived from the same LTR, suggesting that

SVAs can act in a similar way to LTR elements (Ono et al. 1987; Wang et al. 2005). This topic is addressed in Chapter 4.

1.2.4 Relationships between initial and long-term distributions of TEs

It has been observed that for some elements, for example Alus, the final genomic distribution does not correspond to that of newly-inserted active elements of the same type (International Human Genome Sequencing Consortium 2001) (See also Chapter 2). In the case of Alus, their consensus TTAAAA insertion site is expected to be found more often in regions of high AT sequence content and, indeed, recently inserted Alus tend to be located in AT-rich regions, similar to the pattern seen for L1 elements (Smit 1999; International Human Genome Sequencing Consortium 2001). However, the vast majority of Alus are found in GC-rich regions. Several theories have attempted to explain this disparity.

Pavlicek et al (2001) noted that, in spite of using the same target site, the observed long term distributions of Alus and L1s were biased to different GC content isochores. In their data set, the presumed slightly older AluYb8 subfamily was more skewed to higher GC isochores than the slightly younger AluYa5 subfamily. Therefore, they proposed that the GC-richness of Alu elements makes them more stable in regions where the surrounding GC content is similar to that of the Alu consensus, and that excision by an unknown mechanism happens more frequently in high-AT isochores (Pavlicek et al. 2001).

Others have hypothesized that Alu elements are selectively retained in GC-rich regions because of a functional benefit to genes, which reside in high-GC regions. For example, Britten cites individual Alu elements from earlier literature conferring functional binding sites for various proteins to nearby genes (Britten 1997). Schmid

hypothesized that Alus might be involved in chromatin remodeling and signaling of double-stranded RNA-dependent protein kinase in response to cell stress, conferring a selective advantage for having Alus in transcribed regions (Schmid 1998). While these studies are tantalizing and suggest selective advantage of individual Alus, no study since that time has demonstrated an overall selective advantage for accumulation of Alus in GC-rich regions. Instead, the developmentally critical HoxD gene cluster is almost devoid of retroelements (International Human Genome Sequencing Consortium 2001), suggesting that some classes of genes may need to exclude such sequences from their environment to ensure proper function or regulation.

A more neutralist third hypothesis proposes that Alus are maintained in GC-rich regions because deletions are unlikely to be precise and deletions of these elements in gene-rich regions would likely also involve important adjacent regulatory sequence (Brookfield 2001). While more satisfying in that it attempts to explain the Alu distribution without invoking a functional role, rates of deletion in gene-rich vs. gene-poor regions in primates have not been conclusively assessed. Furthermore, given that only five percent of mammalian genomes is estimated to be under purifying selection (Mouse Genome Sequencing Consortium 2002) and therefore vulnerable, it is unlikely that this explanation alone can account for the massive accumulation of Alus in gene-rich regions.

A fourth mechanism theorized to contribute to the relative paucity of Alus in gene-poor, high-AT regions is Alu-Alu recombination. Closely-spaced Alu pairs are found only occasionally in the human genome (Lobachev et al. 2000; Stenger et al. 2001), possibly because of clearance of these elements through the mechanism of inverted repeat (IR)-mediated recombination (Leach 1994). Later studies have linked this

phenomenon to sister chromatid exchange (Nag et al. 2005). In addition, unequal homologous recombination results in loss of the sequence separating closely spaced directly repeated Alus (Stenger et al. 2001). Even Alus up to 20% divergent have been found to recombine efficiently (Lobachev et al. 2000). That this process is ongoing is evidenced in its observed involvement in human disease, discussed above. This mechanism provides an additional explanation for the enrichment of Alus in GC rich regions without requiring a functional role.

While the above theories address likelihood of fixation of inserted elements, some interesting recent work motivated by the use of retroviral vectors in gene therapy has focused instead on mechanisms involved at the time of insertion. For example, work using unselected *in vitro* integrations of HIV and HIV-based vectors consistently demonstrated a propensity to integrate in transcribed regions, with higher frequency in highly transcribed genes (Schroder et al. 2002; Mitchell et al. 2004; Barr et al. 2005). MLV, on the other hand, has a marked preference to integrate into the start sites of more active genes (Wu et al. 2003), and avian leukosis virus (ALV) demonstrated a weaker, though highly significant, preference for transcribed regions (Mitchell et al. 2004; Barr et al. 2005).

The question arises why the different viruses target active genes, and then with varying locations within genes. Early experiments in Swiss mouse cells found that the majority of MLV insertion sites were sensitive to a micrococcal nuclease and DNAse I, suggesting that viral integrations chiefly target accessible DNA, which is found in regions of open chromatin (Panet and Cedar 1977). While access to targets may well partially explain integration targeting, it fails to explain the additional marked preference of MLV for 5' regions of genes. Instead, a tethering model whereby interactions between

the viral preintegration particle and host factors bound to DNA facilitate targeting of integrations to specific regions within open chromatin has been proposed (Bushman 2003; Bushman et al. 2005). In such a model, MLV might be tethered by transcription factors, while HIV might interact with proteins that bind within transcription units. One candidate tethering factor has been identified for HIV (Ciuffi et al. 2005). Additional evidence for this phenomenon comes from the recent discovery of symmetrical base pair profiles around sites of insertion of HIV-1, ALV, and MLV (Holman and Coffin 2005).

It should be noted, in any case, that studies of *in vitro* insertions represent insertions before they have a chance to be tested during organismal development. Only after escaping purifying selection during development and the lifetime of an organism do germline mutant alleles have a chance to segregate in the population and attain a small chance of spreading to fixation.

1.3 Neutral or beneficial roles for TEs in the transcriptome

1.3.1 Regulatory motifs donated by TEs

In addition to a potent role as mobile genomic mutagens, there has been a growing appreciation for the positive roles TEs can fulfill (Kidwell and Lisch 1997). One early role elucidated for TEs was as the parotid-specific enhancer of the human amylase gene (Ting et al. 1992). In that investigation, a 700 bp fragment approximately 300 bp upstream of the transcription start and derived entirely from a human endogenous retrovirus E (HERV-E) LTR was sufficient to confer salivary expression on a reporter gene in transgenic mice. Curiously, an LTR element has also been found to act in the opposite role as a strong upstream repressor of annexin A5 transcription (Carcedo et al.

2001).

Many more examples of LTRs acting as alternative promoters of cellular genes exist. A growing body of literature describes genes in many roles under the control of ERV LTRs (For reviews, see Leib-Mosch et al. 2005; Medstrand et al. 2005)(see Chapter 3). An early investigation identified an ERV9 LTR that acts as a tissue-specific promoter of the zinc finger gene ZNF80 and drives expression in several hematopoietic cell lineages (Di Cristofano et al. 1995). An ERV9 element in the human globin locus control region has also been shown to participate in expression of downstream genes, likely by participation in chromatin remodeling (Yu et al. 2005). Several examples of HERV-E LTRs acting as alternative promoters, including involvement in MID1, apolipoprotein C1 and endothelin B receptor expression, have been identified as well (Medstrand et al. 2001; Landry et al. 2002). More recently, an LTR of the ERV-L superfamily has been shown to form the dominant promoter of the human beta1,3-galactosyltransferase 5 gene in humans (Dunn et al. 2003). This promoter was found to be more conserved than expected by chance, suggesting purifying selection preserving these retroviral sequences (Dunn et al. 2005). This example is instructive, as the orthologous mouse gene is also expressed in colon despite the lack of an LTR promoter, leading to the conclusion that the insertion of this ERV has been co-opted because it provided useful motifs in the approximately correct position rather than because it is essential. This example may also suggest external control by a separate colon-specific enhancer.

Finally, a genome-wide bioinformatic survey showed that most TE types have some basal capacity to form 5' transcriptional start sites, presumably by contributing pre-existing promoter-related sequences or by evolving them *in situ* (Dunn et al. 2005, See also Chapter 3). However, LTR elements upstream of genes in the same transcriptional

orientation as the gene form the transcriptional start sites of genes far more often than other TE types, relative to their local genomic density (Figure 1.7). LTRs that are antisense to the gene transcriptional direction form the 5' ends of genes next most often, suggesting a significant, though secondary, role for transcription factor binding sites donated by the LTR as a nearby enhancer or downstream promoter.

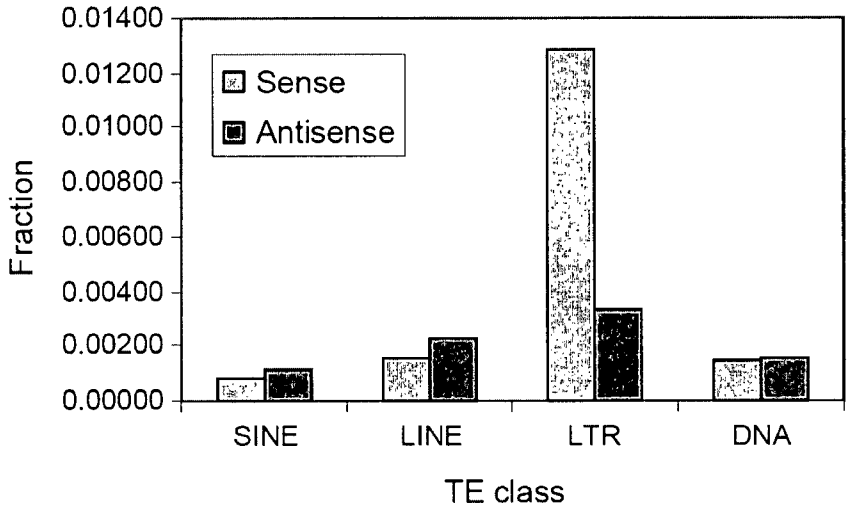

Figure 1.7 Orientation of TEs that contain human gene transcriptional start sites. TEs within the genomic region 5kb upstream and 5kb downstream of human RefSeq gene transcriptional start sites were grouped by class and orientation with respect to the direction of gene transcription. The fractions of sense and antisense TEs that contain transcriptional start sites are depicted for each class by gray and black bars, respectively. Taken from Dunn et al (2005).

Most of the above-discussed roles for TEs involve LTRs in control of transcription, but this balance likely reflects some bias in the current research interests of the groups involved. TEs have also been shown to perform other neutral or beneficial roles. As mentioned above, L1s have been shown to have antisense promoter activity. This has been shown to involve several cellular transcripts (Nigumann et al. 2002). Furthermore, LTRs have been shown to donate polyadenylation signals to cellular genes. In one example, ERV-H LTRs provide polyadenylation signals to the HHLA2 and 3 genes (Mager et al. 1999). The absence of these LTRs in baboon was associated with use

of alternate polyadenylation signals. Another example involves polyadenylation of receptor tyrosine kinase FLT4 and other transcripts by human ERV-K LTRs (Baust et al. 2000). Lastly, an array of several retroelements has been shown to exert a modulatory effect on transcription and translation efficiency of the zinc finger gene ZNF177 (Landry et al. 2001).

The above considerations, taken together, are strong evidence of a neutral or positive role for some TEs, particularly LTRs, and motivated a bioinformatic survey of the contributions of TEs to the transcripts of protein-coding genes, as reported in Chapter 3.

1.3.2 Protein domains donated by TEs

In addition to roles in transcription control, TEs have also been demonstrated to have contributed to mammalian proteomes. Perhaps the best known example is that of the recombination activating (RAG) genes involved in V(D)J recombination, which are derived from DNA transposons (Reviewed in Brandt and Roth 2004; Schatz 2004). The RAG genes are found in jawed vertebrates and mediate precise, site-specific combinatorial joining of gene segments making up antigen receptors in T and B cells. In addition, ERV sequences have also been found to perform useful functions. In two known and completely unrelated cases, the fusogenic properties of coding-competent retroviral *env* genes from ERV-W and ERV-FRD have been co-opted as two different syncytin genes, both with a role in placenta formation (Mi et al. 2000; Renard et al. 2005). These and other potential roles of retroviral-derived proteins in health and disease are discussed by Bannert and Kurth (2004).

1.4 Stability of repetitive sequence

1.4.1 Assumptions of stability and use of retrotransposed sequence as population markers

Insertions of retroelements have undergone intense scrutiny as a destabilizing influence in mammalian genomes. In particular, insertions of recent families of SINEs and LINEs in primates have been shown to be associated with deletions and other rearrangements upon insertion (Gilbert et al. 2002; Symer et al. 2002; Callinan et al. 2005). Furthermore, as mentioned above, high copy number genomic elements may cause disease by homologous recombination with each other (Deininger and Batzer 1999).

New insertions of retroelements that are at worst mildly deleterious segregate as Mendelian alleles in the population and have a small chance of reaching fixation. Once fixed, elements are generally assumed to be stable, except when involved in infrequent rearrangements. As a result, assessments of relative activity of retroelements have assumed that presence of an element at a given site is an unambiguous indication of insertion (Liu et al. 2003).

A corollary of this assumed unidirectionality of retroelement insertions with no known mechanism for precise deletion is that genomic retroelement loci may be assumed identical by descent rather than identical by state. This property makes active families of Alu and L1 elements ideal for use in studying relationships between human populations (Perna et al. 1992; Sheen et al. 2000; Carroll et al. 2001; Roy-Engel et al. 2001). For example, 20 to 30 percent of insertion loci of some active Alu families are polymorphic for presence or absence of the element between human populations (Carroll et al. 2001; Roy-Engel et al. 2001). Similar analyses have also been used in the resolution of primate phylogeny. For example, one study used evidence from the AluYe5 family to show

strong support for a sister grouping of chimpanzees and humans (Salem et al. 2003b).

1.4.2 . A role for DNA double-strand break (DSB) repair in creating de novo insertions, deletions, and tandem duplications

While insertion of pol-II transcribed retroelements has the potential to explain many genomic regulatory changes, many more sequence differences between genomes may be explained by the paradigm of DNA DSB repair. The association of single gene mutations in this pathway with known diseases (for review, see Thompson and Schild 2002) and the feasibility of studying DSB repair in cell culture systems has made it possible to investigate many of the most important proteins and their mechanisms.

Double strand breaks in DNA, induced by gamma rays or free radical insult, are repaired by one of two main mechanisms. One is slower and involves similarity between the sequences to be joined, while the other is faster and homology-independent. The interactions between these pathways have been characterized as competitive in eukaryotic cells (Prudden et al. 2003), and resolution of a DSB sometimes involves both mechanisms. Before repair begins, however, broken and damaged DNA ends are trimmed by an endonuclease complex (Helleday 2003).

The fast, homology-independent mechanism of DSB repair, commonly termed non-homologous end joining (NHEJ), initiates with the binding of a Ku heterodimer in a ring around each broken DNA end, stabilizing it (Walker et al. 2001). The DNA-bound Ku heterodimers are then bound together and the DNA ends trimmed (Helleday 2003). Finally, ligases rejoin the two ends.

The slower, homology-driven mechanism of DSB repair is believed to begin with a 5' to 3' peeling back or resectioning of the DNA by an unknown exonuclease, exposing a 3' single-stranded DNA (ssDNA) end. Frequently, exposure of several hundred base

pair repeats such as Alus in two 3' ssDNA ends in mammalian cells results in repair by a mechanism known as single-strand annealing (SSA) (Elliott et al. 2005). SSA is an error-prone mechanism which results in loss of one of the flanking repeats and the intervening sequence. In the absence of long similar sequences, shorter sequences are also used, but less frequently, and the precise mechanisms are uncertain (Sankaranarayanan and Wassom 2005).

As an alternative to SSA, homologous recombination (HR), which involves BRCA1, BRCA2, and RAD51 proteins, may occur. In this case, RAD51/ssDNA filaments invade nearby DNA duplexes and pair with homologous DNA (Helleday 2003). The sister chromatid, available during the S and G2 phases of the cell cycle, is the homologous DNA duplex most often used as a template for repair (Johnson and Jasin 2000), rather than the homologous chromosome, use of which could result in loss of heterozygosity (Moynahan and Jasin 1997). Upon strand invasion, a Holliday junction is formed and DNA synthesis occurs, often continuing beyond the original site of DNA breakage. In any case, final resolution of the break may occur by NHEJ, secondary SSA, or reinvasion of the original duplex by the nascent DNA strand at the homologous location beyond the site of the original break. Resolution by secondary NHEJ or SSA has the potential of causing deletions or tandem DNA insertions, while secondary SSA or reinvasion of the original duplex may result in error-free repair. Finally, aberrant template choice, for example in the case of DNA breakage at Alu elements or other ubiquitous motifs, may result in ectopic sequence duplication (for review, see Helleday 2003).

Chapter 5 discusses observations of the absolute rate with which retroelement insertions actually revert or undergo precise deletion, presumably by a DNA DSB

homologous repair mechanism. The relative paucity of these events shows that identity by descent is a relatively safe assumption for retroelement loci compared across species. It also addresses the role of homologous direct repeats in sequence removal from primate genomes.

1.5 Repeat finding methods

While the goal of this thesis was to understand the interaction of repetitive elements, it depended heavily on repeat annotation provided by RepeatMasker (A.F. A. Smit and P. Green, unpublished) as tracks in the UCSC Genome Browser. RepeatMasker makes use of the Repbase libraries (Jurka 2000) to iteratively mask genomic sequence, excising fully embedded repeats and allowing more confident identification of the targeted repeat (A.F. A. Smit and P. Green, unpublished).

MaskerAid (Bedell et al. 2000) is a suite of perl scripts that adapts WU-BLAST (W. Gish, unpublished) for use with RepeatMasker, functionally replacing the cross_match aligner. MaskerAid increases the speed of RepeatMasker analysis dramatically, by 40 fold at most RepeatMasker sensitivity settings, while identifying repeats with nearly identical sensitivity and specificity (Bedell et al. 2000). This reflects the fact that both methods rely on pairwise sequence alignment of genomic sequence with consensus sequences, typically from Repbase Update (Jurka 2000), to identify repetitive regions in genomes.

RECON, by contrast, uses a tunable, heuristic analysis of results of a self-BLAST to perform *de novo* definition of repeat boundaries and thus repeat families (Bao and Eddy 2002). The algorithm is tunable on at least two levels, that of the sensitivity of the BLAST search it is based on and the 'willingness' of the heuristic analysis to split elements up (Bao and Eddy 2002; Holmes 2002). This and related methods of

constructing repeat family consensus elements can make a valuable contribution to analysis of newly sequenced genomes.

RepeatScout represents a further development on the RECON concept (Price et al. 2005). RepeatScout uses overrepresented short sequences as seeds for local pairwise alignments, which then are reconstructed into longer repeat consensus sequence. In testing, this *de novo* repeat finding method identified a more complete set of rodent repeat families than had been included in Repbase Update (Jurka 2000). However, repeat identification by masking with RepeatScout-generated human libraries found less repeats than masking with Repbase libraries. This observation was attributed to the fact that the human genome has been better studied, allowing for many years of curation of human repeat families.

In summary, all these methods are limited in that they rely on pairwise alignment methods to find repeats. Repeats that undergo many short insertions or deletions, despite being recognizable by eye, are unduly penalized for being broken up. It seems clear that a better model of sequence evolution is required before repeat finding can reach optimal sensitivity. In the interim, repeats found by RepeatMasker and Repbase libraries have provided an opportunity to assess the nature of the effects of transposable elements on their host genome.

1.6 Thesis objectives and chapter summaries

Repetitive sequence is ubiquitous in mammalian genomes and has an array of consequences for the organism, both positive and negative. In some cases, positive roles for repetitive sequence are related to their mutagenic role. For example, binding sites donated by ERV LTRs can function as transcriptional enhancers, repressors, and polyadenylation signals. While these functions can interfere with the function of nearby

genes, in many cases they can also provide regulatory diversity. While a few cases of neutral or positive roles have been identified for LTRs, there are likely many more. Nonadjacent repeated sequences, on the other hand, can have a different role. Whether in the form of high-copy repeats or short random nonadjacent segments of identity, these sequences can function as substrates for DNA repair. While the potential for deletion of tumor suppressors and other genes is obvious, nonadjacent repeats may also have a positive function in genome size attenuation.

The main goal of this thesis work was to use automated sequence analysis and global statistical analyses of mammalian, primarily human, genomic repeat distributions to gain understanding of the roles of repetitive sequence in defining organisms at the genetic and hopefully also phenotypic levels. As the project progressed, more and more information became available that increased our ability to ask fundamental questions about these roles. Time was spent at the outset of the project developing methods and algorithms; however, as more data became available, the data formats and repositories evolved in response. The availability of a draft of the human genome sequence provided the positive stimulus for the development of the various genome browsers as well as long term choices being made on data formats. The information infrastructure and tools made available with the data have considerably aided in this analysis. Subsequent availability of a draft of the mouse genome sequence enabled us to use comparative approaches to assess TE involvement in the human and mouse transcriptomes. At the same time, availability of raw sequencing traces from multiple human sources enabled others to assess the levels of repeat polymorphism present in the human. Using these published data sets, we were able investigate detrimental impacts of retroelement families active in humans and compare them to the projected detrimental impacts of older, inactive

families. Finally, availability of sequence traces from Rhesus monkey allowed us to assess precise deletion of retroelements in the human and chimpanzee lineages.

Chapter 2 describes our initial analyses of the draft sequence of the human genome. The analyses performed were essentially collations of repetitive sequence and assessment of their location with respect to genomic features like sequence composition and genic positions. This analysis showed that TEs of different ages localize to different regions of the human genome and that LTR elements demonstrate orientation bias up to 5kb upstream and downstream of transcribed regions.

Chapter 3 describes our analysis of TEs in transcripts. Mapping of human and mouse protein coding transcripts and TEs were compared to find transcripts that contained TE sequence. Databases of gene classification information were then developed and classification of human and mouse genes performed to determine which classes of genes had TEs as part of their transcripts more often. The same classes of genes in both human and mouse, such as those with more organism-specific functions, tended to be permissive for the presence of TE sequence in transcripts. Highly conserved genes and genes with important housekeeping or developmental functions were less permissive.

Chapter 4 reviews our work on TE directional biases in human and mouse transcribed regions. Different families of LTR elements showed widely varying retention patterns within transcribed regions. SVA retroelements, which retain a partial LTR in their consensus, also demonstrated an LTR-like pattern, suggesting that TEs transcribed by RNA polymerase II (pol II) exert varying effects upon insertion, depending on the sequence of the element.

Chapter 5 is concerned with analysis of nonadjacent repeated sequence and its

role in recombination and deletion. Retroelement insertions, normally assumed to be stable with no known mechanism for precise deletion, are shown to be deleted precisely in some cases, most likely due to a mechanism involving DNA double strand break repair and the flanking short tracts of identical sequence. Analysis of random deletions showed that a large fraction of these events are also mediated by short nonadjacent tracts of identical sequence.

Chapter 2: Retroelement distributions in the human genome

A version of this chapter has been published:
Medstrand, P.*, L.N. van de Lagemaat*, and D.L. Mager. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12:** 1483-1495.

* these authors contributed equally to this work

I performed all data analysis and wrote sections of the paper.
P. M. and D. L. M. performed postprocessing of the data in Figure 2.2 and wrote a significant share of the paper.

2.1 Introduction

Since Barbara McClintock discovered transposable elements (TEs) in maize (McClintock 1956), it has become well established that such elements are universal. While there are examples of both loss and increase of host fitness due to the activity of transposable elements, their population dynamics are far from being understood, and the forces underlying their genomic distributions and maintenance in populations are a matter of debate (Biemont et al. 1997; Charlesworth et al. 1997). The prevailing view is that TEs are essentially selfish DNA parasites with little functional relevance for their hosts (Doolittle and Sapienza 1980; Orgel and Crick 1980; Yoder et al. 1997). According to this hypothesis, the interaction of TEs with the host is primarily neutral or detrimental and their abundance is a direct result of the ability to replicate autonomously. It is generally accepted that selection is the major mechanism controlling the spread and distribution of TEs in natural populations of model organisms (Charlesworth and Langley 1991). While the exact mechanisms through which selection acts are controversial, the processes controlling transposition involve selection against the deleterious effects of TE insertions close to genes (Charlesworth and Charlesworth 1983; Kaplan and Brookfield 1983) and selection against rearrangements caused by unequal recombination (ectopic exchange) in meiosis (Langley et al. 1988). More recently, the ubiquitous nature of TEs has gained increasing attention and it is now becoming accepted that TEs give rise to selectively advantageous adaptive variability which contributes to evolution of their hosts (McDonald 1995; Brosius 1999). However, the mechanisms responsible for maintenance, dispersion, fixation and genomic clearance of TEs remain largely unknown.

While most work on TEs has focused on model organisms, sequencing of the human genome has revealed that nearly half of our DNA is derived from ancient TEs,

mainly retroelements (Smit 1999; International Human Genome Sequencing Consortium 2001). The wealth of human genomic information now allows comprehensive explorations into the evolutionary history and genomic distribution patterns of transposable elements with a view to increasing our understanding of the forces that have shaped our genome and its mobile inhabitants. The retroelements present in the human genome are divided in two major types, the non-LTR and LTR retroelements (International Human Genome Sequencing Consortium 2001). The non-LTR retroelements are represented by the autonomous L1 and L2 elements (LINE repeats) and the non-autonomous Alu and MIR (SINE) repeats and have been extensively studied (Smit 1999; International Human Genome Sequencing Consortium 2001; Ostertag and Kazazian 2001; Batzer and Deininger 2002) but appreciation of the heterogeneous collection of LTR retroelements is more limited. These sequences make up 8% of the human genome (International Human Genome Sequencing Consortium 2001) and include defective endogenous retroviruses (ERVs) (Wilkinson et al. 1994; Sverdlov 2000; Tristem 2000), related solitary LTRs, and sequences with LTR-like features for which no homologous proviral structure has been found. Over 200 families of LTR retroelements are defined in Repbase (Jurka 2000) but they can be grouped into six broad superfamilies (see Methods). While some of the LTR retroelement families, particularly members of class I and II ERVs, presumably entered the primate germline as infectious retroviruses and then amplified via retrotransposition (Wilkinson et al. 1994; Sverdlov 2000; Tristem 2000), other LTR families likely represent ancient retrotransposons that amplified at different stages during mammalian evolution (Smit 1993).

The vast majority of human retroelements were actively transposing at various stages prior to and during the radiation of mammals and are now deeply fixed in the

primate lineage. Essentially only the youngest subtypes of Alu (Batzer and Deininger 2002) and L1 elements (Ostertag and Kazazian 2001) are still actively retrotransposing in humans. Some ERVs belonging to the Class II HERV-K family are human-specific (Medstrand and Mager 1998) and a few are polymorphic (Turner et al. 2001) but no current activity of human ERVs has been documented. Here we show that genomic densities of human retroelements vary with distance from genes and that their distributions with respect to surrounding GC content also shift as a function of their age.

2.2 Methods

2.2.1 Description of retroelements

Human retroelements are classified into two major classes: non-LTR and LTR retroelements. The former contain the LINEs, represented by the L1 and L2 elements, whereas the Alu and MIR elements belong to SINEs. For this analysis LTR retroelements were divided into the following six groups (Smit 1999; Jurka 2000; International Human Genome Sequencing Consortium 2001; Mager and Medstrand 2003) : class I ERVs, which are similar to type C or gamma retroviruses such as murine leukemia virus; class II ERVs, which are similar to type B or beta retroviruses like mouse mammary tumor virus; class III ERVs (also called ERV-L), which have limited similarity to spuma retroviruses; MER4 elements, which are non-autonomous class I related ERVs; and MST (named for a common restriction enzyme site MstII) and MLT (mammalian LTR transposon) elements, which are both part of the large non-autonomous Mammalian apparent LTR Retrotransposon (MaLR) superfamily. Solitary LTRs outnumber LTR elements with internal sequences by approximately 10 fold.

2.2.2 *Data sources*

Genomic sequence and annotated gene data for all figures were derived from the August 6, 2001 draft human genome assembly at http://genome.ucsc.edu. Retroelement locations derived from RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html), GC-content calculated in non-overlapping windows of 20-kb, sequence gap data, and known gene data from the Reference Sequence database were all downloaded from this site. After compilation, data points were included in graphs only if supported by more than 100 retroelements. Element count was calculated to reflect as nearly as possible the number of individual integrations of the element. That is, nearby repeat segments (within 20 kb of each other) having the same family name and RepeatMasker alignment parameters (alignment score, substitution and gap levels) were combined and treated as a single element. Subfamily assignments and divergence values were taken directly from RepeatMasker output files. Internal sequences of LTR elements were excluded from the analysis. Data was further conditionally discarded in figures where retroelement divergence is used as a measure of age. In some cases where element length was short (below 150 bp), it was noted that RepeatMasker assigned an artificially low divergence value due to the alignment method used in finding repeats. This was a particular problem for the old MIR and L2 sequences. An attempt was therefore made to ensure that relative divergence indeed represented age by plotting element length versus assigned divergence values. Since repeats in general grow shorter as they age (e.g. see Figure 2.5), retroelement divergence cohorts were considered anomalous and discarded if they did not follow this trend.

2.2.3 *Density analysis*

The retroelement data were compiled by repeat superfamily, divergence from consensus,

and surrounding genomic GC content. The density function in Figures 2.1, 2.4 and 2.6 was calculated as follows: the fraction of the retroelement base pairs in a given GC bin divided by the fraction of the genome in that GC bin. Thus, it affords a measure of preference of a particular age class for different GC contents. When an age class of an element had a significant presence in only some of the GC bins, the effective genome size for that age class was calculated from the sizes of only those GC bins. Thus for the Figure 2.6 genomic data, the 'whole genome' is that fraction of the genome with GC content less than 46%. In Figure 2.2, the bin considered was an individual chromosome. With these considerations in mind, the calculations of density are identical.

For Figure 2.2 (retroelement density versus GC content on each chromosome), correlation coefficients (r) and level of significance (p-values) were calculated for each data set. The graphs of chromosomal retroelement density as a function of gene density are not shown, but are almost identical because of the highly significant correlation between GC content and gene density (International Human Genome Sequencing Consortium 2001).

For Figure 2.3, a script divided the chromosomes into eleven segment types or bins: within the transcript start and end positions of known (annotated) genes and 0-5, 5-10, 10-20, 20-30, and >30 kb upstream and downstream of genes. The majority of the genome was located either within genes (22% of the total) or at distances greater than 30 kb from genes (63% of the total). In each segment, the script determined the base pair contribution of each retroelement type and noted the orientation of the element with respect to the nearest gene. The GC content of each segment was calculated and then the density data from Figure 2.1 was used to predict the base pair contribution by each retroelement type in the segment. Predictions done within genes or at distances >30 kb

from genes were compiled from predictions made from 10 kb sub-segments. Half of the predicted retroelement base pairs were assumed to be in the sense orientation and half in antisense. Finally, the observed base pairs in each bin were divided by the cumulative predicted base pairs for each retroelement type.

P values shown in Tables 2.1, 2.2, and 2.3 and variability of the data in Figures 2.3, 2.4 and 2.6 were calculated as follows. The sequence segments comprising the whole genome were divided up into four 'subgenomes' of equal composition. The retroelement distributions were calculated in each subgenome, and the means and standard deviations of retroelement distributions were calculated. After appropriate normalization, the significance (p-value) of the difference between different retroelement distributions was tested by the one-tailed unpaired t-test.

2.3 Results and Discussion

2.3.1 Distributions of retroelements in different GC domains

To begin our analysis, we measured the density of various retroelements with respect to GC content in 20-kb windows across the human genome sequence. As reported previously (Smit 1999; International Human Genome Sequencing Consortium 2001), L1 elements are predominantly found in the AT-rich regions, L2 elements are more uniformly distributed whereas Alu and MIR repeats reside in the higher GC fractions of the genome (Figure 2.1A) in comparison to the entire genome which has an average GC content of 40% (International Human Genome Sequencing Consortium 2001). For the different LTR superfamilies, an uneven distribution in GC occupancy is also observed. The relatively young Class I ERVs and the non-autonomous MER4 sequences, which

may have been propagated by Class I elements, have very similar broad distributions that peak in regions of medium GC. Class II ERVs, which include the youngest known HERVs (Medstrand and Mager 1998; Turner et al. 2001), have a distribution more skewed toward higher GC regions (Figure 2.1B). Distributions of the older Class III ERVs and their distantly related MLT and MST elements are generally biased toward low GC regions, except for MLT elements which are spread more uniformly (Figure 2.1C).

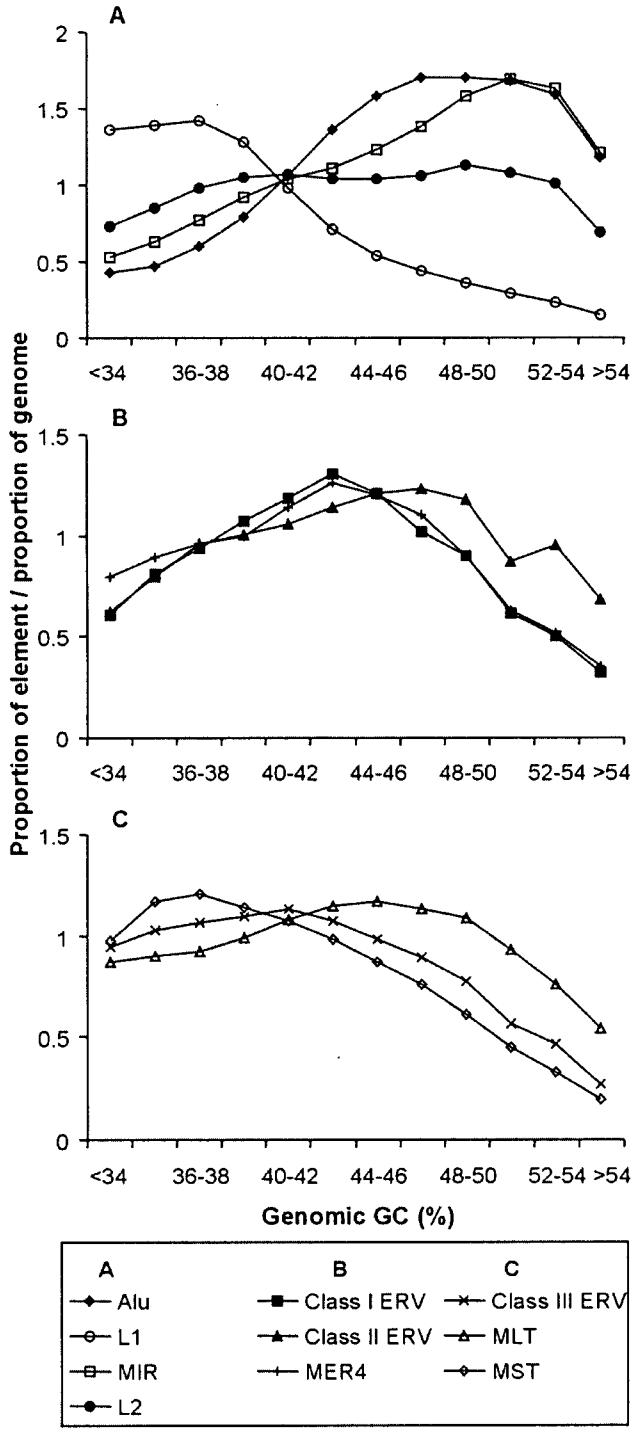

Figure 2.1 Density of retroelements in different GC fractions in the human genome, calculated over 20-kb windows across the genome sequence. Panels A to C show the density of various retroelement classes and those represented in each panel are indicated in the box below the graphs. The bins from left to right correspond to an increasing 2% GC fraction.

To determine if retroelement densities on each chromosome agree with overall densities shown in Figure 2.1, we plotted densities against estimated gene (data not shown) or average GC content of each chromosome (Figure 2.2). As expected, the two distribution profiles are almost identical because of the strong correlation between GC content and gene density (International Human Genome Sequencing Consortium 2001). The density of Alu elements increases as a strict function of increasing GC content and MIR elements also generally follow this trend (Figure 2.2A, C). In contrast, there is generally a negative or no correlation between the density of L1, L2 or LTR elements and gene density or GC content (Figure 2.2). The Class II ERVs and the MLT elements show little, if any, bias for GC-poor chromosomes, while the L1, Class I, III and MST groups are overrepresented on these chromosomes. Class I-II elements are dramatically over represented on chromosome Y, as noted before (Kjellman et al. 1995; Smit 1999; International Human Genome Sequencing Consortium 2001), and also somewhat on 19. Abundance of the youngest ERVs on chromosome Y may be due to recombination isolation and absence of major recent rearrangements on much of this chromosome (Graves 1995; Lahn et al. 2001), and since chromosome 19 is much more gene-dense than the other chromosomes (International Human Genome Sequencing Consortium 2001), one possible explanation for the over-representation of the same ERVs on this autosome is that these elements had an initial integration preference for regions near genes or gene-related features such as CpG islands. We also noted an under representation on Y of the old L2, MIR and MLT retroelements which is consistent with major rearrangements and deletions of Y during mammalian evolution (Lahn et al. 2001). Similar trends are observed for MER4 distributions and their autonomous class I counterparts (over representation on Y and 19), and for the non-autonomous MaLR

(MLT and MST) elements and their apparent autonomous class III ERVs (over representation on 21). Alu, L1, MER4, class I and II ERV sequences represent the younger elements which have actively amplified during the last 40 MYR of primate evolution, whereas other element types were already inactivated for transposition by this time (International Human Genome Sequencing Consortium 2001). All younger retroelements except Alu sequences are over represented on Y. Even though some of the LTR superfamilies show a stronger negative correlation than others, the distribution profiles demonstrate that various retroelement families cluster preferentially in different genomic landscapes and are in agreement with the general trends observed in Figure 2.1.

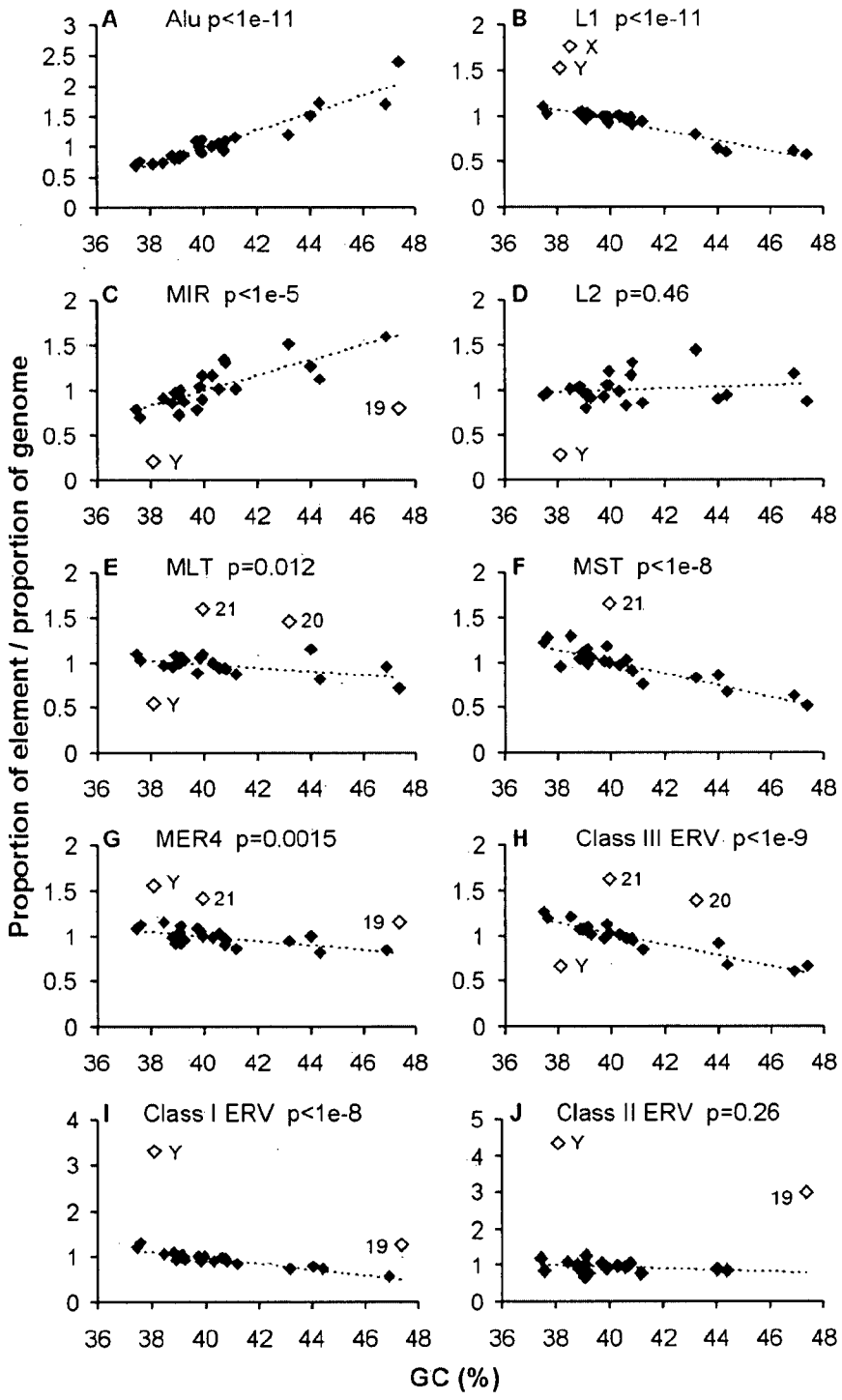

Figure 2.2 Density of retroelements as a function of average GC content of each human chromosome. The line connecting solid diamonds indicates the general correlation trend between retroelement and GC content of individual chromosomes. The level of significance (p-values) of the correlation for each data set is indicated. Open diamonds were excluded from the correlation analysis and indicate over or under representation of retroelement density on a particular chromosome. Chromosomes 20, 21 and 22 were excluded from the Class II graph (panel J) due to having less than 100 supporting elements.

2.3.2 Arrangements of retroelements with respect to genes

Given the results in Figures 2.1 and 2.2, we looked in more detail at the distribution of retroelements by locating all elements in the human genome relative to annotated genes. While it is reasonable to assume that locations with respect to genes affect retroelement dispersal and fixation patterns, the aim of this analysis was to obtain a measure of this effect. Our strategy was to determine how closely retroelement densities with respect to genes could be predicted based on the surrounding GC content. DNA regions located upstream of each gene's transcriptional start site and downstream of the polyadenylation site were divided into segments of various size fractions (see Methods) and the density of each retroelement class in either transcriptional orientation with respect to the gene was determined. Regions within the boundaries of a gene, including the introns, were assigned a single segment. The local GC content of each segment was also calculated and used to determine an expected retroelement density based on the whole genome distributions indicated in Figure 2.1 (see Methods) and the results shown in Figure 2.3. To obtain estimates of the variation associated with this type of analysis, we divided the genome into four 'subgenomes' as detailed in Methods and performed the analysis independently for each. The points in the graphs represent the mean and standard deviation derived from values obtained for each subgenome.

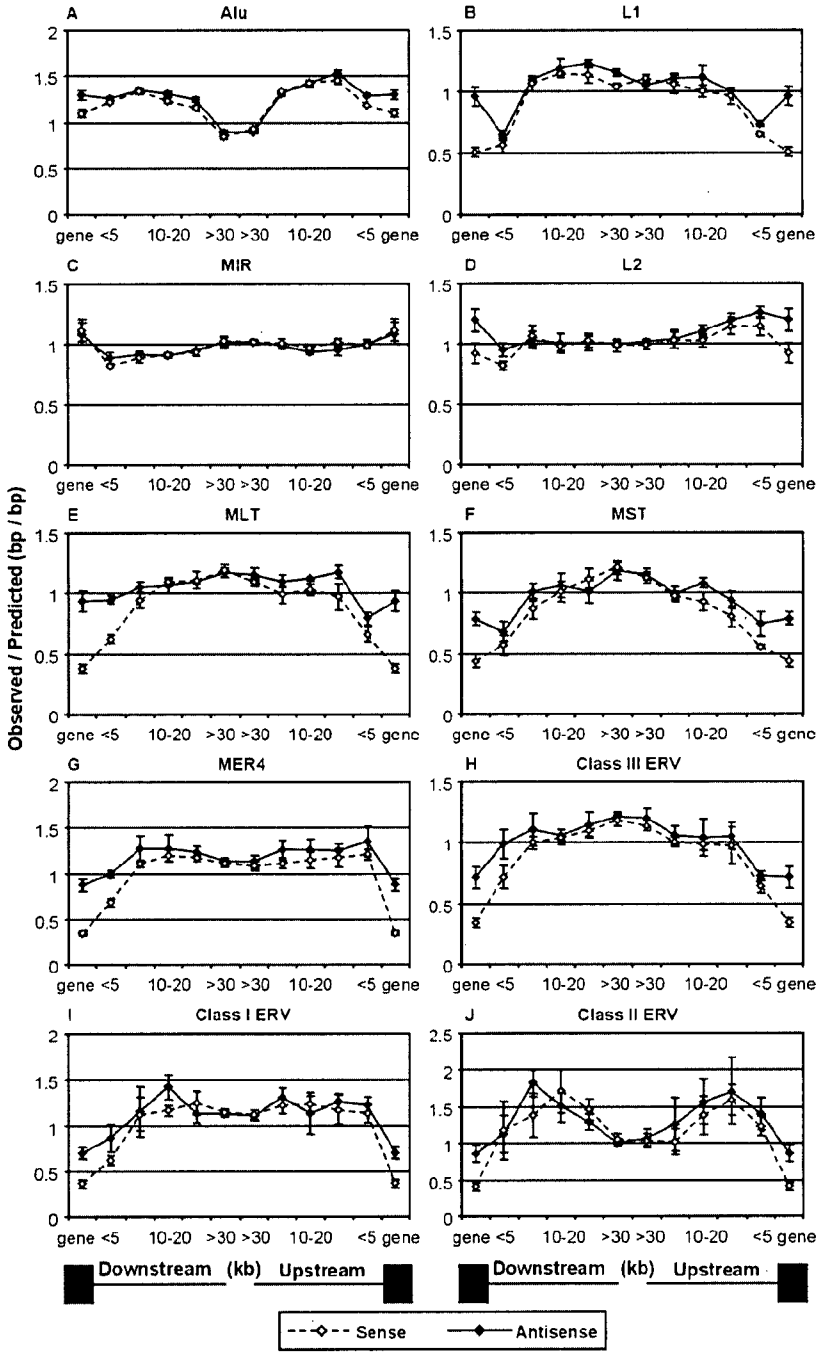

Figure 2.3 Ratios of observed to predicted retroelement densities with respect to genes in the human genome. The points above 'gene' and '<5' of each graph indicate the density in gene regions, and in the first 5 kb either 5' or 3' of genes. The other bins are 5-10, 10-20, 20-30, and >30 kb either upstream or downstream of genes. Open symbols and dashed lines indicate elements in the same or sense orientation with respect to the nearest gene and solid symbols and lines indicate elements in the reverse direction. Standard deviation error bars, which are too small to see in some cases, were determined as described in Methods. Solid boxes below the graphs represent gene regions and the lines indicate the distance bins of the intergenic regions. It should be noted that the vast majority of retroelements within genes are located in introns.

Dividing the genome based on proximity to genes revealed several intriguing patterns. First, densities of the relatively old MIR and L2 elements in intergenic regions generally conform to that predicted from the GC content of each region. That is, the ratio of observed to expected density is close to one (Figure 2.3C, D). Second, for the SINE (Alu and MIR) elements, densities within genes are close to that predicted or are overrepresented based on average GC content of gene regions (Figure 2.3A, C). In contrast, L1 elements and all six LTR classes, particularly those in the same transcriptional direction, are underrepresented within genes (Figure 2.3B, E-J). L1 sequences and the older MLT, MST and Class III elements are also underrepresented in the 0-5 kb regions both upstream and downstream of genes, while the younger class I and MER4 elements are underrepresented in the downstream region only. The higher tendency for LTR elements and L1s within genes to be oriented in the antisense direction has been noted previously (Smit 1999) and likely reflects lower fixation rates resulting from interference by retroelement regulatory motifs, such as polyadenylation signals, when genes and elements are located in the same transcriptional direction. However, this is the first study to demonstrate lower densities of LTR and L1 elements within genes relative to that predicted based on the surrounding GC content. In addition, the fact that an orientation bias for some elements extends to significant distances away from genes has not been reported previously. Moreover, our analysis indicates that the densities of most LTR elements and L1s are highest in regions furthest from genes. These patterns suggest that L1 and LTR elements are excluded from genes and nearby regions by selection. Interestingly, the density distribution of Alu elements with respect to genes is opposite to that observed for L1 and most LTR elements in that the density is lowest in

regions most distant from genes and they are overrepresented (as predicted by GC content) in regions within and near genes. It is also noteworthy that densities of the relatively young LTR class II elements peak in the region 5-20 kb 5' or 3' of genes and, indeed, are overrepresented in these areas compared to the expected densities based on regional GC content (Figure 2.3J). Such a pattern may reflect a preference for this class of elements to integrate near genes.

The statistical significance of these results is shown in Table 2.1 which lists the resulting p-values for three sets of comparisons. The top part of the table compares the sense versus antisense distributions and confirms the significance of the orientation biases discussed above. MIR elements are the only group to show no significant orientation bias. In contrast, an orientation bias extends up to 20 kb 5' of genes for MLT and MST elements. The bottom two panels in Table 2.1 compare densities of retroelements in each orientation at each intergenic location to the densities of retroelements in regions most distant (>30 kb) from genes. These latter comparisons illustrate that the retroelement density differences plotted relative to gene location are highly significant. For example, the densities of Alu sequences at all locations are highly significantly different from their density in regions >30 kb from genes.

Table 2.1 Significance (p-values) of retroelement locations with respect to genes

Sense vs. antisense

	Alu	L1	MIR	L2	MLT	MST	MER4	Class III	Class I	Class II
in[a]	0.001	4.9E-05	0.34	0.005	2.9E-05	9.7E-05	9.2E-06	3.6E-04	1.6E-04	3.7E-04
0-5 dnst[b]	0.051	0.041	0.042	0.007	9.1E-06	0.069	3.9E-05	0.011	0.02	0.44
5-10 dnst	0.48	0.22	0.17	0.32	0.03	0.034	0.054	0.13	0.44	0.037
10-20 dnst	0.014	0.19	0.43	0.32	0.24	0.26	0.2	0.35	0.012	0.18
20-30 dnst	0.002	0.034	0.29	0.37	0.37	0.092	0.089	0.26	0.14	0.078
>30 dnst	0.019	0.003	0.4	0.39	0.35	0.29	0.3	0.21	0.41	0.2
>30 upst[c]	0.035	0.047	0.41	0.24	0.057	0.29	0.21	0.14	0.38	0.34
20-30 upst	0.25	0.16	0.23	0.45	0.037	0.27	0.018	0.14	0.13	0.17
10-20 upst	0.42	0.053	0.067	0.054	0.008	0.012	0.11	0.28	0.25	0.23
5-10 upst	0.043	0.17	0.066	0.23	0.014	0.042	0.15	0.25	0.23	0.35
0-5 upst	6.0E-05	0.001	0.29	0.024	0.015	0.009	0.096	0.03	0.15	0.15
in	0.001	4.9E-05	0.34	0.005	2.9E-05	9.7E-05	9.2E-06	3.6E-04	1.6E-04	3.7E-04

Antisense vs. >30 kb from genes

	Alu	L1	MIR	L2	MLT	MST	MER4	Class III	Class I	Class II
in	8.9E-06	0.015	0.13	0.007	0.003	1.2E-04	0.001	6.9E-05	4.3E-05	0.02
0-5 dnst	9.9E-07	2.3E-06	0.006	0.1	1.3E-04	1.0E-04	0.005	0.012	0.015	0.27
5-10 dnst	7.1E-07	0.45	0.009	0.29	0.006	0.018	0.08	0.12	0.42	1.8E-04
10-20 dnst	4.9E-07	0.058	0.003	0.49	0.026	0.097	0.07	0.002	0.005	0.006
20-30 dnst	1.1E-06	0.002	0.074	0.5	0.011	0.03	0.032	0.18	0.41	0.005
20-30 upst	4.2E-07	0.38	0.2	0.27	0.06	0.007	0.033	0.01	0.013	0.17
10-20 upst	1.5E-07	0.38	0.011	0.012	0.045	0.064	0.061	0.054	0.46	0.014
5-10 upst	2.1E-07	0.004	0.083	0.001	0.38	0.004	0.022	0.029	0.028	0.022
0-5 upst	3.0E-07	3.0E-06	0.34	7.7E-05	2.4E-05	4.8E-04	0.033	1.2E-06	0.06	0.016
in	8.9E-06	0.015	0.13	0.007	0.003	1.2E-04	0.001	6.9E-05	4.3E-05	0.02

Sense vs. >30 kb from genes

	Alu	L1	MIR	L2	MLT	MST	MER4	Class III	Class I	Class II
in	1.7E-04	8.6E-07	0.069	0.14	4.4E-07	2.2E-06	1.3E-07	1.6E-07	1.5E-07	1.3E-05
0-5 dnst	1.0E-06	6.9E-06	9.3E-05	0.002	2.5E-06	2.3E-05	1.0E-05	1.9E-04	2.0E-06	0.3
5-10 dnst	9.0E-07	0.45	0.003	0.12	0.003	0.001	0.39	0.003	0.48	0.055
10-20 dnst	2.9E-06	0.007	0.003	0.39	0.15	0.019	0.034	0.02	0.18	0.005
20-30 dnst	3.9E-06	0.096	0.011	0.24	0.28	0.22	0.03	0.068	0.078	0.002
20-30 upst	1.8E-07	0.4	0.36	0.24	0.012	0.002	0.44	0.002	0.073	0.4
10-20 upst	1.4E-07	0.053	0.066	0.24	0.015	0.003	0.23	0.002	0.04	0.041
5-10 upst	4.7E-07	0.02	0.41	0.012	0.026	4.6E-04	0.14	0.045	0.33	0.002
0-5 upst	6.8E-07	6.8E-07	0.13	0.012	2.1E-05	1.3E-06	0.017	8.3E-06	0.48	0.043
in	1.7E-04	8.6E-07	0.069	0.14	4.4E-07	2.2E-06	1.3E-07	1.6E-07	1.5E-07	1.3E-05

Shaded regions are significant ($p < 0.05$)

[a]within a gene

[b]dnst: kb downstream of the nearest gene

[c]upst: kb upstream of the nearest gene

2.3.3 Shifting retroelement distributions with age

It is apparent that the retroelement distributions in genes and intergenic regions (Figure 2.3) do not fully conform to the genome-wide distribution patterns of elements observed in Figures 2.1 and 2.2. Furthermore, for Alu repeats, it has been reported previously that young elements (< 1 Myr) have a preference for AT-rich regions whereas older Alus show an increasing density in GC-rich DNA (Smit 1999; International Human Genome Sequencing Consortium 2001) (see Figure 2.4A) and hypotheses to explain this phenomenon have been proposed (Schmid 1998; Brookfield 2001; International Human Genome Sequencing Consortium 2001; Pavlicek et al. 2001)(see Section 1.2.4). Transposition into AT-rich regions might be expected to lead to accumulation of TEs in this gene poor part of the genome (e.g. the heterochromatin) where recombination is strongly reduced and element interference with genes is less pronounced. However, the observed density differences of the youngest Alu elements (present in AT rich regions) as opposed to older elements (in GC rich regions) do not follow this expectation. A possible explanation for the age-related Alu density differences is that these retroelements are removed preferentially from their initial integration sites in the AT rich regions of the genome prior to fixation. However, because there is a gradual density increase of Alu elements by age in the GC rich fraction, it is possible that already fixed elements are gradually lost from the AT rich region while they are maintained in GC rich regions.

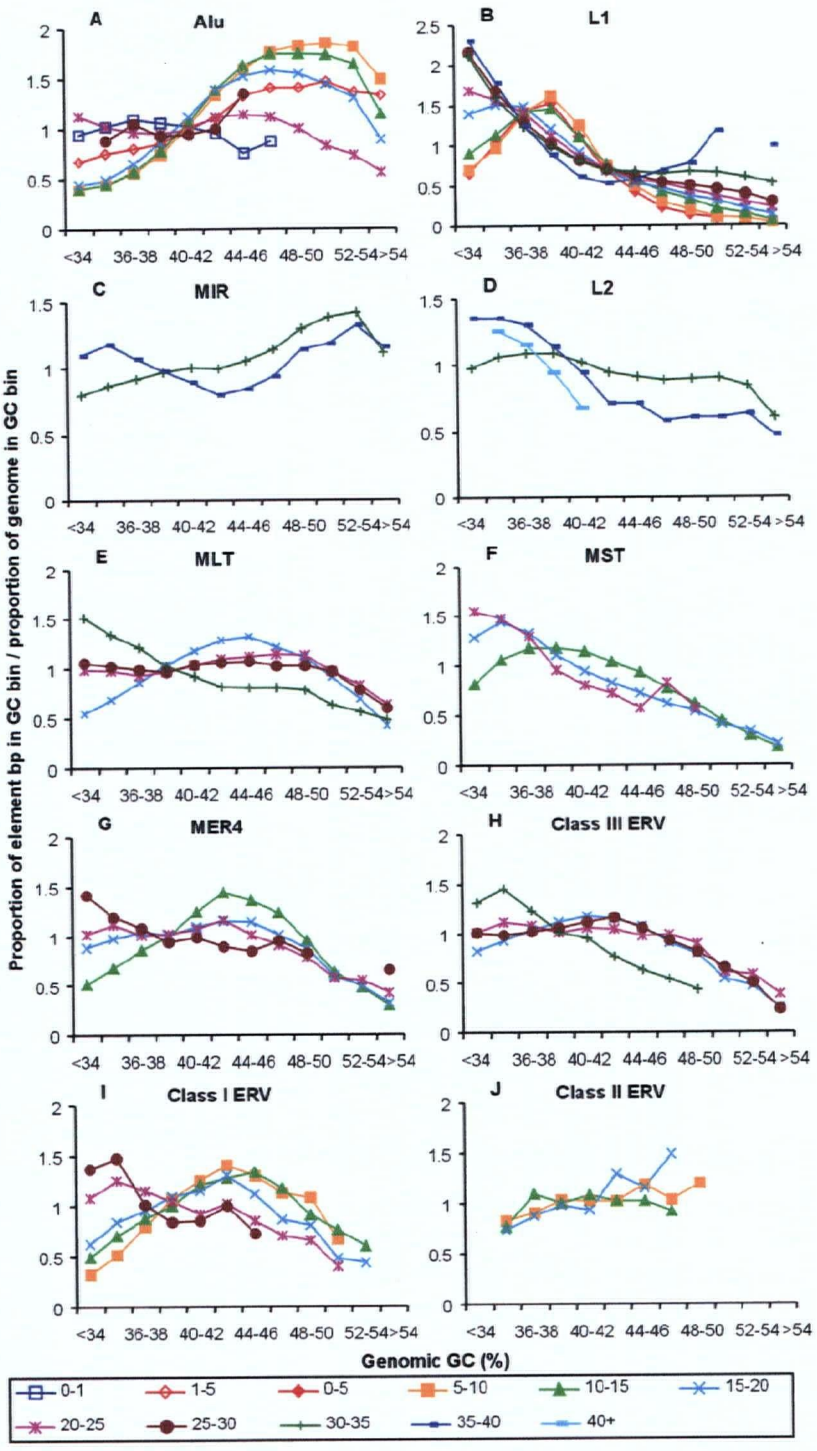

Figure 2.4 Retroelement densities of different divergence classes in various GC fractions of the human genome. The density distribution of each retroelement divergence cohort was plotted in GC bins as indicated in the legend to Figure 2.1. The divergence classes are indicated in % divergence from the consensus sequence below the graphs. Data points missing in traces are due to GC bins containing less than 100 elements. Standard deviations were calculated (see Methods) but are not shown in the interest of clarity.

To investigate if other retroelements also change their genomic distribution with age, we determined the distribution patterns of LTR elements, SINEs and LINEs of different ages as a function of GC content (Figure 2.4). As discussed above, it is apparent that the youngest Alu elements (0-1% divergent), many of which are polymorphic insertions (Carroll et al. 2001; Batzer and Deininger 2002), are distributed differently than the next youngest (fixed) Alus of the 1-5% divergence group and that the densities of the next two Alu age cohorts (5-15% divergent) are skewed even further to GC-rich regions (Figure 2.4A). Notably, this figure also reveals that the oldest Alu repeats are less prevalent in GC-rich domains and, indeed, have a density distribution closer to that of the youngest age class. This density pattern of the oldest Alu elements was not evident in a similar analysis reported previously (International Human Genome Sequencing Consortium 2001). In that study, Alu elements were divided by subfamily instead of divergence and the density of the oldest subfamily, AluJ, was still highly skewed to GC rich regions. However, the AluJ subfamily was considered as a single large cohort, the members of which have divergences ranging from less than 10% to greater than 25%. When the more divergent AluJ members of 15-20% and 20-25% divergence are separated into their own groups, their densities are essentially identical to the patterns presented in Figure 2.4A (data not shown). Thus, the different methods for separating Alu elements accounts for the differences between our analysis and that in the genome consortium study.

Results of similar analyses conducted for the other retroelements reveal some provocative trends. As noted before (Smit 1999) and as shown in Figure 2.4B, young L1 elements are preferentially found in the AT-rich fraction in the genome and older

elements tend to be found in the most AT-dense part of the genome. Analysis of the ancient L2 and MIR repeats was hampered by the short average length of most elements which prevented an accurate determination of their divergence from a consensus sequence (age) (see Methods for details). However, for the two divergence classes that could be reliably determined, the oldest L2 and MIR sequences also show an increased density in the less GC rich sections of the genome compared to their younger counterparts (Figure 2.4C, D).

For most of the LTR elements, we observe a trend similar to that seen for the L2 and MIR sequences. For elements belonging to the MLT, MST, MER4, Class I and III ERV groups, densities of the youngest members of these superfamilies peak in regions of higher GC compared to their older relatives (Figure 2.4E-I). That is, the highest concentrations of these elements appear to gradually shift to regions of lower GC with increasing age. This tendency is not evident for the Class II ERVs (Figure 2.4J). Potential explanations for this trend will be discussed below.

To determine if the shifting patterns observed in Figure 4 are statistically significant, we again divided the genome into four subgenomes and repeated the analysis for each of these. Each point in the graphs could then be assigned a mean and standard deviation based on values obtained for each subgenome. The t-test was used to determine if the density distribution of a particular age cohort was significantly different when compared to the next oldest cohort. Table 2.2 lists the p values resulting from this analysis. For all retroelements except the Class II ERVs, the majority of the density points are significantly different ($p < 0.05$) for at least one comparison between adjoining age cohorts. Indeed, for the most numerous elements, Alu and L1, almost all comparisons are statistically significant. If the youngest and oldest age cohorts of each

superfamily are compared, all except the Class II ERVs are highly significant (data not shown).

Table 2.2 Significance (p-values) of distributional differences between divergence cohorts

	Alu						L1						
	[b]0-1: 1-5	1-5: 5-10	5-10: 10-15	10-15: 15-20	15-20: 20-25	20-25: 25-30	0-5: 5-10	5-10: 10-15	10-15: 15-20	15-20: 20-25	20-25: 25-30	25-30: 30-35	30-35: 35-40
<34[a]	0.20	0.10	0.50	0.36	0.038		0.49	0.26	0.18	0.34	0.30	0.48	0.43
34-36	0.002	2.5E-06	0.43	7.9E-05	1.6E-07	0.07	0.06	0.002	2.0E-07	0.08	0.021	0.007	0.006
36-38	0.001	1.8E-04	0.25	1.3E-04	5.9E-07	0.31	0.004	0.27	0.031	0.002	0.001	0.016	0.47
38-40	4.4E-04	2.1E-04	0.002	9.2E-05	0.003	0.39	0.41	0.006	1.5E-04	9.2E-05	2.7E-05	0.13	0.020
40-42	0.007	0.17	0.013	0.002	0.006	0.24	0.029	0.010	8.0E-05	0.005	0.011	0.28	0.001
42-44	0.06	4.2E-05	0.009	0.46	0.001	0.14	0.37	0.36	0.003	0.042	0.18	0.12	0.05
44-46	4.1E-04	4.0E-05	0.025	0.002	1.4E-05	0.016	0.08	0.002	0.043	0.023	0.001	0.032	0.019
46-48	2.3E-04	6.1E-05	0.29	4.9E-04	9.1E-07		0.023	6.2E-05	0.005	0.002	0.09	0.003	0.25
48-50		4.0E-04	0.022	2.9E-04	3.5E-05		0.020	3.9E-05	0.010	0.021	0.009	3.9E-05	0.048
50-52		4.1E-04	0.040	1.2E-04	1.4E-06		0.10	0.001	0.001	0.008	0.017	0.013	0.001
52-54		6.8E-05	4.0E-04	7.7E-05	4.2E-06			0.002	0.002	0.001	0.001	3.9E-04	
>54		0.017	3.2E-07	3.1E-07	0.001			2.6E-04	6.4E-06	1.4E-05	1.1E-04	1.8E-06	0.001

	MIR	L2		MLT			MST		MER4		
	30-35: 35-40	30-35: 35-40	35-40: 40+	15-20: 20-25	20-25: 25-30	25-30: 30-35	10-15: 15-20	15-20: 20-25	10-15: 15-20	15-20: 20-25	20-25: 25-30
<34	0.23	0.24		0.11	0.45	0.22	0.16	0.30	0.12	0.36	0.22
34-36	5.7E-05	3.8E-06	0.05	7.6E-05	0.12	0.001	1.5E-05	0.10	1.5E-04	0.029	0.10
36-38	9.2E-05	0.001	0.20	0.010	0.020	8.9E-05	3.7E-04	0.24	0.005	0.30	0.08
38-40	0.50	0.026	0.48	0.001	0.34	0.006	0.015	0.13	0.35	0.27	0.07
40-42	0.001	0.019	0.07	3.6E-04	0.23	0.014	1.2E-04	0.024	0.012	0.31	0.044
42-44	0.001	3.2E-05		8.4E-06	0.08	1.7E-04	0.001	0.14	2.8E-04	0.45	0.004
44-46	0.001	2.3E-04		9.1E-05	0.16	0.001	0.001	0.014	0.010	0.023	0.06
46-48	0.036	0.001		0.043	0.023	0.002	0.001	0.026	0.016	0.11	0.38
48-50	0.029	3.4E-04		0.25	0.039	0.004	0.031	0.38	0.21	0.17	0.29
50-52	0.06	1.6E-04		0.045	0.46	1.1E-04	0.011		0.40	0.38	
52-54	0.43	0.009		0.032	0.30	0.019	0.14		0.45	0.23	
>54	0.016	0.029		6.9E-05	0.22	0.004	0.16		0.15	0.002	0.014

	Class III			Class I				Class II	
	15-20: 20-25	20-25: 25-30	25-30: 30-35	5-10: 10-15	10-15: 15-20	15-20: 20-25	20-25: 25-30	5-10: 10-15	10-15: 15-20
<34	0.33	0.48	0.26	0.12	0.29	0.09	0.33		
34-36	0.005	0.029	0.003	0.022	0.05	0.001	0.06	0.30	0.30
36-38	0.21	0.08	4.1E-05	0.020	0.18	0.041	0.06	0.050	0.028
38-40	0.008	0.13	0.39	0.07	0.038	0.39	0.13	0.32	0.20
40-42	0.016	0.16	0.026	0.13	0.07	0.010	0.29	0.21	0.07
42-44	0.023	0.020	2.8E-04	0.035	0.21	0.011	0.28	0.46	0.08
44-46	0.036	0.08	0.001	0.48	0.06	0.030	0.25	0.001	0.18
46-48	0.10	0.36	1.4E-06	0.38	0.006	0.003		0.18	0.001
48-50	0.06	0.21	0.011	0.004	0.08	0.031			
50-52	0.033	0.24		0.30	0.031	0.15			
52-54	0.041	0.19			0.05				
>54	0.003	0.001							

[a]GC content (%); [b]Divergence cohorts compared

One qualification regarding this data concerns the method used to identify retroelements of different ages. Elements were classified as belonging to divergence cohorts based on percent substitution from their consensus sequence (Jurka 2000). The consensus sequence corresponds to the approximate sequence at the time of integration in the genome, where retroelements in higher divergence cohorts indicate an older time of integration relative to the retroelements of lower divergence values (Li and Graur 1991; Shen et al. 1991; Smit et al. 1995; International Human Genome Sequencing Consortium 2001). Therefore, the validity of this method is highly dependent on having accurate consensus sequences for all subfamilies. It is quite possible and even likely that some elements have been assigned an incorrect age due to extreme heterogeneity of some of the retroelement classes, particularly among the LTR groups. However, if this was a major problem, one would not expect to observe a consistent shift in density in one direction – namely toward lower GC regions with increasing divergence.

2.3.4 Length differences do not account for the shifting patterns

To investigate potential mechanisms that may underlie the age related distribution differences, we used two different methods to try to determine if differential rates of retroelement deletions in different genomic GC regions account for the shifting patterns observed in Figure 2.4. First, we examined the relative length of elements in different GC fractions. The results of this analysis indicated that retroelements gradually become shorter as they age, presumably due to small deletions or loss of recognition of diverged segments by RepeatMasker, but the shortening is largely independent of the surrounding GC content (data not shown). The two exceptions to this general observation are represented by L1 elements and older Alu sequences (Figure 2.5). The average length of

younger L1 elements (<10% divergence) peaks in the 38-42% GC fractions which might explain the abundance of L1 base pairs in this region (Figure 2.4B). In the case of Alu elements in the 20-30% divergence cohorts, there is a slight decrease in apparent length with increasing GC content (Figure 2.5B) but this is not enough to account for the density pattern of this age group (Figure 2.4A). In addition, the small degree of shortening as measured here does not explain the rapid enrichment of younger Alu elements in higher GC fractions.

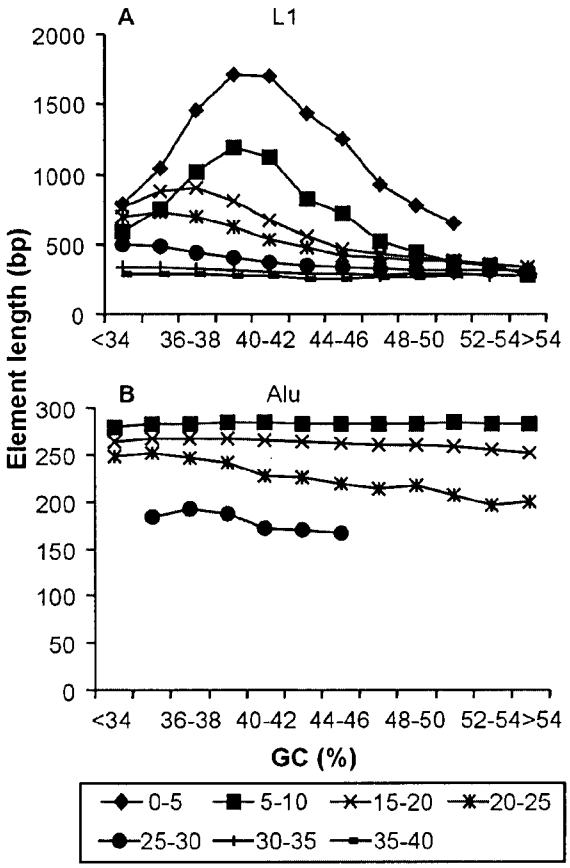

Figure 2.5 Length distribution of retroelements with respect to surrounding GC content. Retroelements of each group were classified as belonging to divergence cohorts as described in the text. The average length in base pairs (bp) of each retroelement divergence cohort contained within each GC bin (see legend to Figure 2.1) is shown for L1 and Alu elements (panels A and B). GC bins containing less than 100 elements were excluded from the graphs.

2.3.5 *Delay of Alu density changes on the Y chromosome*

As another way of investigating the change in distribution of younger Alus toward GC-rich regions, we analyzed Alu density patterns on the Y chromosome, much of which does not recombine (Graves 1995), and detected a major difference on this chromosome compared to the whole genome (Figure 2.6). Alu elements on chromosome Y less than 5% divergent are not numerous enough to include in this analysis. However, the density pattern of Alus in the 5-10% divergence class is strikingly opposite to that observed in the whole genome in that they are much more prevalent in AT-rich regions compared to GC-rich regions (Figure 2.6C). The distributions of older Alu elements (>10% divergent from the consensus) with respect to GC content are consistent with the patterns seen in the entire genome (Figure 2.6D-F). Table 2.3 shows the p values resulting from this analysis. This finding suggests that the density shift of Alus from AT-rich to GC-rich regions during evolution was significantly delayed on the Y chromosome and, therefore, that the ability to recombine with a homologous chromosome greatly facilitated this shift.

Table 2.3 Significance (p-values) of distributional difference between Alus on the Y chromosome versus the whole genome

	[b]5-10	10-15	15-20	20-25
34-36[a]	0.0012	0.023	0.13	0.022
36-38	4.5E-04	0.014	0.0022	0.28
38-40	0.021	0.022	0.28	0.14
40-42		0.039	0.41	0.27
42-44		0.10	0.24	0.34
44-46		0.11	0.08	

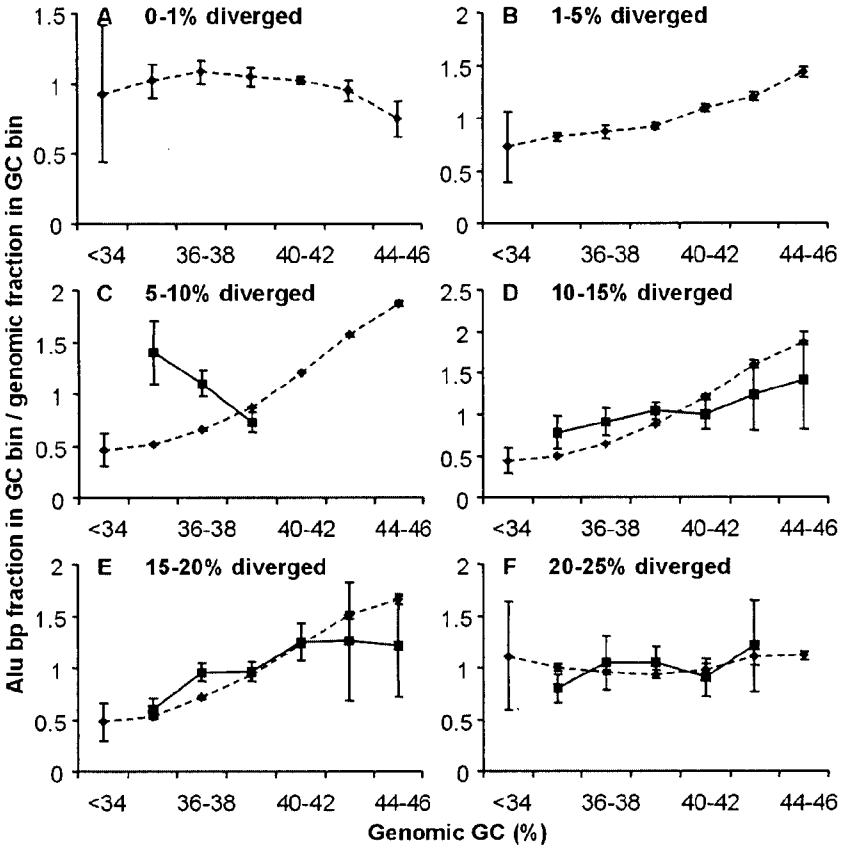

Figure 2.6 Density of Alu divergence cohorts in different GC fractions on chromosome Y compared to the whole genome. Solid lines indicate Alu elements on chromosome Y whereas dashed lines represent the Alu density in the whole genome. Parts A to F indicate the density of a specific divergence class which is indicated on the top of each panel. There were insufficient numbers of Alu elements on the Y chromosome in the first two divergence cohorts to be plotted in panels A and B. The density distribution of each Alu divergence class is plotted against the local 20-kb genome GC content. Standard deviations were calculated as described in Methods.

2.3.6 Potential explanations for Alu distribution patterns

The density patterns of Alu elements do not conform to trends observed for other retroelements. These elements integrate into the AT-rich part but accumulate in GC-rich DNA (International Human Genome Sequencing Consortium 2001) (Figure 2.4A) and at least three hypotheses have been proposed to account for this phenomenon. One proposed explanation is that the GC-rich Alu elements are more stable in regions where the surrounding GC content is similar (Pavlicek et al. 2001). However, we have observed that partial deletions or apparent shortening of various Alu age groups are uniformly

distributed irrelevant of GC occupancy (Figure 2.5B). This finding does not seem to support such a hypothesis although it is possible that the tendency of retroelements to remain in regions of matching GC content does play some role. A second hypothesis proposes that Alu elements are selectively retained in GC-rich regions because having these elements close to genes is of functional benefit (Britten 1997; Kidwell and Lisch 1997; Schmid 1998). Figure 2.3A shows that the Alu density near genes is higher than predicted based on GC content. That is, the tendency of Alu elements to be located near genes is not fully explained by the general GC-richness associated with coding regions and such a pattern may therefore reflect a functional role for these elements. However, other observations appear discordant with this view. For example, it is known that the developmentally critical HoxD gene cluster is almost devoid of retroelements (International Human Genome Sequencing Consortium 2001). A recent study has also found that SINEs (Alu and MIR elements) are less frequently associated with imprinted than non-imprinted genomic regions (Greally 2002). Certain classes of genes may therefore need to exclude such sequences from their environment to ensure proper function or regulation.

A third hypothesis proposes that the maintenance of Alus in GC rich regions may be due to the adverse effects that deletions and unequal recombinations could have in gene-rich regions (Brookfield 2001). Indeed, due to the vast numbers of Alu elements in the genome, it is likely that specific recombinational mechanisms have been a major force in shaping the distribution of Alus in the genome. It has recently been demonstrated that the efficiency of Alu-Alu recombination in yeast increases as a pair of elements are placed closer together (Lobachev et al. 2000). Such closely spaced Alu pairs are found only occasionally in the human genome (Lobachev et al. 2000; Stenger et al.

2001), possibly because of clearance of these elements through the mechanism of inverted repeat (IR)-mediated recombination (Leach 1994). Alu elements seem quite promiscuous for recombination because two elements up to 20% divergent are still able to recombine efficiently (Lobachev et al. 2000). Furthermore, there are many examples of Alu-mediated recombination resulting in mutations in humans (Batzer and Deininger 2002). These findings suggest a possible explanation for the changing Alu distribution profiles shown in Figure 2.4A and their enrichment near genes. Considering the high number of genomic Alu elements and the fact that they preferentially target AT-rich regions, these domains must have suffered a massive build up of Alu integrations. Such accumulation likely resulted in increased recombination as the occurrence of closely spaced, highly related Alus increased which could have led to loss of both newly integrated and fixed Alu elements in the AT rich fraction of the genome. In regions close to genes, it is possible that Alu-Alu recombination events are less likely to be allowed or become fixed because of an increased chance of simultaneously removing gene regulatory domains (Brookfield 2001). This could help explain the over-representation of Alu elements near genes without invoking a functional role. The fact that we observe no increased density in GC- or gene-rich regions for the oldest Alus could be explained by the fact that Alus in these age cohorts are much less numerous and therefore would have been less subject to loss via recombination in AT-rich regions. Alu elements of 20-30% divergence are present in only ~25,000 copies whereas younger Alus in the 5-10, 10-15 and 15-20% divergence classes are present in ~300,000, 480,000 and 210,000 copies respectively. Furthermore, due to their higher divergence values, the oldest Alus would also have been less able to recombine with their younger, more numerous relatives when the latter populated the genome.

Differences in recombination are likely also responsible for the fact that Alu elements are not over represented on chromosome Y as are other younger retroelements such as Class I and II ERVs (International Human Genome Sequencing Consortium 2001) (Figure 2.2). This finding suggests that Alus are lost more readily than the LTR elements. However, loss of Alu elements on the Y appears delayed compared to on the autosomes (Figure 2.6), likely because only intrachromosomal/IR recombination can operate on most of the Y. IR recombination seems to work more efficiently when two elements are closely located (Lobachev et al. 2000) and it is likely that this is true also for intrachromosomal recombination in general. Thus we postulate that LTR elements are removed less efficiently than Alu elements due to their much lower copy number and, therefore, larger average inter-element distance.

2.4 Concluding remarks

One view of transposable elements considers them to be selfish DNA of no use to the host (Doolittle and Sapienza 1980; Orgel and Crick 1980; Yoder et al. 1997), while others hypothesize that their fixation reflects functional interactions with the host (McDonald 1995; Brosius 1999). Our data support the idea that retroelements have a general negative impact on the host because of a gradual accumulation of most retroelement superfamilies in the AT rich fraction and on the Y chromosome (which is predicted to occur according to the selfish DNA hypothesis) (Charlesworth et al. 1997). However, these findings also support a concept in which retroelements gradually are cleared (or maintained) from the host genome, a relationship that seems dependent on the age of their association. (Di Franco et al. 1997; Junakovic et al. 1998; Torti et al. 2000; Kidwell and Lisch 2001). The fact that densities of old MIR and L2 retroelements near genes are close to that predicted by average GC content suggests a relatively benign

relationship between these retroelements and genes. In contrast, retroviral elements may have interfered more often with gene function due to initial integration site preference into gene rich regions. The density pattern of the relatively young class II ERVs (Figure 2.3J) supports this suggestion. Of those LTR elements which have been fixed in the population (i.e. almost all those in humans), our analyses have revealed that the highest densities of the older elements gradually shift with age to AT-rich or gene-poor DNA. Furthermore, we have shown that all types of LTR retroelements are significantly underrepresented within genes. Since LTRs carry transcriptional regulatory signals very similar to those in cellular genes (Majors 1990), it seems reasonable that insertion of an LTR close to or within a gene would frequently be disadvantageous unless it is efficiently silenced by methylation or other mechanisms (Yoder et al. 1997; Whitelaw and Martin 2001). Such insertions with a marked negative impact will be selected against with no chance to spread to fixation. However, it is known that a mutation with a selective disadvantage can still be fixed through genetic drift, especially if the effective population size is small (Li and Graur 1991). It is possible that some LTR elements, despite being fixed in the species, had a slight negative impact and were gradually eliminated with time. Alternatively, mechanisms unrelated to selection, such as differential rates of recombination in different GC domains, may also explain the shifting density patterns of LTR retroelements. The fact that the youngest Class II ERVs do not show the same density pattern shifts as seen for most of the LTR superfamilies could be because there has not been sufficient evolutionary time for their distribution to be shaped by selective forces and/or recombination.

Once fixed in the population, it is not possible for an insertion to be eliminated unless insert-free alleles are re-created. While unequal crossing-over between

homologous chromosomes may be the main mechanism responsible for elimination of retroelements in GC rich regions, which have higher rates of recombination (Fullerton et al. 2001), intrachromosomal deletions and IR-mediated recombination might enhance this effect, especially in regions of high retroelement density. Such processes could regenerate insert-free alleles and again provide an opportunity for the original insertion to be lost from the population through natural selection or drift.

While these studies have attempted to address some of the potential mechanisms or forces that have shaped the genomic distributions of human retroelements, further studies are warranted to elucidate the complex evolutionary and functional relationships between these sequences and their host genome.

Chapter 3: Analysis of transposable elements in the human and mouse transcriptomes

A version of this chapter has been published:
van de Lagemaat, L.N., J.R. Landry, D.L. Mager, and P. Medstrand. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19:** 530-536.

I performed all bioinformatic analyses, except that in Table 3.2, and wrote sections of the paper.
J. R. L. created Table 3.1 and Figure 3.2.
D. L. M. and P. M. are senior authors on the paper.

3.1 Introduction

TEs (primarily retroelements) comprise at least 45% of the human genome and 40% of the mouse genome and ancient elements which have diverged beyond recognition have also undoubtedly contributed to the composition of mammalian chromosomes (Deininger and Batzer 2002; Mouse Genome Sequencing Consortium 2002). While the negative effects of TEs in causing mutations in individuals are well recognized (Ostertag and Kazazian 2001; Deininger and Batzer 2002; Mouse Genome Sequencing Consortium 2002), their major impact may be their ability to induce changes in gene regulation (Murnane and Morales 1995; Brosius 1999; Hamdi et al. 2000; Medstrand et al. 2001; Nigumann et al. 2002; Jordan et al. 2003; Kashkush et al. 2003) or coding potential (Murnane and Morales 1995; Nekrutenko and Li 2001) without destroying existing gene functions. The primary goal of this study was to test the hypothesis that TEs foster variation in some gene classes while being excluded from others.

3.2 Methods

3.2.1 Prevalence of TEs in human and mouse gene transcripts

Genomic coordinates of TEs and mRNAs contained in the RefSeq database were downloaded from the human June 2002 and mouse February 2003 Genome Browser at the University of California Santa Cruz (http://genome.ucsc.edu). Genes were defined by their mRNAs, which were required to have non-zero-length 5′ and 3′ UTRs as mapped to the sequence assemblies. We excluded 1998 human and 2557 mouse mRNAs contained in RefSeq from the analysis because they lacked either a 5′ or a 3′ UTR or both UTRs.

For each genome, we constructed mySQL databases containing the mapping data and Perl scripts that conducted automated queries to determine genomic overlaps between TEs and UTR exons. Overlaps of greater than one bp were allowed but only 2% of detected TEs had an annotated overlap of <5 bp. To eliminate false-positive TEs (which can occur in regions where the local GC content differs from the surroundings), genomic sequences of all putative repeats in UTRs were remasked using RepeatMasker (http://repeatmasker.genome.washington.edu) with -s (sensitive) and -gccalc (expected GC content equal to that of the repeat itself) settings. Finally, all 490 human Alu repeat elements, identified in the sense orientation at the 3′ end of transcripts, were moved to the 3′ 'internal' UTR category because many appear to represent oligo-dT mispriming on the A-rich Alu terminus during cDNA synthesis.

3.2.2 Variation of TE prevalence with gene class or function

The Gene Ontology (GO) analyses of RefSeq transcripts was carried out as follows: the RefSeq database was downloaded from the online NCBI repository and each record was parsed to obtain the accession numbers of the nucleotide source records of each RefSeq transcript, and these links were recorded in a database table. Further, the SwissProt database was downloaded and parsed to obtain the links between SwissProt identifiers and the accession numbers in the nucleotide database upon which each SwissProt record was based. We then also constructed a table of links between GO terms and SwissProt identifiers parsed from a table downloaded from the GO online repository (http://www.geneontology.org). With these tables in a large mySQL database, a three-way table join allowed us to make a classification of most RefSeq transcripts. We then also downloaded the database of GO terms and links between them and used a mySQL database system to translate the assigned GO terms to more general ones within the

molecular function and biological process classification trees.

The conservation-based analyses of Ka/Ks were done using basic assumptions. Alignments of all mouse and all human RefSeq transcripts were constructed by finding the best human-mouse hit using ungapped BLASTn (Altschul et al. 1990). The aligned results were parsed and analyzed in three reading frames. The optimal reading frame was chosen by minimization of the sum of stop codons and non-synonymous substitutions. The Ka/Ks ratio was calculated in this reading frame.

3.3 Results and Discussion

3.3.1 Prevalence of TEs in human and mouse gene transcripts

We first determined the overall prevalence of TEs in the UTRs of human and mouse genes in the RefSeq database (http://www.ncbi.nlm.nih.gov/RefSeq/). This analysis revealed that 27.4% of 12179 human RefSeq loci with annotated UTRs (referred to from now on as human genes) have at least one mRNA with TE-derived sequence within the 5' or 3' UTR. The percentages of genes with TEs in different UTR locations are shown in Figure. 3.1a. These data are in general agreement with a recent survey of the Mammalian Gene Collection (http://mgc.nci.nih.gov/) which showed that close to 20% of human genes in that dataset contain TE sequences in a 5' or 3' UTR (Jordan et al. 2003). Analysis of the mouse Refseq database in the same way revealed that 18.4% of 10064 mouse RefSeq loci (or mouse genes) contain at least one TE within their UTRs. The lower TE coverage in mouse genes is likely to be more apparent than real due to an incomplete rodent repeat database and to the higher nucleotide substitution rate in the mouse lineage resulting in fewer detectable ancient TEs in mouse compared to human

(Mouse Genome Sequencing Consortium 2002).

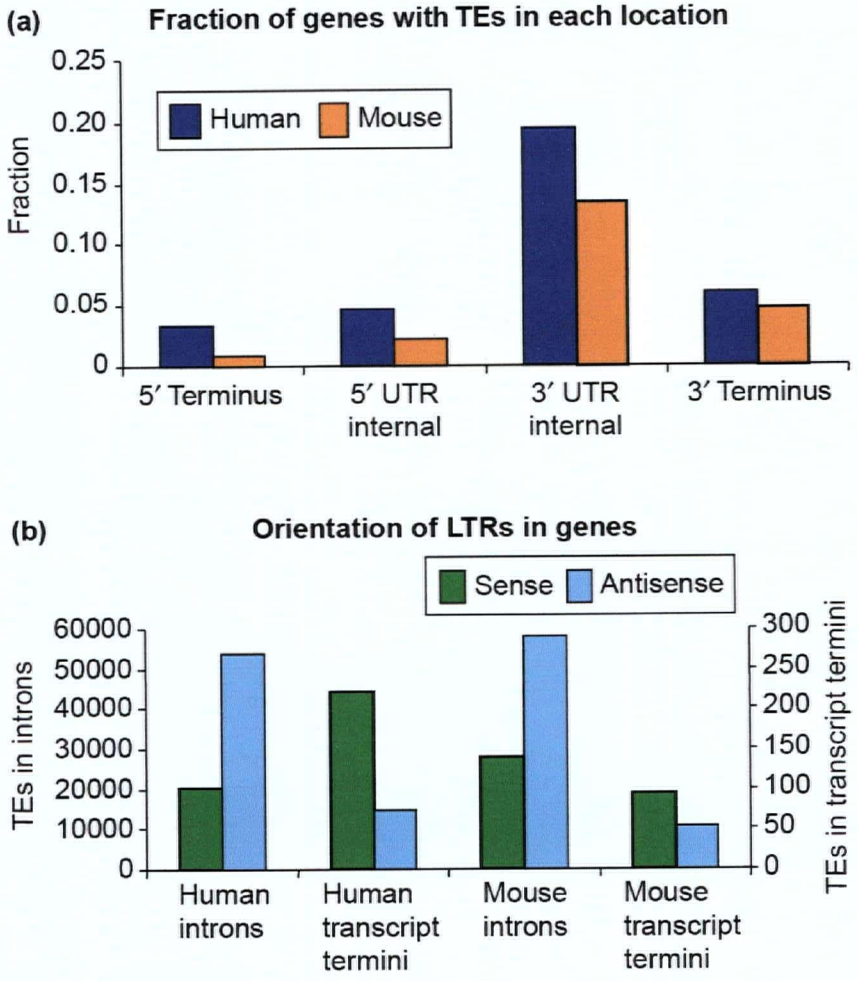

Figure 3.1 TEs in genes by species and orientation. (a) Fraction of human and mouse genes with TEs in UTRs. The fraction of genes with one or more RefSeq mRNA having at least one TE extending across the 5' or 3' end of the transcript ('terminus'), or totally within ('internal')the 5' or 3' UTR is shown. (b) Transcriptional orientation of LTR elements in introns compared to those spanning mRNA termini. The left scale is the absolute numbers of LTR elements within introns of genes and the right hand scale shows numbers overlapping mRNA termini. Note that over 99% of TEs within genes are in introns. The orientation bias of all four categories shown is significant ($p<0.01$).

3.3.2 TEs serve as alternative promoters of many genes

A search of the Human Promoter Database (http://zlab.bu.edu/~mfrith/HPD.html) has previously shown that close to 25% of analyzed promoter regions contain some TE-

derived sequence (Jordan et al. 2003) and several individual cases showing a role for TEs in human gene transcription have been reported (Murnane and Morales 1995; Brosius 1999; Hamdi et al. 2000; Medstrand et al. 2001; Nekrutenko and Li 2001; Nigumann et al. 2002) Many of these cases were detected by our method and we also found numerous new examples of apparent usage of a TE-derived promoter where TE involvement has not been reported previously (see Table 3.1 for a partial list). Genes are candidates for having a TE-derived promoter if the 5' end of their 5' UTR resides in a TE sequence and those in Table 3.1 were examined in more detail. This analysis illustrated several ways in which TE-derived promoters might contribute to gene expression, including examples of a) different expression patterns in human and mouse orthologs that correlate with the TE insertion (CYP19, TMPRSS3, HYAL4, ENTPD1, CASPR4, MKKS); b) the same TE insertion correlating with a tissue-specific promoter in both species (CA1, SPAM1, KLK11); c) presence of the TE in both species but apparent usage as a promoter in human only (MSLN) d) presence of the TE only in human but similar overall expression patterns as in the mouse (BAAT, SIAT1, CLDN14, MAD1L1) and e) a human multi-gene family where the member with a TE has a different expression pattern compared to other family members (FUT5, ILT2).

Table 3.1 RefSeq transcripts beginning within a previously unrecognized TE

Gene	Full name	Functional role (Disease)	Probable TE involvement in human	TE in mouse[a]	Human expression[b]	Mouse expression[c]
CYP19	Aromatase	Estrogen synth (repro abnormal)	LTR one of at least 6 promoters	No	LTR drives very high placental expression.	No placental exp.
TMPRSS3	Transmembrane protease, serine 3	Serine protease (deafness)	LTR/Alu as alternate promoter	No	TE form exp. primarily in PBLs. Other forms widespread	Inner ear, kidney, stomach, testis
HYAL-4	Hyaluronidase 4	Hyaluronan catabolism	Antisense L1/Alu as only known promoter	No	Primarily placenta	Primarily skin
ENTPD1 /CD39	Ectonucleoside triphosphate diphosphohydrolase 1	Lymphoid cell activation antigen	LTR as 1 of 2 promoters & results in HERV-derived N-terminus	No	LTR drives exp. in placenta & melanoma. Overall expression widespread	Widespread
CASPR4	Contactin associated protein-like 4	Brain cell adhesion	LTR is one of 3 promoters & donates protein N-terminus	No	LTR form exp. in brain, testis, tumors. High exp. in brain & sp. cord	Brain
MKKS	McKusick-Kaufman syndrome gene	Chaperonin (Mckusick-Kaufman)	LTR/L2 as alternate promoter	No	TE form in testis and fetal tissues. Overall expr. widespread	Widespread
CA1	Carbonic anhydrase 1	Carbon metabolism	LTR one of 2 major promoters	Yes	LTR drives erythroid exp.	LTR drives erythroid exp.
SPAM1 /PH20	Sperm adhesion molecule 1	Sperm-egg adhesion	Antisense ERV as only known promoter	Yes	Primarily testis	Primarily testis
KLK11	Kallikrein 11	Serine protease	MIR one of 3 promoters & leads to alt. N-terminus	Yes	Widespread	Brain & prostate
MSLN /MPF	Mesothelin	Megakaryocyte potentiating factor	LTR as 1 of 2 promoters & part of 5' UTR of other transcript form; alt. promoter is an MIR	Yes for both	Widespread	Widespread but not from LTR or MIR
BAAT	Bile acid CoA	Bile metab. (hypercholanemia)	LTR only known promoter	No	Liver	Liver
MAD1L1	Mitotic arrest-deficient 1 like 1	Cell cycle regulation (cancer)	LTR as 1 of 2 promoters & part of 5' UTR of other form	No	LTR form in tumors. Other form widespread	Widespread
CLDN14	Claudin-14	Tight junct component (deafness)	LTR as 1 of 2 promoters	No	LTR form: melanoma /skin and kidney, other form in liver	Widespread
SIAT1	Sialyltrasferase 1	Humoral immunity	ERV one of at least 3 promoters	No	ERV form in mature B cells. Other forms in various tissues	B- cell & liver exp. from multi-promoters
FUT5	Fucosyltransferase-5	Cell adhesion	Antisense Alu/L1 as only known promoter	N/A	Colon, liver. Much lower exp. compared to related FUTs	No mouse ortholog
ILT2/ LIR1/ LILRB1	Ig-like transcript -2	Immune inhibitory receptor	ERV as alternate promoter	N/A	Only ILT known to be expressed in natural killer cells	ILTs expanded after human-mouse split

[a]Presense of TE in mouse was determined by Genome Browser annotation, BLAST and dotplot alignments.
[b]Information from literature, where available, or expression databases. Expression pattern of the TE-initiated form is given if known.
[c]Information from literature or databases with attention to patterns that differ from human.

One of the most striking examples of a TE insertion involved in new tissue-specific expression is the CYP19 gene (Table 3.1 and Figure 3.2a). CYP19 encodes aromatase P450, the key enzyme in estrogen biosynthesis and is expressed only in the gonads and brain of most mammals but the primate gene is also expressed at high levels in the syncytiotrophoblast layer of the placenta (Kamat et al. 2002). Placental-specific transcription of CYP19 is driven by a well-characterized alternative promoter located ~100 kb upstream of the coding region (Kamat et al. 2002) and our analysis has revealed that this promoter is actually an endogenous long terminal repeat (LTR), a fact that has escaped previous notice (Figure 3.2a). Using genomic PCR, we found this LTR present in Old World monkeys and in one New World monkey (marmoset) (data not shown). Therefore, this insertion early during primate evolution appears to have provided a placental-specific promoter that assumed an important role in transcription of CYP19 and, consequently, in controlling estrogen levels during pregnancy.

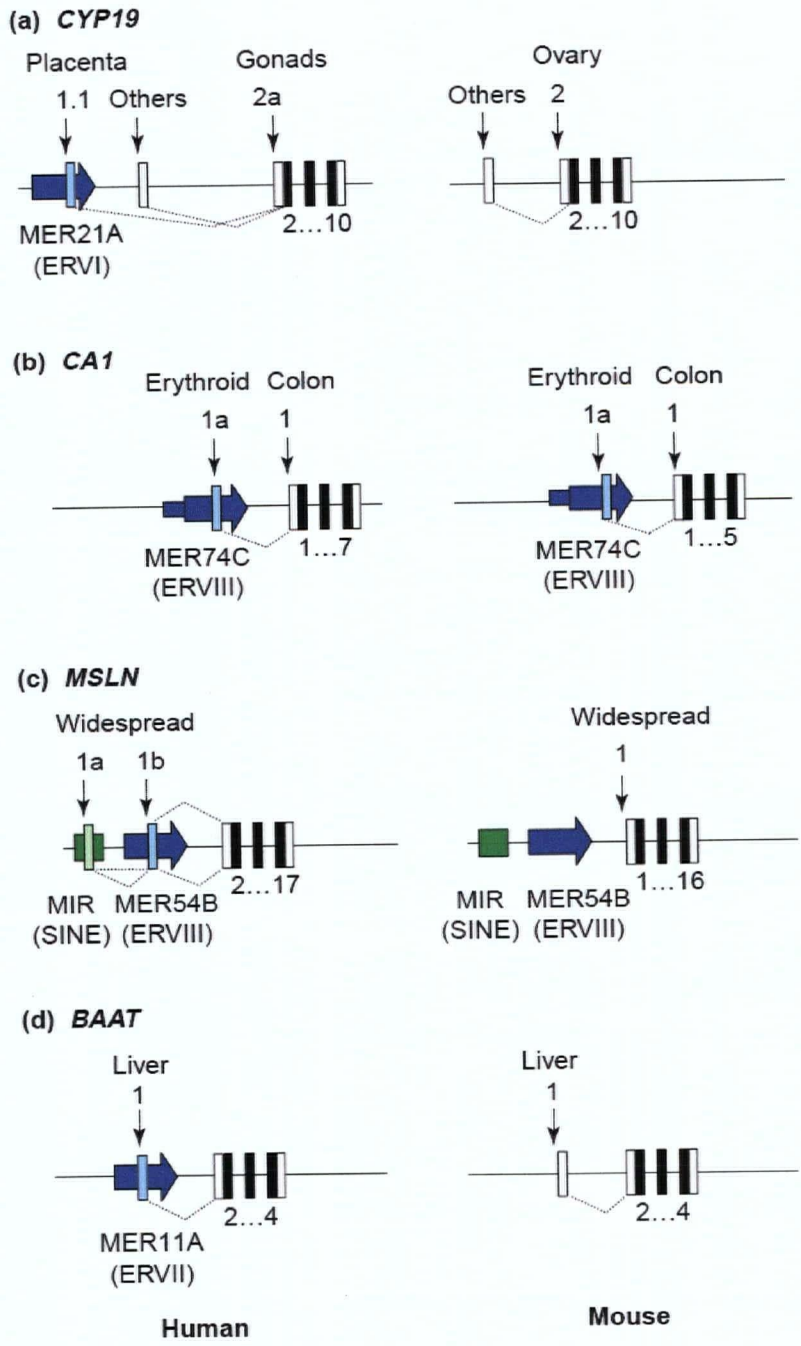

Figure 3.2 Examples of genes with apparent TE-derived promoters. Exon sequences derived from TEs are depicted by boxes shaded in a similar color as the element they are derived from. SINE elements are shown in green and retrovirus-like (ERV) sequences are in blue, where the thick arrows represent LTRs. The specific type of element is indicated below. Protein coding exons are represented by black boxes and non TE-derived UTRs are indicated by white boxes. Splicing of the alternative first exons is represented by broken lines. For the MSLN gene (part c), transcripts with the 1a exon either splice or read-through to exon 1b. Alternative transcription start sites are illustrated by black vertical arrows and tissue-specific expression patterns are indicated above each promoter. Gene expression patterns were deduced from the literature and/or database sources. See Supplementary Table of van de Lagemaat et al. (2003). The figure is not drawn to scale.

Many other instances of putative TE-derived promoters or polyadenylation signals are worthy of mention. The high levels of carbonic anhydrase in human and mouse red blood cells (Brady et al. 1989) appear to be due to an LTR-derived promoter of the CA1 gene (Figure 3.2b). SPAM1 and HYAL4, closely linked members of the same hyaluronidase gene family (Csoka et al. 2001), have different putative TE promoters giving rise to different expression patterns. For SPAM1, the ERV element is mostly deleted in mouse compared to human but the promoter region is retained, suggesting functional conservation of this segment. The human MSLN (or MPF) gene (Urwin and Lake 2000) appears to have two promoters, both of which are TE-derived. Both TE insertions are present in the mouse and rat genomes but we found no transcripts in the databases initiating from either TE in rodents (Figure 3.2c). The only apparent promoter of the liver-specific BAAT gene, which has recently been implicated in familial hypercholanemia (Carlton et al. 2003), is an ancient LTR in human but not in mouse (Figure 3.2d). FUT5 and ILT2 belong to gene families that amplified after the mouse-human divergence (Cameron et al. 1995; Martin et al. 2002). As shown in Table 3.1, these genes have putative TE-derived promoters and have acquired an expression pattern distinct from other family members, although it is not known if the TE is the cause of the differential expression. We also found many examples of TEs serving as polyadenylation sites. Disease-associated genes with primary transcripts terminating in a TE include the F8 (factor 8) gene, which is polyadenylated in an LTR and the ING1 (or p33ING1) tumor suppressor gene, which ends in a DNA TE.

We (Chapter 2, Medstrand et al. 2002) and others (Smit 1999) have observed that some classes of TEs found within introns of genes are more likely to be oriented in the

antisense transcriptional direction. This is particularly true for LTR elements and L1 sequences and is thought to reflect the fact that regulatory motifs such as polyadenylation signals within these elements are more likely to be detrimental by, for example, leading to truncated proteins (and thus less likely to be fixed) if oriented in the same direction as the gene. This study revealed that, in contrast to their intronic antisense orientation bias, LTR elements located at the 5' and 3' termini of both human and mouse UTRs are significantly more likely to be oriented in the same transcriptional direction as the gene transcript (Figure 3.1b). This observation supports the concept that LTR elements at transcript termini are not merely tolerated by the gene but actually participate in transcript formation by providing promoters and polyadenylation signals which function only in the sense direction.

3.3.3 TE prevalence varies with gene class or function

To ascertain if certain types of genes were more or less likely to have TE-containing mature transcripts, we used several methods to classify genes. First we used the Gene Ontology database (http://www.geneontology.org/) to classify genes according to their biological process or molecular function and determined the fraction of human and mouse genes containing TEs in transcripts. A remarkably similar pattern in the two species emerged from this analysis. For several classes of genes, the fraction of TE-containing transcripts was significantly less than the overall average (Figure 3.3a, b). Members of these categories are involved in basic housekeeping functions and many are evolutionarily conserved with few identified paralogs (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). In contrast, genes involved in functions such as defense, stress response and response to external stimuli were more likely to have TEs than most of the other gene classes (Figure 3.3a, b).

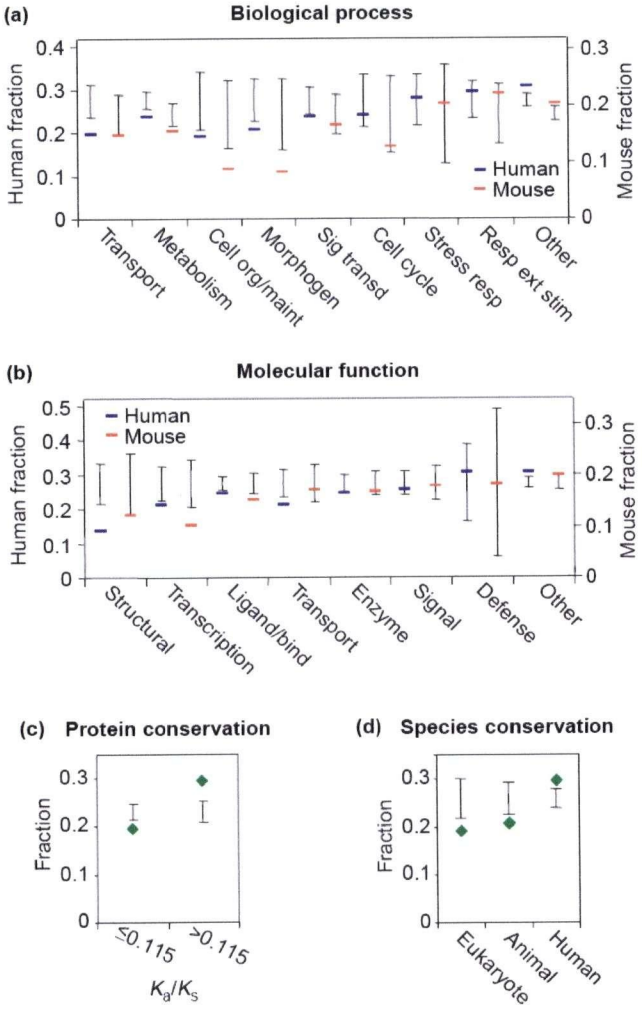

Figure 3.3 Prevalence of TEs in mRNAs of various gene classes. Symbols show the observed fraction and vertical bars represent the expected fraction of genes containing a TE sequence in the UTR. This expected fraction was determined by assuming a random distribution of TEs in UTRs of all genes and by considering the number of genes belonging to each class. The expected range or length of the bar is shown for $p<0.01$. (a) Gene classification by 'biological process' using the Gene Ontology (GO) database (http://www.geneontology.org). (b) Gene classification by GO 'molecular function'. For parts a and b, the scale for human genes (blue symbols) is indicated on the left and the mouse scale (orange symbols) is shown on the right of each panel. The 'other' category includes genes of unknown function as well as those of other functional groups. (c) Human and mouse genes separated into those having a Ka/Ks ratio less than or greater than 0.115 - the median ratio reported previously for all known human-mouse orthologues (Mouse Genome Sequencing Consortium 2002). Ka/Ks values were calculated for 7296 mouse-human gene pairs in our RefSeq dataset. (d) Genes grouped using the KOG database (euKaryotic Clusters of Orthologous Groups; http://www.ncbi.nlm.nih.gov/COG/new/shokog.cgi). Human: genes found only in human (as the sole mammalian representative in the KOG database) plus gene groups conserved in other eukaryotes but expanded to 10 or more members in human. Animal: genes conserved in *C. elegans, D. melanogaster*, and *H. sapiens*. Eukaryote: genes conserved in animal, *Arabidopsis thaliana* (mustard weed), *Saccharomyces cerevisiae* (budding yeast), *Schizosaccharomyces pombe* (fission yeast), and *Encephalitozoon cuniculi* (Microsporidia).

We next classified genes using the InterPro (IPR) database of functional protein domains (http://www.ebi.ac.uk/interpro/) and again found good agreement between human and mouse. TEs were either significantly enriched or reduced in genes containing 33 of the 80 most abundant IPR domains. Table 3.2 shows a list of those IPR domains with at least 20 genes in the human RefSeq database and where the observed fraction of genes with TEs in UTRs differs from the expected fraction, based on a random distribution of TEs ($p<0.01$). We found that transcripts of genes encoding Ig/MHC, C-type lectin or some cytokine domains were significantly enriched for TEs as were genes with KRAB/ Zn-finger transcription factor domains In contrast, genes important in development, transcription and replication and those with some enzymatic domains were much less likely to include TEs in their mRNAs (Table 3.2b).

Table 3.2 Domains associated with TE enrichment or exclusion in mRNAs

(a) Domains associated with TE enrichment

InterPro		Human		Mouse		Percentage of total genes[d]		
ID	Name	TE-free genes[b]	TE-genes[c]	TE-free genes	TE-genes	Human	Mouse	Fly
IPR001909	**KRAB box[a]**	36	79 (29) **	15	13 (4) **	**1.7**	**0.7**	0
IPR007087	Zn-finger, C2H2 type	202	149 (88) **	93	27 (18) **	**5.0**	3.0	3.6
IPR003006	**Immunoglob./ major histocompatibility complex**	157	102 (65) **	92	30 (18) **	**3.2**	**3.2**	1.7
IPR000276	Rhodopsin-like GPCR superfamily	80	60 (35) **	75	21 (14) **	**4.6**	**3.9**	1.1
IPR000315	Zn-finger, B-box	26	23 (12) **	6	7 (2) ** !	**0.5**	**0.3**	0.1
IPR001304	**C-type lectin**	34	27 (15) **	33	15 (7) **	0.5	**0.8**	0.5
IPR003877	SPla/RYanodine receptor SPRY	32	24 (14) **	6	6 (2) ** !	**0.5**	**0.3**	0.1
IPR002996	Cytokinereceptor, common beta/gamma chain	10	12 (6) **	12	5 (3) !	**0.2**	**0.3**	0

(b) Domains associated with TE exclusion

InterPro		Human		Mouse		Comparative domain coverage[f]		
ID	Name	TE-free genes	TE-genes	TE-free genes	TE-genes	Human	Mouse	Fly
IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	143	23 (42) **	55	10 (10)	1.6	1.6	1.7
IPR001356	Homeobox	93	13 (27) **	84	6 (13) **	1.4	**1.8**	1.2
IPR004046	Glutathione S-transferase, C-terminal	22	0 (6) **	12	1 (2) !	0.2	0.3	0.4
IPR000629	ATP-dependent helicase, DEAD-box	23	0 (6) **	9	0 (1) !	0.2	0.2	0.3
IPR000387	Tyrosine specific protein phosphatase and dual specificity protein phosphatase	53	6 (15) **	28	5 (5)	**0.6**	**0.7**	0.4
IPR000225	Armadillo repeat	28	1 (7) **	7	0 (1) !	0.2	0.3	0.2

[a]Domains in bold are under reduced purifying selection or increased diversifying selection in mammals (Mouse Genome Sequencing Consortium 2002).
[b]Observed number of genes without a TE in the UTR.
[c]Observed number of genes with a TE in the UTR and, in parenthesis, expected number of genes with a TE assuming a random distribution of TEs in UTRs of all genes.
**Indicates cases where the observed number differs from the expected with $p < 0.01$ (chi-squared). ! Indicates less than 20 genes in RefSeq.
[d]Percentage of total genes with the domain in the genomes of the species listed using the Ensembl gene classification (http://www.ensembl.org/). Significant domain expansions ($p<0.01$ for chi-squared considering the number of genes in each species) for human vs. fly and mouse vs. fly are indicated in bold.

3.3.4 *TEs are more prevalent in mRNAs of rapidly evolving and mammalian-specific genes*

TE prevalence in UTRs was next determined in genes separated according to their sequence conservation, measured by their Ka/Ks value, which is the ratio of the rate of nonsynonymous to synonymous change in coding sequences (Hurst 2002). A median Ka/Ks value of 0.115 for mouse-human orthologous gene pairs was recently determined (Mouse Genome Sequencing Consortium 2002) and, relative to this median, we found that genes with low Ka/Ks values are significantly less likely to have TEs in UTRs compared to those with values above the median (Figure 3.3c). These results indicate that genes with rapidly-evolving coding sequences are, in general, more likely to have TEs in their UTRs. Earlier analysis showed that at least eight functional domains are under increased positive diversifying selection or reduced purifying selection based on their high Ka/Ks ratio of >0.15 (Mouse Genome Sequencing Consortium 2002). Three of these domains are also significantly associated with genes with TE overrepresentation in their UTRs and these are bolded in Table 3.2a. Furthermore, it is noteworthy that 6 out of 8 domains associated with enrichment of TEs in genes (Table 3.2a) are represented at significantly higher numbers in mammalian genomes compared to the fruit fly (*Drosophila melanogaster*), whereas four of the six domains associated with a reduced TE-content in UTRs are equally represented in human and mouse vs. fly. Taken together, these data suggest that TEs are preferentially found in mRNAs containing rapidly diversifying domains, many of which have expanded during vertebrate evolution.

Finally, we examined TE prevalence in genes divided into three categories (eukaryotic, animal and human) based on presence of orthologous proteins in different

species using the euKaryotic Clusters of Orthologous Groups (KOG) database (http://www.ncbi.nlm.nih.gov/COG/new/shokog.cgi). This analysis showed that mammalian (human)-specific gene mRNAs are significantly enriched in TEs compared to transcripts of genes with orthologs in other animals and/or all eukaryotes. This enrichment was most apparent when we expanded our definition of mammalian-specific genes to include all genes with an ancient origin but which have expanded in humans (and likely other mammals) (Figure 3.3d). Transcripts of 'old' gene classes that are not expanded in mammals have a low prevalence of TEs, while genes specific to mammals or those associated with mammalian expansions are significantly more likely to harbor TE sequences in their mRNAs.

We considered two simple reasons for the above patterns. First, we addressed the possibility that TE prevalence in mRNAs is fully or partially dependent on genomic features rather than on gene function. For example, TEs are rarely found in UTRs of genes with homeobox domains (Table 3.2b), an expected observation given that the genomic regions encompassing the human and mouse homeobox gene clusters are nearly devoid of TEs (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002). We grouped genes by functional class and found that genomic parameters, such as number of TEs available, TE density, gene size, number of exons, length of introns, and local GC content, were insignificantly different between the functional groups. Correlation analysis confirmed this observation, demonstrating that a gene's functional category and its genomic surroundings independently influence the number of TEs within UTRs (data not shown). Second, we considered the possibility that transcripts enriched in TEs are more likely to be derived from non-functional but expressed pseudogenes. To address this possibility, we

separated RefSeq loci into two categories, 'reviewed' and 'provisional', which represent relatively well documented genes, and 'predicted', for which the support is less strong (http://www.ncbi.nlm.nih.gov/RefSeq/). We obtained similar TE frequencies with these two gene sets. Thus, neither genomic features nor the presence of pseudogenes fully account for the observed patterns shown in Figure 3.3 and Table 3.2. Rather, the data suggest that gene function and conservation act as independent variables in determining TE prevalence in mRNAs.

3.4 Conclusions

A growing appreciation for the role of TEs in genome evolution and gene regulation is evident from a number of recent studies (Murnane and Morales 1995; Sverdlov 1998; Brosius 1999; Hamdi et al. 2000; Makalowski 2000; Kidwell and Lisch 2001; Medstrand et al. 2001; Nekrutenko and Li 2001; Ostertag and Kazazian 2001; Deininger and Batzer 2002; Mouse Genome Sequencing Consortium 2002; Nigumann et al. 2002; Jordan et al. 2003; Kashkush et al. 2003). Here we have shown that highly conserved genes, such as those with essential functions in metabolism, development or cell structure, have a low prevalence of TEs in their mRNAs. This finding suggests, as might be predicted, that changes in expression of fundamental genes due to TE insertions cannot be allowed by the host and are strongly selected against. In contrast, younger or mammalian-specific genes, such as those involved in immunity and those that have expanded during mammalian evolution, are enriched for TEs in their mRNAs. It is possible that due to functional redundancy, such genes are initially more tolerant of TE insertions, some of which may then evolve a role in gene expression. These results suggest the TEs have had a major impact on the rapid evolution and functional diversification of gene families in humans and other mammals.

Chapter 4: Analysis of genic distributions of endogenous retroviral long terminal repeat families in humans

A version of this chapter has been submitted for publication:

van de Lagemaat, L.N., P. Medstrand, and D.L. Mager. 2006. Insertion patterns of endogenous retroviruses and SVA elements: Insights into their initial effects on genes.

I planned and performed all analyses in this paper and wrote the paper
P. M. was involved in planning this research and writing the paper
D. L. M. helped plan research and is the senior author

4.1 Introduction

Transposable elements (TEs), including endogenous retroviruses (ERVs), have profoundly affected eukaryotic genomes (Kidwell and Lisch 1997; Deininger and Batzer 2002; Kazazian 2004). Similar to exogenous retroviruses, ERV insertions can disrupt gene expression by causing aberrant splicing, premature polyadenylation, and oncogene activation resulting in pathogenesis (Boeke and Stoye 1997; Rosenberg and Jolicoeur 1997; Maksakova et al. 2006). While ERV activity in modern humans has apparently ceased, about 10% of characterized mouse mutations are due to ERV insertions (Maksakova et al. 2006). In rare cases, elements that become fixed in a population can provide enhancers (Ting et al. 1992), repressors (Carcedo et al. 2001), alternative promoters (Di Cristofano et al. 1995; Medstrand et al. 2001; Dunn et al. 2003; Jordan et al. 2003; van de Lagemaat et al. 2003, Chapter 3; Bannert and Kurth 2004; Leib-Mosch et al. 2005) and polyadenylation signals (Mager et al. 1999; Baust et al. 2000) to cellular genes due to transcriptional signals in their long terminal repeats (LTRs). It has been previously shown that LTRs/ERVs fixed in gene introns are preferentially oriented antisense to the gene's transcriptional direction (Smit 1999; Medstrand et al. 2002; Cutter et al. 2005). In contrast, studies on initial insertion patterns of exogenous retroviruses or retroviral vectors *in vitro* have not found any such bias for these unselected insertions (Schroder et al. 2002; Barr et al. 2005). Therefore, the antisense bias exhibited by fixed ERVs/LTRs in genes strongly suggests that retroviral elements found in the same transcriptional orientation within a gene are much more likely to have a negative effect and be eliminated from the population by selection. In this study, we closely examined genic distribution patterns of individual ERV families in the human genome and find

significant differences. These differences provide clues to the original activity profile of each element type, helping to explain nascence of biases in patterns of insertion.

4.2 Methods

4.2.1 Directional bias of insertions in transcribed regions in mice

We assessed the transcriptional orientation of Early Transposon (ETn) LTR retroelements in mouse transcribed regions. Retroelement and gene annotation from the April 2004 UCSC Mouse Genome Browser (Karolchik et al. 2003) was used to assess insertion frequency and orientation of insertions within the longest RefSeq transcribed regions of mouse genes. ETn LTR elements were represented by the RLTRETN family of ETn/MusD LTRs, and pairs of elements within 10 kb of each other and in the same orientation were assumed to belong to the same original insertion. The antisense bias observed in the genic ETn LTR population was then compared to genic orientation bias in a data set of documented mutagenic ETn/MusD LTR insertions coming from earlier studies (Baust et al. 2002; Mouse Genome Sequencing Consortium 2002; Maksakova et al. 2006).

4.2.2 Directional bias of retroelements in the human genome

RepeatMasker annotations of solitary LTRs and LTRs plus internal sequences of endogenous retroviral elements from the July 2003 UCSC Human Genome Browser were compiled and compared to annotated transcribed region start and end points of the per-chromosome longest transcript of each RefSeq gene, defined by its HUGO gene name.

Annotations for internal elements were matched with their respective LTRs as follows: HERVE or Harlequin internal sequences were matched with LTR2, LTR2B, or LTR2C; HERVK (HML-2) with LTR5, LTR5_Hs, LTR5A, or LTR5B; HERV17 with

LTR17 (where HERV17 represents HERV-W); HERVH with LTR7, LTR7A, or LTR7B; HERV9 with LTR12, LTR12B, LTR12C, LTR12D, LTR12E, or LTR12_; MLT2x with ERVL-x (where x is a unique identifier); MST* with MST*-int (where * represents a wildcard); THE1* with THE1*-int; and MLT1* with MLT1*-int. Groups of LTR element segments of the same type, with internal sequence all in the same orientation, and occurring within 10 kb were deemed part of the same composite element. Manual checks confirmed the validity of this criterion. Names of consensus elements occurring in each composite element were recorded, as well as names of the ERV type. Composite elements without internal sequence were deemed LTR-only, and elements with contributions from at least two consensus elements were deemed to contain LTR and internal sequence and therefore were considered full-length. Again, manual checking confirmed the validity of this criterion.

The assigned genomic position of each LTR or full-length element was computed as the average of the beginning and end coordinates of each composite element and compared against the positions of longest transcribed regions of each gene. Each transcribed region was divided into ten equal bins and the TE location within a gene was specified by which of these bins it fell into. In addition to the ten intragenic bins, two bins upstream and two bins downstream of the gene, of the same size as the intragenic bins, were also considered. Counts of elements for each orientation were computed for each bin.

A similar approach was used for computing SVA and Alu distributions across genes as for LTR elements.

4.3 Results and Discussion

4.3.1 Opposite orientation bias of fixed versus mutation-causing retroviral insertions

As mentioned above, *in vitro* studies of *de novo* retroviral insertions within gene introns have not detected any bias in proviral orientation with respect to the transcribed direction of the gene (Schroder et al. 2002; Barr et al. 2005). The fact that integrations that have not yet been tested for deleterious effect during organismal development show no directional bias indicates that the retroviral integration machinery itself does not distinguish between DNA strands in transcribed regions. Presumably, then, any orientation biases observed for endogenous retroviral elements must reflect the forces of selection. In support of this premise is a recent study by Bushman's group that was the first to directly compare genomic insertion patterns of exogenous avian leukosis virus (ALV) after infection *in vitro* with patterns of fixed endogenous elements of the same family (Barr et al. 2005). Endogenous elements in transcriptional units were four times more likely to be found antisense to the transcriptional direction, suggesting strong selection against ALV in the sense direction. We reasoned that, if the marked orientation biases of LTRs/ERVs were a reflection of detrimental impact by sense-oriented insertions, then we would also expect a dominant sense orientation among insertions with known detrimental effects. While no mutagenic or disease-causing ERV insertions are known in humans, significant numbers have been studied in the mouse. We analyzed element orientation of fixed, non-pathogenic insertions of the still active ETn/MusD family of ERVs in the mouse (Figure 4.1), and found similar degrees of antisense bias as seen for human and chicken ERVs (data not shown). However, as expected, 15/18 new mouse germ-line ETn/MusD insertions in transcribed regions that are associated with mutations are in the sense orientation (Maksakova et al. 2006) (Figure 4.1). Moreover, in

most of these cases, the predominant effect of the ERV was to cause premature polyadenylation of the gene through use of LTR polyA signals, accompanied by aberrant splicing (Maksakova et al. 2006).

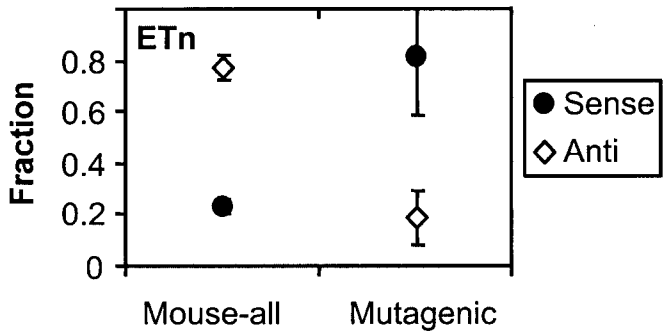

Figure 4.1 Directional bias of retroelements in mouse transcribed regions. ETn elements were those annotated as RLTRETN in the UCSC May 2004 mouse genome repeat annotation. The mutagenic population of ETn elements was reported in earlier reviews (Baust et al. 2002; Mouse Genome Sequencing Consortium 2002; Maksakova et al. 2006). Expected variability in the data was calculated from Poisson statistics, which describe randomized gene resampling.

4.3.2 Variation in density of genic insertions of different HERV families

ERVs/LTR elements in the human genome actually comprise hundreds of distinct families of different ages and structures, many of which remain poorly characterized (Gifford and Tristem 2003; Mager and Medstrand 2003). Thus, grouping such heterogeneous sequences together, as has been done for previous studies on orientation bias (Smit 1999; Medstrand et al. 2002), may well mask variable genomic effects of distinct families. To investigate genic insertion patterns of different human ERV families, we chose nine well studied Repbase-annotated (Jurka 2000) families or groups of related families to analyze in more detail. These families, their copy numbers and their approximate evolutionary ages are listed in Table 4.1.

Table 4.1 Annotated copy numbers and evolutionary ages of various ERV familes

Name	*copy number*[a]	*full length*[b]	*evolutionary age(Myr)*	*Reference*
MLT1	160,000	36,000	>100	(Smit 1993)
MST	34,000	5175	75	(Smit 1993)
THE1	37,000	9019	55	(Smit 1993)
HERV-L	25,000	4777	>80	(Cordonnier et al. 1995)
HERV-W	675	242	40-55	(Blond et al. 1999)
HERV-E	1138	294	25	(Taruscio et al. 2002)
HERV-H	2508	1284	>40	(Jern et al. 2004)
HERV9	4837	697	15	(Costas and Naveira 2000)
HERV-K (HML2)	1206	178	30	(Bannert and Kurth 2004; Belshaw et al. 2005)

[a]Including LTRs with no internal sequence and LTRs with associated internal sequence
[b]Elements including both LTR and internal sequence

As a first step, we plotted the fraction of total elements in either orientation found within RefSeq (http:// www.ncbi.nlm.nih.gov/RefSeq/) transcriptional units (see Methods) and the results are shown in Figure 4.2. To put our results in context, we considered a model of random initial integration throughout the genome. Since 34% of the sequenced genome falls within our analyzed set of RefSeq transcriptional units, we would expect 34% of ERV insertions, 17% in either direction, to be found in these regions. This is a conservative model since initial integration patterns of most exogenous retroviruses are biased toward genic regions (Panet and Cedar 1977; Schroder et al. 2002; Mitchell et al. 2004; Barr et al. 2005) (see also below). An earlier study (Medstrand et al. 2002) noted that, for large superfamilies of LTR elements, elements in either orientation are less prevalent in genic regions than expected by chance. In the present study, we were particularly interested in the fact that, for many LTR types, there are fewer antisense LTRs than expected by chance. This may be a reflection of enhancer effects by these

elements, an effect often seen in mice and reviewed elsewhere (Rosenberg and Jolicoeur 1997). Closer analysis of our chosen individual families revealed significant variation in the magnitude of this effect (Figure 4.2). For example, antisense LTRs of the MLT1 and HERV-K (HML-2) families are relatively more prevalent in genes, suggesting that the presence of these sequences in the antisense direction is less likely to negatively affect the enclosing gene. An alternative explanation is that the initial integration preference of these families was biased more heavily to transcriptional units, compared to most other families. The density patterns of most of the other families are qualitatively similar to each other, with a moderate, though significant, under-representation of antisense elements and a further 2 to 3 fold reduction in sense elements. Similarly, a recent study of a chimpanzee-specific family of gammaretroviruses by Yohn et al (2005) showed that all 13 of these ERVs found within transcribed regions were oriented antisense to the direction of gene transcription. The exception to this pattern is HERV9 (ERV9), which will be discussed further below.

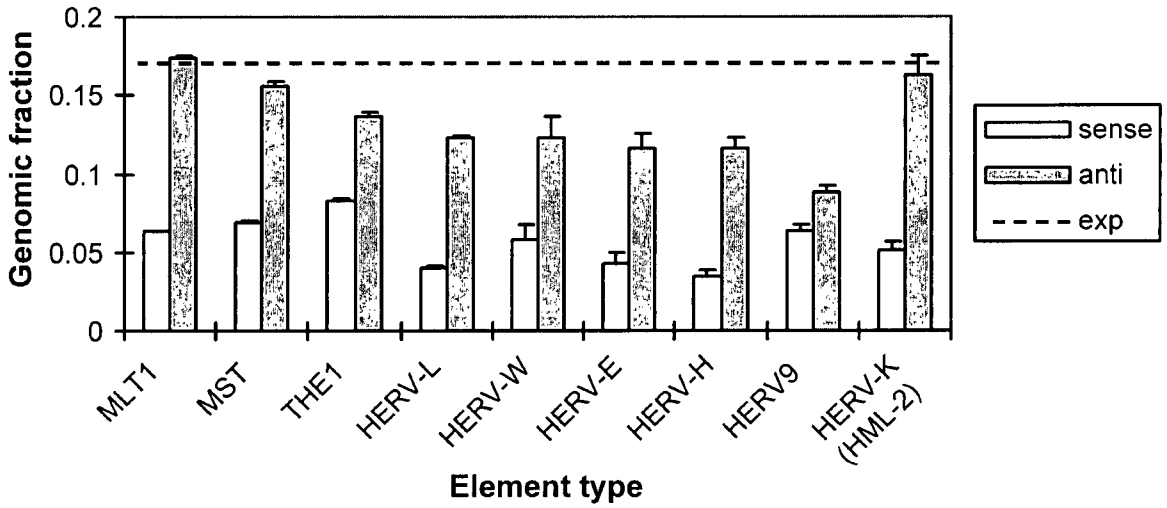

Figure 4.2 Orientation bias of various full length ERV sequences in genes. ERV families are as annotated by RepeatMasker in the human genome and are listed in Table 4.1. Fraction of all genomic elements actually found in genes in the sense and antisense orientations is presented, with neutral prediction (dotted line) based on fraction of total genomic elements expected in sense and antisense directions in genes under assumption of uniform random insertion.

4.3.3 Density profiles of ERVs across transcriptional units

At least three factors could account for the antisense bias exhibited by most ERV families. First, the sense-oriented polyadenylation signal in the LTR could cause premature termination of transcripts and therefore be subject to negative selection. Gene transcript termination within LTRs commonly occurs in ERV-induced mouse mutations (Maksakova et al. 2006) and this effect has been proposed as the most likely explanation for the orientation bias (Smit 1999).

Second, splice signals within the interior of proviruses could induce aberrant RNA processing, a phenomenon also frequently observed in mouse mutations (Maksakova et al. 2006). To test this second possibility, we plotted graphs similar to Figure 4.2 separately for solitary LTRs, which comprise the majority of retroviral elements in the genome (International Human Genome Sequencing Consortium 2001; Mager and Medstrand 2003), and for composite elements containing LTR and internal sequence (data not shown). While the numbers of the latter are much lower than for solitary LTRs for most families, we detected no significant differences in the density patterns, suggesting that signals within the LTRs, and not interior splice signals, are the primary determinant of the orientation bias.

A third factor that could result in orientation bias is the presence of the LTR transcriptional promoter with a potential to cause ectopic expression of the gene, resulting in detrimental consequences, as occurs in cases of oncogene activation by retroviruses (Rosenberg and Jolicoeur 1997). If introduction of an LTR promoter was the primary target of negative selection, one would predict that sense-oriented LTRs located just 5' or 3' to a gene's native promoter would be equally damaging and therefore subject to similar degrees of selection.

To look for evidence of these effects, and to more closely examine the distributions of ERVs within genes, we measured the absolute numbers of ERVs/LTRs of different families in bins across the length of RefSeq transcriptional units and in equal-sized bins upstream and downstream (see Methods). The results of this analysis (Figure 4.3) revealed density profiles that shift dramatically at gene borders. Specifically, for most ERV families, we found that the prevalence of sense-oriented elements drops markedly inside the 5' terminus of a gene, remains relatively low across the gene and then jumps just as markedly 3' of the gene. This type of pattern does not indicate a strong detrimental effect of LTR sense-oriented promoter motifs. Rather, these profiles suggest that presence of the LTR polyadenylation signal, which would generally only affect a gene if located within its transcriptional borders, is the regulatory signal primarily responsible for the resulting lack of sense-oriented LTR elements within introns.

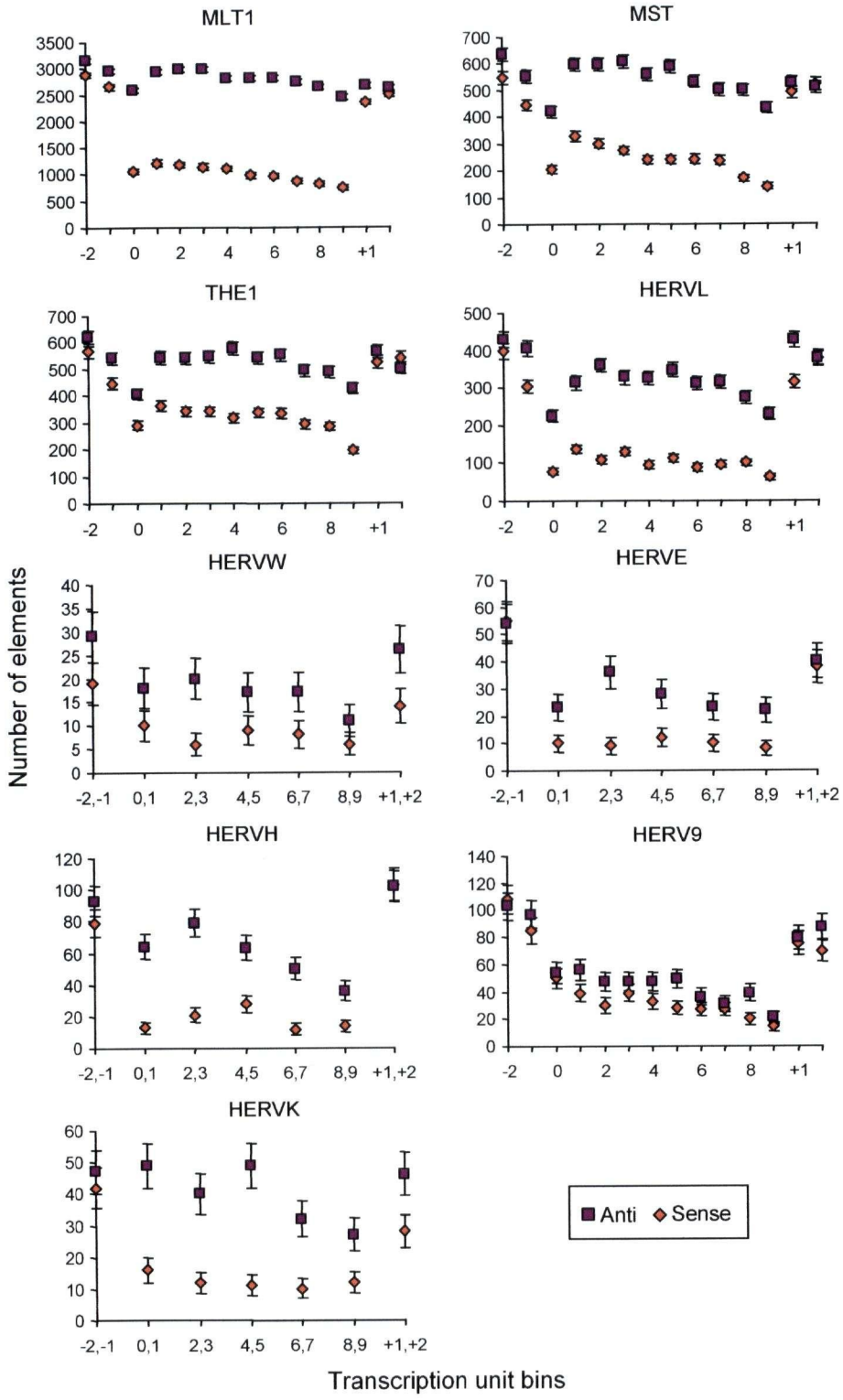

Figure 4.3 Patterns of annotated ERV presence in equal-sized bins across transcriptional units. Ten bins, numbered 0-9, were considered within transcribed regions. Four bins, two in either direction outside gene borders and equal in length to intragenic bins, were considered, and are shown as bins -2 and -1 upstream and +1 and +2 downstream.

4.3.4 Distinct pattern of HERV9 elements with respect to genes

By separating ERVs into different families, we uncovered a unique genic distribution pattern of HERV9 elements. As shown in Figure 4.2, HERV9 has a more significant deficit of antisense elements and little orientation bias compared to other ERV families. The distinct HERV9 density profile across gene regions illustrates the same point in greater detail (Figure 4.3H). These data suggest that HERV9 elements within introns are nearly equally likely to adversely affect the gene, regardless of orientation. This is the only HERV family we have analyzed that displays this type of distribution pattern and is likely the result of the complex structure of HERV9 LTRs. These LTRs are 0.7-1.5 kb long and extraordinarily rich in the CpG dinucleotide – examination of the consensus elements from RepBase (Jurka 2000) reveals that all HERV9 LTR (LTR12, in RepBase nomenclature) subfamily members are CpG rich, with approximately 90 CpGs spread over the consensus of the most-abundant LTR12C LTR. A relatively CpG-poor tract in the LTR's U3 region contains 5-17 repeats of a sequence rich in transcription factor binding sites, which have recently been shown to bind a transcription factor complex involved in the regulation of the beta globin locus (Yu et al. 2005).

HERV9s are under-represented in both orientations in all bins within transcribed regions compared to the nearest regions upstream and downstream of genes. These results suggest exclusion of these ERVs from genic regions in both orientations, likely due to a transcription defect similar to simple polyadenylation. However, analysis for polyadenylation signals using DNAFSMiner (Liu et al. 2005) reveals the presence of polyadenylation signals in all LTR12 family consensus sequences and the absence of a polyadenylation signal on the opposite strand, suggesting that polyadenylation alone cannot account for this distribution pattern. One possible explanation comes from recent

work by Lorincz et al (2004), which showed that methylated intragenic CpG dinucleotides were associated with transcriptional elongation defects, likely due to induction of a closed chromatin formation. This observation, coupled with the fact that HERV9 LTRs are CpG rich, has the potential to explain strong selection against intragenic insertions of HERV9 LTRs in either orientation.

4.3.5 SVA SINE elements display distribution patterns similar to LTRs

SVA elements are composite SINE sequences composed of a tandem hexamer repeat, a partial Alu element, a variable number of tandem repeats, a partial ERV-K LTR, and a poly-A tail (Ono et al. 1987; Zhu et al. 1992; Shen et al. 1994; Wang et al. 2005). These elements contain an internal promoter and the LTR-derived poly-A signal. Like Alu elements (Dewannieux et al. 2003), SVA elements are thought to utilize L1-encoded machinery to retrotranspose (Wang et al. 2005). SVA elements are a relatively young and actively transposing family in humans and have caused several mutations (Ostertag et al. 2003; Chen et al. 2005). These elements have a stronger antisense bias in genic regions, compared to AluY elements (Figure 4.4A, C), suggesting interference with pol II transcriptional machinery. Similar to *in vitro* insertions of exogenous viruses (discussed above), antisense SVAs are found in transcribed regions more frequently than expected by random chance, suggesting that genic regions are favored targets for SVA insertions. Analysis of the profiles of sense and antisense SVA elements across transcription units revealed that, as with LTR elements, there was a drastic change in antisense bias at the boundaries of transcribed regions, mostly due to a sudden drop in density of sense-oriented insertions. This low density of sense-oriented insertions persisted across transcriptional units (Figure 4.4B), again suggesting that polyadenylation plays a significant role in deleterious consequences of germline SVA insertions.

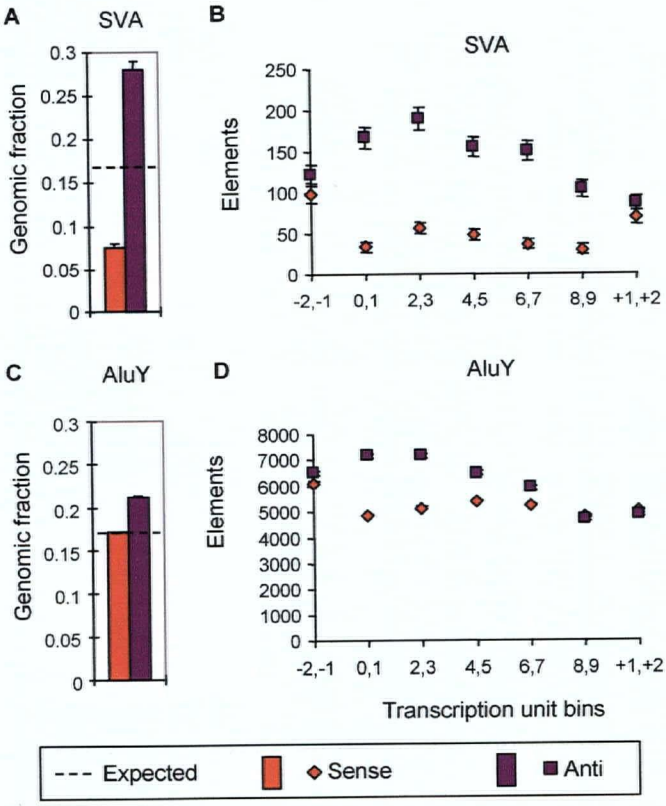

Figure 4.4 Insertion pattern of SVA and AluY retroelements across transcriptional units. A,C. Fraction of annotated genomic SVA and AluY elements found in the sense and antisense directions in transcribed regions. Dotted line shows expected fraction (17%) assuming initially uniform random insertion. B, D. Cumulative insertions of SVA and AluY elements in ten bins, numbered 0-9, across transcriptional units. Similar to Figure 4.3, four extra bins were considered, two in either direction upstream and downstream of genes, and denoted -2, -1, +1 and +2.

In contrast to SVAs, the AluYs, which comprise the youngest superfamily of Alu elements (International Human Genome Sequencing Consortium 2001), have a significantly smaller antisense bias, both overall, and across transcriptional units (Figure 4.4C, D). Given the similar mechanism by which SVAs and Alus have likely inserted, these results suggest that intronic insertions of Alus in both orientations have been significantly less likely to be selected against than sense-oriented insertions of SVAs. An intriguing feature of both the SVA and Alu distribution profiles across transcriptional

units is the slightly higher prevalence of sense-oriented elements in the 5' regions of genes (bins 0-3) compared to more 3' regions. This pattern could reflect original insertion site preferences favoring the 5' parts of genes. The biochemical mechanisms are unclear but could be analyzed using *in vitro* retrotransposition assays.

It is interesting to note that SVAs, which also have a large numbers of CpG dinucleotides, do not show a similar pattern to HERV9 LTRs. CpG dinucleotides in genomic SVA elements do appear to be methylated, given their accelerated mutation rate relative to non-CpG sites (Wang et al. 2005). Why these elements fail to cause a potential elongation defect is unclear and requires further analysis, perhaps using experimental approaches.

4.4 Concluding remarks

The preferential antisense orientation of LTRs/ERVs fixed in gene introns has been shown before (Smit 1999; Medstrand et al. 2002; Cutter et al. 2005). However, patterns of *in vitro* insertions by exogenous retroviruses have not demonstrated this bias (Schroder et al. 2002; Barr et al. 2005). Using patterns of insertions across transcribed regions for individual families, we have shown that individual ERVs have had differing impacts upon original insertion. A feature that most share, however, is deleterious impact by the polyadenylation signal, evidenced by the sharp increase in antisense bias immediately downstream of the start of transcription, corresponding to sharp decrease in the density of sense-oriented insertions. The anomalous insertion pattern of HERV9 in transcribed regions suggests an adverse impact on genes regardless of orientation, perhaps due to induction of closed chromatin as a result of methylation of its many CpG dinucleotides. However SVA elements, which are similarly rich in the CpG dinucleotide, show a robust antisense bias. In conclusion, although some functions of LTRs, primarily their

promoters, may have been inactivated or repressed, modern patterns of insertions can be used to deduce original selective forces acting on these elements. Furthermore, LTR sequence, presumably mobilized as part of the still active SVA SINE, continues to have an LTR-like impact on the human genome.

Chapter 5: Analysis of repeats and genomic stability

A version of this chapter has been published:
van de Lagemaat, L.N., L. Gagnier, P. Medstrand, and D.L. Mager. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Res* **15:** 1243-1249.

I performed all bioinformatic analysis and wrote sections of the paper.
L. G. performed PCR assays.
P. M. and D. L. M. discussed research and wrote sections of the paper.

5.1 Introduction

Current genome size in mammals and other eukaryotes has been greatly affected by massive amplifications of transposable elements (TEs) or retroelements throughout evolution (Brosius 1999; Kidwell 2002; Liu et al. 2003). In mammals, close to 50% of the genome is recognizably TE-derived (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002; Rat Genome Sequencing Project Consortium 2004) and in some plant species, the figure is nearly 80% (SanMiguel et al. 1998; Li et al. 2004). The various classes of TEs and their distributions in genomes have been widely studied in many species (Adams et al. 2000; International Human Genome Sequencing Consortium 2001; Aparicio et al. 2002; Kidwell 2002; Mouse Genome Sequencing Consortium 2002; Yu et al. 2002; Kirkness et al. 2003; Baillie et al. 2004; Ma and Bennetzen 2004; Rat Genome Sequencing Project Consortium 2004). In contrast, much less is known about mechanisms that attenuate genome size. Studies in plants have shown that retroelement-driven genome expansion is counteracted by deletions within retroelements, likely mediated by illegitimate recombination between short flanking segments of identity (Devos et al. 2002). Comparison of related rice genomes has also revealed that illegitimate recombination has deleted both retroelement-derived sequences as well as unique nuclear DNA (Ma and Bennetzen 2004).

A number of studies have documented the prevalence of small deletions and insertions (indels) in primate genomes (Britten et al. 2003; Liu et al. 2003; Watanabe et al. 2004) but there has been no genome-wide analysis to determine the molecular mechanisms which generate these events. Recent availability of the chimpanzee draft

sequence has afforded the opportunity to analyze the spectrum of genomic deletions that have occurred in the last 5-6 million years of primate evolution. Moreover, a large-scale comparison of the human and chimpanzee genomes allows examination of the genomic stability of retroelement insertions, which are generally considered to be irreversible with no known mechanism for precise excision from the genome (Hamdi et al. 1999; Roy-Engel et al. 2001; Batzer and Deininger 2002; Salem et al. 2003a; Salem et al. 2003b). Due to this 'unidirectional' property, retroelements, particularly Alu elements, are widely viewed as ideal markers for human population genetic studies (Carroll et al. 2001; Roy-Engel et al. 2001; Batzer and Deininger 2002; Salem et al. 2003a) and elucidation of primate phylogenetic relationships (Hamdi et al. 1999; Salem et al. 2003b; Gibbons et al. 2004). In primates, Alu sequences are the most abundant family of retroelements, comprising over 10% of the human genome (International Human Genome Sequencing Consortium 2001; Batzer and Deininger 2002). While most of the one million Alu elements retrotransposed over 40 million years ago, several thousand have integrated into the human genome since divergence from the great apes and close to a thousand of the youngest Alus are polymorphic (Carroll et al. 2001; Roy-Engel et al. 2001; Batzer and Deininger 2002; Salem et al. 2003a; Bennett et al. 2004). Most are associated with flanking direct repeats or target site duplications (TSDs) of 10-20 bp (Jurka 1997). In this study, we have obtained evidence that Alu elements can be precisely deleted from the genome via recombination between these flanking repeats. Similarly, a significant fraction of 200-500 bp deletions of non-repetitive sequence have likely taken place due to recombination between short regions of identical sequence flanking the deleted fragment. We demonstrate that this fraction is much greater than expected if blunt-end joining were responsible for generating all these deletions. Our results are in agreement with a model

of genomic deletion occurring both by non-homologous and error-prone homology-driven mechanisms of DNA double strand break repair (Helleday 2003).

5.2 Methods

5.2.1 Direct assessment of retroelement deletion rate

Putative retroelement insertions were obtained from the chimpanzee scaffold alignments to the UCSC July 2003 human genome (Kent et al. 2002) using RepeatMasker (A.F.A. Smit & P. Green, unpublished data), MaskerAid (Bedell et al. 2000), and libraries from the RepBase Update (Jurka 2000). Pseudogenes were detected using BLAT (Kent 2002) and the human RefSeq mRNA records. Insertions were defined as having a single retroelement (including pseudogenes) filling all but up to 90 bp of the indel and not extending beyond the indel by more than 10 bp on either side. Search queries were then constructed of the 50-bp sequences upstream and downstream of each putative retroelement insertion location. Scripted discontiguous megablast searches of the relevant NCBI trace archive were then carried out using perl scripts and the QBLAST application programming interface (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi; http://www.ncbi.nlm.nih.gov/BLAST/Doc/urlapi.html) (Altschul et al. 1990; McGinnis and Madden 2004). Our BLAST queries used a non-coding template of size 21 and required only one seed hit per high-scoring segment pair. To minimize false positives and ensure non-redundant hits, we required that 75% of the query be free of known human repeats. Further, the accepted hits were required to match the query at least 30 bp on either side of the putative breakpoint. All traces not fulfilling these requirements were ignored. Deletions in human relative to chimpanzee or vice versa were diagnosed by the presence in Rhesus of an insertion at least 80% of the size expected and no traces with

less than this amount of sequence. A site was considered an insertion if one or more Rhesus traces matched the empty site and no traces had extra sequence. Putative deletions in human or chimpanzee were further individually aligned with their Rhesus counterpart using ClustalW version 1.82 (Higgins et al. 1996) and the alignments were edited using Jalview (Clamp et al. 2004) to check for the presence of the same element in the expected position in the Rhesus trace. The alignments are provided in Appendix A.

5.2.2 Detection of deletions internal to Alu elements

All indel loci in the chimpanzee scaffold alignments to the UCSC July 2003 human genome, masked as described above, were reanalyzed. Deletions occurring entirely within Alu elements were analyzed for involvement of the approximately 80 bp and 50 bp internal homologies. Putative deletions occurring between the 80 bp similar regions were detected by having one deletion endpoint occurring within positions 1-84 of the consensus and the other endpoint within positions 136 to 219. Similarly deletions between positions 85-135 and 219 to the end of the consensus were considered as occurring between the 50-bp similar regions.

5.2.3 Assessment of deletion frequency due to illegitimate recombination

Human chromosomal sequence files with human repeats pre-masked to lower case by RepeatMasker were used. These files were further masked using Tandem Repeats Finder 3.21 (Benson 1999). All repetitive sequence was excised, including human repeats and tandem repeats. We then constructed a C++ program that used an alignment method to find all nonredundant nonadjacent identical segments up to 20 bp long between 200 and 500 bp apart in the nonrepetitive genome. These were tallied by length, giving the expected distribution of potential sites for illegitimate recombination in this distance

range.

We then analyzed all fully-sequenced insertions and deletions (indels) 200-500 bp long present in the alignments of the UCSC July 2003 human sequence to the chimpanzee scaffolds (Kent et al. 2002) for the presence of flanking identical segments beginning at the deletion breakpoints. Specifically, indels with 50 bp flanking sequence on each were analyzed, and those containing putative new retroelement insertions, tandem duplications, or flanking homologous retroelements corresponding to the indel breakpoints were removed from consideration, as were all indels occurring inside transposable elements. The remaining indels were classified as deletions.

Retroelement insertions were detected as described above. Other insertions were diagnosed if the indel internal sequence was found by BLAT elsewhere in the human genome. Tandem duplications were diagnosed by running Tandem Repeats Finder (Benson 1999) on a sequence including the indel extra sequence and equivalent lengths of sequence flanking the indel upstream and downstream. An indel was considered to be a case of tandem duplication if a tandem duplication was found covering one of the breakpoints and extending to within one bp of the other. Manual checks confirmed the validity of this criterion. After disqualifying putative insertions, tandem duplications, and indels with flanking homologous repeats, the remaining 1927 indels were termed random deletions and were analyzed for the distribution of flanking repeat sizes. Flanking repeats were considered to begin at the breakpoint positions and consist of a tract of identical sequence. No mismatches in the flanking repeat or offsets of the identical segment from the indel breakpoints were allowed.

5.2.4 Genomic PCR and sequencing

Primate genomic DNA was isolated from various cell lines as described

previously (Goodchild et al. 1993). Additional chimpanzee DNA samples were kindly provided by Dr. Peter Parham (Stanford University). 150 ng of human or primate genomic DNA was amplified in a 50μl reaction with 200μM each dNTP, 200nM each primer (see Appendix A), 1.5mM MgCl2, and 1 unit of Platinum Taq (Invitrogen) in 1X PCR buffer (Invitrogen). The conditions for the PCR were 94°C for 1 min followed by 30 cycles of the amplification step (94°C for 30 s, 48-60°C for 30 s, and 72°C for 30s-1 min). The annealing temperature and extension time varied for different primer combinations. Sequencing was performed directly on PCR products using the BigDye Terminator v3.1 Cycle Sequencing Kit (ABI) in an ABI PRISM® 3730XL DNA Analyzer system at the McGill University sequencing facility.

5.3 Results and Discussion

5.3.1 Direct assessment of retroelement deletion frequency

During an analysis to identify transposable element (TE) insertions that occurred after divergence of human and chimpanzee, we detected some apparent insertional differences involving Alu elements of older subfamilies. The AluY subfamily is the only family known to have been active in the last few million years of human evolution (Batzer and Deininger 2002). However, we identified 187 Alu elements from older families such as AluS and AluJ (98 in human and 89 in chimpanzee) which appeared to be insertional differences. This finding raised the possibility that at least some of these cases represent deletions in one species rather than new insertions in the other. To explore this possibility, scripted blast searches of the Rhesus macaque whole-genome shotgun trace archive were used to assess the ancestral state of apparent retroelement insertional

differences in humans and chimpanzees (see Methods). It should be noted that our requirement that 75% of the totally 100 bp flanking sequence be free of known repeats resulted in only 8389 of 14765 retroelement loci being tested, and therefore we expect that our findings represent an underestimate of the overall level of precise deletion of retroelements.

Of 7120 human-chimp indel sites with accepted Rhesus trace matches, 7010 were identified as insertions by our criteria (see Methods). That is, the retroelement was absent in Rhesus. The other 110 sites were examined more closely. Fifty-two of these cases appeared to be rearrangements or multi-copy regions in the Rhesus genome due to the existence of multiple Rhesus traces covering the region, some with and some without the retroelement. Three further cases with partial poor trace alignments were likely genomic rearrangements. The remaining 55 cases were subjected to more detailed analysis to confirm that the indel was a case of deletion in human or chimpanzee and not an insertion or other rearrangement.

Multiple sequence alignments of the human, chimpanzee, and Rhesus sequences were done in each of the 55 cases (reproduced in Appendix A). Only one (# 23) resulted from poor sequence quality in the chimpanzee assembly. Another (#51) was a tandemly-duplicated L2 element. Four other cases (# 5, 13, 18, and 31) showed evidence of independent insertions in the same site or in sites only several base pairs apart. Independent insertions at the same site have been reported before (Conley et al. 2005).

The remaining 49 cases appeared to be retroelement deletions. Twelve cases, six in humans and six in chimpanzees, were imprecise deletions, removing sequence from older retroelements such as L2 and MIR. A similar case of imprecise Alu deletion has been previously reported (Edwards and Gibbs 1992). In each case, our 12 imprecise

deletions had little or no similarity at the deletion breakpoints, suggesting a non-homologous deletion mechanism as an explanation for these events.

Thirty-seven cases represented apparent precise deletion of previously retrotransposed sequence and all cases but one were Alu elements. The one anomaly (case #6) was a polyadenylated sequence flanked by apparent target site duplications. This is a fragment of a ~340 bp sequence with ~20 copies mutually ~6-10% divergent in the human genome, suggesting possible earlier mobilization as a retrotransposable element.

We found 36 cases of apparent precise deletions of Alu elements. The loss of the Alu was also associated with loss of one copy of the TSD, leaving behind the original, pre-integration site only. This observation raised the possibility that these deletions were mediated by recombination between the flanking identical regions. A possible example of precise Alu deletion on human chromosome 21 has been reported recently by Hedges et al (Hedges et al. 2004) but the authors considered Alu excision to be a remote possibility, instead favoring other explanations. Unfortunately, there is no coverage of this region in the chimpanzee scaffolds. Furthermore, recent PCR analysis of human-chimpanzee indels on chimpanzee chromosome 22 revealed two precisely deleted Alu elements; however, sequences and positions of these events were not given. These deletions resulted in loss of the Alu and deletion of one of the TSD copies, leading the authors to speculate that a homology-dependent recombination mechanism might be responsible for these deletions (Watanabe et al. 2004).

We reasoned that under a null hypothesis of deletions mediated by non-homologous mechanisms, very few should be flanked by short identical segments. Instead, the majority of the 49 deletions (37 with flanking identical segments and 12

without) had identical regions of 10 bp or more. Compared to the null hypothesis, this association between deletion and flanking identical DNA was highly significant ($p < 1e-100$; Chi squared test). The skeptical reader could argue that we were only looking at deletions with breakpoints near retroelements, and therefore we would be more likely to find breakpoints located within TSDs, even with a non-homologous deletion mechanism. However, the likelihood of locating the breakpoints precisely at the same location within the TSD in the vast majority of the cases by random chance alone remains extremely small. Our findings strongly suggest that short, nonadjacent identical segments recombine, likely during double-strand break repair, to mediate deletion of these sequences. Consistent with this notion is the fact that at least 20-fold more deletions that involve Alus are actually internal to Alu elements and have occurred between the roughly 80-bp and 50-bp homologous regions internal to intact Alu elements (Figure 5.1A; See Methods). These findings suggest that double strand DNA breaks internal to Alus are repaired using the internal Alu homologies, obviating use of the flanking TSDs as repair templates and thus retaining remnants of the Alu element. The proposed mechanism of double-strand break repair is illustrated in Figure 5.1B, which shows a specific non-Alu small deletion in chimpanzee.

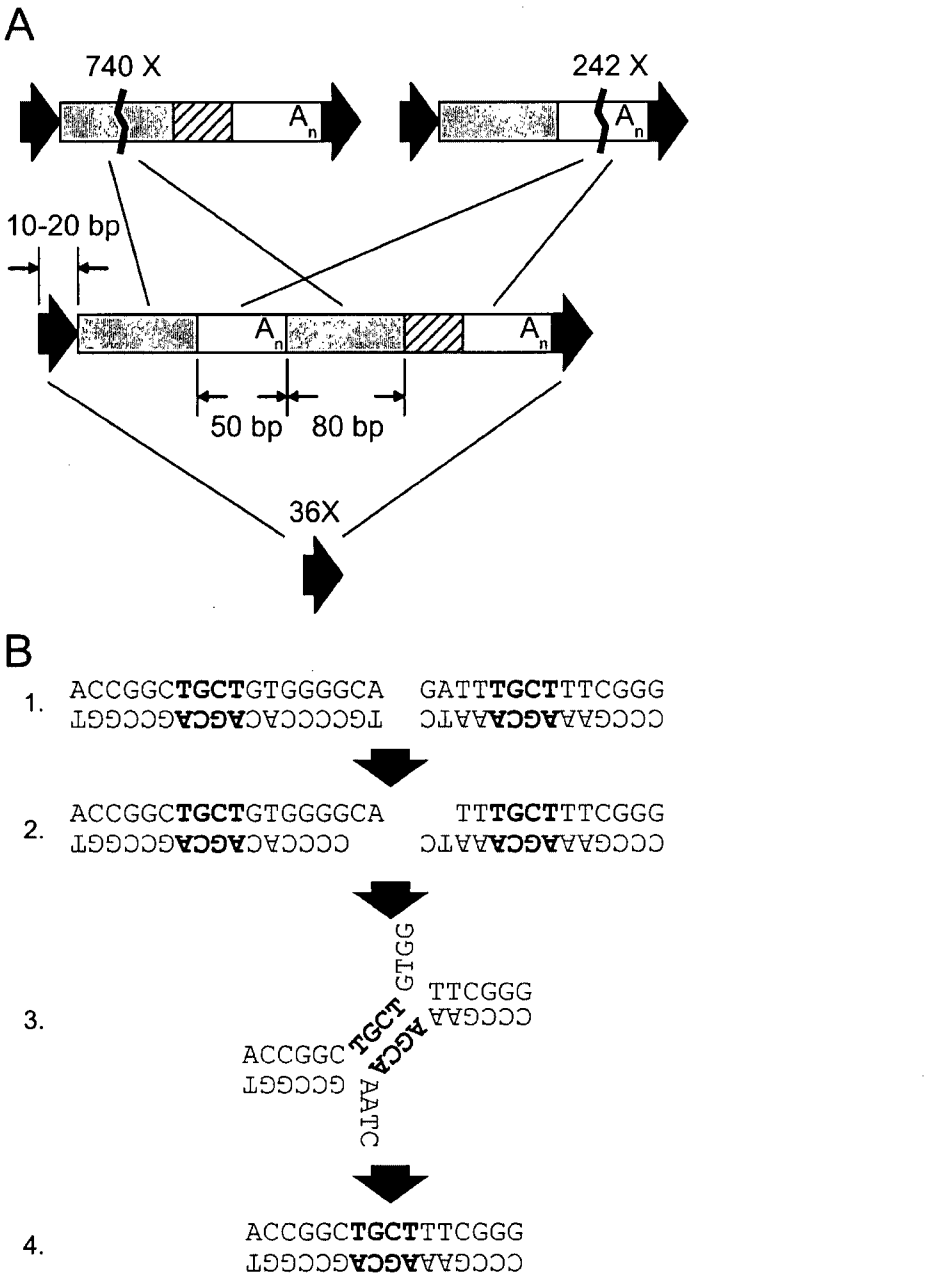

Figure 5.1 Deletions due to DNA double strand break repair. A) Whole and partial Alu element deletions. A full-length Alu is shown in the middle and black arrows represent target site duplications. Shaded and white internal regions represent internal ~70% identical regions. Deletions involving the 84 bp internal Alu homologies (shaded regions) were found 740 times in the human-chimpanzee alignments (top left). Alu internal deletions occurring between the other homologies (white regions) were found 242 times (top right). Precise deletion of entire Alu elements, likely involving the target site duplication (black arrows) was found in 36 cases (lower) in relatively repeat-free regions since human-chimpanzee divergence. B) A non-Alu deletion in chimpanzee at human chr1:1448280-1448311. Precise deletions of Alu elements, internal deletions within Alus, and other deletions are explained by an error-prone homology-dependent repair mechanism, involving 1. a double-strand DNA break, 2. resection of DNA and exposure of 3' tails, 3. homology search, and 4. ligation. In this case, a 4-bp homology mediated a 16-bp deletion.

Several of the apparent deletions from chimpanzee corresponded in human to human-specific Alu families, such as AluYa5 and AluYb8. However, in each case the corresponding element in Rhesus monkey shared identical TSDs and was also an AluY. Two explanations can account for this observation: multiple independent insertions at the identical site, or recent gene conversion in human which converted an existing older AluY insertion into an apparent human-specific family. Although we cannot rule out independent insertions as an explanation in these cases, we believe gene conversion, reported previously to occur between Alu elements (Salem et al. 2003a), is more likely. It should be noted that both deletion in the chimpanzee lineage and gene conversion in the human lineage, rather than controverting one another, are dual lines of evidence suggesting elevated recombinational or double-strand DNA break repair activity in these loci in recent evolutionary time.

We further noticed a relative paucity of precise deletions in human vs. chimpanzee (only 9/37 occurred in the human lineage). Without further study, it is unclear what this might mean. However, further BLAT alignments confirmed that, with the exception of two events (case #25, deleted in human, and case #43, deleted in chimpanzee), these events have all occurred in single-copy regions of the human and chimpanzee genomes. Furthermore, we used discontiguous megablast against the chimpanzee sequence trace database at NCBI to check for the possibility that some of the putative deletions in chimpanzee were a result of anomalous assembly, in which an Alu-containing trace at a locus was over-ruled by traces not containing the Alu. No such cases were found. By comparison, the numbers of random deletions between 200 and 500 bp long, discussed below, were more similar between human and chimpanzee (1011 and 916, respectively).

5.3.2 Analysis of random genomic deletion by illegitimate recombination

To further investigate the genomic prevalence of deletions that might be mediated by short repeats during the last few million years of primate evolution, we examined all length differences of 200-500 bp (thus approximating the 300-bp size of Alu elements) between human and chimpanzee and looked for flanking repeats at the breakpoints. After eliminating cases of tandem duplications, insertions (including sequence having additional copies elsewhere in the human genome), indels within transposable elements, and deletions between homologous transposable elements (see Methods), 1927 indels remained, and we termed these random deletions. It should be noted that our method did not exclude genomic deletions having one or both breakpoints within repetitive sequence, as long as the repetitive sequence at the endpoints did not belong to homologous repeats. We found that the endpoints of 367, or 19.0% of 200-500 bp random deletions in the human and chimpanzee lineages, are associated with flanking identical repeats of at least 10 bp.

To put this observation in the context of non-random sequence composition in primate genomes, we attempted to measure the background density of nonadjacent homologies 200-500 bp apart occurring in nonrepetitive human genome sequence. Therefore, repetitive sequence recognized by RepeatMasker (A.F.A. Smit & P. Green, unpublished data; http://www.repeatmasker.org) and tandem repeats found by Tandem Repeats Finder 3.21 (Benson 1999) were excised from the genome. This left 1.58 Gbp, or 55.6% of the human genome. A C++ program was constructed that computed alignments between all genomic positions 200 to 500 bp apart. From the banded alignments, the program directly calculated the length distribution of randomly-occurring identical segments flanking sequence tracts 200-500 bp long. We then extrapolated the

observed homology counts to compute the expected random homology occurrence in a complete genome. This method projected that 1.62 million random homologies of 10+ bp would exist 200-500 bp apart in the full size 2.84 Gbp human genome. The 376 random deletions that we observe with 10+ bp flanking repeats therefore account for 0.0226% of all such homologies available in the genome. This observation again fits well within the paradigm of deletion-prone homology-driven DNA double strand break repair, known as single-strand annealing (Karran 2000; Helleday 2003). In that model, DNA breakage results in binding of complexes that initiate peeling back of DNA, followed by a stochastic homology search in regions adjacent to the broken ends. In this type of DNA repair, many local homologies may be bypassed before fortuitous matching occurs. Exonucleases break down loose DNA ends, followed by ligation of the broken ends (Figure 5.1B). This mechanism accounts for deletion sizes over several orders of magnitude (data not shown), and for varying flanking repeat sizes (Figure 5.2).

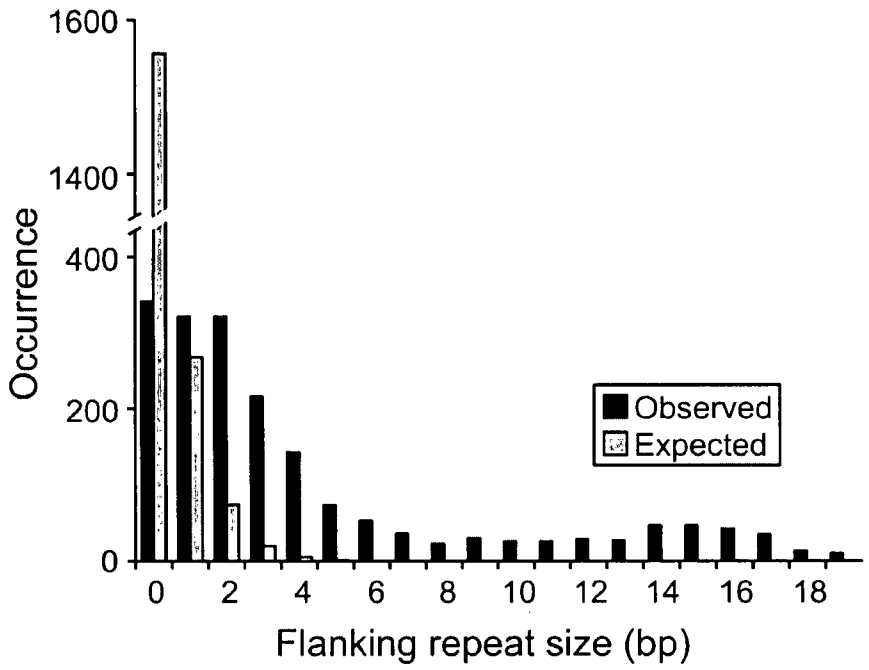

Figure 5.2 Prevalence of direct repeats at deletion boundaries. 1927 random deletions 200-500 bp in length were observed in the UCSC chimpanzee scaffold alignments to the July 2003 human genome. Observed flanking repeat occurrence (black bars) and expected occurrence if these deletions occurred by nonhomologous end joining alone (grey bars) are displayed. Flanking repeats 7 bp in size and above are expected to occur in <1/1927 cases.

As observed with Alu deletions, the observed association of random deletions with 10+ bp flanking repeats appeared much greater than would occur if homology played no role. Indeed, the suggestion that nonadjacent homologies play a role in genomic deletions has also been made based on studies in plants, although no statistical analysis has been done (Devos et al. 2002; Ma and Bennetzen 2004). To statistically confirm a strong association between flanking repeats and deletion, our results were compared to what would be expected in a process of purely random breakage followed by blunt-end rejoining (Figure 5.2). We reasoned that, under the hypothesis of no association between homology at breakpoints and deletion occurrence, homology occurrence at breakpoints of 200-500 bp deletions should mirror that observed 200-500 bp apart in the nonrepetitive genome. Using the data described above without extrapolation, 0.903 million randomly-occurring homologies occur in the nonrepetitive

genome, wherein there exist 300 times as many, or 0.474 trillion position combinations 200-500 bp apart. Thus 10+ bp homologies occur randomly at a frequency of 1.9×10^{-6} of any two positions 200-500 bp apart. Therefore, if homology plays no role in these deletions, we would expect much less than one occurrence of 10+ bp homology in our set of 1927 deletions ($1927 * 1.9 \times 10^{-6} = 0.0036$ occurrences, precisely), compared to the observed 367 occurrences ($P << 1 \times 10^{-100}$; Chi-squared test). Furthermore, by plotting the observed number of deletions associated with different lengths of flanking identity, we found that flanking repeats as short as two base pairs were overrepresented in the data set (Figure 5.2). This strong association of short flanking identities with deletion further confirms that illegitimate recombination between such short sequences has played a highly significant role in sequence deletion during primate evolution.

5.3.3 Direct confirmation of Alu element deletions

Finally, to confirm our findings, we chose 9 cases of AluS elements present in human but absent in the draft chimpanzee sequence to examine in more detail. These loci were chosen within and at varying distances from genes. To avoid regions of poor or anomalous alignments, we only investigated cases where the percentage identity between human and chimpanzee sequence surrounding the Alu is very high (>98%) and the Alu is a complete element with recognizable target site duplications (TSDs). Five of the cases (#14, 33, 42, 43, and 52; see Appendix A) were predicted to be deletions in chimpanzee, and as a control we selected four cases expected to be insertions in human (# C1-C4). The presence or absence of each of these Alus in a range of primate species was then determined using genomic PCR and the results summarized in Table 5.1.

Table 5.1 AluS indels assayed in primates by PCR and BLAST

#[1]	Fam.	Position[2]	TSD[3]	H[4]	C[4]	G[4]	O[4]	Gi[4]	B[4]	R[5]	Location/nearest genes
C1	Sx	20:11512274	13	Y[6]	N[6]	N	N	N	N	N	~354 kb 5' of BTBD3 (BTB/POZ domain containing-3)
C2	Sg	15:83819720	16	Y	N	N	N	N	N	N	In intron of AKAP13 (A-kinase anchor protein)
C3	Sg	7:104197804	19	Y	N	N	N	I[7]	N	N	~17 kb 5' of MLL5 (Myeloid/lymphoid leukemia 5)
C4	Sg	20:18254452	15	Y	N	N	N	N	N	N	~9.7 kb 5' of ZNF133 (Kruppel Zn-finger protein)
14	Sg	3:127318836	17	Y	N	Y	Y	?	?	Y	~63 kb 3' of KLF15 (Kruppel-like factor 15)
33	Sx	12:48585272	16	Y	N	Y	Y	Y	Y	Y	~1.3 kb 5' of FAIM2 (Fas apoptotic inhibitory molecule 2)
42	Sq	16:69279114	16	Y	N	Y	Y	Y	Y	Y	~5.2 kb 3' of CYB5-M (cytochrome b5)
43	Sx	16:74232245	17	Y	Y/N[8]	Y	Y	Y	?	Y	In intron of LOC348174 (secretory protein)
52	Sq	22:45658137	15	Y	N	Y	?	Y	?	Y	In intron of C22orf4 (putative GTPase activator)

[1] Case number. Cases Cn are controls, and others refer to case number in Supplementary information.
[2] Chromosome and position in July 2003 Human Genome Browser (http://genome.ucsc.edu)
[3] Size of Target Site Duplication (bp)
[4] H- human; C-chimpanzee; G- gorilla; O- Orangutan; Gi- Gibbon; B- Baboon; assayed by PCR
[5] R- Discontiguous MegaBLAST results from Rhesus monkey trace archive.
[6] Y= Alu is present; N= Alu is absent (as determined by PCR or Discontiguous MegaBLAST); ?= primers did not amplify or product of unexpected size
[7] I = Independent Alu insertion in same region in Gibbon
[8] Alu #43 is 'polymorphic' in all chimpanzees tested. Region is triplicated in human with all 3 having the Alu in human and one region lacking the Alu in chimpanzee.

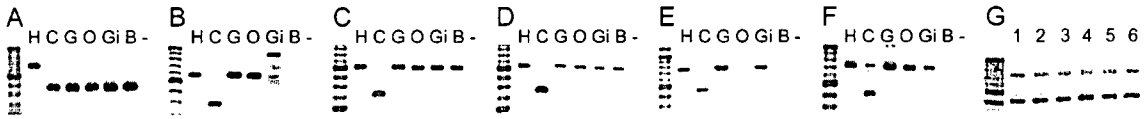

Figure 5.3 PCR and sequence evidence for precise Alu element deletion. A-F) cases C4, 14, 33, 42, 52, 43 from Table 5.1; lanes are human (H), chimpanzee (C), gorilla (G), orangutan (O), gibbon (Gi), baboon (Ba), and no-template control (-) G) case 43, genomic PCR in 6 additional chimpanzees, labeled 1-6.

As expected, our four controls demonstrate AluS presence only in human and no other primate, consistent with insertion in the human lineage after divergence from chimpanzee (Table 5.1, Figure 5.3A). In accord with this finding is a study suggesting that some AluSx elements may still be active (Johanning et al. 2003). Therefore, some of the non-AluY differences between human and chimpanzee may reflect recent low levels of retrotranspositional activity of AluS elements. An alternative explanation is that young AluY elements inserted in these locations, followed by gene conversion templated by older AluS elements. We therefore more carefully examined these Alu sequences to look for nucleotide positions diagnostic of young AluY subfamilies (Batzer and Deininger 2002). Although we found no convincing evidence for partial gene conversion, this mechanism cannot be ruled out. Interestingly, in control #3, gibbon has an independent AluY insertion at this locus, offset by 4 bp (NCBI Accession no. AY953324). Independent parallel retroelement insertions at or near the same genomic site have been previously noted (Salem et al. 2003a; Conley et al. 2005).

In the remaining five cases, PCR evidence confirms deletion in chimpanzee rather than lineage-specific insertion in human (Table 5.1). In four cases (#14, 33, 42, and 52), the Alu element was found to be uniformly present in 10 of 10 humans and absent in 10 of 10 chimpanzee DNA samples (data not shown). These four regions are apparently unique in the human genome with no evidence of segmental duplication. Insertion of these Alu elements could be verified by PCR in orangutan, which diverged from the higher apes 12-15 mya (Glazko and Nei 2003), or in even more distantly related primates (Figure 5.3B-E). (For case #14 in gibbon, the PCR product was of unexpected size, Figure 5.3E, suggesting rearrangement or other insertions in the region.) Given these long periods of time, it is unlikely that these loci reflect lineage sorting of ancestral

polymorphisms, proposed previously to explain unexpected Alu presence/absence relationships in the great apes (Salem et al. 2003b; Hedges et al. 2004). Rather, these results suggest that pre-existing fixed Alu elements have been deleted in the chimpanzee lineage. To verify that the loci in other primates contain the same Alu insertion, we sequenced the region in gorilla for cases #33 and 52 (NCBI Accession nos. AY953323 and AY953322) and compared to the human, chimpanzee, and Rhesus macaque genomic sequences from the databases (Figure 5.4A,B). In both cases, the gorilla and Rhesus loci are occupied by the same ancestral Alu as in human with the same target site duplication (TSD). Moreover, the sequence in chimpanzee has the expected structure of the pre-integration locus with only one copy of the TSD generated upon Alu insertion.

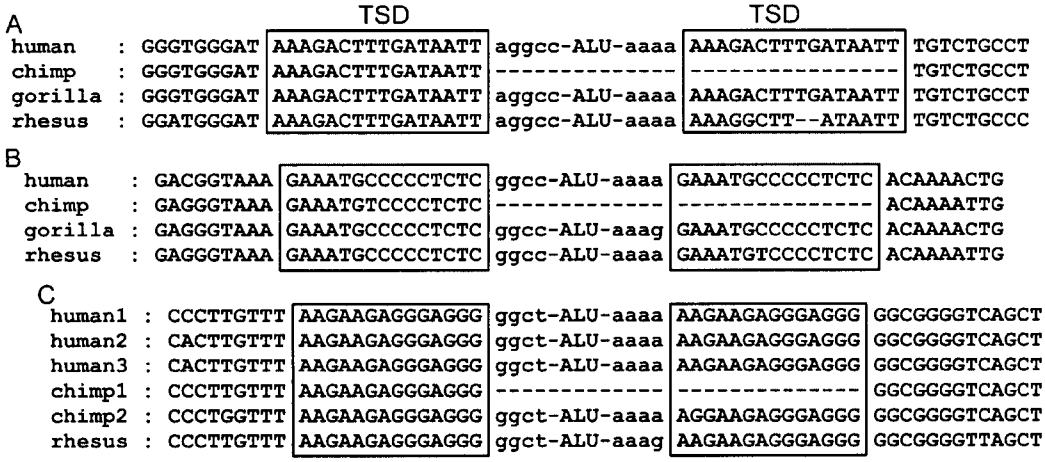

Figure 5.4 Sequence evidence for precise Alu element deletion. A,B) cases 33 & 52, sequenced in gorilla and compared to the database sequences of human, chimpanzee, and Rhesus macaque C) case 43, showing available human, chimpanzee, and Rhesus loci. Target site duplications are boxed.

The final case (#43) is more complex in that chimpanzee appears to have both occupied and unoccupied alleles or loci (Figure 5.3F). This pattern was seen in DNA from 6 of 6 additional chimpanzees tested (Figure 5.3G), suggesting that it does not reflect allelic polymorphism. Indeed, database analysis revealed that this locus is part of

complex segmental duplications that resulted in three copies in the human genome, all of which have the Alu insertion. The draft chimpanzee sequence has two copies, one of which lacks the Alu insertion. We cannot determine if a third copy exists in chimpanzee because of gaps and poor sequence coverage in these regions. An alignment of the three human and two chimpanzee sequences, as well as one Rhesus sequence is depicted in Figure 5.4C and shows that the chimpanzee locus without the Alu has the expected structure of a pre-integration allele. We confirmed the database entries by sequencing the two loci in chimpanzee (NCBI Accession nos. AY953325 and AY953326). The most probable explanation for this finding is that the Alu integrated prior to duplication of the region followed by loss of the Alu in one chimpanzee copy.

5.4 Concluding remarks

In summary, our analysis strongly suggests an important role for short nonadjacent segments of DNA identity in genomic deletions. In rare cases, even retroelement insertions deeply fixed in the primate lineage can apparently be precisely excised from the genome in a manner involving the flanking TSDs, leaving behind no footprint of their insertion. We believe that illegitimate recombination between short identical stretches of DNA, likely involving a DNA double-strand break repair mechanism, is the most likely and simplest molecular mechanism to explain the findings reported here. This conclusion is supported by the fact that a large fraction of non-TE associated deletions distinguishing human and chimpanzee have short repeats at the breakpoints. Furthermore, this study provides new insights into genomic attenuation and contradicts a rigid view that all insertions of retroelements represent unidirectional events. On the other hand, this study demonstrates that, for Alu elements in particular, homoplasy freedom is a mostly valid assumption and implicates internal homologous regions as preventing wholesale deletion

of Alus.

Finally, an aspect of Alu biology that has provoked interest is the slight preferential localization of younger elements in AT-rich regions but higher density of older elements in more GC-rich DNA (International Human Genome Sequencing Consortium 2001). Several theories have been proposed to explain the differences in Alu distributions with element age (Schmid 1998; Brookfield 2001; Pavlicek et al. 2001; Medstrand et al. 2002; Jurka 2004). While our findings indicate that precise deletion of Alu elements makes reversal of retroelement insertions possible, the phenomenon is nevertheless quite rare (~0.5% of length polymorphisms) and is likely insufficient to explain the shifts in Alu distribution. However, ectopic illegitimate recombination not involving TSDs may help to explain overall Alu sequence loss and distribution patterns.

Chapter 6: Summary and conclusions

6.1 Summary

The purpose of this thesis work was to use global analyses of populations of repeated sequences of various kinds in mammalian genomes to understand the interactions of these sequences and their host genomes. The methods developed in this pursuit were almost exclusively computational and involved analysis of sequenced human, chimpanzee, and mouse genomes and related sequence data. Trends identified in our analyses have provided insight into the global effects of transposable elements and their impact on the host. Below I briefly discuss several considerations arising out of this work.

6.2 Initial and long-term genomic localization of mobile elements are related in complex ways

Appropriate normalization of the currently observed distributions of repetitive elements allows comparison of the distributions of elements with widely varying total population sizes. While most element types seem simply to be lost from high-GC regions over time, the Alu distribution is particularly intriguing, in that very young and very old Alus seem to have less of a high-GC sequence composition preference, while Alus of intermediate ages seem to be strongly clustered in regions of high-GC content, which, in many cases, coincide with gene-rich regions.

Several explanations have been advanced that may account for these observations, each perhaps explaining some part of the nascence of the observed mobile element distributions. Our own work pointed to a role for recombination in shaping the distributions of Alu elements (Chapter 2). However, a complete understanding of the localization of different retroelements in regions of varying sequence composition requires knowledge of all processes taking place at and after the time of insertion. The

recent discoveries that transcripts, particularly those of exogenous viruses such as HIV, can be tethered in genic regions resulting in an altered propensity to insert there (Ciuffi et al. 2005; Lewinski and Bushman 2005) suggests that a similar mechanism may have accounted for the Alu accumulation in GC-rich regions. This theory is particularly pleasing in that it could explain the differential GC-content localization of the different Alu families based on their consensus sequence alone, without invoking any functional role or any more significant interactions with the genome. Furthermore, the recent discovery of base pair preferences in the vicinity of exogenous retroviral insertion sites (Holman and Coffin 2005) further suggests binding by tethering factors. Not least, differences in insertion patterns relative to genes for L1, SVA, and Alu retroelements, all of which are presumed to insert using L1-encoded machinery, strongly suggest the presence of auxiliary factors influencing the localization of these elements.

6.3 Some TEs interact strongly with genes, leading to population biases in regions surrounding genes

Although the localization of insertions of mobile elements is apparently determined, at least in part, by the sequence composition of the region, a separate, gene-specific interaction may also be measured. It must be noted that, given the strong association between high-GC content and high gene density, some of the apparent genomic sequence composition effect on TE localization may be due to the presence of genes. In any case, mapping of TEs with respect to genes and comparison of this mapping with that expected by consideration of sequence composition alone allowed a conservative assessment of the effect of genes on TE localization (Chapter 2). In short, TEs whose life cycle involves transcription by the pol-II machinery, which therefore contain internal active pol-II transcriptional signals in their sequence, seem to be at least partially disallowed from

entry into pol-II transcribed regions, likely due to pathogenicity upon insertion there. This has been suggested by others (Smit 1999). However, disallowance of TEs in the same transcriptional direction as genes extends upstream and downstream of genes as well, especially for LTR-containing elements (Chapter 2), suggesting that the transcriptional signals provided by these elements are detrimental at some distance from genes.

The fact that pol-II signal-containing TEs are not totally excluded from transcribed regions raises several interesting questions. If pol-II driven TEs are so pathogenic, why have they been allowed to remain at some level in transcribed regions? Are they disabled in some way? To what extent, if any, does methylation silencing of TE pol-II promoters affect the permissiveness of genic regions for insertion of these elements? Do they have weaker promoters or polyadenylation signals? Are there adjacent cellular sequences that override these dangerous motifs in some way? Are there perhaps binding sites for tethering proteins that help keep the pol-II holoenzyme on track in spite of these polyadenylation signals? These and other questions suggest a fruitful avenue for further research. For example, it might be interesting to search for association between multi-species conserved sequences, which have recently been shown to occur at higher density within long introns (Sironi et al. 2005a; Sironi et al. 2005b), and sense-oriented pol-II driven TEs found in genic regions.

6.4 Many gene UTRs are associated with TE-derived sequence

Analysis of transcripts revealed that 27.4% of human genes have permitted inclusion of TE sequence in the UTRs of one or more of their transcripts (Chapter 3). The corresponding figure for mouse is 18.4%. This lower figure may reflect the higher mutation rate in the mouse lineage, by which repetitive sequence becomes undetectable

by alignment methods. Indeed, mutually aligning genomic regions for which an older repeat is found in humans often have no annotated repeat in mouse, reflecting an inability of current alignment methods to find these repeats in mouse. In addition, high-copy repetitive sequence has, in general, been less well studied in the mouse lineage and the lower detected genomic coverage by repeats in rodent genomes is a reflection of this fact (Mouse Genome Sequencing Consortium 2002; Baillie et al. 2004).

A synthesis of the mapping data we have in both humans and mouse results in a perhaps-unsurprisingly consistent picture of both systems. In both human and mouse, TEs appear to affect expression of many genes through donation of transcriptional regulatory signals. Furthermore, recently expanded gene classes, such as those involved in immunity or response to external stimuli, have transcripts enriched in TEs, whereas TEs are excluded from mRNAs of highly conserved genes with basic functions in development or metabolism. These results could support one of two views. On one hand, one might argue that TEs have played a significant role in the diversification and evolution of mammalian genes. On the other hand, a more neutralist conclusion might be that permissive genes are so because their product is less dosage-critical, and therefore interference with expression due to TE donation of signals or sequence is less likely to ill-affect such genes. In this regard, it would be interesting to study inclusion of TE sequence in transcripts in the context of variations in expression level. One might expect that genes whose expression level is critical and conserved across species would most strongly exclude TEs. A first attempt to study this might entail compilation of a list of haploinsufficient genes followed by assessment of TE content in their transcripts.

6.5 Short repeated sequences are involved in genomic deletions

Insertion of transposable elements is a major cause of genomic expansion in eukaryotes.

Less is understood, however, about mechanisms underlying contraction of genomes. A combination of global bioinformatic analyses and PCR-based approaches showed that retroelements can, in rare cases, be precisely deleted from primate genomes, most likely via recombination between 10-20 bp TSDs flanking the retroelement (Chapter 5). The deleted loci are indistinguishable from pre-integration sites, effectively reversing the insertion. It is estimated that 0.5 to 1% of apparent retroelement insertions distinguishing humans and chimpanzees actually represent deletions. Furthermore, 19% of genomic deletions of 200-500 bp that have occurred since the human-chimpanzee divergence are associated with flanking identical repeats of at least 10 bp. A large number of deletions internal to Alu elements are also flanked by similar sequence. These results suggest that illegitimate recombination between short direct repeats has played, and likely continues to play, a significant role in human genomic deletion processes and is likely implicated in DNA deletion syndromes such as cancer. In short, while this study lends support to the view that insertions of retroelements are mostly irreversible, it is the first to conclusively demonstrate precise reversion of these events and estimate the rate of precise deletion of these elements. The data presented also suggested that the same mechanism is responsible for a large number of random genomic deletions.

In addition to the insights it provided, our study of short direct repeats and their role in deletion processes suggests further questions. Perhaps we can learn more about the frequency of error-prone modes of operation of ideally error-free DNA DSB repair pathways. To answer this question, one might perform further study of putative deletions of all sizes, using other primate genomes to determine the ancestral state of indels. This approach would help to elucidate in further detail the contributions of the various DNA DSB repair mechanisms to genomic sequence change. Such insights can help us

understand the etiologies of cancer and similar diseases caused by DNA breakage and repair.

6.6 Conclusions

Repeated sequences, primarily TEs, make up a large fraction of mammalian genomes. Bioinformatic analysis of genomic data is a uniquely powerful technique that has allowed us to conduct global genomic analyses of repeats and address their involvement in genomic-scale phenomena. The theme that has emerged repeatedly is the familiar one where the survival of the organism is the ultimate determinant of whether sequence change is accepted or not. Within this paradigm, apparently selfish sequence, such as that of apparently viral origin, may be co-opted by a genome to perform a useful function. However, it remains subservient to the needs of the organism, only rarely persisting when its overall impact is negative, and then only if the negative impact is slight. In rare cases, positive effect has been argued with convincing evidence, such as in the case of salivary expression of a digestive enzyme under the control of an enhancer of viral origin (Ting et al. 1992). As more detailed genomic analyses are done of insertions and deletions and genes affected by such events, it may be expected that more of these events will come to light. Especially interesting in this regard is the recent availability of the chimpanzee genome and its alignment with the human genome. Availability of more sequenced genomes promises to shed additional light on the role of repetitive DNA of all kinds in making the broad array of organisms what they are.

Although their global nature makes bioinformatic analyses powerful, it is also limiting. Genome-wide analyses, though they survey all genes, cannot address every factor governing each individual case. This is largely because many influences on gene expression, for example chromatin remodeling and RNA interference to name just two,

though characterized on some level for individual genes, are at present far from being well-enough understood in terms of their regulation and their regulatory effect on other genes to apply that knowledge on a genomic scale. It is exciting to think that, as more data on the various phenomena become available, we may gain sufficient understanding to map such information to genomes and conduct global analyses of them. This and an increasing trove of sequencing and phenotypic data in the public databases promise to provide grist for bioinformatic analyses addressing more and more sophisticated biological questions as time progresses. At no time, however, must the researcher lose sight of the critical importance of complementation of bioinformatic studies by wet laboratory approaches. Rather, bioinformatics and the wet laboratory are envisioned as partners in an iterative process, in which the wet lab provides fundamental understandings which are then used in global bioinformatic analyses. The goal of those analyses is to synthesize diverse data into a more systems-level picture of biology, offering a whole new round of hypotheses amenable to testing in wet-lab environments.

References

Adams, M.D., S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle et al. 2000. The genome sequence of Drosophila melanogaster. *Science* **287:** 2185-2195.

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

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J.M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297:** 1301-1310.

Athanikar, J.N., R.M. Badge, and J.V. Moran. 2004. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* **32:** 3846-3855.

Baillie, G.J., L.N. van de Lagemaat, C. Baust, and D.L. Mager. 2004. Multiple groups of endogenous betaretroviruses in mice, rats, and other mammals. *J Virol* **78:** 5784-5798.

Bannert, N. and R. Kurth. 2004. Retroelements and the human genome: new perspectives on an old relation. *Proc Natl Acad Sci U S A* **101 Suppl 2:** 14572-14579.

Bao, Z. and S.R. Eddy. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12:** 1269-1276.

Barr, S.D., J. Leipzig, P. Shinn, J.R. Ecker, and F.D. Bushman. 2005. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J Virol* **79:** 12035-12044.

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

Baust, C., G.J. Baillie, and D.L. Mager. 2002. Insertional polymorphisms of ETn retrotransposons include a disruption of the wiz gene in C57BL/6 mice. *Mamm Genome* **13:** 423-428.

Baust, C., W. Seifarth, H. Germaier, R. Hehlmann, and C. Leib-Mosch. 2000. HERV-K-T47D-Related long terminal repeats mediate polyadenylation of cellular transcripts. *Genomics* **66:** 98-103.

Bedell, J.A., I. Korf, and W. Gish. 2000. MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **16:** 1040-1041.

Belshaw, R., A.L. Dawson, J. Woolven-Allen, J. Redding, A. Burt, and M. Tristem. 2005. Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *J Virol* **79:** 12507-12514.

Bennett, E.A., L.E. Coleman, C. Tsui, W.S. Pittard, and S.E. Devine. 2004. Natural genetic variation caused by transposable elements in humans. *Genetics* **168:** 933-951.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27:** 573-580.

Biemont, C., A. Tsitrone, C. Vieira, and C. Hoogland. 1997. Transposable element distribution in Drosophila. *Genetics* **147:** 1997-1999.

Blond, J.L., F. Beseme, L. Duret, O. Bouton, F. Bedin, H. Perron, B. Mandrand, and F. Mallet. 1999. Molecular characterization and placental expression of HERV-W, a

new human endogenous retrovirus family. *J Virol* **73:** 1175-1185.
Boeke, J.D. and J.P. Stoye. 1997. Retrotransposons, Endogenous Retroviruses, and the Evolution of Retroelements. In *Retroviruses* (eds. J.M. Coffin S.H. Hughes, and H.E. Varmus), pp. 343-436. Cold Spring Harbor Laboratory Press, Plainview, New York, USA.
Brady, H.J., J.C. Sowden, M. Edwards, N. Lowe, and P.H. Butterworth. 1989. Multiple GF-1 binding sites flank the erythroid specific transcription unit of the human carbonic anhydrase I gene. *FEBS Lett* **257:** 451-456.
Brandt, V.L. and D.B. Roth. 2004. V(D)J recombination: how to tame a transposase. *Immunol Rev* **200:** 249-260.
Britten, R.J. 1997. Mobile elements inserted in the distant past have taken on important functions. *Gene* **205:** 177-182.
Britten, R.J., L. Rowen, J. Williams, and R.A. Cameron. 2003. Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci U S A* **100:** 4661-4665.
Brookfield, J.F. 2001. Selection on Alu sequences? *Curr Biol* **11:** R900-901.
Brosius, J. 1999. Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* **107:** 209-238.
Bushman, F., M. Lewinski, A. Ciuffi, S. Barr, J. Leipzig, S. Hannenhalli, and C. Hoffmann. 2005. Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol* **3:** 848-858.
Bushman, F.D. 2003. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell* **115:** 135-138.
Callinan, P.A., J. Wang, S.W. Herke, R.K. Garber, P. Liang, and M.A. Batzer. 2005. Alu retrotransposition-mediated deletion. *J Mol Biol* **348:** 791-800.
Cameron, H.S., D. Szczepaniak, and B.W. Weston. 1995. Expression of human chromosome 19p alpha(1,3)-fucosyltransferase genes in normal tissues. Alternative splicing, polyadenylation, and isoforms. *J Biol Chem* **270:** 20112-20122.
Carcedo, M.T., J.M. Iglesias, P. Bances, R.O. Morgan, and M.P. Fernandez. 2001. Functional analysis of the human annexin A5 gene promoter: a downstream DNA element and an upstream long terminal repeat regulate transcription. *Biochem J* **356:** 571-579.
Carlton, V.E., B.Z. Harris, E.G. Puffenberger, A.K. Batta, A.S. Knisely, D.L. Robinson, K.A. Strauss, B.L. Shneider, W.A. Lim, G. Salen et al. 2003. Complex inheritance of familial hypercholanemia with associated mutations in TJP2 and BAAT. *Nat Genet* **34:** 91-96.
Carroll, M.L., A.M. Roy-Engel, S.V. Nguyen, A.H. Salem, E. Vogel, B. Vincent, J. Myers, Z. Ahmad, L. Nguyen, M. Sammarco et al. 2001. Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* **311:** 17-40.
Charlesworth, B. and D. Charlesworth. 1983. The population dynamics of transposable elements. *Genet Res* **42:** 1-27.
Charlesworth, B. and C.H. Langley. 1991. Population genetics of transposable elements in Drosophila. In *Evolution at the molecular level* (eds. R.K. Selander A.G. Clark, and T.S. Whittam), pp. 150-176. Sinauer Associates, Sunderland, MA, USA.
Charlesworth, B., C.H. Langley, and P.D. Sniegowski. 1997. Transposable element distributions in Drosophila. *Genetics* **147:** 1993-1995.

Chen, J.M., P.D. Stenson, D.N. Cooper, and C. Ferec. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* **117:** 411-427.

Ciuffi, A., M. Llano, E. Poeschla, C. Hoffmann, J. Leipzig, P. Shinn, J.R. Ecker, and F. Bushman. 2005. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11:** 1287-1289.

Clamp, M., J. Cuff, S.M. Searle, and G.J. Barton. 2004. The Jalview Java alignment editor. *Bioinformatics* **20:** 426-427.

Conley, M.E., J.D. Partain, S.M. Norland, S.A. Shurtleff, and H.H. Kazazian, Jr. 2005. Two independent retrotransposon insertions at the same site within the coding region of BTK. *Hum Mutat* **25:** 324-325.

Cordonnier, A., J.F. Casella, and T. Heidmann. 1995. Isolation of novel human endogenous retrovirus-like elements with foamy virus-related pol sequence. *J Virol* **69:** 5890-5897.

Costas, J. and H. Naveira. 2000. Evolutionary history of the human endogenous retrovirus family ERV9. *Mol Biol Evol* **17:** 320-330.

Csoka, A.B., G.I. Frost, and R. Stern. 2001. The six hyaluronidase-like genes in the human and mouse genomes. *Matrix Biol* **20:** 499-508.

Cutter, A.D., J.M. Good, C.T. Pappas, M.A. Saunders, D.M. Starrett, and T.J. Wheeler. 2005. Transposable element orientation bias in the Drosophila melanogaster genome. *J Mol Evol* **61:** 733-741.

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

Deininger, P.L. and M.A. Batzer. 2002. Mammalian retroelements. *Genome Res* **12:** 1455-1465.

Devos, K.M., J.K. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res* **12:** 1075-1079.

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

Dewannieux, M. and T. Heidmann. 2005. L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. *J Mol Biol* **349:** 241-247.

Di Cristofano, A., M. Strazullo, L. Longo, and G. La Mantia. 1995. Characterization and genomic mapping of the ZNF80 locus: expression of this zinc-finger gene is driven by a solitary LTR of ERV9 endogenous retroviral family. *Nucleic Acids Res* **23:** 2823-2830.

Di Franco, C., A. Terrinoni, P. Dimitri, and N. Junakovic. 1997. Intragenomic distribution and stability of transposable elements in euchromatin and heterochromatin of Drosophila melanogaster: elements with inverted repeats Bari 1, hobo, and pogo. *J Mol Evol* **45:** 247-252.

Doolittle, W.F. and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284:** 601-603.

Dunbar, C.E. 2005. Stem cell gene transfer: insights into integration and hematopoiesis from primate genetic marking studies. *Ann N Y Acad Sci* **1044:** 178-182.

Dunn, C.A., P. Medstrand, and D.L. Mager. 2003. An endogenous retroviral long terminal repeat is the dominant promoter for human beta1,3-galactosyltransferase 5 in the colon. *Proc Natl Acad Sci U S A* **100:** 12841-12846.

Dunn, C.A., L.N. van de Lagemaat, G.J. Baillie, and D.L. Mager. 2005. Endogenous

retrovirus long terminal repeats as ready-to-use mobile promoters: The case of primate beta3GAL-T5. *Gene* **364:** 2-12.
Edwards, M.C. and R.A. Gibbs. 1992. A human dimorphism resulting from loss of an Alu. *Genomics* **14:** 590-597.
Elliott, B., C. Richardson, and M. Jasin. 2005. Chromosomal translocation mechanisms at intronic alu elements in mammalian cells. *Mol Cell* **17:** 885-894.
Esnault, C., J. Maestre, and T. Heidmann. 2000. Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **24:** 363-367.
Friedrich, G. and P. Soriano. 1991. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev* **5:** 1513-1523.
Fullerton, S.M., A. Bernardo Carvalho, and A.G. Clark. 2001. Local rates of recombination are positively correlated with GC content in the human genome. *Mol Biol Evol* **18:** 1139-1142.
Furano, A.V. 2000. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Prog Nucleic Acid Res Mol Biol* **64:** 255-294.
Gibbons, R., L.J. Dugaiczyk, T. Girke, B. Duistermars, R. Zielinski, and A. Dugaiczyk. 2004. Distinguishing humans from great apes with AluYb8 repeats. *J Mol Biol* **339:** 721-729.
Gifford, R. and M. Tristem. 2003. The evolution, distribution and diversity of endogenous retroviruses. *Virus Genes* **26:** 291-315.
Gilbert, N., S. Lutz-Prigge, and J.V. Moran. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110:** 315-325.
Glazko, G.V. and M. Nei. 2003. Estimation of divergence times for major lineages of primate species. *Mol Biol Evol* **20:** 424-434.
Goodchild, N.L., D.A. Wilkinson, and D.L. Mager. 1993. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology* **196:** 778-788.
Goodman, M., C.A. Porter, J. Czelusniak, S.L. Page, H. Schneider, J. Shoshani, G. Gunnell, and C.P. Groves. 1998. Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* **9:** 585-598.
Graves, J.A. 1995. The origin and function of the mammalian Y chromosome and Y-borne genes--an evolving understanding. *Bioessays* **17:** 311-320.
Greally, J.M. 2002. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc Natl Acad Sci U S A* **99:** 327-332.
Gregory, T.R. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc* **76:** 65-101.
Hacein-Bey-Abina, S., C. Von Kalle, M. Schmidt, M.P. McCormack, N. Wulffraat, P. Leboulch, A. Lim, C.S. Osborne, R. Pawliuk, E. Morillon et al. 2003. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302:** 415-419.
Hamdi, H., H. Nishio, R. Zielinski, and A. Dugaiczyk. 1999. Origin and phylogenetic distribution of Alu DNA repeats: irreversible events in the evolution of primates. *J Mol Biol* **289:** 861-871.
Hamdi, H.K., H. Nishio, J. Tavis, R. Zielinski, and A. Dugaiczyk. 2000. Alu-mediated phylogenetic novelties in gene regulation and development. *J Mol Biol* **299:** 931-939.

Han, J.S. and J.D. Boeke. 2004. A highly active synthetic mammalian retrotransposon. *Nature* **429:** 314-318.

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

Harada, F., N. Tsukada, and N. Kato. 1987. Isolation of three kinds of human endogenous retrovirus-like sequences using tRNA(Pro) as a probe. *Nucleic Acids Res* **15:** 9153-9162.

Hassoun, H., T.L. Coetzer, J.N. Vassiliadis, K.E. Sahr, G.J. Maalouf, S.T. Saad, L. Catanzariti, and J. Palek. 1994. A novel mobile element inserted in the alpha spectrin gene: spectrin dayton. A truncated alpha spectrin associated with hereditary elliptocytosis. *J Clin Invest* **94:** 643-648.

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

Helleday, T. 2003. Pathways for mitotic homologous recombination in mammalian cells. *Mutat Res* **532:** 103-115.

Higgins, D.G., J.D. Thompson, and T.J. Gibson. 1996. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol* **266:** 383-402.

Holman, A.G. and J.M. Coffin. 2005. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc Natl Acad Sci U S A* **102:** 6103-6107.

Holmes, I. 2002. Transcendent elements: whole-genome transposon screens and open evolutionary questions. *Genome Res* **12:** 1152-1155.

Hurst, L.D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet* **18:** 486.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Jern, P., G.O. Sperber, G. Ahlsen, and J. Blomberg. 2005. Sequence variability, gene structure, and expression of full-length human endogenous retrovirus H. *J Virol* **79:** 6325-6337.

Jern, P., G.O. Sperber, and J. Blomberg. 2004. Definition and variation of human endogenous retrovirus H. *Virology* **327:** 93-110.

Jiang, N., Z. Bao, X. Zhang, S.R. Eddy, and S.R. Wessler. 2004. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431:** 569-573.

Jiang, N., Z. Bao, X. Zhang, H. Hirochika, S.R. Eddy, S.R. McCouch, and S.R. Wessler. 2003. An active DNA transposon family in rice. *Nature* **421:** 163-167.

Johanning, K., C.A. Stevenson, O.O. Oyeniran, Y.M. Gozal, A.M. Roy-Engel, J. Jurka, and P.L. Deininger. 2003. Potential for retroposition by old Alu subfamilies. *J Mol Evol* **56:** 658-664.

Johnson, R.D. and M. Jasin. 2000. Sister chromatid gene conversion is a prominent double-strand break repair pathway in mammalian cells. *Embo J* **19:** 3398-3407.

Jordan, I.K., I.B. Rogozin, G.V. Glazko, and E.V. Koonin. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19:** 68-72.

Junakovic, N., A. Terrinoni, C. Di Franco, C. Vieira, and C. Loevenbruck. 1998. Accumulation of transposable elements in the heterochromatin and on the Y chromosome of Drosophila simulans and Drosophila melanogaster. *J Mol Evol*

46: 661-668.
Jurka, J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* **94:** 1872-1877.
Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16:** 418-420.
Jurka, J. 2004. Evolutionary impact of human Alu repetitive elements. *Curr Opin Genet Dev* **14:** 603-608.
Kamat, A., M.M. Hinshelwood, B.A. Murry, and C.R. Mendelson. 2002. Mechanisms in tissue-specific regulation of estrogen biosynthesis in humans. *Trends Endocrinol Metab* **13:** 122-128.
Kaplan, N.L. and J.F. Brookfield. 1983. The effect of homozygosity of selective differences between sites of transposable elements. *Theor Popul Biol* **23:** 273-280.
Karolchik, D., R. Baertsch, M. Diekhans, T.S. Furey, A. Hinrichs, Y.T. Lu, K.M. Roskin, M. Schwartz, C.W. Sugnet, D.J. Thomas et al. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res* **31:** 51-54.
Karran, P. 2000. DNA double strand break repair in mammalian cells. *Curr Opin Genet Dev* **10:** 144-150.
Kashkush, K., M. Feldman, and A.A. Levy. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* **33:** 102-106.
Kazazian, H.H., Jr. 2004. Mobile elements: drivers of genome evolution. *Science* **303:** 1626-1632.
Kazazian, H.H., Jr., C. Wong, H. Youssoufian, A.F. Scott, D.G. Phillips, and S.E. Antonarakis. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332:** 164-166.
Kent, W.J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12:** 656-664.
Kent, W.J., C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, and D. Haussler. 2002. The human genome browser at UCSC. *Genome Res* **12:** 996-1006.
Kidwell, M.G. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115:** 49-63.
Kidwell, M.G. and D. Lisch. 1997. Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A* **94:** 7704-7711.
Kidwell, M.G. and D.R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution Int J Org Evolution* **55:** 1-24.
Kiem, H.P., S. Sellers, B. Thomasson, J.C. Morris, J.F. Tisdale, P.A. Horn, P. Hematti, R. Adler, K. Kuramoto, B. Calmels et al. 2004. Long-term clinical and molecular follow-up of large animals receiving retrovirally transduced stem and progenitor cells: no progression to clonal hematopoiesis or leukemia. *Mol Ther* **9:** 389-395.
Kim, H.S., O. Takenaka, and T.J. Crow. 1999. Cloning and nucleotide sequence of retroposons specific to hominoid primates derived from an endogenous retrovirus (HERV-K). *AIDS Res Hum Retroviruses* **15:** 595-601.
Kimberland, M.L., V. Divoky, J. Prchal, U. Schwahn, W. Berger, and H.H. Kazazian, Jr. 1999. Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. *Hum Mol Genet* **8:** 1557-1560.
Kirkness, E.F., V. Bafna, A.L. Halpern, S. Levy, K. Remington, D.B. Rusch, A.L.

Delcher, M. Pop, W. Wang, C.M. Fraser et al. 2003. The dog genome: survey sequencing and comparative analysis. *Science* **301:** 1898-1903.
Kjellman, C., H.O. Sjogren, and B. Widegren. 1995. The Y chromosome: a graveyard for endogenous retroviruses. *Gene* **161:** 163-170.
Lahn, B.T., N.M. Pearson, and K. Jegalian. 2001. The human Y chromosome, in the light of evolution. *Nat Rev Genet* **2:** 207-216.
Landry, J.R., P. Medstrand, and D.L. Mager. 2001. Repetitive elements in the 5' untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics* **76:** 110-116.
Landry, J.R., A. Rouhi, P. Medstrand, and D.L. Mager. 2002. The Opitz syndrome gene Mid1 is transcribed from a human endogenous retroviral promoter. *Mol Biol Evol* **19:** 1934-1942.
Langley, C.H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth. 1988. On the role of unequal exchange in the containment of transposable element copy number. *Genet Res* **52:** 223-235.
Lavie, L., M. Kitova, E. Maldener, E. Meese, and J. Mayer. 2005. CpG methylation directly regulates transcriptional activity of the human endogenous retrovirus family HERV-K(HML-2). *J Virol* **79:** 876-883.
Leach, D.R. 1994. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *Bioessays* **16:** 893-900.
Leib-Mosch, C., W. Seifarth, and U. Schon. 2005. Influence of Human Endogenous Retroviruses on Cellular Gene Expression. In *Retroviruses and Primate Genome Evolution* (ed. E.D. Sverdlov), pp. 123-143. Landes Bioscience, Georgetown, Texas, USA.
Lev-Maor, G., R. Sorek, N. Shomron, and G. Ast. 2003. The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons. *Science* **300:** 1288-1291.
Lewinski, M.K. and F.D. Bushman. 2005. Retroviral DNA integration--mechanism and consequences. *Adv Genet* **55:** 147-181.
Li, W., P. Zhang, J.P. Fellers, B. Friebe, and B.S. Gill. 2004. Sequence composition, organization, and evolution of the core Triticeae genome. *Plant J* **40:** 500-511.
Li, W.H. and D. Graur. 1991. *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA, USA.
Liu, G., S. Zhao, J.A. Bailey, S.C. Sahinalp, C. Alkan, E. Tuzun, E.D. Green, and E.E. Eichler. 2003. Analysis of primate genomic variation reveals a repeat-driven expansion of the human genome. *Genome Res* **13:** 358-368.
Liu, H., H. Han, J. Li, and L. Wong. 2005. DNAFSMiner: a web-based software toolbox to recognize two types of functional sites in DNA sequences. *Bioinformatics* **21:** 671-673.
Lobachev, K.S., J.E. Stenger, O.G. Kozyreva, J. Jurka, D.A. Gordenin, and M.A. Resnick. 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. *Embo J* **19:** 3822-3830.
Lorincz, M.C., D.R. Dickerson, M. Schmitt, and M. Groudine. 2004. Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. *Nat Struct Mol Biol* **11:** 1068-1075.
Ma, J. and J.L. Bennetzen. 2004. Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* **101:** 12404-12410.
Mager, D.L., D.G. Hunter, M. Schertzer, and J.D. Freeman. 1999. Endogenous retroviruses provide the primary polyadenylation signal for two new human genes

(HHLA2 and HHLA3). *Genomics* **59:** 255-263.
Mager, D.L. and P. Medstrand. 2003. Retroviral repeat sequences. In *Nature Encyclopedia of the Human Genome, Volume 5*, pp. 57-63. Macmillan Publishers Ltd., London, U. K.
Majors, J. 1990. The structure and function of retroviral long terminal repeats. *Curr Top Microbiol Immunol* **157:** 49-92.
Makalowski, W. 2000. Genomic scrap yard: how genomes utilize all that junk. *Gene* **259:** 61-67.
Maksakova, I.A., M.T. Romanish, L. Gagnier, C.A. Dunn, L.N. van de Lagemaat, and D.L. Mager. 2006. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. *PLoS Genetics* **2:** e2.
Martin, A.M., J.K. Kulski, C. Witt, P. Pontarotti, and F.T. Christiansen. 2002. Leukocyte Ig-like receptor complex (LRC) in mice and men. *Trends Immunol* **23:** 81-88.
McClintock, B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36:** 344-355.
McClintock, B. 1956. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol* **21:** 197-216.
McDonald, J.F. 1995. Transposable elements: possible catalysts of organismic evolution. *Trends Ecol Evol* **10:** 123-126.
McGinnis, S. and T.L. Madden. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* **32:** W20-25.
Medstrand, P., J.R. Landry, and D.L. Mager. 2001. Long terminal repeats are used as alternative promoters for the endothelin B receptor and apolipoprotein C-I genes in humans. *J Biol Chem* **276:** 1896-1903.
Medstrand, P. and D.L. Mager. 1998. Human-specific integrations of the HERV-K endogenous retrovirus family. *J Virol* **72:** 9782-9787.
Medstrand, P., L.N. van de Lagemaat, C.A. Dunn, J.R. Landry, D. Svenback, and D.L. Mager. 2005. Impact of transposable elements on the evolution of mammalian gene regulation. *Cytogenet Genome Res* **110:** 342-352.
Medstrand, P., L.N. van de Lagemaat, and D.L. Mager. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12:** 1483-1495.
Meunier, J., A. Khelifi, V. Navratil, and L. Duret. 2005. Homology-dependent methylation in primate repetitive DNA. *Proc Natl Acad Sci U S A* **102:** 5471-5476.
Mi, S., X. Lee, X. Li, G.M. Veldman, H. Finnerty, L. Racie, E. LaVallie, X.Y. Tang, P. Edouard, S. Howes et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature* **403:** 785-789.
Miskey, C., Z. Izsvak, K. Kawakami, and Z. Ivics. 2005. DNA transposons in vertebrate functional genomics. *Cell Mol Life Sci* **62:** 629-641.
Mitchell, R.S., B.F. Beitzel, A.R. Schroder, P. Shinn, H. Chen, C.C. Berry, J.R. Ecker, and F.D. Bushman. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2:** E234.
Morgan, H.D., H.G. Sutherland, D.I. Martin, and E. Whitelaw. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet* **23:** 314-318.
Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.
Moynahan, M.E. and M. Jasin. 1997. Loss of heterozygosity induced by a chromosomal

double-strand break. *Proc Natl Acad Sci U S A* **94:** 8988-8993.
Muratani, K., T. Hada, Y. Yamamoto, T. Kaneko, Y. Shigeto, T. Ohue, J. Furuyama, and K. Higashino. 1991. Inactivation of the cholinesterase gene by Alu insertion: possible mechanism for human gene transposition. *Proc Natl Acad Sci U S A* **88:** 11315-11319.
Murnane, J.P. and J.F. Morales. 1995. Use of a mammalian interspersed repetitive (MIR) element in the coding and processing sequences of mammalian genes. *Nucleic Acids Res* **23:** 2837-2839.
Nag, D.K., M. Fasullo, Z. Dong, and A. Tronnes. 2005. Inverted repeat-stimulated sister-chromatid exchange events are RAD1-independent but reduced in a msh2 mutant. *Nucleic Acids Res* **33:** 5243-5249.
Nakaya, S.M., T.C. Hsu, S.J. Geraghty, M.J. Manco-Johnson, and A.R. Thompson. 2004. Severe hemophilia A due to a 1.3 kb factor VIII gene deletion including exon 24: homologous recombination between 41 bp within an Alu repeat sequence in introns 23 and 24. *J Thromb Haemost* **2:** 1941-1945.
Nekrutenko, A. and W.H. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* **17:** 619-621.
Nigumann, P., K. Redik, K. Matlik, and M. Speek. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79:** 628-634.
Ono, M., M. Kawakami, and T. Takezawa. 1987. A novel human nonviral retroposon derived from an endogenous retrovirus. *Nucleic Acids Res* **15:** 8725-8737.
Orgel, L.E. and F.H. Crick. 1980. Selfish DNA: the ultimate parasite. *Nature* **284:** 604-607.
Ostertag, E.M., J.L. Goodier, Y. Zhang, and H.H. Kazazian, Jr. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* **73:** 1444-1451.
Ostertag, E.M. and H.H. Kazazian, Jr. 2001. Biology of mammalian L1 retrotransposons. *Annu Rev Genet* **35:** 501-538.
Panet, A. and H. Cedar. 1977. Selective degradation of integrated murine leukemia proviral DNA by deoxyribonucleases. *Cell* **11:** 933-940.
Pavlicek, A., K. Jabbari, J. Paces, V. Paces, J.V. Hejnar, and G. Bernardi. 2001. Similar integration but different stability of Alus and LINEs in the human genome. *Gene* **276:** 39-45.
Perepelitsa-Belancio, V. and P. Deininger. 2003. RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet* **35:** 363-366.
Perna, N.T., M.A. Batzer, P.L. Deininger, and M. Stoneking. 1992. Alu insertion polymorphism: a new type of marker for human population studies. *Hum Biol* **64:** 641-648.
Pinsker, W., E. Haring, S. Hagemann, and W.J. Miller. 2001. The evolutionary life history of P transposons: from horizontal invaders to domesticated neogenes. *Chromosoma* **110:** 148-158.
Price, A.L., N.C. Jones, and P.A. Pevzner. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1:** i351-i358.
Prudden, J., J.S. Evans, S.P. Hussey, B. Deans, P. O'Neill, J. Thacker, and T. Humphrey. 2003. Pathway utilization in response to a site-specific DNA double-strand break in fission yeast. *Embo J* **22:** 1419-1430.

Rabson, A.B. and B.J. Graves. 1997. Synthesis and Processing of Viral RNA. In *Retroviruses* (eds. J.M. Coffin S.H. Hughes, and H.E. Varmus), pp. 205-261. Cold Spring Harbor Laboratory Press, Plainview, New York, USA.

Rat Genome Sequencing Project Consortium. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493-521.

Renard, M., P.F. Varela, C. Letzelter, S. Duquerroy, F.A. Rey, and T. Heidmann. 2005. Crystal structure of a pivotal domain of human syncytin-2, a 40 million years old endogenous retrovirus fusogenic envelope gene captured by primates. *J Mol Biol* **352:** 1029-1034.

Rosenberg, N. and P. Jolicoeur. 1997. Retroviral Pathogenesis. In *Retroviruses* (eds. J.M. Coffin S.H. Hughes, and H.E. Varmus), pp. 475-586. Cold Spring Harbor Laboratory Press, Plainview, New York, USA.

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

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

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

Sankaranarayanan, K. and J.S. Wassom. 2005. Ionizing radiation and genetic risks XIV. Potential research directions in the post-genome era based on knowledge of repair of radiation-induced DNA double-strand breaks in mammalian somatic cells and the origin of deletions associated with human genomic disorders. *Mutat Res* **578:** 333-370.

SanMiguel, P., B.S. Gaut, A. Tikhonov, Y. Nakajima, and J.L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* **20:** 43-45.

Schatz, D.G. 2004. Antigen receptor genes and the evolution of a recombinase. *Semin Immunol* **16:** 245-256.

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

Schroder, A.R., P. Shinn, H. Chen, C. Berry, J.R. Ecker, and F. Bushman. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110:** 521-529.

Sheen, F.M., S.T. Sherry, G.M. Risch, M. Robichaux, I. Nasidze, M. Stoneking, M.A. Batzer, and G.D. Swergold. 2000. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. *Genome Res* **10:** 1496-1508.

Shen, L., L.C. Wu, S. Sanlioglu, R. Chen, A.R. Mendoza, A.W. Dangel, M.C. Carroll, W.B. Zipf, and C.Y. Yu. 1994. Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. *J Biol Chem* **269:** 8466-8476.

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

Sironi, M., G. Menozzi, G.P. Comi, N. Bresolin, R. Cagliani, and U. Pozzoli. 2005a. Fixation of conserved sequences shapes human intron size and influences

transposon-insertion dynamics. *Trends Genet* **21:** 484-488.
Sironi, M., G. Menozzi, G.P. Comi, R. Cagliani, N. Bresolin, and U. Pozzoli. 2005b. Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum Mol Genet* **14:** 2533-2546.
Smit, A.F. 1993. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res* **21:** 1863-1872.
Smit, A.F. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr Opin Genet Dev* **9:** 657-663.
Smit, A.F. and A.D. Riggs. 1995. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res* **23:** 98-102.
Smit, A.F., G. Toth, A.D. Riggs, and J. Jurka. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246:** 401-417.
Sorek, R., G. Ast, and D. Graur. 2002. Alu-containing exons are alternatively spliced. *Genome Res* **12:** 1060-1067.
Sorek, R., G. Lev-Maor, M. Reznik, T. Dagan, F. Belinky, D. Graur, and G. Ast. 2004. Minimal conditions for exonization of intronic sequences: 5' splice site formation in alu exons. *Mol Cell* **14:** 221-231.
Stenger, J.E., K.S. Lobachev, D. Gordenin, T.A. Darden, J. Jurka, and M.A. Resnick. 2001. Biased distribution of inverted and direct Alus in the human genome: implications for insertion, exclusion, and genome stability. *Genome Res* **11:** 12-27.
Sverdlov, E.D. 1998. Perpetually mobile footprints of ancient infections in human genome. *FEBS Lett* **428:** 1-6.
Sverdlov, E.D. 2000. Retroviruses and primate evolution. *Bioessays* **22:** 161-171.
Swergold, G.D. 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* **10:** 6718-6729.
Symer, D.E., C. Connelly, S.T. Szak, E.M. Caputo, G.J. Cost, G. Parmigiani, and J.D. Boeke. 2002. Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110:** 327-338.
Taruscio, D., G. Floridia, G.K. Zoraqi, A. Mantovani, and V. Falbo. 2002. Organization and integration sites in the human genome of endogenous retroviral sequences belonging to HERV-E family. *Mamm Genome* **13:** 216-222.
Tchenio, T., J.F. Casella, and T. Heidmann. 2000. Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res* **28:** 411-415.
Temin, H.M. 1982. Function of the retrovirus long terminal repeat. *Cell* **28:** 3-5.
Thompson, L.H. and D. Schild. 2002. Recombinational DNA repair and human disease. *Mutat Res* **509:** 49-78.
Ting, C.N., M.P. Rosenberg, C.M. Snow, L.C. Samuelson, and M.H. Meisler. 1992. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. *Genes Dev* **6:** 1457-1465.
Torti, C., L.M. Gomulski, D. Moralli, E. Raimondi, H.M. Robertson, P. Capy, G. Gasperi, and A.R. Malacrida. 2000. Evolution of different subfamilies of mariner elements within the medfly genome inferred from abundance and chromosomal distribution. *Chromosoma* **108:** 523-532.
Tournier, I., B.B. Paillerets, H. Sobol, D. Stoppa-Lyonnet, R. Lidereau, M. Barrois, S. Mazoyer, F. Coulet, A. Hardouin, A. Chompret et al. 2004. Significant contribution of germline BRCA2 rearrangements in male breast cancer families.

Cancer Res **64:** 8143-8147.
Tristem, M. 2000. Identification and characterization of novel human endogenous retrovirus families by phylogenetic screening of the human genome mapping project database. *J Virol* **74:** 3715-3730.
Turner, G., M. Barbulescu, M. Su, M.I. Jensen-Seaman, K.K. Kidd, and J. Lenz. 2001. Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr Biol* **11:** 1531-1535.
Ullu, E. and C. Tschudi. 1984. Alu sequences are processed 7SL RNA genes. *Nature* **312:** 171-172.
Urwin, D. and R.A. Lake. 2000. Structure of the Mesothelin/MPF gene and characterization of its promoter. *Mol Cell Biol Res Commun* **3:** 26-32.
van de Lagemaat, L.N., J.R. Landry, D.L. Mager, and P. Medstrand. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19:** 530-536.
Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li, R.J. Mural, G.G. Sutton, H.O. Smith, M. Yandell, C.A. Evans, R.A. Holt et al. 2001. The sequence of the human genome. *Science* **291:** 1304-1351.
von Melchner, H., J.V. DeGregori, H. Rayburn, S. Reddy, C. Friedel, and H.E. Ruley. 1992. Selective disruption of genes expressed in totipotent embryonal stem cells. *Genes Dev* **6:** 919-927.
Walker, J.R., R.A. Corpina, and J. Goldberg. 2001. Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature* **412:** 607-614.
Wang, H., J. Xing, D. Grover, D.J. Hedges, K. Han, J.A. Walker, and M.A. Batzer. 2005. SVA Elements: A Hominid-specific Retroposon Family. *J Mol Biol* **354:** 994-1007.
Ward, B.D., B.C. Hendrickson, T. Judkins, A.M. Deffenbaugh, B. Leclair, B.E. Ward, and T. Scholl. 2005. A multi-exonic BRCA1 deletion identified in multiple families through single nucleotide polymorphism haplotype pair analysis and gene amplification with widely dispersed primer sets. *J Mol Diagn* **7:** 139-142.
Watanabe, H., A. Fujiyama, M. Hattori, T.D. Taylor, A. Toyoda, Y. Kuroki, H. Noguchi, A. BenKahla, H. Lehrach, R. Sudbrak et al. 2004. DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429:** 382-388.
Wei, W., N. Gilbert, S.L. Ooi, J.F. Lawler, E.M. Ostertag, H.H. Kazazian, J.D. Boeke, and J.V. Moran. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21:** 1429-1439.
Whitelaw, E. and D.I. Martin. 2001. Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat Genet* **27:** 361-365.
Wilkinson, D.A., D.L. Mager, and J.C. Leong. 1994. Endogenous Human Retroviruses. In *The Retroviridae* (ed. J.A. Levy), pp. 465-535. Plenum Press, New York, NY.
Wu, X., Y. Li, B. Crise, and S.M. Burgess. 2003. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300:** 1749-1751.
Yang, N., L. Zhang, Y. Zhang, and H.H. Kazazian, Jr. 2003. An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res* **31:** 4929-4940.
Yoder, J.A., C.P. Walsh, and T.H. Bestor. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* **13:** 335-340.
Yohn, C.T., Z. Jiang, S.D. McGrath, K.E. Hayden, P. Khaitovich, M.E. Johnson, M.Y.

Eichler, J.D. McPherson, S. Zhao, S. Paabo et al. 2005. Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biol* **3:** e110.

Yu, J., S. Hu, J. Wang, G.K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science* **296:** 79-92.

Yu, X., X. Zhu, W. Pi, J. Ling, L. Ko, Y. Takeda, and D. Tuan. 2005. The long terminal repeat (LTR) of ERV-9 human endogenous retrovirus binds to NF-Y in the assembly of an active LTR enhancer complex NF-Y/MZF1/GATA-2. *J Biol Chem* **280:** 35184-35194.

Zhang, Y., N. Zeleznik-Le, N. Emmanuel, N. Jayathilaka, J. Chen, P. Strissel, R. Strick, L. Li, M.B. Neilly, T. Taki et al. 2004. Characterization of genomic breakpoints in MLL and CBP in leukemia patients with t(11;16). *Genes Chromosomes Cancer* **41:** 257-265.

Zhu, Z.B., S.L. Hsieh, D.R. Bentley, R.D. Campbell, and J.E. Volanakis. 1992. A variable number of tandem repeats locus within the human complement C2 gene is associated with a retroposon derived from a human endogenous retrovirus. *J Exp Med* **175:** 1783-1787.

Appendix A

Human-chimpanzee indel loci assayed in Rhesus monkey trace archive.

Record structure:
1: Header, showing human chromosome: chromosome start position in July 2003 UCSC genome browser: chromosome end position (if necessary): chimpanzee scaffold name : scaffold start position: scaffold end position (if necessary): orientation of scaffold/chromosome alignment.
2: short description of case
3: Multiple sequence alignment of human, chimpanzee, and Rhesus sequences. Underlining highlights identical segments flanking the indels

---

chr1:144298621:144298777:scaffold_37562:1315854:+
...imprecise deletion of AluSx fragment in chimpanzee

CLUSTAL

```
case1h/1-256              GAAGAAATTTGGAAGAATTGCCACATGTGGAGCTATCTCTATATATAACATATACATATC
gnl|ti|518618788/314-567  GAAGAAATTTGGAAGAATTGCCATATATAGAGCTATCTCTGCATATAACATATACATATC
case1c/1-101              GAAGAAATATGGAAGGATTGCCATATGTGGAGCCATCTCTACTTATAACAGA--------

case1h/1-256              TACATATAACACCTGTAATCCCAGTACTTTGGGAGAATGAGGCAGGTGGATCACCTGAGG
gnl|ti|518618788/314-567  TACATATAATGCCTGTAATCCCAACACTTTGGGAGGCTGAGGCAGGTGGATCACCTGAGG
case1c/1-101              ------------------------------------------------------------

case1h/1-256              TCAGGAGTTTGAGACCAGCCTGGCCAACATGGTGAAACCCCCGTCTCTACCAAAAATACA
gnl|ti|518618788/314-567  TCAGGAGTTTGAGACCGGCCTGGCCAAGATGGTGAAACCCCCTTCTGTACTAAAAATACA
case1c/1-101              ------------------------------------------------------------

case1h/1-256              AAAATTAGCCGGGTGTGGTGGTACCAGCCCACTTCCCCCAGGCCCAATCCCAGAGATTGT
gnl|ti|518618788/314-567  AAAGTGAGCCAGGCATGGTGGCACCAGCCCACTTCCCCCAGGCCCAACCCCAGAGATTGT
case1c/1-101              ----------------------ACTGGCCCACCTCCCCCAGGCCCACCCCCAGATATTAT

case1h/1-256              TAGATGTATCAGGAGC
gnl|ti|518618788/314-567  TA--TCTATCAGGAGC
case1c/1-101              -----CTATAAGGAGC
```

chr2:50683419:50683739:scaffold_37688:19585334:-
...AluY in Rhesus, gene conversion to AluYa5 in human, precise deletion in chimpanzee

CLUSTAL

```
case2h/1-482              TGTAAGTTTCAGAATCATACTTTAAAAAA------ATCTTTTTGGGGGCAGTTTTGTTTC
gnl|ti|508398655/154-589  TGTAAGCTGCAGAATCATACTTTAAAAAAAAAAAAAAATTTTGGGGGGCAGTTTTGTTTC
case2c/1-163              TGTAAGTTTCAGAATCATACTTTTAAAAAA-----ATCTTTTTTGGGGCAGTTTTGTTTC

case2h/1-482              AAAATTTAAATAAGAAACAAAAGTTCGGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAG
gnl|ti|508398655/154-589  AAAATTTAAATAAGAAACAAAAGTTC---------------ATCACGCCTGTAATCCCAG
case2c/1-163              AAAATTTAAATAAGAAACAAAAGTTC---------------------------------

case2h/1-482              CACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTA
gnl|ti|508398655/154-589  CACTTTGGGAGGCCGAGCCGGGCGGATCATGAGGTCAAGAAATCAAGACCATCCTGGCTA
case2c/1-163              ------------------------------------------------------------

case2h/1-482              AAACGGTGAAACCCCGTCTCTACTAAAAATACAAAAAAAAAATTAGCCGGGCGTAGTGGCG
gnl|ti|508398655/154-589  ACATGGTGAAACCCTGTCTCTACTGAAAACACAAAAAA---TTAGCTAGGCGTGGTGGCG
case2c/1-163              ------------------------------------------------------------

case2h/1-482              GGCGCCTGTAGTCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATAGCGTGAACCCGGGAG
gnl|ti|508398655/154-589  GGCGCCTGTA----------CTCGGGAGACTGAGGCAGGAGAATGGCGTGAACCCGAGAG
case2c/1-163              ------------------------------------------------------------

case2h/1-482              GCGGAGCTTGCAGTGAGCCGAGATCCCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAG
gnl|ti|508398655/154-589  GCGGAGCTTGCAGTGAGCAGTGATTGTGCCACTGCACTCCATCCTGGGTGACAGTGCAAG
case2c/1-163              ------------------------------------------------------------
```

```
case2h/1-482              ACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAGAAACAAAAGTTCAGAATATTGAAAGA
gnl|ti|508398655/154-589  ACTCTGTCTTAAAAAAAAAAA----------TTCAAAATATCGAAAGA
case2c/1-163              ----------AGAATATTGAAAGA

case2h/1-482              ATTATATGTTCTGTTGATATGGATAATAACAATATTAACTCCCCAAAACTCACTACAGGA
gnl|ti|508398655/154-589  ATTATCTGTTCTGTTCATATGGATAA---CAATATTAACTCCCCAACACTCACTACAGGA
case2c/1-163              ATTATATGTTCTGTTGATATGGATAATAACAATATTAACTCCCCAAAACTCACTACAGGA

case2h/1-482              AAACGTTA
gnl|ti|508398655/154-589  AAACTTTA
case2c/1-163              AAACGTTA
```

chr2:69566745:scaffold_37688:564230:564543:-
...precise deletion of AluY in human

CLUSTAL

```
case3c/1-464              AAAAAAAAAAAAAAAAAGAAGTAGCTGATTCTTA-----TTTTTCATATAAGCTATCTTT
gnl|ti|507941191/90-555   AAAACAAACAAACAAAAAAAGTAACTGGTTCTTAATTTATTTTTCATATAAGCTATCTTT
case3h/1-150              AAAAAAAAAAAAAAAAAGAAGTAGCTGATTCTTA-----TTTTTCATATAAGCTATCTTT

case3c/1-464              TCTTGGGGCTTCTTAAAAAAAATGGACAGCTCCAGGCCGGGCGCAGTGGCTCACGCCTGT
gnl|ti|507941191/90-555   TCTTGTGGCTTCTTAAAAAAAAAAAAGAAC----GGCCGGGCACGGTGGCTCAAGCCTGT
case3h/1-150              TCTTGGGGTTTCTTAAAAAAA-TGGACAGCTCCA----------

case3c/1-464              AATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCAT
gnl|ti|507941191/90-555   AATCCCAGCACTTTGGGAGGCCGAGACAGGCGGATCACGAGGTCAGGAGATCGAGACCAT
case3h/1-150              ----------

case3c/1-464              CCTGGCTAACACGGTGAAACCCCGTCTCTACTAAAAA-TACAAAA-AATTAGCCGGGCGT
gnl|ti|507941191/90-555   CCTGGCTAACACGGTGAAACCCCGTTTTTATTAAAAAATACAAAACAACTAGCCGGGGGA
case3h/1-150              ----------

case3c/1-464              GGTAGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAAC
gnl|ti|507941191/90-555   GGTGGGGGGTGCCTGTAGTCCCAGCTACTCGGGAGGGTGAGGCAGGAGAATGGGGTAAAC
case3h/1-150              ----------

case3c/1-464              CCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACA
gnl|ti|507941191/90-555   CCGGGAGGGGGAGCTTGCAGTGAGCTGAGATCCGGCCACTGCACTCCAGCCTGGGCAACA
case3h/1-150              ----------

case3c/1-464              GAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAATGGACAGCTCCAGTTATTAACTA
gnl|ti|507941191/90-555   GAGCAAGACTCCGTTTCAAAAAAAAAAAAAAAAAAAA-GGACAGCTCCAGTTATTAACTA
case3h/1-150              ----------GTTATTAACTA

case3c/1-464              GTAATGCTCTTAATTTCCTAAATATAAAATTAATTTGGCTAAGAACCCAGA
gnl|ti|507941191/90-555   GTAATGCTCTTAATTTCCTAAATATAAAATTCATTTAGCTAAGAACCCAGA
case3h/1-150              GTAATGCTCTTAATTTCCTAAATATAAAATTAATTTGGCTAAGAACCCAGA
```

chr2:84195010:scaffold_36190:2002525:2002866:-
...precise deletion of AluY in human

CLUSTAL

```
case4c/1-491              GTGTACACATATGAATCTCAAAGCTGACATCTTTGTAACTAACATCTTAAAAAGCCTGAA
gnl|ti|332419397/459-927  AGGTACACATATGAATCTCAAAGCTGACATCTTTGCAACTAAAATCTTAAAAAGCCTGAA
case4h/1-150              GTGTACACATATGAATCTCAAAGCTGACATCTTTGTAACTAACATCTTAAAAAGCCTGAA

case4c/1-491              ATCTTAAAAATCAACATCTTGGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTT
gnl|ti|332419397/459-927  ATCTTAACAATCAGCATGTT----------GGTGGCTCACGCCTGTAATCCTAGCACTTT
case4h/1-150              ATCTTAAAAATCAACATCTT----------

case4c/1-491              GGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGG
gnl|ti|332419397/459-927  GGGAGGCCGGGACTGGTGGATCACGAGGTCAAGAGATGCAGACCATCGTGGCTAACATGG
case4h/1-150              ----------

case4c/1-491              TGAAACCCCGTCTCTACT----------AAAAATACAAAAAATTAGCCGGGCGTGGTA
```

```
gnl|ti|332419397/459-927  TGAAAACCCGTCTCTTCTTAAAAAAAAAAAAAAAAAAAAAAAAAAAATAGCCAGGGGAGGTG
case4h/1-150               ------------------------------------------------------------

case4c/1-491               GCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGG
gnl|ti|332419397/459-927  GTGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCCGAGGCAGGAGAATGGTGTGAACCCGG
case4h/1-150               ------------------------------------------------------------

case4c/1-491               GAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGC
gnl|ti|332419397/459-927  GAGGCGGAGCTTGCAGTGAGCCGAGATCGCACCACTGCACTCCAGCCTGGGCGACAGAGC
case4h/1-150               ------------------------------------------------------------

case4c/1-491               GAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
gnl|ti|332419397/459-927  CAGACTCCATCTCAAACAAACAAACAAATGAACAAAAAA--------------------
case4h/1-150               ------------------------------------------------------------

case4c/1-491               AATTCAACTTCTTAGTCATTTAAAAATCTGATATCTCACCTTCATGAAACAAAAAAGAGT
gnl|ti|332419397/459-927  ---TCAACATCTTAATCATTTAAAAATCTGATATCTCAGCTTCATGAAACAGAAAAGAGT
case4h/1-150               ------------AGTCATTTAAAAATCTAATATCTCAACTTCATGAAACAGAAAAGAGT

case4c/1-491               AGGATCAGTGCAGTAAAAAGAAG
gnl|ti|332419397/459-927  AGGATCAGTGCAGTAGAAAGAAG
case4h/1-150               AGGATCAGTGCAGTAGAAAGAAG
```

chr2:157669833:scaffold_37694:13104195:13104426:-
...rhesus sequence filling this indel is found multiple times in the human genome. Chimp sequence is an L1PA2 without a discernable TSD. Likely multiple independent insertions.

CLUSTAL

```
case5h/1-100               AGATTTAAAATCATTTTGATGATTTCTCTAAGATTATGTTGAGTTTCT------------
gnl|ti|540008912/212-794  AGATTTTAAATCATTTTGATGATTTCTCTAAAATTATGTTGGGTTCCTTTTTTTTTTTTT
case5c/1-331               AGATTTTAAATCATTTTGATGATTTCTCTAAGATTATGTTGAGTTTTT------------

case5h/1-100               ------------------------------------------------------------
gnl|ti|540008912/212-794  TTTTGAATAGCAAATAGTTTATTGGTAAGTACATGGTTTCAACAAGAGTAATAAATTCAC
case5c/1-331               ------------------------------------------------------------

case5h/1-100               ------------------------------------------------------------
gnl|ti|540008912/212-794  ATGAAAAGGAGACAATAATCAAGTCAAAAGAATAAATGCTTACTAATCATCAGAAAATCT
case5c/1-331               ------------------------------------------------------------

case5h/1-100               ------------------------------------------------------------
gnl|ti|540008912/212-794  GTGGCCATTAGGGCTGGCACGTAAAAATCCAAAATCACTCAAAGGCCAAATCTTAAAGAA
case5c/1-331               ------------------------------------------------------------

case5h/1-100               ------------------------------------------------------------
gnl|ti|540008912/212-794  GATTCGTCCTCTTATTAGTCCATATGGAATAGGTCCATAGTACACAGAATCTGTAGAATT
case5c/1-331               ------------------------------------------------------------AA

case5h/1-100               ------------------------------------------------------------
gnl|ti|540008912/212-794  CTGTAGATTATCACCTTCTAACCAAACATGACCCATTGGCACCATAGGTCTGTTCTACAA
case5c/1-331               AATTGAACAATGAGATCACATGGACACATGAAGGGGAATATCACACTCTGGGGACTGTGG

case5h/1-100               ------------------------------------------------------------
gnl|ti|540008912/212-794  TTCATTCAATTTTTACGGAAGTCACAGGCCTTCCAGAAAAAAAAAAAAAATTAAAGCATGT
case5c/1-331               TGGGGTGGGGGGAGGGGGGAGGGATAGCATTGGGAGATATACCTAAGGCTAGATGACGAG

case5h/1-100               ------------------------------------------------------------
gnl|ti|540008912/212-794  TTCAAAGCATAGGTGGATGATCCATGATTTCAGGAATCCTCGGGCTCCAAAGAACCCTGA
case5c/1-331               TTAGTGGGTGCAGCGCACCAGCATGGCACATGTATACATATGTAACTAACCTGCACAATG

case5h/1-100               ------------------------------------------------------AAAACTTGG
gnl|ti|540008912/212-794  AGACCCTCAACCAGGACACAGGTGGGCCTTTCTCACCTATGTTGGATTTTTAAAAGTTGG
case5c/1-331               TGCACATGTACCCTAAAACTTAAAGTATAAAAACAAAAACAAAAAAA--AAATAACTTGG

case5h/1-100               TTTGTGTTAAATCTGTTCATTGATTTGGAGACTGACAGATATA
gnl|ti|540008912/212-794  TTTGCATTAAATCTGTTCATTGATTTGGAGACTGACAGATATA
case5c/1-331               TTTGTGTTAAATCTGTTCATTGATTTGGAGACTGACAGATATA
```

chr2:183829238:scaffold_37634:13637014:13637174:+
...non-transposable element sequence, likely previously inserted with TSD, also found on human chromosome 5, precise deletion in human

CLUSTAL

```
case6c/1-260              GTCAGAGGGGTAAGAAAGCCAAGAGAGAGTAGTGATAAATGCTAAAAAAGAAATAAGAAA
gnl|ti|513283760/530-796  GTCAGAGGGGTAAGAAAGCCAAGAGAGAGTAGTGATAAATGCTAAAAAAGAAATAAGAAA
case6h/1-100              GTCAGAGGGGTAAGAAAGCCAAGAGAGAGTAGTGATAAATGCTAAAAAAGA---AAGAAA

case6c/1-260              GGAGTGGTCAGCCTCACACAGTCCTTCTGAGATTGTTAAGCAGATTACTTCCACCAGTAT
gnl|ti|513283760/530-796  GGAGTGGTCAGCCTCGCACAGTCCTTCTGAGATTGTTAAGCAGATTACTTCCATCAGTAT
case6h/1-100              GGAGTGGTCAGC----------------------------------------------

case6c/1-260              TGAGCCAGGAGTTGAGGTGGAAGTCACCATTGCAGATGCTTAAGTCAACTATTTTAATAA
gnl|ti|513283760/530-796  TGAGCCAGGAGTTGAGGTGGAAGTCACCATTGTAGATGCTTAAGTCAACTATTTTAATAA
case6h/1-100              ------------------------------------------------------------

case6c/1-260              ATTGATTACCAATTGTTTTAAAAAAAAAAAA-------------AAGAAAGGAGTGGTCA
gnl|ti|513283760/530-796  ATTGATTACCAGTTGTTTTAAAAAAAAAAAAAAAAAAAAAAAAAA------GAGTGGTCA
case6h/1-100              ------------------------------------------------------------

case6c/1-260              GCAGTAAACATCAGTTGCGCAATAAGAAGACCA
gnl|ti|513283760/530-796  GCAGTAAACATCAGTTGCCCAATAAGATCAAAA
case6h/1-100              --AGTAAACATCAGTTGCGCAATAAGAAGACCA
```

chr2:187720209:187720500:scaffold_37634:17562752:+
...AluY in Rhesus, partial deletion and gene conversion to AluYb8 in human, precise deletion in chimpanzee

CLUSTAL

```
case7h/1-438              TGATATGTATGAGAAAGATACTGGTTTCTACATTTTGCTTTTAAATACTGTGAAGTAAAG
gnl|ti|507795985/79-552   TGATATGTATGAGAAAGATACTCGTTTCTACATTTTGCTTTTAAATACTATGAGGTAAAG
case7c/1-146              TGATATGTATGAGAAAGATACTGGTTTCTACATTTTGCTTTTAAATACTGTGAAGTAAAG

case7h/1-438              CACGAGACAACTTAAAAAAATATCTATAATG-----------------------------
gnl|ti|507795985/79-552   GATGAGACAACTTAAAAAA-TATCTATAATGGGCCGGGCGCGGTGGCTCAAGCCTGTAAT
case7c/1-146              CACGAGACAACTTAAAAAA-TATCTATAGTG-----------------------------

case7h/1-438              ---AGCACTTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGATCGAGACCATCCT
gnl|ti|507795985/79-552   CCCAGCACTTTGGGAGGCCAAGACGGGCAGATCATGAGGTCAGGAGATCGAGACCATCAT
case7c/1-146              ------------------------------------------------------------

case7h/1-438              GGCTAACAAGGTGAAACCCCGTCTCTACTAAAA--ATACAAAAAATTAGCCGGGCGTGTT
gnl|ti|507795985/79-552   GGCTAACACGGTGAAACCCCGTCTATACTAAAAAAATACAAAAAACTAGCCAGGCGAGGT
case7c/1-146              ------------------------------------------------------------

case7h/1-438              GGTGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCG
gnl|ti|507795985/79-552   GGTGGGCGCCTGTAGCCCCAGCT---TGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCG
case7c/1-146              ------------------------------------------------------------

case7h/1-438              GGAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCTGCAGTCCGGCCTGGGC
gnl|ti|507795985/79-552   GGAGGCGGAGCTTGCAGTGAGCTGAGATCCGGCCACTGCACTC-------CAGCCTGGGT
case7c/1-146              ------------------------------------------------------------

case7h/1-438              GACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAA-------------AAAAAAAAT
gnl|ti|507795985/79-552   GACAGAGCGAGACTCCATCTCAAAAAAAAAAAAAAAAAAAAATATATATATATATATATAT
case7c/1-146              ------------------------------------------------------------

case7h/1-438              CTATAATGAAAATAGAGCAAAAATTACTGAAGTAGCATACAAATGACAAAGATGAGAGAA
gnl|ti|507795985/79-552   ATAAAATGAAAATAGAGCAAAAATTACTGAAGTACCATACAAATGACAAAGATGAGAGAA
case7c/1-146              --------AAAATAGAGCAAAAATTACTGAAGTAGCATAAAA-TGACAAAGATGAGAGAA

case7h/1-438              AGAAT
```

```
gnl|ti|507795985/79-552  AGAAT
case7c/1-146             AGAAT
```

chr2:192411818:192412134:scaffold_37634:22298194:+
...AluY in Rhesus, gene conversion to AluYa5 in human, precise deletion in chimpanzee

CLUSTAL

```
case8h/1-476             AAAGAACTTTGCCTGTGTTGACTTTGTAAATGTTGTCTTAACCAAGCTTGGGCAACTATC
gnl|ti|498009529/103-591 AAAGAACTTTGCCTGTCTTGACTTTGTAAATGTTGTCTTAACCGAGCTTGAGCAACTATT
case8c/1-160             AAAGAACTTTGCCTGTGTTGACTTTGTAAATGTTGTCTTAACCAAGCTTGAGCAACTATT

case8h/1-476             TTTTTAAGAATCAAAAACACA--GCCGGGCGTGGTGGCTCACGCCTGTAATCCCAGCACT
gnl|ti|498009529/103-591 TTTTTAAGAATCAAAAACACAAAGCCGGGCGCGGTGGCTCAAGCCTGTAATCCCAGCACT
case8c/1-160             TTTTTAAGAATCAAAAACACAA----------------------------------------

case8h/1-476             TTGGGAGGCCGAGGGGGGTGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAAC
gnl|ti|498009529/103-591 TTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACAT
case8c/1-160             ------------------------------------------------------------

case8h/1-476             GGTGAAACCCCGTCTCTACTAAAAATACAAAAAA-----TTAGCCGGGCGTAGTGGCGGG
gnl|ti|498009529/103-591 GGTGAAACCCCGTCTCTACTAAAAATACAAAAAAAAAAAACTAGCCGGGCGTGGTGGCGGG
case8c/1-160             ------------------------------------------------------------

case8h/1-476             CGCCTGTAGTCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGC
gnl|ti|498009529/103-591 CACCTGTAGTCCCAGCTACTCGG-AGACTGAGGCAGGAGAATGGCGTGAACCTGGGAGGC
case8c/1-160             ------------------------------------------------------------

case8h/1-476             GGAGCTTGCAGTGAGCCGAGATCCCGCCGCTGCACTCCAGCCTGGGCGACAGAGCG--AG
gnl|ti|498009529/103-591 GGAGCTTGCAGTGAGCCGAGATCGCACCACTGCACTCCAGCCTGGGTGACACAGCGCGAG
case8c/1-160             ------------------------------------------------------------

case8h/1-476             ACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAGAATC-----AAAAACACGAAAAAGCCA
gnl|ti|498009529/103-591 ACTCCGTCTCAAAAAAAAAAAAACAAAAACAAAAACAAAACAAAAAACACAAAAAAGCCA
case8c/1-160             ----------------------------------------------------AAAAGCCA

case8h/1-476             CTCCACACAAAATATAATGCATCAAGTGTGGCTGCTGAATTACCGGAGTTAACATAGGTA
gnl|ti|498009529/103-591 CCCCATATAAAATACAGTGCATCAGGTGTGGCTGCTAAATTACCAGAGTTAACATATGTA
case8c/1-160             CTCCACACAAAATATAATGCATCAAGTGTGGCTGCTGAATTACTGGAGTTAACATAGGTA

case8h/1-476             AGCACACACC
gnl|ti|498009529/103-591 AACATACACC
case8c/1-160             AGCACACACC
```

chr2:210695342:210695583:scaffold_36996:2672593:+
...precise deletion of AluSg in chimpanzee

CLUSTAL

```
case9h/1-363             TAACACATGTACTTCTAAATGTGTTTGAAGTTTGAGTCTTACTAACTTATGTAGAATTCA
gnl|ti|562549776/185-546 TAATACATGCAGTTCTAAATGTGTTTGAAGTTTGAGTCTTACTGACTTATATAGAATTCA
case9c/1-122             TAACACATGTACTTCTAAATGTGTTTGAAGTTTGAGTCTTACTAACTTATGTAGAATTCA

case9h/1-363             CATATTTTTCGGCCGGGCACGGTGACTCACGCC-TGTAATCCCAG--CACTTTGGGAGGC
gnl|ti|562549776/185-546 CATATTTTTCGGCCAAACATGGTGAC--ACACC-TGTAATCACAG--AACTTTGGGAGGC
case9c/1-122             CATATTTTTC----------------------------------------------------

case9h/1-363             CGAGGCAGACAGATCACAAGGTCAGGAGTTTGAGACCAGCCTGACCAACATGGTGAAACC
gnl|ti|562549776/185-546 CGAGGCGGGCAGATCACAAGGTCAGGAGTTTGAGACCAGTCTGACCGACATGGTGAAACC
case9c/1-122             ------------------------------------------------------------

case9h/1-363             CCCGTCTCTACTAAAAATA-AAAAATTAGCCGGGCATGGTGGCACGTGCCTGTAATCCCA
gnl|ti|562549776/185-546 TCCATCTCTACTAAAAATACAAAAATTAGCCAGGCTTGGTGGCAGGCATCTGTAATCCCA
case9c/1-122             ------------------------------------------------------------

case9h/1-363             GCTACTCAGGAGGCTGAAGAAGGAGAATCGCTTGAACCCGGGAGACGGAGGTTGCAGTGA
```

```
gnl|ti|562549776/185-546  GCTACTCAGGAGGCTGAGGCAGGAGAATCGCTGGAACCCAGGAGGCAGAGGTTGCAGTGA
case9c/1-122              ------------------------------------------------------------

case9h/1-363              GCCGAGATATTTTTCACACAAAAACCTTAAATATTACAGTGCAGTGGTTAGCAACTCTGA
gnl|ti|562549776/185-546  GCCAAGATATTTTTCACACGAAAACCTTAAATATTACAGTGCAGTGGTTAGCAACTCTGA
case9c/1-122              ---------------ACACAAAAACCTTAAATATTACAGTGCAGTGATTAGCAACTCTGA

case9h/1-363              AAGTTTG
gnl|ti|562549776/185-546  AAGTTTG
case9c/1-122              AAGTTTG
```

chr2:231119183:231119489:scaffold_34265:2315868:+
...AluY in Rhesus, gene conversion to AluYg6 in human, precise deletion in chimpanzee

CLUSTAL

```
case10h/1-461             TCAGTGCA-TTTATTTTATATATTATTTCTATAATGTGCATATTTAATAACAAATAACAT
gnl|ti|540018070/562-1046 TCAATGCA-TTTATTTTATATATTATTTCTATAATGTGCATATTTAATAACAAATAACAT
case10c/1-156             TCAGTGTAGTTTATTTTATATATTATTTCTATAATGTGCATATTTAATAACAAATAACAT

case10h/1-461             TTATATAAAGTATGTTTA--------------GGCCGGGCGCGGTGGCTCACGCCTGTAA
gnl|ti|540018070/562-1046 TTATATAAAGTATGTTTAATATATTAAGTCTAGGCCGGGCGCGGTGGCTCACGCCTGTAA
case10c/1-156             TTATATAAAGTATGTTTAATATATTAAGTCTA----------------------------

case10h/1-461             TCCCAGCACTTTGGGAGGCCGAGACGGGCGGATCACGAGGTCAGGAGATCGAGACCATCC
gnl|ti|540018070/562-1046 TCCCAGCACTTTGGAAGGCCGAGACAGGCAGATCATGAGGTCAGGAGATCGAGATCATCC
case10c/1-156             ------------------------------------------------------------

case10h/1-461             TGGCTAACACGGTGAAACCCCGTCTCTACTAAAAATACAAAAA------TTAGCCAGGCA
gnl|ti|540018070/562-1046 TGGCTAACACGGTGAAACCCCCTCTCTACTAAAAATACAAAAAAAAAAAATTAGCCGGGCG
case10c/1-156             ------------------------------------------------------------

case10h/1-461             TGGTGGCGTGCGCCTGTAGTCCCAGCTACACGGGAGGCTGAGGCAGGAGAATGGCGCGAA
gnl|ti|540018070/562-1046 TGGTGGTGGGCGCCTGTAGTCCCAGCTACACGGGAGGCTGAGGCAGGAGAATGGCATAAA
case10c/1-156             ------------------------------------------------------------

case10h/1-461             CCCGGGAGGCGGAGCTTGCAGTGAGTCGAGATCGCGCCACTGCACTCCAGCCTGGGCAAC
gnl|ti|540018070/562-1046 CCCGGGGAGCGGAGCTTGCAGTGAGCTGAGATCGCGCCACTGCACTCCAGCCTGGGCAAC
case10c/1-156             ------------------------------------------------------------

case10h/1-461             AGAGCTAAACTCCGTCTCAAAAAAAAAAAAAA----AAAGTATGTTTAATATATTAAGTC
gnl|ti|540018070/562-1046 AGAGCGAGACTCCGTTTCAAAAAAAAAAAAAAAAATATATATATATATATATATAAAGTC
case10c/1-156             ------------------------------------------------------------

case10h/1-461             TATTCTAAGGAGTTAATTTACTGAATTATTTCCTGTAATGCAACAGAGTTAATTAAATTA
gnl|ti|540018070/562-1046 TATTCTAAGAAGTTGCTTTACTGAATTATTTCCTGTAATGCAGCAGAGTTAATTATATTA
case10c/1-156             --TTCTAAGGAGTTAATTTACTGAATTATTTCCTGTAATGCAACAGAGTTAATTAAATTA

case10h/1-461             ATAAGT
gnl|ti|540018070/562-1046 ATAAGT
case10c/1-156             ATAAAT
```

chr2:234434078:234434368:scaffold_37338:1917127:-
...AluY in Rhesus, gene conversion to AluYb8 in human, precise deletion in chimpanzee

CLUSTAL

```
case11h/1-437             GAAACTAGTTATTGTCAGAAGAGACCATTGGGATGAAAAGTGTGTATCTTCGAAACCACT
gnl|ti|538232282/352-812  GAAACTAGTTATTGTCAGAAGAGACCATTGGGATGAAAAGTGTGTATCTTCAAAACCACT
case11c/1-147             GAAACTAGTTATTGTCAGAAGAGACCATTGGGATGAAAAGTGTGTATCTTCAAAACCACT

case11h/1-437             AAAGAAAAACTC--------------------------------ACACTTTGGGAGGCC
gnl|ti|538232282/352-812  AAAGAAAAACTCGGCCGGGCACGGTGGCTCAAGCCTGTAATCCCAGCACTTTGGGAGGCC
case11c/1-147             AAAGAAAAACTC------------------------------------------------

case11h/1-437             GAGGCGGGTGGATCATGAGGTCAGGAGATCAAGACCATCCTGGCTAACAAGGTGAAACCC
```

```
gnl|ti|538232282/352-812  GAGATGGGTGGATCACGAAGTCAGGAGATCGAGACCATCCTGGCTAACAGGGTGAAACCC
case11c/1-147             ------------------------------------------------------------

case11h/1-437             TGTCTCTACTAAAAA-TACAAA----AAATTAGCCGGGCGCGGTGGCGGGCGCCTGTAGT
gnl|ti|538232282/352-812  CGTCTCTACTAAAAAATACAAAAAAAAAATTAGCCGGGCGTGGTGGCGGGTGCCTGTAGT
case11c/1-147             ------------------------------------------------------------

case11h/1-437             CCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAGCGAAGCTTGCA
gnl|ti|538232282/352-812  CCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCCTGAACCCGGGAGGCAGAGCTTGCA
case11c/1-147             ------------------------------------------------------------

case11h/1-437             GTGAGCCGAGATTGCACCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGAGCGAGACTCC
gnl|ti|538232282/352-812  GTGACCTGAGATCCGGCCACTGCACTCCA-------GCCTGGGTGACAGAGCAAGACTCC
case11c/1-147             ------------------------------------------------------------

case11h/1-437             GTCTCAAAAAAAAAAAAAAGAAAAAAAGAAAGAAAAACTCAGTCTGTACCTTCTATAATA
gnl|ti|538232282/352-812  GTCTCAAAAAAAAAAAAAAAAAA-------GAAAAACTCAGTCTGTACCTTCTATAATA
case11c/1-147             ----------------------------------------AGTCTGTACCTTCTATAATA

case11h/1-437             ATAATTAGTAATGCCAGTAACAGCAGGTAACATTTATTGAATGTATACAGTGTGT
gnl|ti|538232282/352-812  ATAATTAGTAATGCCAGTAACAGAAAGTAATATTTATTGAATGTGTACAATGTGT
case11c/1-147             ATAATTAGTAATGCCAGTAACAGCAGGTAACATTTATTGAATGTATACAGTGTGT
```

chr3:3802176:scaffold_32934:3840390:3840497:+
...imprecise deletion of L2 fragment in human, no similarity at breakpoint

CLUSTAL

```
case12c/1-207             TAATTCCTCATGTCCCCACCTTAGCCCCCATATGATCCTCTCAGGCCTTGCTAAAGCTCA
gnl|ti|583293545/557-763  TCATCCCTCATGTCCCCACCTCTGCCCCCACATGATCCTCTCAGGCCTTGCTAAAGCTGA
case12h/1-101             TAATTCCTCATGTCCCCACCTTAGCCCCCATATGATCCTCTCAGGCCTTGTT--------

case12c/1-207             GAGCCAAAAAAACAGTTCTTAGAACACAGTAGAAACTCAACAAATATTTGCTGAATTAGTA
gnl|ti|583293545/557-763  GAGCCAAAAAACAGTTCTTAGAACACAGTAGAAACTCAACAAATATTTGCTGAATTAGTG
case12h/1-101             ------------------------------------------------------------

case12c/1-207             TCCATGTTCGTCTAGTCTCCCAGTTCATAGGCCATCATGTTTTACATGTAGCTGATGTTT
gnl|ti|583293545/557-763  TCCATGTTTGTCTAGTCTCCCAGTTCATAGGCCATCATGTTTTACAAGTAGCTGATGTTT
case12h/1-101             ----------------------------------------GTTTTACATGTAGCTGATGTTT

case12c/1-207             GTCGATGTCTACTGAGTCAAATATAGA
gnl|ti|583293545/557-763  GTCGATGTCTACTGAGTCAAATATAGA
case12h/1-101             GTCGATGTCTACTGAGTCAAATATAGA
```

chr3:34069680:scaffold_37683:15182979:15183196:-
...independent insertions of an AluY in Rhesus and L1PA2 in chimpanzee in the same site

CLUSTAL

```
case13c/285-601           AAGCAGATGGGGGTCAGGAAGCAAGAAGAGGTAGGTTAGAAAGCCTTTACGGAAGGGGAA
gnl|ti|583406125/478-897  AAG-GGATGGGGGTCAGGAAGCAAGAAGAGGTAGATTAGAAAGGCTTTACCAGCCGGGCG
case13h/1-99              AAG-AGATGGGGGTCAGGAAGCAAGAAGAGGTAGATTAGAAAGGCTTTAC----------

case13c/285-601           TA----------CCATACTCTGGGGACTGTGGTGGG--GAGGGGGGAGG----------
gnl|ti|583406125/478-897  CGGTGGCTCACGCCTGTAATCCCAGAACTTTGGCAGGCCGAGGCGGGCGGATCACGAGGT
case13h/1-99              ------------------------------------------------------------

case13c/285-601           --GGGGAGGGATAGCAT--TGGGAGATATACCTAA----------TGCTAGA--------
gnl|ti|583406125/478-897  CAGGAGATCGAGACCATCCTGGCTAACACAGTGAAACCCCGTCTCTACTAAAAATACAAA
case13h/1-99              ------------------------------------------------------------

case13c/285-601           ------------------------TGACGAGT--TAGTGGGTGCAGCGCACCAGCATGGC
gnl|ti|583406125/478-897  AAGAAAAAAATTAGCCGGGCATGGTGGCGAGCGCCTGTAGTCGCAGC-TACTCGGGAGGC
case13h/1-99              ------------------------------------------------------------

case13c/285-601           ACATGTATACATATGTAACTAACCTG------------CACAATGTGCACA---------
gnl|ti|583406125/478-897  TGAGGCAGGAGAATGGCGTGAACCTGGGAGGCGGAGCTTGCAGTGAGCTGGATCGCGCCA
```

```
case13h/1-99              ------------------------------------------------------------

case13c/285-601           -TGTACCCTAA---AACTTAAAGTATAATAATA------AAAAAAAAAAAAAAAAAAAGAA
gnl|ti|583406125/478-897  CTGCACTCCAGCCTGGGTGAAAGAGCGAGACTCCTTCTCAAAAGAAAGAAAAAAAAAAGAA
case13h/1-99              ------------------------------------------------------------

case13c/285-601           AGCCTTTACCAAAAAAAA---CATGGTGTTTGAACTGATACTTAAACTGGTTCTTAAGCT
gnl|ti|583406125/478-897  AGGCTTTACCAAAAAAAAAACCACAGTGTTTGAACTGATACTTAAGCTGGTTCTTAAGCT
case13h/1-99              ---------CAAAAAAAA---CATGGTGTTTGAACTGATACTTAAACTGGTTCTTAAGCT

case13c/285-601           GG
gnl|ti|583406125/478-897  GG
case13h/1-99              GG
```

chr3:127318836:127319143:scaffold_36943:880680:-
...precise deletion of AluSg in chimpanzee

CLUSTAL

```
case14h/1-462             CGGTGAGTATTCACTTAGCACCCAGGATCTT------------TAAACCAACCAGTGCAT
gnl|ti|501884074/122-632  CGGTGAGTATTCACTTAGCATCCGGGATCTTCACACTCTGCCCTAAACCAACCAGTGCAT
case14c/1-155             CGGTGAGTATTCACTTAGCACCTAGGATCTT------------TAAACCAACCAGTGCAT

case14h/1-462             TCCCTTCTGAAGAGCCTTCAACTTCACTTTAAAAAATCCTGTGATGGGGCGGGCGTGGTG
gnl|ti|501884074/122-632  TCACTTCTGAGGATCCTTCAACTTCACTTAAAAAAATCCTGTGACGGGGCGGGCATGGTG
case14c/1-155             TCCCTTCTGAAGAGCCTTCAACTTCACTTTAAAAAATCCTGTGATGG-------------

case14h/1-462             GCTCATGCCTGCAATCCCAGCACTTTGGGAGGTCAAGGTGGGCAGATCACGAGGTCAGGA
gnl|ti|501884074/122-632  GCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGCGGGCGGATCACCAGGTCAGGA
case14c/1-155             ------------------------------------------------------------

case14h/1-462             GTTCGACACCAGCCTTACCAACATGGTGAAACCCTGTCTCTACTAAAAATACAAAAATTA
gnl|ti|501884074/122-632  GCTCAACACCAGCCTTACCAACATGGTGAAACCCTGTCTTTACTAAAAATACAAAAATTA
case14c/1-155             ------------------------------------------------------------

case14h/1-462             GCCGGGTGTGGTGACACGCGCCTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAAT
gnl|ti|501884074/122-632  GCCAGGTGTGGTGGAGTGCGCCTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAAT
case14c/1-155             ------------------------------------------------------------

case14h/1-462             CACTTGAATCCGGGAGGTGGAGGTTGCAGTGAGCCGAAATCATGCCACTGCACTCCAGCC
gnl|ti|501884074/122-632  CACTTCAACCCGGGAGGTGGAGGTTGCAGTGAGCCGCGATCATGCCACTGCACTCCAGCC
case14c/1-155             ------------------------------------------------------------

case14h/1-462             TGGGCAACAGAACGAGACTTTGTCTAAAAAAAAAAAAA----------------------
gnl|ti|501884074/122-632  TGGGCGACAGAGTGAGATTCTGTCAAAAAAAAAAAAAAAAAAAAAAAAAACACACACACAC
case14c/1-155             ------------------------------------------------------------

case14h/1-462             --------------AAAAATCCTGTGATGGCAAAGACGTTCT-CAGGCTAAATTCAACTC
gnl|ti|501884074/122-632  ACAACAAAACAAAATAAAATGCTGTGATGGCAAAGACGTTTTTCAGGCTAAATTCAACTC
case14c/1-155             -----------------------------CAAAGACGTTCT-CAGGCTAAATTCAACTC

case14h/1-462             ATGTATTTTTTACATACATAATATTTGAAGG
gnl|ti|501884074/122-632  ATGTATTTTTTACATAGATAATATTTGAAGG
case14c/1-155             ATGTATTTTTTACATACATAATATTTGAAGG
```

chr4:62743877:62744205:scaffold_37623:10184651:+
...AluY in Rhesus, partial TSD deletion and gene conversion to AluYb8 in human, precise deletion in chimpanzee

CLUSTAL

```
case15h/1-494             AGTGACAAAGTGCAGGTCTACAGAACAAAGGGCACAAATAAAACTAGGCAATTTTGTGTG
gnl|ti|523759330/256-733  ACTGACAAAGTGCAGGTCTACAGAAAAAAAGGCACAAATAAAACTAGGCAATTTTGTGTG
case15c/1-166             AGTGACAAAGTGCAGGTCTACAGAACAAAGGGCACAAATAAAACTAGGCAATTTTGTGTG

case15h/1-494             TTGACTATAAAAGAGCATTTTG------------------------GGCCGGGCGCGGTG
gnl|ti|523759330/256-733  TTGACTATAAAAGAGCATTTTGATGTTGTTGAAAATATTTTACACAGGCCGGGCGCGGTG
```

```
case15c/1-166             TTGACTATAAAATAGCATTTTGATGTCATTGAAAATATTTTACACA----------

case15h/1-494             GCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGA
gnl|ti|523759330/256-733  GCTCAAGCCTGTAATCCCAGCACTTTGGGAGGCCAAGACGGGCGGATCACGAGGTCAGGA
case15c/1-166             ------------------------------------------------------------

case15h/1-494             GATCGAGACCATCCTGACTAACAAGGTGAAACCCCGTCTCTACTAAAAA--TACAAAAAA
gnl|ti|523759330/256-733  GATCGAGACCATCCTGGCTAACACAGTGAAACCCCGTCTCTACTAAAAAAATACAAAAAA
case15c/1-166             ------------------------------------------------------------

case15h/1-494             TTAGCCGGGCGCGGTGGTGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAG
gnl|ti|523759330/256-733  CTAGCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCAGGAGGCTGAGGCAAGAG
case15c/1-166             ------------------------------------------------------------

case15h/1-494             AATGGCGTGAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCG
gnl|ti|523759330/256-733  AATGGCGTAAATCCGGGAGGCGGAGCTTGCAGTGAGCCGACATCCGGCCACTGCACTCCA
case15c/1-166             ------------------------------------------------------------

case15h/1-494             CAGTCCGGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAA
gnl|ti|523759330/256-733  -------GCCTGGGCAACAGAGCGAGACTCCGCCTCAAAAAAAAAAAAA-----------
case15c/1-166             ------------------------------------------------------------

case15h/1-494             AAAAAAGAGCATTTTGATGTCGTTGAAAATATTTTACACAAATGAAAAAGTGGTGATGGT
gnl|ti|523759330/256-733  ----------------------GAAAATATTTTACACAAATGAAAAAGTGGTGATGGT
case15c/1-166             ----------------------------------------AATGAAAAAGTGGTGATGGT

case15h/1-494             TATCACGTGAGGGGGTTCTGTGGTCTCCCCTTCTCTTAGT
gnl|ti|523759330/256-733  TATCATGTGAAGGGGTTCTGTGGTCTCCCTTTCTCTTAGT
case15c/1-166             TATCACATGAGGGGGTTCTGTGGTCTCCCCTTCTCTTAGT


chr4:110847016:110847328:scaffold_37491:1921071:+
...precise deletion of AluY in chimpanzee

CLUSTAL

case16h/1-470             TCATTTGCTTTATTTTAGAAAAGCCTATTAGTACATTATAATTCAGTACTAGAACTTCAA
gnl|ti|540783035/118-571  TCATTTGCTTTATTTTAGAAAAGCCTATTAGTACATTATAATTCAGTACTAGAACTTCAA
case16c/1-158             TCATTTGCTTTATTTTAGAAAAGCTTATTAGTACATTATAATTCAGTACTAGAACTTCAA

case16h/1-470             TTATTTTATTAAAAAGTCCACTCCAG-CCGGGCGCGGTGGCTCACGCCTGTAATCCCAGC
gnl|ti|540783035/118-571  TTATTTTATTAAAAAGTCCACTCCGGGCTGGGCATGGTGGCTCACGCCTGTAATCCCAGC
case16c/1-158             TTATTTTATTAAAAAGTCCACTCCA----------------------------------

case16h/1-470             ACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAA
gnl|ti|540783035/118-571  ACTTTGGGAGGCTAAGGCAGGCAGATCATGAGGTCAGGAGATCGAGACCATCCTGGCAAA
case16c/1-158             ------------------------------------------------------------

case16h/1-470             CACGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGTGGTAGCGGGCG
gnl|ti|540783035/118-571  CACGGTGAAACCCCGTCTCTACTAAATATACAAAAAATTAGCTGGGCAAGGTGGCGGGCG
case16c/1-158             ------------------------------------------------------------

case16h/1-470             CCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGG
gnl|ti|540783035/118-571  CCTGTAGTCCCAGCTTCTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGG
case16c/1-158             ------------------------------------------------------------

case16h/1-470             AGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTC
gnl|ti|540783035/118-571  AGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCCAGCCTGGGGGACAGAGCGAGACTC
case16c/1-158             ------------------------------------------------------------

case16h/1-470             CGTCTCAAAAAAAAAAAAAAAAAAAAAAAGTCCGCTCCAAGTATGTATGTTTTGAGTGTTA
gnl|ti|540783035/118-571  CGTCTCAAAAA----------------GCCCACTCCAAGTATGTATGTTTTGAGTGTTG
case16c/1-158             ------------------------------------AGTATGTATGTTTTGAGTGTTA

case16h/1-470             CATTAGTTGTTAAAGTTGGTTGCACTTTTGGCTAGTGTTTAAAAGGTGTCA
gnl|ti|540783035/118-571  CATTAGTTGTTAACTTAGGTTGCACTTCTGGCTAGTGTTTAAAAGGTGTCA
case16c/1-158             CATTAGTTGTTAAAGTTGGTTGCACTTTTGGCTAGTGTTTAAAAGGTATCA
```

chr4:113908780:113908932:scaffold_37491:5060691:+
....AluY in Rhesus, partial deletion/gene conversion to AluYb9 in human, precise deletion in chimpanzee

CLUSTAL

```
case17h/1-252             TATGAAATTTGTGTTGTGTCTTCAGGTGATTTAAAAAAATATATGACAT----------
case17c/1-100             TATGAAATTTGTGTTGTGTCTTCAGGTGATTTAAAAAAATATATGACAT----------
gnl|ti|541340560/355-730  TATGAAATTTGTGTTGTGTCTTCAGG----------------------GGCGCGGTGGC

case17h/1-252             ------------------------------------------------------------
case17c/1-100             ------------------------------------------------------------
gnl|ti|541340560/355-730  TCAAGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACAAGGTCAGGAGA

case17h/1-252             ------------------------------------------------------------
case17c/1-100             ------------------------------------------------------------
gnl|ti|541340560/355-730  TCGAGACCACAGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGCGGTG

case17h/1-252             -----------------------------------------------------------CGG
case17c/1-100             ------------------------------------------------------------
gnl|ti|541340560/355-730  ACGGGCGCCTGTAGTCCCAGCGACTCAGGAGGCTGAGGCAGGAGAATGGCGGGAACCCGG

case17h/1-252             GAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCAGCCTGGGCG
case17c/1-100             ------------------------------------------------------------
gnl|ti|541340560/355-730  GAGGCGGAGCTTGCAGTGAGCCGAGATCGCGCCACTGCACTCC-------AGCCTGTGCA

case17h/1-252             ACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAATATAT
case17c/1-100             ------------------------------------------------------------
gnl|ti|541340560/355-730  ACAGCGTGAGACTCCGTCTCACACGNCCNNNNNNNACATNAAAAAAA-------------

case17h/1-252             ATATATATATATATATATATATATATGACATCTATCCTGTCAAGTTGATGTTAATTTGG-AT
case17c/1-100             ------------------------------CTATCCTGTCAAGTTGATGTTAATTTGG-AT
gnl|ti|541340560/355-730  ----------------------TGACATCTATCCCGTTAAGTTGATGTTGATTTTGCAT

case17h/1-252             AGAATTGGC-TTTATAGTTGTA
case17c/1-100             AGAATTGGC-TTTATAGTTGTA
gnl|ti|541340560/355-730  GGAATTGCCCTTTCCATTTGTA
```

chr4:127295460:127295786:scaffold_34705:5896785:-
...independent insertions of L1PA2 in human, and AluY in Rhesus. Imprecise localization due to microdeletion here?

CLUSTAL

```
case18h/209-634           ATAGTAATCTAAATTGCTCCAGACAAATATTTTCTGTTTTAAAAATAGTA----------
gnl|ti|558352008/293-768  ATAGTCGTCTAAACTGATCTCGACAAAGATTTTCTGTTTTAAAAATAGTATTAACTACGT
case18c/1-114             ATAGTAATCTAAATTGATCCAGACAAATATTTTCTGTTTTAAAAATAGTATTAACTACAT

case18h/209-634           ----------------------------------------AGGACATGGATGAAATTGGAA
gnl|ti|558352008/293-768  TATAAGAAAATTTAGCAGGGTGCAGTCGTTCGTGCCTGTAATCCCGCACTTTGGGAGGGC
case18c/1-114             TATTAGAAAATTT------------------------------------------------

case18h/209-634           ATCATCATTCTCAGTAAACTATCGCAAGAACAAAAAACCAAACACCGCATATTCTCACTC
gnl|ti|558352008/293-768  GAGGCGGGTGGATCACGAAGTCAGGAGATCAAGACCACCCTGGCTAACACAGTGAAACCC
case18c/1-114             ------------------------------------------------------------

case18h/209-634           ATAGGTGGGAATTGAACAATGAGATCACATGGACACAGGAAGGGGAATATCACACTCTGG
gnl|ti|558352008/293-768  CATCTTTCCTAAAAATACAAAAAAATTAACCAGGCATTGTGGCGGGCGCCTGTAGTTCCA
case18c/1-114             ------------------------------------------------------------

case18h/209-634           GGACTGTGGTGGGGTGGGGGGAGGGGGGAGGGATAGCATTGGGAGATATACCTAATGCTA
gnl|ti|558352008/293-768  GCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCTGGGAGGCAGAGCTTGCGGTGA
case18c/1-114             ------------------------------------------------------------

case18h/209-634           GATGACGAGTTAGTGGGTGCAGCGCA-CCAGCATGGCACATGTATACATATGTAACTAAC
gnl|ti|558352008/293-768  GCCAAGATCACACCACTGCACTCCAGCCTGGGAGGGAAAGGTGTTTTTGGGAGACAGAGC
case18c/1-114             ------------------------------------------------------------

case18h/209-634           CTGCACAATGTGCACATGTACCCTAAAACTTAAAGTATAATAAAAAAATAAATAAATAAA
```

```
gnl|ti|558352008/293-768  AAAACTCCGTATCAAAAAAAAAAAAAAAAAAAAACCACACAAAAAACTCCGTCTCAAAAAA
case18c/1-114             ------------------------------------------------------------

case18h/209-634           TAAAAAAGAAAATTTAAATAGGTACAATGACCATTTATAGAAACATAATTCACTAG
gnl|ti|558352008/293-768  AAAAAAAAAAAAATTACATAGGTATAATGACCATTTATAGAAAAATAATTCACTGG
case18c/1-114             ---------------AAATAGGTACAATGACCATTTATAGAAACATAATTCACTAG
```

chr4:180115809:scaffold_31924:12267927:12268281:+
...imprecise deletion of AluY in human, but involving TSD and fortuitous upstream match

CLUSTAL

```
case19c/1-504             TGATCCTGGGCAGCAACATAAGAAGGGTTGGGTCATGGGGCCAGGATGCTTTTGTAATGG
gnl|ti|458957840/166-649  TGATCTTGGGCGGCAACATGAGAAGGGTTGGGTCATGGGGCCAGGATGCTTTTCTAAAGG
case19h/1-150             TGATCCTGGGCAGCAACATAAGAAGGGTTGGGTCGTGGGGCCAGGATGCTTTTGTAATGG

case19c/1-504             AGTCTTCTCCAAAAGAAGCATGAGCCTGAGGAAAAGAAAAGCCCCTTTTAGGCCAGGCAC
gnl|ti|458957840/166-649  GGTCTTCTCCAAAAGAAGCATGAGCCTGAGGAAAAGAAAAGTCCCCTCTAGGCCAGGCGC
case19h/1-150             AGTCTTCTCCAAAAG----------------------------------------------

case19c/1-504             TGTGGCTCATGCCTTTAATCCCAGCACTTTGGGAAACCCAGGTGGGCGGATCACCTGAGG
gnl|ti|458957840/166-649  GGTGGATCATGCCTTTAATCCCAGCACTTTGAGAAACCCAGGTGGGCAGATCACCTCAGG
case19h/1-150             ------------------------------------------------------------

case19c/1-504             TCAGGAGTTCGAGACAAAACTGGCCAACGTGGCAAAACCTCATCTCTACTAAAAATACAA
gnl|ti|458957840/166-649  TCAGGAGTTCCAGACCAAACTGGCCAACATGGCAAAACCTCATCTTTACTAAAAATACAA
case19h/1-150             ------------------------------------------------------------

case19c/1-504             AAATTAAGCGGGCATGGTGGCTTATGCTGGTAAACCCAGCTACTCGGGAGGCTGATGCAT
gnl|ti|458957840/166-649  AAATTAACCGGGCATGATGGCTCATGTCGGTAAACGCAGCTACTCGGGAGGCTGACGCAT
case19h/1-150             ------------------------------------------------------------

case19c/1-504             GAGAATCGCGTGAACCAGGGTGGCAGAGGCTGCAGTGAGCCGAGAACATGCCACAGCACT
gnl|ti|458957840/166-649  GAGAATCGCTTGTACCAGGGTGGCGGAGTTTGCAGTGAGCCGTGAACACGCCACTGCACT
case19h/1-150             ------------------------------------------------------------

case19c/1-504             CCAGGCTGGGCCACAGAGTGAGACTCCATCTGAAAAAAAAAAAAAAAAGAAAAAGAAAAAG
gnl|ti|458957840/166-649  CCAGCCTGGGCGACAGAGTGAGACTCCGTCTGAAAAAAAAAAAA----------------
case19h/1-150             ------------------------------------------------------------

case19c/1-504             AAAAAAAAGCCCCCTCTAGCATAAACTTATTTCAGAAATAACACTAGTGAACAAGTAACC
gnl|ti|458957840/166-649  ----AAAAGCCCCCTCTAGCGTAAACTTATTTCAGAAATAACACTAGCGAACAAGTAACC
case19h/1-150             ---------CCCCCTCTAGCATAAACTTATTTCAGAAATAACACTAGTGAACAAGTAACC

case19c/1-504             CTTTCCAAATTATTTTGAATCTAA
gnl|ti|458957840/166-649  CTTTCCACATTATTTTGAATCTAA
case19h/1-150             CTTTCCAAATTATTTTGAATCTAA
```

chr5:29115431:scaffold_37078:3005315:3005552:-
...imprecise deletion of HAL1 fragment in human, no flanking identity

CLUSTAL

```
case20c/1-337             AATACTGAAGAGACTGTCAGTGAGGCCCCTCCTTAATGCACCATTCCTATCAACTTTTGC
gnl|ti|509178888/301-662  AATACTGAAGAGACTATCAGTGAGGCCTCTCCTTTATGCACCATTCCTGTCAACATTCGT
case20h/1-100             AATACTGAAGAGACTGTCAGTGAGGCCCCTCCTTAATGCACCATTCCTAT----------

case20c/1-337             TCTTTGGAAACAATAACAGAAAACAGAAAGTTAATGAAAATATTGACTCCATGATATAAA
gnl|ti|509178888/301-662  TCTTTGGAAACAATAATGGAAAACAGAAAGTTAATGAAAATATTGACTCTATGATATAAA
case20h/1-100             ------------------------------------------------------------

case20c/1-337             GCAGGATACTATAAAAAGGGATATTTAAAGAAAAAA-TAGAAATTAAAAACATGATTGT-
gnl|ti|509178888/301-662  GCAGGATACTATAAAAAGGGATATTTAGAGAAAAAAATGGAAATTAAAAACATGATAGTT
case20h/1-100             ------------------------------------------------------------

case20c/1-337             -----GATATATATATATATATATATATAT------------------TTGAATGTTT
gnl|ti|509178888/301-662  TTATAAATATATATATATATAAATATATATATATATAAATATATATAATTTTGAATGTTT
```

```
case20h/1-100             ------------------------------------------------------------

case20c/1-337             GGAAGACAAAGTTGTAGACATAACTTAGAGAGTCAGAGACGTGGAGAATAGAAGACAAAG
gnl|ti|509178888/301-662  GGAAGACAAAGTTGTGGACATAACTTAGAGAGTGAGAGACATGGAGAACAGAAGACAAAG
case20h/1-100             ------------------------------------------------------------

case20c/1-337             GGAGGAGCAACTGAACAGATTAAGTATAAACATATACATTGTTTCCAAAAAGAAAACAGT
gnl|ti|509178888/301-662  GGAGGACCAACTGACCACATTAAGTATAAAGATATAGATTGCTTCCAAAAAGAAGACAGT
case20h/1-100             ------------GAACAGATTACGCATAAACCTATACATTGTTTCCAAAAAGAAAACAGT

case20c/1-337             GT
gnl|ti|509178888/301-662  GT
case20h/1-100             GT
```

chr5:169226842:scaffold_37615:12572677:12572918:+
...imprecise deletion of a MIRb element, 3 bp identity flanking the deletion, probably blunt-end deletion

CLUSTAL

```
case21c/1-341             TTACGTACCCTTTGATGAAGTATAAGCAAAAAGTTTATATTTGGACAAATTAAATTCTGG
gnl|ti|572959929/194-538  TTACGTACCCTTTAATGAAGTATAAGCAAAAAGTTTAGATTTGGACAAATTAAATTCTGG
case21h/1-100             TTACGTACCCTTTGATGAAGTATAAGCAAAAAGTTTATATTTGGACAAAT----------

case21c/1-341             CCCTGCCACTAACTTGCTCTGTAGCCTGAGTTTACTTATTTCAACTCACATAAGCCTCAA
gnl|ti|572959929/194-538  CCCTGCCACTAACTTGCTCTGTAGCCTGAGTTTACTTATTTCAACTCATATAAGCCTCAG
case21h/1-100             ------------------------------------------------------------

case21c/1-341             TTTGCTCATCAGTAACATGGAGATGATAACACCTTACTCAAAGAATTGTGGTAGAAATAA
gnl|ti|572959929/194-538  TTTGCTCAACAGTAACATGGAGATGATAACAGCTTACTCAAAGAATTGTGGTAGAAATAA
case21h/1-100             ------------------------------------------------------------

case21c/1-341             ACTGACTTCTGAATATAAAGTGCTTAGCACAGAGTTGGGCTTATAGCA----TTCATTAA
gnl|ti|572959929/194-538  ACTGACTTCTGAATATAAAGGGCTTAGGACAGAGTTGGGCTTATAGCAAGCATTCATTAA
case21h/1-100             ------------------------------------------------------------

case21c/1-341             CATGAATAACATTATGTCAGTATTTTTAAAACAAGTACCCACCATGAATTAGAATATAAA
gnl|ti|572959929/194-538  CATGAATAACATTATGTCAATATTTTTAAAACAAGTATCCACCATGAATTATAATATAAA
case21h/1-100             ------------------------------------------------------ATAAA

case21c/1-341             GTATGCCTGAATAATTAAGATGAAACATAAGACTGAATTCAAATA
gnl|ti|572959929/194-538  GTATGCATGAATAATTAAGATAAAACATAAGACTGAATTCAAATA
case21h/1-100             GTATGCATGAATAATTAAGATGAAACATAAGACTGAATTCAAATA
```

chr5:176660475:176660732:scaffold_34495:399572:-
...precise deletion of AluJo in chimpanzee

CLUSTAL

```
case22h/1-387             CCTGATTTCTTCACTGTTTACATGCTGTAACATCTACACATCATGCTAAGAAAAAAAAAA
gnl|ti|537031087/370-730  CCTGAGTTCTTCACTGTTTACATGCTGTAACATCTACATATCATGCTTAAAAAAAAAAAA
case22c/1-130             CCTGATTTCTTCACTGTTTACATGCTGTAACATCTACACATCATGCTAAGAAAAAAAAAA

case22h/1-387             AAAAAAAGGAGAGAGAGAGCGAGAGAACACTTTCCAGCCTGGGCAACATAGTAAGACCCC
gnl|ti|537031087/370-730  GAGAGA--GAGAGAGAGAGTGAGAGAACACTTTCCAGCCTGGGCAACATAGTTAAGACCC
case22c/1-130             AAAAA------------------------------------------------------

case22h/1-387             CTTTCTCAACAAAAAATAAAAAAA-ATTGCCCACGCATGGTGGCAAGTGGCTGCAGTCCC
gnl|ti|537031087/370-730  CCATCTCAATAAAAAATAAAAAAATATTACCCATGCATGGTGGCAAGTGGCTGTAGTCCC
case22c/1-130             ------------------------------------------------------------

case22h/1-387             AGCTACTTGGGAAGCTGAGTTGGGAGGATTGCTTGAGCCCAGGAGCTCAAGACTACAATG
gnl|ti|537031087/370-730  AGCTACTTGGGAAGCTGA----------TTGCTTGAGCCCAGGAGCTCAAGACTACAATG
case22c/1-130             ------------------------------------------------------------

case22h/1-387             AGCTATGATCACGCCACTGTACTCAAACCTGGACAACAAGACCTCATCTCTTATTAAAAA
```

```
gnl|ti|537031087/370-730  AGCTACGATCACGCCACTGCACTCCAACCTGGACAA---GACCTCATCTCTTATTTAAAA
case22c/1-130             ------------------------------------------------------------

case22h/1-387             AAAAAAAAAAAAAAAAAAAAAAAGAACACTTTACTTTAGCGCAGCTTTATGAAGTGCTTT
gnl|ti|537031087/370-730  AC-----------AAAAAAAAAAGAACACTTTACTT-AGCACAGCTTTATGAAGTGCTTT
case22c/1-130             ----------------------GAACACTTTACTTTAGCACAGCTTTATGAAGTGCTTT

case22h/1-387             ACCAGCTGGTCTCAGCTGACTATTCCTA
gnl|ti|537031087/370-730  ACCAGCTGGTCTCAGCTGACTATCCCTA
case22c/1-130             ACCAGCTGGTCTCAGCTGACTAGTCCTA
```

```
chr6:129040862:scaffold_37501:5172375:5172650:-
...bad match....chimpanzee and rhesus loci do not match

>case23c
TAAGAAATACATCGAGAATATAAAAAGATAATTTAAACATCCTGAAAATA
TTGGGAGGCCGAGACGGGCGGATCACGAGGTCAGGAGATCGAGACCATCT
TGGCTAACACGGTGAAACCCCGTTTCTACTAAAAATACAAAAAATTAGCC
GGGCGTGTTGGCGGGCGCCTGTAGTCCCAGCTACTTGGGAGGCTGAGGCA
GGAGAATGGCATGAACCTGGGAGGCGGAGCTTGCAGTGAGCCGAGATCGC
GCCACTGCACTCCAACCTGGGAGACACAGCGAGACTCCGTCTCAAAAAAA
AAAAAAAAAAAACATCCTGAAAATAAGAAATAATGCCATTAAAAGAAAAT
TCAATACATATGATTAATATAAAAA
>case23h
TAAGAAATACATCGAGAATATAAAAAGATAATTTAAACATCCTGAAAATA
AGAAATAATGCCATTAAAATAAAATTCAATACATATGATTAATATAAAAA
>gnl|ti|540771525
NNNAAAACCTTTATCNNNNNGTGGGCCCGCGTTGTCATCAGATGGGGGGG
CTCATAACACTCGTTCCAGCTACCCAGGTGTTTAGGTCGGAGAATCACCT
AAGCCCAGGTGGTCATGGCTGCACCGAGTCACGATCGCATCACCGCATTC
GAAGCCTGGGTGACAGAGTGAGACCCCGTCTGGGGGTAAAAAAAAAAAAA
AGTAAGTAAGTAAAGTAAAGTAGTGAAGTAAAGAAAAAAAATAGAGAATC
TTAAGAAATCCATCGAGAATATAGAAAGATCGTTTTAAAATAAAAACAAT
GCCATTAAGGGAGGAGACCACCCCTCATATCATCTTATGCCCAAGTTCTG
CCTCCAAAGAATGAAGAAGTAAAAACTAAAAGGCAGAAATGAAGTCTACA
GGCAGACAGCCCGGCGCTGCACCCTGGGCCTGGTAGTTAAAGATCTATCT
AATCGGTTCTGTTATCTGTAGATTACAGACATTGTATAGAAATGCACTGT
GAAAATTCCTATCTTGTTTTGTTCCAATTACCGGTGCATGCAGCCCCCAG
TCACGTACCCCCTCCTTGCTCAATCAATCATGACCCTCTCACATGCACCC
CCTCAGAGTTGGGAGCCCTTAAAAGGGACAGGAATTGCTCACTCAGGGGG
CTGGGCTCTTGAGACAGGAGTCTTGTGGACACCCCCAGCCGAATAAACCC
CTTCCTTCTTTAACTTGGTGTCTGAGGGGTTTTGTCTGCAGCTCATCATG
CTATACCATTAAAATAATATTCAATACATATGATTAATATAAAAATGGTA
ACAACTACAGAATGACTAGGGAGATGACATAAGGATGGAACTCATCTTGC
ATACAGTTCAGAGAAGCTAGAATAAACATAGAAATGTAATTGAGAATATA
GAAGGTAAGGTTGAAAGCTAAAACTCTTTCTTNATGNNGAAGACTGGTAG
AGATAGAATATTTGAGCGATACTGCTTTGAGATGCTCCAGAATTTTCATG
GGTTCCCATCTGTGNTTTCCCTTATGCTAGNCACTGCGTTTCAAAATTAA
ATGGAAATTCCNGAATAAAATTCNTTTTAATTTGTTAGGAATACTGCTCC
AAACAGGGTTCCCTTATTCTGTTCCCACCAAAAAGGGTTCAAAATCTGTT
GTAAAAAAAAGGTTTCAATATCTGTTTCATAAAAAGGTGGGGGCTGTTTT
TTATTTAATCACAATCGGAAGTAAAAACAAGTTTCCTGACTTCACAATAA
AAAGAAGCTCGCCCCTTTTCTTTTTTTTTTTTTTTACCCCCACGCCCAAA
AATTTATGGTTTGTTATTATTAAATTNNNNNNNNNNNNNNNNNNNNTTTT
ATTTTTTTTTAACCACAAAAAAAAAACAACCTTTTTTTTTTTTTTTNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCCCCNN
NNNT
```

```
chr7:5779663:scaffold_37557:3099237:3099555:+
...precise deletion of AluSg in human

CLUSTAL

case24c/1-468             TGAACTGCAACTGTGATACTTATTAGCATGTCACCTGGTGTTTTGTTTTCATTTCACTTT
gnl|ti|567316288/263-732  TGCACTGCAATAGTGATACTTATCAGCACGTCACCTGGGGTTTTGTTTTCATTTCACTTT
```

```
case24h/1-150             TGAACTGCAATAGTGATCCTTATCAGCATGTCACCTGGTGTTTTGTTTTCATTTCACTTT

case24c/1-468             AAGAATGTGCCATGTTGGCCGGGTGTGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGA
gnl|ti|567316288/263-732  AAGAATGTGCCATGTTGGCCGGGCGTGGTGGCTCATGCCTGTAATCCCAGCACTTTGGGA
case24h/1-150             AAGAATGTGCCATGT---------------------------------------------

case24c/1-468             GGCCGAGGCGGGAGGATCACAAGGTCAGGAGTTTGAAACCAGCCTGACCAACATAGTGAA
gnl|ti|567316288/263-732  GGCCGAGGCGGGCGGATAACAAGGTCACAAGTTCAAAACCTGCCTGACCAACATGGTGAA
case24h/1-150             ------------------------------------------------------------

case24c/1-468             ACTCCATCTCTACTAAAAATACAAAAAAATAAAAAATTAGCCAGACATGGTGGCGCATGC
gnl|ti|567316288/263-732  ACACCATCTCTACTAAATATACAAA------AAAAGTCAGCCAGGCGTGGTGGCGCATGC
case24h/1-150             ------------------------------------------------------------

case24c/1-468             CTGTAATCCCAGCTACTCTGGAGGCTGAGGCAGGAGAATCGCTTGAACCTGGGAGGCAGA
gnl|ti|567316288/263-732  CTGTAATCGCAGCTACTGGGGAGGCTGAGGCAGGAGAATCGCTTGAACCTGGGAGATGGA
case24h/1-150             ------------------------------------------------------------

case24c/1-468             GGTTGCAGTGAGCCGAGATCGTGCCATTGCACTCCAGCCTGGGCAATGAAAGTGAAACTC
gnl|ti|567316288/263-732  GGTTGCAGTGAGCCGAGATCGTGCCATTGCACTCCAGCCTGGGCAATAAGAGTGAAACTC
case24h/1-150             ------------------------------------------------------------

case24c/1-468             CATCTCAAAAAAAAAAAG--------AAGAATGTGCCGTGTGATCGTATTTCTAATCCCT
gnl|ti|567316288/263-732  CATCTCAAAAAAAAAAAAAAAAAAAAAAAGAATTTGCCATGTGATCGTATTTCTAATCCTT
case24h/1-150             ------------------------------------------GATCGTATTTCTAATCCCT

case24c/1-468             TTCACTCTGGAATCCTGCTCTTACCATATTAATGTTGATTAGCATCTCAGGTTTCA
gnl|ti|567316288/263-732  TTCACTCTGGAATCCTGTCCTCACCATATTAATGTTGAATAGCATCTCAGGTTTCA
case24h/1-150             TTCACTCTGGAATCCTGCCCTTACCATATTAATGTTGAATAGCATCTCAGGTTTCA
```

chr7:24969020:scaffold_36484:641689:641996:+
...precise deletion of AluSc in human: 2 of these sites in human and chimpanzee genomes; both filled with AluSc in chimpanzee, one empty and one filled site in human

CLUSTAL

```
case25c/1-461             AGGCCCGTTCTGTGTCCGGAGGGGCTGTGATCCTATCAGGACAGGAATCCAGCTCGGAGC
gnl|ti|541662238/247-704  AGGCCCGTTCTGTGTCCGGAGGGGCTGTGATCTTATCAGGACAGGAATGCAGCTCGGATC
case25h/1-150             TGGCCTGTTCTGTGTCCAGAGGGACTGTGA-----TCAGGACAGGAACCCAGCTCGGAGC

case25c/1-461             TCCTATTA-AAGATGACTGTTGGCCGGGTGCAGTGGCTCCCGCCTGTAATCCCAGCACTT
gnl|ti|541662238/247-704  TCCTGTGGTAAGATGACTGTTGGCCGG-TGCGGTGGCTCCCGCCTGTAATCCCAGCACTT
case25h/1-150             TCCTGTGATAAGATGACTGTT-----------------------------------------

case25c/1-461             CAGGAGGTCGAAGCGGGCAGATCATGAGGTCAAGAGATCGAGACCGTCATGGCCAACATC
gnl|ti|541662238/247-704  CAGGAGGCCGATGAGGGCAGATCACAAGGTCAAGAGATTGAGACCATCATGGCCAACATG
case25h/1-150             ------------------------------------------------------------

case25c/1-461             GTGAAACCCCGTCTCTACTAAAAATACAAAAATTAGCTGGGTGTGGTGGCAGGCACCTGT
gnl|ti|541662238/247-704  GTGAAACCCCGTCTCCACTGAAAATAGAAAAATTAGCTGGATGTGGTGGCAGGCGCCTGT
case25h/1-150             ------------------------------------------------------------

case25c/1-461             AGTCACAGCTATTTGGGAGGCTGAGGCAGGAGAATCGCTTGAACCCAGGAGGTGGAGGTT
gnl|ti|541662238/247-704  AGTCCCAGAAACTTGGGAGGCCAAGGCAGGAGAAACGCTTAAACCTGGGAGGTGGAGGTT
case25h/1-150             ------------------------------------------------------------

case25c/1-461             GCAGTGAGCCAAGATCGCGCCACTGCACTCCAGCCTGGTGACAGAACGAGACTACGTCTA
gnl|ti|541662238/247-704  GCAGTGAGCCAAGATTGCGCCACTGTACTCCAGCCTGGTGACAGAATGAGACTCCATCTC
case25h/1-150             ------------------------------------------------------------

case25c/1-461             AAAAAAAAAAAAAAAAAAAAAGACTGTAGATACTAATATAAACCCCACCTTCCCCAAATC
gnl|ti|541662238/247-704  AAAAAAAAAAAAAAAGAT----GACTGTTGATATTAATATAAACCCCACCTTCCCAAAATT
case25h/1-150             ---------------------------GATATTAATATAAACCCCACCTTCCAAAAACC

case25c/1-461             TGTTTATCCTGATCTTAAGATACGGC-ACATGTTAATGAGTTG
gnl|ti|541662238/247-704  TATTTATCCTGATCTTAAGATATGGCTACGTGTTAAAGAGTGG
case25h/1-150             TATTTATCCTAATCCTAAGATAAGAC-ACATTTTAATGAGTTT
```

chr7:128710174:scaffold_37671:707170:707496:-
...precise deletion of AluSx in human

CLUSTAL

```
case26c/1-476             TTGCTGTGAAGCTATGGGGGCTTTTTATGAGGAAGGCTCTATGCAAGGGTGAGGATGGAA
gnl|ti|513419099/206-664  TTGCTGTGAAACTATGGGGGCTTTTTGTGAGGAAGGCTCTATGCAAGGGAGAGGATGGAT
case26h/1-150             TTGCTGTGAAGCTATGGGGGCTTTTTATGAGGAAGGCTCTATGCAAGGGTGAGGATGGAT

case26c/1-476             ATGGGGTTTGATGGTTAGAAAGTGGGAGAGAAGGCCAGGTGCAGTGGCTCACACCTGTAT
gnl|ti|513419099/206-664  ACGGGACTTGATGGTTAGAAAGTGGGAGAGAAGTCCAGGTGCAGTGGCTCACACCTGTAA
case26h/1-150             ATGGGGCTTGATGGTTAGAAAGTGGGAGAGAA----------------------------

case26c/1-476             TCCCAGCACTTTGGGAGGCCGAAGTGGGTGGATCACCTGAAGTCAGGAGTTCAAGACCAG
gnl|ti|513419099/206-664  TCCCAGCACTTTGGGAGGCCAAAATGGGTGGATCACCTGAGGTCAAGAGTTCGAGACCAG
case26h/1-150             ------------------------------------------------------------

case26c/1-476             CCTGGCCAACATGGTGAAACCCTGTCTCTACTAAAAATAAAAAAATTAGCCGGGCATGGC
gnl|ti|513419099/206-664  CCTGGCCAATGTGGTAAAACCCCGTCTCTCCTAAAAATAAAAAAATTAGCTGGGCATGGT
case26h/1-150             ------------------------------------------------------------

case26c/1-476             GGCGTACACCTGTAATCCCAGCTACTCGGAAGGCCGAGGC---AGAAGAATTGCTTGTAC
gnl|ti|513419099/206-664  GGCAGGCGCCTGTAGTCCCAGCTACTCGGGAGGCCAAGGCCAAGGGAGAATGGATTGAAC
case26h/1-150             ------------------------------------------------------------

case26c/1-476             CTGGGAGGTAGATGTTGCAGTGAGCCAAGATTGCACCATTGCACTCCAGCCTGGGTGACA
gnl|ti|513419099/206-664  CTGGGAGCTTGAGCTTGCAGTGAGCCAAGATCACGCCACTGTACTCCAGCCTGGGTGACA
case26h/1-150             ------------------------------------------------------------

case26c/1-476             GGGGGAGACTCCGTCTCAAAAAAAAAAAAAAAAAGAAAAAGAAAAAGAAAGTGGGAGAGA
gnl|ti|513419099/206-664  GAGTGAGACT--ATCTCAAAAAAAAAAAAAAA----------------GTGGGAGAGA
case26h/1-150             ------------------------------------------------------------

case26c/1-476             AGGGAACAGTGACTGCAGGAATAGAAAAAATGTCCACAGAGGAGGAATGAGGACATGGC
gnl|ti|513419099/206-664  AGGGAACGGTGACTACAGGAATAGAAAAAGTGTCCACTGAGGAGGAAGGAGGACTTGGC
case26h/1-150             -GGGAACGGTGACTGCAGGAATAGAAAAAATGTCCACAGAGGAGGAAGGAGGACATGGC
```

chr7:132523884:132524216:scaffold_32923:964569:-
...AluY in Rhesus, gene conversion to AluYb9 in human, precise deletion in chimpanzee

CLUSTAL

```
case27h/1-500             TTACTACAGAAATTATGGGAAATGCTTACTTTTTAAAAGGAGAAAGGGAAAGAATCTCAT
gnl|ti|496205884/177-647  TTACTGCAGAAATTGTGGGAAATGCTTACTTTTTAAAAGGAGAAAGGGAAAGAATCTCAT
case27c/1-168             TTACTACAGAATTTGTGGGAAATGCTTACTTTTTAAAAGGAGAAAGGGAAAGAATCTCAT

case27h/1-500             GTTGAAAATTGTCATTAGGTAAAGTAATAGAATTAATGGAGCAGGCCGGGCGCGGTGGCT
gnl|ti|496205884/177-647  GTTGAAAATTATC-------------------------------------GCGGTGGCT
case27c/1-168             GTTGAAAATTGTCATTAGGTAAAGTAATAGAATTAATGGAGCA-----------------

case27h/1-500             CACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGAT
gnl|ti|496205884/177-647  CAAGCCTGTAATCCCAGCACTTTGGGAGGCCGAGATGGGCGGATCACGAGCTCAGGAGAT
case27c/1-168             ------------------------------------------------------------

case27h/1-500             CGAGACCATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAA-AAATACAAAAAATTAG
gnl|ti|496205884/177-647  CGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTTTATTAAGAAATACAAAAAAT-AG
case27c/1-168             ------------------------------------------------------------

case27h/1-500             CCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGGAGAATG
gnl|ti|496205884/177-647  CCGGGCGAGGTGGCAG-CGCCTGTAGTCCCAGCTACTCGGGAGACTGAGGCCGGAGAATG
case27c/1-168             ------------------------------------------------------------

case27h/1-500             GCGTTGAACCCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAG
gnl|ti|496205884/177-647  GCAT-GAACCCGGGAGGCGGAGCTTGCAGTGAGCTGAGATCCGGCCACTGCACTCC----
case27c/1-168             ------------------------------------------------------------
```

```
case27h/1-500             TCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAAA--
gnl|ti|496205884/177-647  ---AGCCTGGGCTACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAGAAAAT
case27c/1-168             ------------------------------------------------------------

case27h/1-500             ----------------AATAGAATTAATGGAGCAATCCAAAATGAGCTAAATAAGTTTCT
gnl|ti|496205884/177-647  TATCACTAGGTAAAGTAATAGAATTAACGGAGCAATTCAAAATGAGCTAAATAAGTTTCT
case27c/1-168             ----------------------------------ATCCAAAATGAGCTAAATAAGTTTCT

case27h/1-500             TTGTGAGAATTCTTCTAGAGATTAATTCTAGAATTCTTC
gnl|ti|496205884/177-647  TTGTGAGAATTCTTTTAGAGATTAATTCTAGAATTCTTC
case27c/1-168             TTGTGAGAATTCTTCTAGAGATTAATTCCAGAATTCTTC
```

chr9:32249992:scaffold_37419:6540548:6540685:-
...imprecise deletion of a MIRb element in human, no flanking identity

CLUSTAL

```
case28c/1-237             CTCCTTAAAGGAAAGTGAATGTGAACTGAAGTCTAGACCAACACC-TTATGCACAGTCTA
gnl|ti|511659877/91-328   CTCCTTAAAAGAAAGGAAATGTGAACTGAAGTTGAGAACGACACCCTTATGCACAGTATA
case28h/1-100             CTCCTTAAAGGAAAGTGAATGTGAACTGAAGTCTAGACCAACACCTT-------------

case28c/1-237             AGAGGAAAGAATATGAGACTGAAGCTATTGAGGACTGGGTTTAAGTCACTATCCTATTAT
gnl|ti|511659877/91-328   AGAGGAAAGAATATGAGACTGAAGCTGTTGAGGACTGGATTTAAGTCACTATCCCATTAT
case28h/1-100             ------------------------------------------------------------

case28c/1-237             TTACTATCTTTGCAATCTAGGATAAAGTACTTAACTCCCCTGAACCTTGGATCCTGAAAC
gnl|ti|511659877/91-328   TTACTATCTTTGCAATCTAGGACAAAGTACTTAACTCCCCTGAACCTTGGATTCTGAAAC
case28h/1-100             ------------------------------------------------------------

case28c/1-237             CACACATGGGAAGACATTGTTACCTTTGTAGCACAGGATTGATGTGTTAACAAATGGG
gnl|ti|511659877/91-328   TATACATGGGAAGACACTGTTACCTTTGTAGCACAGGATTGAGGTGTTAACAAATGGG
case28h/1-100             -----ATGGGAAGACATTGTTACCTTTGTAGCACAGGATTGATGTGTTAACAAATGGG
```

chr9:67728861:67729179:scaffold_34695:1245496:+
...inversion of AluY in Rhesus, gene conversion to AluYb8 in human, precise deletion in chimpanzee

CLUSTAL

```
case29h/1-479             CTTTCTACC-CACTTAACAGACTTGTCTTCTCTTCCCCATGGGAATAAAATTTTAGGCAG
gnl|ti|556286157/11-486   GATGGTAAGTAGTTCATC--ACTTGTCTTCTCTTCACCATTGGAATAAAATTTTAGGCAG
case29c/1-161             CTTTCTACC-CACTTAACAGACTTGTCTTCTCTTCCCCATGGGAATAAAATTTTAGGCAG

case29h/1-479             GTCTCTTAGATCTCCCGGTTA----GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGC
gnl|ti|556286157/11-486   GTCTCTCAGATCTCCTGGTTATTXXGGCTGGGCGTGGTGGCTCATGCCTGTAATCCCAGC
case29c/1-161             GTCTCTTAGATCTCCTGGTTATT-------------------------------------

case29h/1-479             ACCTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGATCGAGACCATCCTGGCTAA
gnl|ti|556286157/11-486   CCTTTGGGAGGCCAAGGGGGGCGGATCACAAGGTCAGGAGATGGAGAGCATCCTGGCTAA
case29c/1-161             ------------------------------------------------------------

case29h/1-479             CACGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGCGGTGGCGGGCG
gnl|ti|556286157/11-486   CACAGTGAAACCCTATCTGTACTAAAAATACAAAAAATTAGCTGGGCATGGTGGTGGGCG
case29c/1-161             ------------------------------------------------------------

case29h/1-479             CCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAGCGG
gnl|ti|556286157/11-486   CCTGTAGTCCCAGCTACTGGGGAGGCTGAGGCAGGAGAATGGCATGAACCCAGGAGGCAG
case29c/1-161             ------------------------------------------------------------

case29h/1-479             AGCTTGCAGTGAGCCAAGACAGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAGAGC
gnl|ti|556286157/11-486   AGCTTGCAGTGAACCGAGATCACGCCACTATAC----CACTCCAGCCTCGGCGAAACAGC
case29c/1-161             ------------------------------------------------------------

case29h/1-479             GAGACTCCGTCTCAAAAAAAAAAAAAAAAAAGAT--CTCCCGGTTATTCTTAAGTGGTGAA
gnl|ti|556286157/11-486   AAGACTCGGTCTCAAAATAATAATAATAATGATXXCT---GGTTAGTCTTAAGTGGTGAA
case29c/1-161             ------------------------------------------------CTTAAGTGATGAA
```

```
case29h/1-479            AAGTTTGACTACTAGTCACATAATACATCCAATGTGATAATAAATAACTCAAATTCTCAT
gnl|ti|556286157/11-486  GAGTTTGACTACTAGTCACATAATACATCCAGTATAATAATAAA-AACTCAAATTCTCAT
case29c/1-161            AAGTTTGACTACTAGTCACATAATACATCCAATGTGATAATAAATAACTCAAATTCTCAT

case29h/1-479            CTAATA
gnl|ti|556286157/11-486  CTAATA
case29c/1-161            CTAATA
```

chr9:84428070:84428205:scaffold_36211:2258065:-
...precise deletion of an AluSg/x fragment in chimpanzee

CLUSTAL -
Alu boundary shown by '|'

```
case30h/1-306            AGTAAGACTCTGTCTCAAAAAAAAAAAA----------------TTAAATCCTTAGAAAGA
gnl|ti|542095096/367-694 AGTAGGATTCTGTCTCAAAAAAAAAAAAAAAAAAAAAAAAAAATTTAAATCCTTAGAAAGA
case30c/1-173            AGTGAGACTCTGTCTCAAAAAAAAAAAAA-------------ATTAAATCCTTAGAAAGA

case30h/1-306            TGTATCCTCTGCTGCTGCTCTAGAACTCTAGGT|GGAGGCAAAGGTTGCAGCCAGCGGAGA
gnl|ti|542095096/367-694 TGTATCCTCTGCTGCTACTCTAGAACTCTAGGT|GGAGGCGGAGGTTGCAGTGAGTGGAGA
case30c/1-173            TGTATCCTCTGCTGCTGCTCTAGAACTCTAGAT|GGAGGCAGA------------------

case30h/1-306            TCATGCCACTGCACTCCAGCCTGGGCAACAGAGTGGCACTCCATCTCAAAAAAAAAAAAAA
gnl|ti|542095096/367-694 TCATGCCACTGCCCTCCAGCCTGGGCAACAGAGTGACACTCCATCTCAAAAAAAAAAAAA
case30c/1-173            ------------------------------------------------------------

case30h/1-306            -------AAATCCTTAAAAGTATGTATCCTCTGCTGCTACTCTAGAACTCTAGGTGGAGG
gnl|ti|542095096/367-694 TGTATTCAAATCCTTAAAAAGATGTGTCCTCTACTGCTACTCTAGAACTCTAGGTGGAGG
case30c/1-173            ------------------------------------------------------------

case30h/1-306            CAGATTTCTAAACCAGTTCTTTTGCACAGGAATTACTGGTTTGGGTGGGGCGTGGTGACG
gnl|ti|542095096/367-694 CAGATTTCTAAACCAGTTCTTTTGTACAGGAATTACTGGTTTGGGTGGGGCGTAGTGACG
case30c/1-173            ----TTTCTAAACCAGTTCTTTTGCACAGGAATTACTGGTTTGGGTGGGGCGTGGTGACG

case30h/1-306            CTGCTGTCTCAGAATTAATGTAATCATG
gnl|ti|542095096/367-694 CTGCTGTCTCAGAATTAATGTAATCATG
case30c/1-173            CTGCTGTCTCAGAATTAATGTAATCATG
```

chr10:35597212:35597542:scaffold_37564:16236237:+
...independent insertions of AluYa5 in human and L1PA5 in Rhesus

CLUSTAL

```
case31h/1-430            AAACCTGCAATATGCTAACCAAACCACTTTTAATTAAAAGGAGAAAAAAGGCCGGGAGCG
case31c/1-100            AAACCTGCAATATGCTAACCAAACCACTTTTAATTAAAAGAAGAAAAAA----------
gnl|ti|470892232/187-735  AAACCTGCAATATGCTAACAAAACCACTTTTAATTAAAAGAA----------------

case31h/1-430            GTGGCTCACGCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACGAGGTCA
case31c/1-100            ------------------------------------------------------------
gnl|ti|470892232/187-735  --------TATTGCGGCACTATTCACAATAGCAAAGACTTGGAACCAACCCAAATGTCCA

case31h/1-430            AGAGATCGAGACCATCCC-GGCTAAAACGGTGAAACCCCGTCTCTACTAAAAATACAAAA
case31c/1-100            ------------------------------------------------------------
gnl|ti|470892232/187-735  TCAATGATAGACTGGATTAAGAAAATGTGGCACATATACACCATGGAGTACTGTGCAGCC

case31h/1-430            AAATTAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGGGAGGCTGAGGCAG
case31c/1-100            ------------------------------------------------------------
gnl|ti|470892232/187-735  ATAGAAAAGGATGAGTTCATGTCCTTTGTAGGGACATGGATGAAGCTGGAAACCATCATT

case31h/1-430            GAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGATCCCGCCACTGCACT
case31c/1-100            ------------------------------------------------------------
gnl|ti|470892232/187-735  CTGAGCAAACTATCACAAGGACAGAAAACCAAACACTGCATGTTCTCACTCATAGGTGTG

case31h/1-430            CCAGCCTGGGCGACAGAGCGAGACTCCGT-CTCAAAAAAAAATAAAAAATAAAAAAAAAA
case31c/1-100            ------------------------------------------------------------
```

```
gnl|ti|470892232/187-735  AATCGAACAATGAGAACACTTGGACACAGGAAGGGGAACATCACACACCAGGGCATGTCG

case31h/1-430             AAAAAAT-------------------------------------------------------
case31c/1-100             -------------------------------------------------------------
gnl|ti|470892232/187-735  GGGTGGGGGGGATAGCATTAGGAGATATACCTAATGTAAATGACGAGTTAATGGGTGCAG

case31h/1-430             -------------------------------------------------------------
case31c/1-100             -------------------------------------------------------------
gnl|ti|470892232/187-735  CACACTAATATGGCACATGTATACATATGTAACAAACCTGCAGGTTGTGCACATGTACCC

case31h/1-430             -------------------------------AAAAGGAGAAAAAAGCCTTTTCAAGATCTT
case31c/1-100             ----------------------------------------------GCCTTTTCAAGAACTT
gnl|ti|470892232/187-735  TAGAACTTAAAGTATAATAATTAAAAAAAAAGAAGAAGAAAAAAGCCTTTTCAAGAATTT

case31h/1-430             ACAACGGCTCTTATTAGATTATAAATTGTAACCTC
case31c/1-100             ACAATGGCTCTTATTAGATTATAAATTGTAACCTC
gnl|ti|470892232/187-735  ACAATGGCTCTTATTAGATTATAAAGTGTAACTTC
```

chr11:82826588:scaffold_37428:2205199:2205501:-
...no TSD discernable, imprecise deletion of L1PA13 element in human, 2 bp flanking identity, probably blunt-end deletion

CLUSTAL

```
case32c/1-402             TGCAGCAGAATTTATGAGGTAGTGATTATTATGTTAAGTGCAAGAACT-GAAGTGGGAGC
gnl|ti|529982443/82-484   TGCAGAAAAAGGTATGAGGTAGTGATTATTATGTTAAGTGCAAGAATTAGAAGTGAGAGT
case32h/1-100             TGCAGCAGAATTTATGAGGTAGTGATTATTATGTTAAGTGCAAGAACT-GAAG-------

case32c/1-402             TAAATGATGAGAACACCTGGACACATAGAGGTGAACAACACACACTGGGGCCTATTGGAG
gnl|ti|529982443/82-484   TAAATGATGAGAACACATGGACACATAGAGGGGAGCAACACACACTGGGGCCTGTTGGAG
case32h/1-100             ------------------------------------------------------------

case32c/1-402             GGTAAAGAGTAGGAGGAGGGAAAGGATCAGGAAAAATAGCTAATGGGTACTAGGCTTAAC
gnl|ti|529982443/82-484   GGTAAAGAGTAGGAGGAGAGAGAGGATCAGGAAAAATAGCTAATGGGTTCTAGGCTTAAC
case32h/1-100             ------------------------------------------------------------

case32c/1-402             ACCTGAATGACAAAATAATGTGTACAGCAAACCCCCATGA-----ATTTACCTATCTAAC
gnl|ti|529982443/82-484   ACCTGGGTGACAAAACAATGTGTACAGCAAAACCCCGTGACACAAATTTACCTATATAAA
case32h/1-100             ------------------------------------------------------------

case32c/1-402             AAACCTGCACACGTACCCCTGGAACTTGAAAATTAAACTTTAAAAACCTGAAAATACAGA
gnl|ti|529982443/82-484   AAACTTGCACATGTACCCCTGGAATTTACAAGTTAAATTTTTAAAAACTGAGA-----GA
case32h/1-100             ------------------------------------------------------------

case32c/1-402             ATGTCAAAAAA-TGATTTTCTATCTTTGTAAGTTTGGGATAATGGAAATCCACAGAGACA
gnl|ti|529982443/82-484   ATGTCAAGAAAGTGATGTTCTATCTTTGTAAGTTTGGGATGATGGAAATCCACAGAGATA
case32h/1-100             ------------------------------------------------------------

case32c/1-402             GAATAAATAAGGTTTCAAGATCCCCTACAATCCTTATAGAGAAGCTGAG
gnl|ti|529982443/82-484   GAATAAATAAGGTTTCAAGATCCC-TACAATCCTTATAGAGAAGCTGAG
case32h/1-100             -AATAAATAAGGTTTCAAGATCCCCTACAATCCTTATAGAGAAGCTGAG
```

chr12:48585272:48585595:scaffold_37077:4241894:-
...precise deletion of AluSx in chimpanzee

CLUSTAL

```
case33h/1-486             TTAAGTGTGTTTGTGCAAAATTGGTGGTTATGGAGTGGGGGACAAAGGGCAAAAAGGGGT
gnl|ti|540393548/121-601  TTAAGTGTGTTTGTACAAATTTGGTGGTTATGGAGTGGGGGACAAAGGGCAAAAAAGGAT
case33c/1-163             TTAAGTGTGCTTGTGCAAAATTGGTGGTTATGGAGTGGGGGACAAAGGGCAAAAAGGGGT

case33h/1-486             GGGATAAAGACTTTGATAATTAGGCCAGGCACGGTGGCTCACACCTGTAATCCCAGCACT
gnl|ti|540393548/121-601  GGGATAAAGACTTTGATAATTAGGCCAGGCGTAGTAGTTCATGTCTGTAATCCCAGCCCT
case33c/1-163             GGGATAAAGACTTTGATAATTT-------------------------------------

case33h/1-486             TTGGGAGGCTGAGGAGGGTGGATCACTTGAGGTCAGGAGTTGGAGACCAGCCTGGCCAA-
```

```
gnl|ti|540393548/121-601  TTGGGAGGCTGAGGTGGGCGGATCACTTGAGGTCAGGAGTTGGAGACCAGCCTGGTCCAG
case33c/1-163             ------------------------------------------------------------

case33h/1-486             ---------CATGGTGAAACCCTGTCTCTACTAAAAATACAAAA-TTAGCCGGGCGTGGT
gnl|ti|540393548/121-601  TGCATTTCACATGGTAAAACCCTGTCTCTACTAAAAATACAAAAATTAGCCGGGCGTGGT
case33c/1-163             ------------------------------------------------------------

case33h/1-486             GGTGCGCGCCTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATCACTTGAACCCG
gnl|ti|540393548/121-601  GGTGCATGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATCACTTGAACCCA
case33c/1-163             ------------------------------------------------------------

case33h/1-486             GAAGACAGAGGTTGCAGTGAGCCAAGATTGTGCCACTGCACTCCAGCCTGAGCAACAGAA
gnl|ti|540393548/121-601  GAAGACAGAGGTGGTAGTGAGCCGAGATTGTGCCGCTGCACTCCAGCCTGAGCAACAGAG
case33c/1-163             ------------------------------------------------------------

case33h/1-486             CAAGATTTTCTCTGTAAAAAAAAAAAAAAAAAAAAAAAAAAAAGACTTTGATAATTTGTCT
gnl|ti|540393548/121-601  CAAGACTCTGTCTC-------------AAAAAAAAAAAAAAGGCTT--ATAATTTGTCT
case33c/1-163             ------------------------------------------------------GTCT

case33h/1-486             GCCTACCTGGGCCCTGCTTTCTCAGGGACTTTAGGTGGGTCCCTGGGCAGTGAGAGGGAA
gnl|ti|540393548/121-601  GCCCACC-GGGCCCTGCTTTCTCAAAGTCTTTAGGTGGGTCCCTGGGCAGCGAGAGGGAA
case33c/1-163             GCCTACCTGGGCCCTGCTTTCTCAGGGACTTTAGGTGGGTCCCTGGGCAGTGAGAGGGAA

case33h/1-486             GCAGAAGGTAAGTGGCC
gnl|ti|540393548/121-601  GCAGAAGGTAAGTGGCC
case33c/1-163             GCAGAAGGTAAGTGGCC
```

chr12:55305648:55305956:scaffold_36205:1170504:+
...precise deletion of AluY in chimpanzee

CLUSTAL

```
case34h/1-464             AGGTTAAGGACCCCATTCATGAGGCAGGTTACAGAGTCCAACCTCAAAAGACTAAAAGTA
gnl|ti|460701445/103-576  AGGTTAAGGACCCCATTCGCGAGGCAGGTTACAGAGTCCAACCTCAAAAGACTAAAAGTA
case34c/1-156             AGGTTAAGGACCCCATTCATGAGGCAGGTTACAGAGTCCAACCTCAAAAGACTAAAAGTA

case34h/1-464             GGAAGACCATCTATCTTTTAAAAACTTCT-------------TGGCTCACGCCTGTAATC
gnl|ti|460701445/103-576  GGAAGATGACCTATCTCTTAAAAACTTCTAGGCCAGGCGCGGTGGCTCAAGCCTGTAATC
case34c/1-156             GGAAGACCATCTATCTCTTAAAAACTTCT-------------------------------

case34h/1-464             CCAGCACTTTGGGAGGCCGAGACGGGCGGATCATGAGGTCAGGAGATCGAGACCATCCTG
gnl|ti|460701445/103-576  CCAGCACTTTGGGAGGCCGAGACGGGCGGATCACGAGGTCAGGAGATCAAGACCATCCTG
case34c/1-156             ------------------------------------------------------------

case34h/1-464             GCTAACACGGTGAAACCCCGTCTCTACTAAAAA--TACAAAAAT-TAGCCGGGCATGGTG
gnl|ti|460701445/103-576  GCTAACCCGGTGAAACCCCATCTCTACTAAAAAAATACAAAAATCTAGCCGGGCGAGACG
case34c/1-156             ------------------------------------------------------------

case34h/1-464             GCGCGCGCCTGTAGTCCCAGCTACACGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGG
gnl|ti|460701445/103-576  GCGGGCTCCTGTAGTCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGG
case34c/1-156             ------------------------------------------------------------

case34h/1-464             GAGGCGGAGCTTGCAGTGAGTCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACACAGC
gnl|ti|460701445/103-576  GAGGCAGAGCTTGCAGTGAGCTGAGATCCGGCCACTGCACTCCAGCCTGGGCGACAGAGC
case34c/1-156             ------------------------------------------------------------

case34h/1-464             GAAACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAACAACAACAAAAAACTTCTTCCTTTC
gnl|ti|460701445/103-576  CAGACTCCGTCTCAGGAAAAAAAAAAAAAAAAAAAAAA------AACACTTCTTCCTTTG
case34c/1-156             ------------------------------------------------------TCCTTTC

case34h/1-464             TTTGGTAAATTTTTGGCAAGATTACAGTATTTTAGTCTGCAAAATGCCCCCATAATAACT
gnl|ti|460701445/103-576  TTTGGTAAATTTTTGGCAAGATTACAGTATTTTAGTCTGCAAAATGCCCCCATAATAATT
case34c/1-156             TTTGGTAAATTTTTGGCAAGATTACAGTATTTTAGTCTGCAAAATGCCCCCATAATAACT
```

chr12:122703868:122704186:scaffold_37680:2506194:-
...precise deletion of AluY in chimpanzee

```
CLUSTAL

case35h/1-479             CAACCTCTGGTCTAGACCAACCAATCCAATCATTACCTATCCCACTGGCTATAAACTATC
gnl|ti|526305977/153-631  CAACCTCTGGTCTAGACCAACCAATCCAATCATTACCTATCCCACTGGCTATAAATTATC
case35c/1-161             CAACCTCTGGTCTAGACCAACCAATCCAATCATTACCTATCCCACTGGCTATAAACTATC

case35h/1-479             TGAAATTAAATAAGTGTGTCGG-TGTG--------TCGGTGGCTCACGCCTGTAATCCCA
gnl|ti|526305977/153-631  TGAAATCAAATAAGTGTGTTTGCCATGGGCCGGGCACGGTGGCTCAAGCCTATAATCCCA
case35c/1-161             TGAAATTAAATAAGTGTGTCTGCCATG---------------------------------

case35h/1-479             GCACTTTGGGAGGCTGAGGCGGGCAGATCACGAGCTCAAGAGATCGAGACCATCCTGGCT
gnl|ti|526305977/153-631  GCACTTTGGGAGGCCGAGATGGGCAGATCACGAGGTCAGGAGATCGAGACTATCCTGGCT
case35c/1-161             ------------------------------------------------------------

case35h/1-479             AACACGGTGAAACCCCGCCTCTACTAAAAA-TACAAAAAATTAGCCGGTCGTGGTGGCGG
gnl|ti|526305977/153-631  AACACGGTGAAACCCCGTCTCTACTAAAAAATACAAAAAACTAGCCGGGCGAGCTGGC--
case35c/1-161             ------------------------------------------------------------

case35h/1-479             GCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCATGAACCTGGGAGG
gnl|ti|526305977/153-631  ----CCATAGTCCCAGCTACGCGGGAGGCTGAGGCAGGAGAATGGCGTAAACCCGGGAGG
case35c/1-161             ------------------------------------------------------------

case35h/1-479             CGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCACTCCAGCCTGGGTGACAGAGCGAGA
gnl|ti|526305977/153-631  TGGAGCTTGCAGTGAGCTGAGATCCAGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGA
case35c/1-161             ------------------------------------------------------------

case35h/1-479             CTCCGTCTCCAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGTGTGTCTGCCATGTTTTC
gnl|ti|526305977/153-631  CTCCGTCTC----AAAAAAAAAAAAAAAAAAAAAAAAAAAAGTGTGTCTGCCATGTTTTC
case35c/1-161             ------------------------------------------------------------TTTTC

case35h/1-479             ACTTATTTTGTGCTCAGGGACATGGGCAGGTCAAGGACAAAGTCATCGCCTTCAACTACA
gnl|ti|526305977/153-631  ACTTAATTTGTGCTCAGGGACATAGGCAGGTCAAGGACAAGGTCATCACCTTCGACTACA
case35c/1-161             ACTTAATTTGTGCTCAGGGACATGGGCAGGTCAAGGACAAGGTCATCGCCTTCGACTACA

case35h/1-479             AACATCCCT
gnl|ti|526305977/153-631  AACATCCCT
case35c/1-161             AACATCCCT


chr14:76323357:76323468:scaffold_37670:5922279:-
...imprecise deletion of L2 (element inside a deletion) in chimpanzee, no flanking
identity

CLUSTAL

case36h/6-211             CCTGCCAGCATGGGAGAATGGCACTCCAAGGGGCACAGCCAGGGCGCTCCAGGGGGCAGG
gnl|ti|538237539/116-325  CCTGCCAGCATGGGAGAACGGCACTCCAAGGGGCAGAGCCAGGGCACTCCAGGGGGCAGG
case36c/6-100             CCTGCCAGCATGGGAGAATGGCACTCCAAGGGGCATAGCCAGGGC---------------

case36h/6-211             AACCCCAGGGGCTGGTGTAGTACCTGGCACAGAGGAGGTGCTCAATAAATGCACATGAGT
gnl|ti|538237539/116-325  GACCCCAGGGGCTGGTACAGTACCTGGCACAGAGGAGGTGCTCAGTAAATGCACATGAGT
case36c/6-100             ------------------------------------------------------------

case36h/6-211             AAATAAGAGGAGGATGGGGAGAAGTTGGATCAGTGCTGGAAAGAAG----CAAACAATAT
gnl|ti|538237539/116-325  AAACAAGAGGAGGATGGGGAGAAGTTGGATCAGCGCTGGAAAGAAGGCAGCAAAGAATAT
case36c/6-100             ------------------------------------TGGAAAGAAG----CAAACAATAT

case36h/6-211             GAGAGCCTGGAGGCATCTCTCCCCTGCGGG
gnl|ti|538237539/116-325  GAGAGCCTGGAGGCATCTCTCCCCTGTGGG
case36c/6-100             GAGAGCCTGGAGGCATCTCTCCCCTGCGGG


chr14:81720447:81720771:scaffold_37670:482346:-
...AluY with partial deletion in Rhesus, gene conversion to AluYb8 in human, precise
deletion in chimpanzee

CLUSTAL
```

```
case37h/1-488             CAGTGTATTTGAAAGAAAAAAATAAGTTTCCAAATTTTTCACAGCTTATCTTCTTTATGG
gnl|ti|520934604/378-835  CAGTATATTTGAAAGAAAAAA-TAAGTTTCCAAATTTTTCCCATCTTATCTTCTTTACAG
case37c/1-164             CAGTGTATTTGAAAGAAAAAAATAAGTTTCCAAATTTTTCCCAGCTTATCTTCTTTATGG

case37h/1-488             CTCTTTAAAAGGTCACTTATGGGCCGGGCGCGATGGCTCACGCCTGTAATCCCAGCACTT
gnl|ti|520934604/378-835  CCCTTTAAAAGCTCACTTATG----------------------------------------T
case37c/1-164             CTCTTTAAAAGGTCACTTATG----------------------------------------

case37h/1-488             TGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGATCGAGACCATCCTGGCTAACAAG
gnl|ti|520934604/378-835  TGGGAGGCCGAGATGGGCAGATCACGAGGTCAGGAGATCGAGAGCATCCTGGCTAACACG
case37c/1-164             ------------------------------------------------------------

case37h/1-488             GTGAAACCCCGTCTCTACTAAAAATACAAAAAAAATTAGCCGGGCGCGGTGGCGGGCGCC
gnl|ti|520934604/378-835  GTGAAACCCCATCTCTACTAAAAA-ATACAAAAAACTAGCCAGGCGAGGTGGTGGGCGCC
case37c/1-164             ------------------------------------------------------------

case37h/1-488             TGTAGTCCCAGCTACTCGGGAGGCTGAGGCGGGAGAATGGCGTGAACCCGGGAAGCGGAG
gnl|ti|520934604/378-835  CGCAGTCCCAGCTACTCTGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCAGAG
case37c/1-164             ------------------------------------------------------------

case37h/1-488             CTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCCGCCTGGGCGACAGAGCAA
gnl|ti|520934604/378-835  GTTGCAGTGAGTTGAGATTCGACCACTGCACTCCA-------GCCTGGGAGACAGAGCGA
case37c/1-164             ------------------------------------------------------------

case37h/1-488             GACTCCGTCTCAAAAAAAAAAAAAAAAAAAAA-----------------AGGTCACTGA
gnl|ti|520934604/378-835  GACTCCGTCTCATAAAAAAAAAAAAAAAAAAAAAGAAAAAAAAGAAAAAAAAGCTCACTTA
case37c/1-164             ------------------------------------------------------------

case37h/1-488             TGAAAAAGTTTGAATTTCTAGAGCAGACACAGTTCATAGTCACTGAAGTTGTGACCTCTC
gnl|ti|520934604/378-835  TGAAAAACTTTGAATTTGTAGAGCAGGCACAGTTCATAGTCACTGAAGTTGTTGCCTCTC
case37c/1-164             --AAAAAGTCTGAATTTCTAGAGCAGACACAGTTCATAGTCACTGAAGTTGTGACCTCTC

case37h/1-488             TTGACAGAGAAGAAAGAGAGTAATA
gnl|ti|520934604/378-835  TTGAGAGAGAAGAAACAGAGTAATA
case37c/1-164             TTGAGAGAGAAGAAAGAGAGTAATA
```

chr15:37146781:37147081:scaffold_37412:10461989:-
...partially deleted AluJo existing prior to human-rhesus split, no tsd; imprecise deletion in chimpanzee, no TSDs or flanking identities

CLUSTAL

```
case38h/1-400             TTAGCAATCATTTTGGAGGGAGTGTGCTAGACATTAAAAAAAA-TTATTAACACACATGG
gnl|ti|523788521/312-713  TTGGCAATCATTTTAGAGGGAGGGTGCTAGACATTAAAAAAAAATTATTAACACACATGG
case38c/1-131             TTAGCAATCATTTTGGAGGGAGTGTGCTAGACATTAAAAAAAA-TTATTAA---------

case38h/1-400             TTCTGGCCAGGCATGGTGGTTCATGCATATAATCC-AGCACTTTGGGAGGCCAAGGTGGA
gnl|ti|523788521/312-713  TTCTGGCCAGGCATGGTGGTTCATGCATATAATCCCAGCACTTTGGGAGGCCAAGGTGGA
case38c/1-131             ------------------------------------------------------------

case38h/1-400             AAGATCCCTTGAGTCTCAGAATTTGAGACCAGCCTTGGCAACATAGTGAGGCCCCCATCT
gnl|ti|523788521/312-713  AAGATCCCTTGAGTCTCAGAATTTGGGACCAGCCTTGGCAACATAGTGAGACCACCATCT
case38c/1-131             ------------------------------------------------------------

case38h/1-400             CTACAGAAAATAAAAAAAAATTAGCTGGGCATGATGACACACACCTGTAGTCCCAGTTAC
gnl|ti|523788521/312-713  CTACAGAAAATAAAAAAA--TTAGCTGGGCGTGATGCTACACACCTGTAGTCCCAGTTAC
case38c/1-131             ------------------------------------------------------------

case38h/1-400             TTGGGAAGCTGAGGTGGGAAGGACTGCTTGAGCACAGGAGTTTGAGGCTGCAGTGAGCCA
gnl|ti|523788521/312-713  TCAGGAAGCTAAGGTGG-AAGGACTGCTTAAGCACAGGAGTTTGAGGCTGCAGTGAGCCG
case38c/1-131             ------------------------------------------------------------

case38h/1-400             CGATTGCACTACTGCACTTCAGCCTGGGCAACAGAGTGAGACCTTGTCTTAAAAGTAAAT
gnl|ti|523788521/312-713  TGACTGCACTACTGCACTTCAGCCTGCGCAACAGAGTGAGACCTCGTCTTAAAAATAAAT
case38c/1-131             ----------------------ATAAGTTAAATTATTAAATTATTAAATTATTAAATAAAT

case38h/1-400             AA-----GTAAAGACACGTGGTCCTTCAAAGAGAGAGGTATAAACAA
```

```
gnl|ti|523788521/312-713  AAATAAAGTAAAGACACATGGTCCTTCAAAGAGAGGTG--TAAACAA
case38c/1-131             AA-----GTAAAGACACGTGGTCCTTCAAAGAGAGGTATAAACAA
```

chr15:50365066:scaffold_37399:651713:652024:-
...precise deletion of AluSq in human

CLUSTAL

```
case39c/1-457             TAACACTTAACACTACTCTGAATTCATGAAAGACCAAAGGTAGCTAATTAATATACAATT
gnl|ti|459269000/431-893  TAACACTTAACACTACTGTGAATTCACAAAAGACCAAAGGTAGCTAATGAATATACAATT
case39h/1-150             TAACACTTAACACTACTCTGAATTCATGAAAGACCAAAGGTAGCTAATTAATATACAATT

case39c/1-457             CCTGAAAATAAAAATTATTCAATCTCATCAAAAGTCAAAGAAGGCCAGGCGCAGTGGCTC
gnl|ti|459269000/431-893  CCTGAAAATAAAAATTATTCCGTCTCTTCAAAAGTCAAAGAAGGCCAGGCGCAGTGGCTC
case39h/1-150             CCTGAAAATAAAAAAAA------AAAAAAAAAAGTCAAAGAA----------------

case39c/1-457             ATGCCTGTAATCCCAGCACTTTGGGAGGGCAAGGCAAGTGGGTCACCTGAGGTCAGGAGT
gnl|ti|459269000/431-893  CTGCCTGTAATCCCAGCACTTTGGGAGACCCAGGCCAGTGGATCACCTGAGGTCAGGAGT
case39h/1-150             ------------------------------------------------------------

case39c/1-457             TCGAGAGCAGCCTAGCCAACATCGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAG
gnl|ti|459269000/431-893  TCGAGACGAGCCTAGCCAACATGGCGAAACCCTGTCTCTACTAAAAATACAAAAAATTAA
case39h/1-150             ------------------------------------------------------------

case39c/1-457             CCAAGTGTGGTGGCAGACACCTGTAATCCCAGCTACTCAGGAGGTTGAGGCAGGAGAATT
gnl|ti|459269000/431-893  CCAAGTGTGGTGGCAGGCGCCTGTAATCCCAGCTACTCAGGAGACTGAGGCAGGAGAATT
case39h/1-150             ------------------------------------------------------------

case39c/1-457             GCTTGAACCCAGGAGGCAGAGGTTGCAGTGAGCTGAGATTGTGCCATTGCACTCCAGCCT
gnl|ti|459269000/431-893  GCTTGAACCCAGGAGGTGGAGGTTGCAGTGAGCCGAGATTGCACCACGGCATGCCAGCCT
case39h/1-150             ------------------------------------------------------------

case39c/1-457             AGGCAACAAGAGCAAAACTCGGTCCAA------AAAAAAAGTCAAAGAAATGCAAATTAA
gnl|ti|459269000/431-893  GGGCAACAAGAGCAAAACTCCGTCCAAGAAAAAAAAAAAAAGTCAAAGAAATGCAAATTAA
case39h/1-150             ----------------------------------------------------ATGCAAATTAA

case39c/1-457             ATCAACAGCAAAGTGCCACTTTTGGTCTATTAACTGAGCTAAT
gnl|ti|459269000/431-893  ATCAACAACAAAGTTCCACTTTTGGTCTATTAACTGAGCTAAA
case39h/1-150             ATCAACAGCGAAGTGCCACTTTTGGTCTATTAACTGAGCTAAT
```

chr16:48429275:scaffold_32947:3554158:3554416:+
...precise deletion of AluY in human

CLUSTAL

```
case40c/16-449            AAAATATATATATATATATGTTTATATAGAATATCTTCCATGCCACAGCAATTCCATC
gnl|ti|536342151/145-590  TATATATATATATATATATATATATATAAAAAATATCTTCCATACCACAGCAATTCCATC
case40h/16-150            TATATATATATT-ATATATATATATA----------------------------TTCCATC

case40c/16-449            CAATCACCTTTCTCAAACATGAAGGGGGGTGGAACACGAGGT-CAGGAGATCAAGATCAT
gnl|ti|536342151/145-590  CAATCCCCTTTCTCAAACATGAAGGGGGGCAGATCACCAGGTTCAGGAGATCAAGACCAT
case40h/16-150            CAATCCCCTTTCTCAAACATGAAGGGG---------------------------------

case40c/16-449            CCTGGCTAACACGGTGAAACCCCATCTTTACTAAAAATACAAAAACAAAATTAGCCGGGC
gnl|ti|536342151/145-590  CCTGGCTAACATGGTGAAACCCTGTCTCTACTAAAAAGACAAAAACAAAATTAGCCGGGT
case40h/16-150            ------------------------------------------------------------

case40c/16-449            GTGGTGGCAGGCGCCTATAGTCCCAGCTACCAGGGAGGCTGAGG-CAGGAGAATGGCGTG
gnl|ti|536342151/145-590  GTGGTGGCAGGTGCTTGTAGTCCCAGCTACTCAGGAGGCTGAGG-CAGGAGAATGGTGTG
case40h/16-150            ------------------------------------------------------------

case40c/16-449            AACCCAAGAGGCGGAGCTTGCAGTGAGCCGAGATCGCACCAGTGCACTCCAGCCTAGGTG
gnl|ti|536342151/145-590  AACCCAGGAGACGGAGCTTGCAGTGAGCAGAGATCGCGCCACTGCACTCCAGCCTACGTG
case40h/16-150            ------------------------------------------------------------
```

```
case40c/16-449            ACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAA----------AACCATGAAGGGGCTG
gnl|ti|536342151/145-590  ACAGAGCGAGACTCCATCTCAAAAACAAAAACAAAAACAAAACAAAACATGAAGGGGCTA
case40h/16-150            ------------------------------------------------------------CTG

case40c/16-449            GGCTTC-TCTCGGCATGGTAGCCAGGTTCCAAGTAAGAAAGTAAGACTATTTCACCAGCA
gnl|ti|536342151/145-590  GGCTTCCTCTCAGCATGGNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
case40h/16-150            GGTTTCCTCTTGGCATGGTAGCCAGGTTCCAAGTAA--------GACTATTTCACCAGCA

case40c/16-449            AGTGTTCTTGCTGGAAGTAGAAGGAAG
gnl|ti|536342151/145-590  NNNNNNNNNNNNNNNNNNNNNNNNN
case40h/16-150            AGTGTTCC------ATGAAAAAGGAAG
```

chr16:66575984:66576310:scaffold_37667:3291777:-
...AluY in Rhesus, gene conversion to AluYb8 in human, precise deletion in chimpanzee

CLUSTAL

```
case41h/1-491             CTCTTCACTTCCATTCTATTCATTAACTCCTTTTGTTCCACCTTGAACTATGCTCATTTT
gnl|ti|536041529/83-557   CTCTTTACTCCCATTCTATTCATTAACTCCTTTTGTTCCACCTTAAACTATGCTCGTTTC
case41c/1-165             CTCTTCACTTCCATTCTATTCATTAACTCCTTTTGTTTCACCTTGAACTATGCTCATTTT

case41h/1-491             TCCTCATCTTAAAAAAATACCC-----GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCA
gnl|ti|536041529/83-557   TCCTCATCTCAAAAAAATACCCAATAGAGCCGGGCGCAGTGGCTCACGCCTGTAATCCCA
case41c/1-165             TCCTCATCTTAAAAAAATACCCAATAG----------------------------------

case41h/1-491             GCACTTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGATCGAGACCATCCTGGCT
gnl|ti|536041529/83-557   GCACTTTGGGAGGCCAAGGCGGGCGGATCACAAGGTCAGGAGATCGAGACCAC-------
case41c/1-165             ------------------------------------------------------------

case41h/1-491             AACAAGGTGAAACCCCGTCTCTACTAAAAA-TACAAAAAATTAGCCGGGCGCGGTGGCGG
gnl|ti|536041529/83-557   -----GGTGAAACCCCGTCTCTACTAAAAAATACAAAAAATTAGCCGGGTGCTGTAGCGG
case41c/1-165             ------------------------------------------------------------

case41h/1-491             GCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAAG
gnl|ti|536041529/83-557   GCGCCTGTAGTCCCAG--------GAGGCTGATGTTTGAGAATGGCGTGAACCTGGGAGG
case41c/1-165             ------------------------------------------------------------

case41h/1-491             CGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACAG
gnl|ti|536041529/83-557   CGGAGCTTGCAGTGAGCCAAGATCGCGCCACTGCACTCCA-------GCCTGGGGGACAG
case41c/1-165             ------------------------------------------------------------

case41h/1-491             AGCGAGACTCCGTCTCAAAAAAA-----AAAAAAAAAAAAAAAAAAAAAAAAAATACCCA
gnl|ti|536041529/83-557   AGCGAGACTCCGTCTCAACAACAACAACAACAACAACAAAACAAAACAAAACACAAAAACACCCA
case41c/1-165             ------------------------------------------------------------

case41h/1-491             ATAGTTTGTTTGATTTGCTATCCTTTCAATTATTCATCCACTCCTCCCAGCAGATGCACA
gnl|ti|536041529/83-557   ATAGTTTGTTTGATTTGCTATCCTTTCAAGTATTCATTCACTCCTCCCAACAGATGCAAG
case41c/1-165             ----TTTGTTTGATTTGCTATCCTTTCAATTATTCATCCACTCCTCCCAGCAGATGCACA

case41h/1-491             TATACCATTAGAAAGTAAATGT
gnl|ti|536041529/83-557   TATCCCATTAGATAGTAAATGT
case41c/1-165             TATACCATTAGAAAGTAAATGT
```

chr16:69279114:69279432:scaffold_37667:485635:-
...precise deletion of AluSq in chimpanzee

CLUSTAL

```
case42h/1-479             TTCTGAAACCCACGTCTCTTGACAACTATGGTCTCTGCAACTTATCTGACCTTAAAACAC
gnl|ti|536066149/241-714  TTCTGAAACCCACATATCTTGACAACTATGGTCTCTGCAACTTATCTGACCGTAAAACAC
case42c/1-161             TTCTGAAACCCACGTCTCTTGACAACTATGGTCTCTGCAACTTATCTGACCTTAAAACAC

case42h/1-479             TTGCCTGGGTAATGTCCTTATAAGAGTTCTTCCTTTCTGGCCGGGCGCGGTGGCTTACAC
gnl|ti|536066149/241-714  TTGCCTGGGTAATGTCCTTGTAAGAGTTCTTCCTTTCTGGCCGGATGCGGTGGCTCACGC
```

```
case42c/1-161             TTGCCTGGGTAATGTCCTTATAAGAGTTCTTCCTTTC----------------------

case42h/1-479             CTGGAATCCCAGCACTTTGGGAGGCCGAGGTGGGTGGATCACTTGAGGTCAGGAGTTT-G
gnl|ti|536066149/241-714  CTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCACTTGAGGTCAGGAGTTTTG
case42c/1-161             ------------------------------------------------------------

case42h/1-479             ATACCAGCCTGGCCAACGTGGTGAAACCCTGCCTCTACTAAAAATACAAAAATTAGCTGG
gnl|ti|536066149/241-714  AGACCAGCCTGGCCAACATGGTGAAACCCTGTTTCTATTAAAAATACAAAAGTTATCTGG
case42c/1-161             ------------------------------------------------------------

case42h/1-479             ACCTGGTAGTGCATGCCTGTAATCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATCTCTT
gnl|ti|536066149/241-714  ACGTGGTAGTGCATGCCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGACTCTCTT
case42c/1-161             ------------------------------------------------------------

case42h/1-479             GAACCTGGGAGGTGGAGGTTGCAGTGAGCTGAGATCGCACCATTGCATTCCGGCCTGGGG
gnl|ti|536066149/241-714  GAACCTGGGAGATGGGGATTGCAGTGAACTGAGATCGTGCCACTGCACTCCAGCCTGGGG
case42c/1-161             ------------------------------------------------------------

case42h/1-479             GACAAGAGTGAAACTCCATCTCAAAAAAAAAAAAAAAAAAAAAAGAGTTCTTCCTTTCCAAG
gnl|ti|536066149/241-714  GACAAGAGTGAAACTCCATCTCAGGGAAAAAA--------AAAAGTTCTTCCTTTCCAAG
case42c/1-161             ------------------------------------------------------------CAAG

case42h/1-479             TATACCTGAGTTCCCATAAGACAGAAAGTCATTTTGTTG-CTGTTAATTTTTTGAG-TAA
gnl|ti|536066149/241-714  TATACCTAAGTTCCCATAAGACAGAAAGTCACTTTTTTGGTTGTTAATTTTTTGAGATAA
case42c/1-161             TATACCTGAGTTCCCATAAGACAGAAAGTCATTTTGTTG-TTGTTAATTTTTTGAG-TAA

case42h/1-479             GG
gnl|ti|536066149/241-714  GG
case42c/1-161             GG
```

chr16:74232245:74232552:scaffold_37614:14060425:-
...precise deletion of AluSx in chimpanzee, 3 copies of this region in human genome, all with the Alu, chimpanzee has one with the Alu, one without

CLUSTAL

```
case43h/1-462             CAGCCACACCTCTCTGCCCTAGTCTCCTGCCCCCAGGAGCCTGGCCTCATATGCTCCCCA
gnl|ti|503026621/313-769  CAGCCACGCCTCTCCGCCCTGGTCTCCTGTCCCCGTA---CCCGCCTCATCGGCTCCCTA
case43c/1-155             CAGCCACGCCTCTCTGCCCTAGTCTCCTGCCCCCAGGAGCCTGGCCTCATATGCTCCCCA

case43h/1-462             CCACGCACAGCTGACCCCGCCCCCTCCCTCTTCTTTTTTTTTTTGAGACAGAGTCTCAC
gnl|ti|503026621/313-769  CCACGCACAGCTAACCCCGCCCCCTCCCTCTTCTTCTTTTTTTT-AAGACAGAGTCTCAC
case43c/1-155             CCACGCACAGCTGACCCCGCCCCCTCCCTCTTCTT------------------------

case43h/1-462             CCTGTCGCCCAGGCTGGAGTGCAGTGGAGAAATCCC--GGCTTACTGCAACCTCCGCCTC
gnl|ti|503026621/313-769  CCTGTTGCCCAGTCTGGAGTGCAGTGGAGCAATCTC--GGCTTACTGCAAACTCCGCCTC
case43c/1-155             ------------------------------------------------------------

case43h/1-462             CCAGGTTCAAGCAATTCTCCTGCCTCAGCCTCCCAAGCAGCTGGGATTACAGCCATGTGA
gnl|ti|503026621/313-769  CCAGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGATTAGAGCCATGTGA
case43c/1-155             ------------------------------------------------------------

case43h/1-462             CACCACACCTGGCTAATTTTTGTATTTTTAGTAGAGACGGGGTTTCACCATGTTGGCCAT
gnl|ti|503026621/313-769  CACCACATCTGGCTAATTTTTGTATTTTTAGTAGAGACAGGGTTTCACCACGTTGGCCAT
case43c/1-155             ------------------------------------------------------------

case43h/1-462             TGCTGGTGTCAAACTGCTGACCTTAGGTGATCTGCTTGTGTCAGCCTCCCAGAGTGCTGG
gnl|ti|503026621/313-769  -GCTGGTCTCAAACTGCTGACCTTAGGTGATCTCCTTGCGTCAGCCTCCCAAAGTGCTGG
case43c/1-155             ------------------------------------------------------------

case43h/1-462             GATTACAGGTGTGAGCCACCGTGCCCAGCCCCCTCCCTCTTCTTAAACAAGGGGCCTGGC
gnl|ti|503026621/313-769  GATTACAGGTGTCAGCCACCGCGCCCAGCCCCCTCCCTCTTCTTAAACAAGGGGCCTGGC
case43c/1-155             ------------------------------------------AAACAAGGGGCCTGGC

case43h/1-462             AATCACCACCCCTGGGTGACTTGGTGCAGTCCCCTGATCTCCCG
gnl|ti|503026621/313-769  AATCGCCACCCCTGGGTGACCTGCTGCAGACCCCTGATCTCCCG
case43c/1-155             AATCACCACCCCTGGGTGACTTGGTGCAGTCCCCTGATCTCCCG
```

chr17:30401123:30401435:scaffold_37172:1193530:+
...imprecise deletion of AluSc in chimpanzee (no TSD copy retained)

CLUSTAL

```
case44h/1-470            ACCCATTAGAAGATAACACCATTTGCCTTTTATTTTTGGATAATTCAGGAATAAAAAATG
gnl|ti|513289005/92-572  ACCCATTAGAAGATAACAGCACTTGTCTTTTGTTTTTGGATAATTCAGGAATAAAAAATG
case44c/1-161            TCCCCCTATAAGATAACACCATTTGCCTTTTATTTTTGGATCATTCAGGAATAAAAA-TG

case44h/1-470            GATCTCAAGCTTTATAAAACTTACAATTCTAGGCTGGGCGCTGTGGCTCACACCTGTAAT
gnl|ti|513289005/92-572  GAACCCAAGTTTTATAAAACTTACAATTCTAGGCTGGGCGCCATGGCACACGCATGTAAT
case44c/1-161            GATCTCAAGCTTTATAAAACCC----------------------------------------

case44h/1-470            CCCAGCACTTTGGGAGGCCAAGGCCGGCGGATCACACGGTCAGGAGGTCAAGACCATCCT
gnl|ti|513289005/92-572  CTCAGCACTCTGGGAGGCCAACGCAGGTGGATCACACAGTCAGGAGGTCAAGACCATCCT
case44c/1-161            ------------------------------------------------------------

case44h/1-470            GGCCAACATGGTGAAACCCTGTCTCTACTAAAAATACAAAAATTAGCTGGGCGTGGTGGT
gnl|ti|513289005/92-572  GGCCAACATGGTAAAACCCTGTCTCTACTAAAAATACAAAAATTAGCTGGGCGTGGTGGT
case44c/1-161            ------------------------------------------------------------

case44h/1-470            GCGAGCCTGTAATCCCAGCTACTCGAGAGGCTGAGACAGGAGAATTGCTTGAACCCAGGA
gnl|ti|513289005/92-572  GCGAGCCTGTAATCCCAGCTACTCCGGAGGCTGAGACAGGAGAATTGCTGGAACCCGGGA
case44c/1-161            ------------------------------------------------------------

case44h/1-470            GGCAGAGATTGCAGTGAGCCGAGATTGTGC-CACTGCACTTCAGCCTGGCAACAGAGTGA
gnl|ti|513289005/92-572  GGCAGAGATTACAGTGAGCTAAGATTGTGT-CGCTGCACTCCAGCCTGGCAACCCAGCAA
case44c/1-161            ------------------------------------------------------------

case44h/1-470            AACTCCGTCTCAAAAA------AAAAAAATTATAATTCTATAGAAAAATAACATTTGTAT
gnl|ti|513289005/92-572  AACTCCATCTCAAAAAAAAAGAAAAAAAATTATAATTCTATAGAAAAAGAACATTGGTAT
case44c/1-161            ------------------------------------CATAGAAAAATAACATTTGTAT

case44h/1-470            AAATTTAAC-TTTGGTGTA-----AAAAAGTGAATTTAACTTTGGTATTGCACACTGGTA
gnl|ti|513289005/92-572  AAATTTAAC-TTTAGTGTAGAAGAAAAAAGTGAATTTAACTTTGGTATTGCACACTGGAA
case44c/1-161            AAATTTAACCTTTGGTGTA-----AAAAAGTGAATTTAACTTTGGTATTGCACACTGGTA

case44h/1-470            ATT
gnl|ti|513289005/92-572  AGT
case44c/1-161            ATT
```

chr17:37731003:scaffold_37479:3983548:3983669:-
...imprecise deletion internal to MIRb element in human, no flanking identity

CLUSTAL

```
case45c/1-221            CAGAATCGATCACTAAAAGATGTTAGTGTTTTTACGCCACTGCGGGTCTTTAATTTCTTG
gnl|ti|529975753/582-804 CAGAATAGATCACTAAAAGATGTTAGTGTTTTTACGCTGCTGCGGGTCTTTAATTTCTTG
case45h/1-100            CAGAATCGATCACTAAAAGATGTTAGTGTTTTTACGCCACTGCGGGTCTT----------

case45c/1-221            GTGCCTCAATTTCCTCCTCTGTAAAGTGGACCTAATCCCAATATTTCTGTCATCAGTTGT
gnl|ti|529975753/582-804 GTGCCTCAATTTCCTCCTCTGTAAAGTGGACCTAATCCCAATGTTTCTATCATCAGTTGT
case45h/1-100            ------------------------------------------------------------

case45c/1-221            GGAAATTACGTGAGGTAACGTTTGCAATTAGCAAAGGAAGGCATTAAGACAGCAAGCTCT
gnl|ti|529975753/582-804 GGAAATTACGTGAGGTAACATTTGCAATTAGCAAAGGAAGGCATTAAGACAGCAAGCTCT
case45h/1-100            ----------------------------------------------------GCAAGCTCT

case45c/1-221            TGGAAGGCGGGAACTAGTCTC--AGTCTCATTTGGCTCACAAC
gnl|ti|529975753/582-804 TGGAGGGCGAGAACTAGTCTCGTAGTCTCCTTTGGCTCACAGC
case45h/1-100            TGGAGGGCGGGAACTAGTCTC--AGTCTCATTTGGCTCACAAC
```

chr17:57592534:57592845:scaffold_37659:17472263:+

...AluY in Rhesus, gene conversion to AluYb8 in human, precise deletion in chimpanzee

CLUSTAL

```
case46h/1-468              AGAAAGAATATAGAGCTTAGGTTGGAGTTGAAATGGTGAGGACTAGTATTTAAGAAATCT
gnl|ti|541136674/627-1103  AGAAAGAATATAGGGCTTAGGTTGGAGTTGAAATGGTGAGGACTAGTATTTA-GAAGTCT
case46c/1-157              AGAAAGAATATAGAGCTTAGGTTGGAGTTGAAATGGTGAGGACTAGTATTTAAGAAATCT

case46h/1-468              TTAGTTATCCCAGCATATTAAGAATATGCCA-----GGCCGGGCGCGGTGGCTCACGCCT
gnl|ti|541136674/627-1103  TTAGTTATCCAAGCATATTAAGAGTATGCCAGTGTTGGCCAGGCGCAGTGGCTCACGCCT
case46c/1-157              TTAGTTATCCAAGCATATTAAGAATATGCCAGTGTT-----------------------

case46h/1-468              GTAATCCCAGCACTTTGGGAGGCCGAGGCGGGTGGATCATGAGGTCAGGAGATCAAGACC
gnl|ti|541136674/627-1103  GTAATCCCAGCACTTTGGGAGGCCAAGGCAGGCGGATCATGAGGTCAGGAGATCGAGACC
case46c/1-157              ------------------------------------------------------------

case46h/1-468              ATCCTGGCTAACAAGGTGAAACCCCGTCTCTACTAAAAATACAAAAAATTAGCCGGGCGC
gnl|ti|541136674/627-1103  ATCCTGGCTAACACAGTGAAACCCCGTCTTCACCAAAAATACAAAAAGTTCTCCGGGCGT
case46c/1-157              ------------------------------------------------------------

case46h/1-468              GGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAAC
gnl|ti|541136674/627-1103  GGTGGCGGGTGCCTGTAGTCCTAGCTACTCCGGAGGCTGAGGCAGGAGAACGGCGTGAGC
case46c/1-157              ------------------------------------------------------------

case46h/1-468              CCGGGAAGCGGAGCTTGCAGTGAGCCGAGATTGCGCCACTGCAGTCCGCAGTCCGACCTG
gnl|ti|541136674/627-1103  CTGGGAGGCGGAGCTTGCAGTGAGCCGAGATCACACCACTGTACTCCA-------GCCTG
case46c/1-157              ------------------------------------------------------------

case46h/1-468              GGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAA------------GAATATGCCAG
gnl|ti|541136674/627-1103  GGCAACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAAGAAGAATATGCCAG
case46c/1-157              ------------------------------------------------------------

case46h/1-468              TGTTCATACTGAGGAGAACATGGGTAAAACAGAAACAGTAGAAAGCTAACTTTTAATTAC
gnl|ti|541136674/627-1103  TGTTCACACTGAGGAGAACATGGGTAAAACAGAAACAGTAGAAAGCTAACTTTTAATTAC
case46c/1-157              ----CATACTGAGGAGAACATGGGTAAAACAGAAACAGTAGAAAGCTAACTTTTAATTAC

case46h/1-468              TCTAA
gnl|ti|541136674/627-1103  TCTAG
case46c/1-157              TCTAG
```

chr17:79427256:79427568:scaffold_34699:232697:+
... AluY in Rhesus has polymorphisms at head and tail, gene conversion to AluYg6 in human, precise deletion in chimpanzee

CLUSTAL

```
case47h/1-412              GAACGTCTTCCCATGTCATTAAACACAACAAAATAAGGTTAGGATAGATTAA-GATTGAA
gnl|ti|503733953/300-735   GAACATCTTCCTATGTCATTAAACACAACAAAATAAGGTTAGGATGGATTAAAGATTGAA
case47c/1-101              GAACGTCTTCCCATGTCATTAAACACAACAAAATAAGGTTAGGATAGATTAAAGATTGAA

case47h/1-412              CGTTTA----------------GGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACT
gnl|ti|503733953/300-735   CATTTAAAATAAACCAT-AAGGGGCCAGGCGCGGTGGCTCACGCCTGTAATCACAGCACT
case47c/1-101              CTTTTAAAACAAACCGTTAAG--------------------------------------

case47h/1-412              TTGGGAGGCCGAGACGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACAC
gnl|ti|503733953/300-735   TTGGGAGGCCGAGGCGGGCGGATCACGAGGTCAGGAAATCAAGACCATCCTGGCTAACAC
case47c/1-101              ------------------------------------------------------------

case47h/1-412              GGTGAAACCCCGTCTCTACTAAAAATACAAAAA-TTAGCCGGGCATGGTGGCGTGCGCCT
gnl|ti|503733953/300-735   GGTGAAACCCCTTCTCTACTAAAAATACAAAAAATTAGCTGGGCGTGGTGGCGGGCACCT
case47c/1-101              ------------------------------------------------------------

case47h/1-412              GTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGC
gnl|ti|503733953/300-735   GTAGTCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATGGGGTGAACCCGGGAAGAGGAGC
case47c/1-101              ------------------------------------------------------------

case47h/1-412              TTGCAGTGAGTCGAGATCGCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAAACTCCGT
gnl|ti|503733953/300-735   TTGCAGTGAGCTGAGATCCGGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCAT
case47c/1-101              ------------------------------------------------------------
```

```
case47h/1-412             CTCAAAAAAAAAAAAAAAAA-----------------------AAAGATTGAACGTTTA
gnl|ti|503733953/300-735  CTCAAAAATAATAATAATAATAATAATAATAATATAATAAAATA----------------
case47c/1-101             ------------------------------------------------------------

case47h/1-412             AAACAAACCGTAAGACACCATGAAAGACCTGGTG
gnl|ti|503733953/300-735  -AATAAACCATAAGACACCATGAAAGACCTGGTG
case47c/1-101             --------------ACACCATGAAAGACCTGGTG
```

chr19:8420410:8420717:scaffold_37480:196649:-
...AluY in rhesus, gene conversion to AluYa5 in human, precise deletion in chimpanzee

CLUSTAL

```
case48h/1-462             GTCACCATGTCGTCACTAGGCCTCGGTCCTATGGAGGCTACTACCTACACGCTTACAGCT
gnl|ti|496120749/118-591  GTCACCCTGTTGTCACTAGGCCTCGGTCCTATGGAGGTTACCACCTACACGCTAACAGCT
case48c/1-155             GTCACCATGTCGTCACTAGGCCTCGGTCCTATGGAGGCTACTACCTACACGCTTACAGCT

case48h/1-462             TTAAAAGAGACTCTTAGGCCGGGCGCGGTGGCTCACGCCTGTAATCCCAGCACTTTGGGA
gnl|ti|496120749/118-591  TTAAAAGAGACTCTT-GGCCGGGCACGGTGGCTCAAGCCTGTAATCTCAGCACTTTGGGA
case48c/1-155             TTAAAAGAGACTCTTA-------------------------------------------

case48h/1-462             GGCCGAGGCGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTAAAACGGTGAA
gnl|ti|496120749/118-591  GGCCGAAACGGGCGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTGACACGGTGAA
case48c/1-155             ------------------------------------------------------------

case48h/1-462             ACCCCGTCTCTACT----AAAAATACAAAAAATTAGCCGGGCGTAGTGGCGGGCGCCTGT
gnl|ti|496120749/118-591  ACCCCGTCTCTACTTAAAAAAAATACAAAAAACTAGCCGGGCGAGGTGGCAGGCGCCTGT
case48c/1-155             ------------------------------------------------------------

case48h/1-462             AGTCCCAGCTACTTGGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTT
gnl|ti|496120749/118-591  AGTCCCAGCTACTCGGGAGGCTGAAGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTT
case48c/1-155             ------------------------------------------------------------

case48h/1-462             GCAGCGAGCCGAGATCCCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCT
gnl|ti|496120749/118-591  GCAGTGAGCTGAGATCCGGCCACTGCACTCCAGCCTGGGCAGCAGGGCGAGACTCCGTCT
case48c/1-155             ------------------------------------------------------------

case48h/1-462             CAAAAAAAAAAAA-----AAAAGAGACTCTTAGCACACAGT----GAGGCAATGATGTAT
gnl|ti|496120749/118-591  CAAAAAAAAAAAAAAAAAAAAAAGAGACTCTTGGCACACAGTAACTGAGGCAATGATGTAT
case48c/1-155             ---------------------------------GCACACAGT----GAGGCAATGATGTAT

case48h/1-462             TTTTCTACCAATAATTTTATTGAGAAGGGTAGAGAGGGGGACTGATTCACTCCTA
gnl|ti|496120749/118-591  TTTTCTACCAATAATTTTATCGAGAAGGGTAGAGAGGCGGACTGATTTACTCCTA
case48c/1-155             TTTTCTACCAATAATTTTATTGAGAAGGGTAGAGAGGGGGGCTGATTCACTCCTA
```

>chr19:41499856:41500219:scaffold_37497:2669080:-
....rhesus has partial tandem duplication of AluSq; chimp 3 deletions, including orig AluSq

CLUSTAL

```
case49h/1-463             AACATTTTTGGTAATTATTTT------CTATATATTTTAGATATTTCATTAAACTAATAC
gnl|ti|563501071/127-787  AACAGTTTTGGTAATTATTTTTACTTTCTATATATTTTAGAGTTTCCATTAAACTAATAT
case49c/1-115             AACATTTTTGGTGATTATTTT------CTGTGTATTTTAGGATTTCCATTAAACTA----

case49h/1-463             AGAATTTCATATTCAGGGCCAGGCATAGTGGCTCATGCCTGTAATCCCAGCACTTTGAGA
gnl|ti|563501071/127-787  AGAATTTCATATTCAGGGCCTGGTGCAGAGGCTCATACCTGTAATCC-AGCACTTTGGGA
case49c/1-115             ---------TATTC----------------------------------------------

case49h/1-463             GTCCAAGGCGGGCGCATCACCTGAGGTCAGGGGTTCGAGACCATCCTGGCCAACAAGGGA
gnl|ti|563501071/127-787  AGCCGAGATGGGCGGATCACCTGAGGTCAAGGGTTCGAGACCAGCCTGGCCAACATGGCC
case49c/1-115             ------------------------------------------------------------
```

```
case49h/1-463             AAACCCCATCTCTATTAAAAATACAAAATTAGCCGGGTGTGGTGTTGCACGCCTGTAATC
gnl|ti|563501071/127-787  ACACCCCGTCTCTACTAAAAATTCAAAATTAGTCGGGTGTGGTGGTGCATGCCTGTAATC
case49c/1-115             ------------------------------------------------------------

case49h/1-463             CCAGCTACTTGGAAGGCTGAGGCAGGAGAATC--------------------------
gnl|ti|563501071/127-787  CCAGCTACTTGGTATGCTGAGGCAGGAGAATCCATTGAAACTGCGAGGCGGAGTTTGCAG
case49c/1-115             ------------------------------------------------------------

case49h/1-463             ------------------------------------------------------------
gnl|ti|563501071/127-787  TGAGCCAACATCACGCCATTGTACTCTAGCCTGGGCAACAGTAGTGAAACTCCAGCTCAA
case49c/1-115             ------------------------------------------------------------

case49h/1-463             ---------------------------------------------GCATG----------
gnl|ti|563501071/127-787  AAAAAAAAAAAAAAAAAAAAAAAAAATACAAAACTTAGCTGGGCATGTTGGCGGGTGCTG
case49c/1-115             ------------------------------------------------------------

case49h/1-463             ------------------------------------------ACCCCGGGGGCAGAGA
gnl|ti|563501071/127-787  ATAATCCCAGTCACTCGGTAGGCTGAGGCAGGAGAATTGCTTCAACCCAGGGAGCAGAGG
case49c/1-115             ------------------------------------------------------------

case49h/1-463             TTGCAGTGAGCTGAGATCTTGCCACTTCATTCCAGCCTGGGCCACAGAGCAAGACTCCTT
gnl|ti|563501071/127-787  TTGCAATGAGCCAAGATCTCATGACTTCGCTCCAGCCTGGGGCACAGGGCAAAACTCCTT
case49c/1-115             ------------------------------------------------------------

case49h/1-463             CTCAAAAAAAAAAAAAAAAAAAAA-AAATTCATATTCGCTCATATCAAAAATGAAAATTT
gnl|ti|563501071/127-787  CTCAAAAAAAAAAAGAAAAAAATTCAAATTCATATTCACTCATTTCAAAAATGAAAATTT
case49c/1-115             ----------------------------------------TCATTTC----------

case49h/1-463             ATTTTTGCAAATTTCTA----AGTGATAGAATTATTTTAATGTAGGAAAGG-TTCATCAA
gnl|ti|563501071/127-787  ATTTTTGCAAATTTCTATTTGAGTGATAGAATTATTTTAATTTAGGAATGGCTTCATAAA
case49c/1-115             --------AAATTTCTATCTGAGTGATAGAATTATTTTAATGTAGGAAAGG-TTCATCAA

case49h/1-463             AA
gnl|ti|563501071/127-787  AA
case49c/1-115             AA
```

chr19:46430068:46430397:scaffold_37543:1476014:-
...precise deletion of AluSq in chimpanzee

CLUSTAL

```
case50h/1-495             AGAGTAGGGAATATTCGCTAGAA---GGATATATTACAACCCAGATGAGCTAGACCCAGC
gnl|ti|502904367/30-527   ACAGTAGGGAATATTTGCTAGAATGAGGATATATTACAACCCAGATGAGCTAGACCCAGA
case50c/1-166             AGAGTAGGGAATATTTGCTAGAA---GGATATATTACAACCCAGATGAGCTAGACCCAGC

case50h/1-495             CTCTGCCCTCAAGTTGCTCCTAGAATAAGAAAACCAAAACCAGGCCAGGTGTGGTGGCTT
gnl|ti|502904367/30-527   CTCTGCCCTCAAGTTCCTCCTAGAGTAAGAAAACTAAAACCAGGCCAGCTGTGGTGGCTT
case50c/1-166             CTCTGCCCTCAAGTTGCTCCTAGAATAAGAAAACCAAAACCA------------------

case50h/1-495             ACACCTGTAACCCCAGCACTTTGGGAGGCCAAGGCTGGTGGATCACCTGAGGTCAGGAGT
gnl|ti|502904367/30-527   ACACCTATAACCCCAGCACTTTGGGAGGCCACGGCGGGTGGATCACCTGAGGTCAGGAGT
case50c/1-166             ------------------------------------------------------------

case50h/1-495             TCGAGACCAGCCTGGCTAACATGGTGAAACCCCATTTCTACTAAAAATACAAAAAATTAG
gnl|ti|502904367/30-527   TCGAGACCAGCCTGGCTAACATGGTGAAACCCCATTTCTACTAAAAATACAAAAAATTAG
case50c/1-166             ------------------------------------------------------------

case50h/1-495             CCGGGTGTGGTGGCACACACCTGTAATCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATC
gnl|ti|502904367/30-527   CCAGGTGTGGTGGCACACACCTGTAATCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATC
case50c/1-166             ------------------------------------------------------------

case50h/1-495             CCTTGAACCCTGGAGGCAGAGGTTGCAGTGAGCCGAGATCGTGCCATTGCACTCCAGCTT
gnl|ti|502904367/30-527   ACTTGAACCTGGGAGGCAGAGGTTGCAGTGAGCCAAGATTGTGCCATTGCACTCCAGCTT
case50c/1-166             ------------------------------------------------------------

case50h/1-495             GGGTAACACGAGCGAACTTCCGTCTCAAGAAAAAAAAAAAAAAGAAAGAAAGAAAGAAGAA
gnl|ti|502904367/30-527   GGGTAACAAGAGCGAACGTCCGTCTCAAAAAAAAAAAAAAAAAAAGACAGAAAGAAGAA
case50c/1-166             ------------------------------------------------------------
```

```
case50h/1-495           AACCAAAACCAAAACAAAACTCACAGATCTGTAAATATAAGGCCTAATTCTGGTCTGAAG
gnl|ti|502904367/30-527 AACCAAAACCAAAACCAAATTCACAGATCTGTAAATATAAGGCTGAATTCTGGTCTGAAG
case50c/1-166           ----------AAACAAAACTCACAGATCTGTAAATATAAGGCCTAATTCTGGTCTGAAG

case50h/1-495           TTCTCTGAGATTAGAAAG
gnl|ti|502904367/30-527 TTCTCTGAGATTAGAATG
case50c/1-166           TTCTCTGAGATTAGAAAG
```

chr20:41311768:41311810:scaffold_37443:5951986:-
...L2 element tandemly duplicated region in human, chimp, and rhesus

CLUSTAL

```
case51c/1-100           GCCTTTTCAGGGGACTTGACCTCAGCTGGAGACTTTGCTTCTTCCTT-------------
gnl|ti|545384313/363-522 TCCTTCTCTGGGGACTTGGCCTTCTCGGGGGACTTTGCTTCCTCCTTCGTTGGGGACTTG
case51h/1-142           GCCTTTTCAGGGGACTTGACCTCAGCTGGAGACTTTGCTTCTTCCTT-------------

case51c/1-100           ------------------------------------------------CACTGAGGACTTG
gnl|ti|545384313/363-522 GCCTTTTCAGGGGACTTGGCCTCAGCCGGTGACTTTGCTTCTTCCTTCACTGGGGACTTG
case51h/1-142           -----CACTGGGGACTTGGCCTCAGCTGGTGACTTTGCTTCTTCCTTCACTGGGGACTTG

case51c/1-100           GCCTTCTCTGGAGACTTGGCCTCAGCTGATGATTTTGCCT
gnl|ti|545384313/363-522 GCCTTCTCTGGAGACTTGGCCTCAGCTGGTGACTTTGCCT
case51h/1-142           GCCTTCTCTGGAGACTTGGCCTCAGCTGGTGATTTTGCCT
```

chr22:45658137:45658441:scaffold_37534:1549045:-
...precise deletion of AluSq in chimpanzee

CLUSTAL

```
case52h/1-458           CTGCTTAACCAGATGAGGAAGAACGAGGTTAATGAAAATGCCCAGTGATGGTGACGGTAA
gnl|ti|555960713/344-767 CTGCTTAACCAAATGAGGGAGAACAAGG-------AAATGCCCAGTGATTGTGAGGGTAA
case52c/1-154           CTGCTTAACCAGATGAGGAAGAACGAGGTTAATGAAAATGCCCAGTGATGGTGAGGGTAA

case52h/1-458           AGAAATGCCCCCTCTCGGCCAGGCGCGGTGGCTCATGTCTGTAATCCCAGCACCCTGGGG
gnl|ti|555960713/344-767 AGAAATGCCCCCTCTCGGCCGGGCACGGTGGCTCACACCTGTAATCCCAGCACTTTGGGA
case52c/1-154           AGAAATGTCCCCTCTC----------------------------------------------

case52h/1-458           GGCCGAGGCGGGCGGATCACTTGAGGTCAGGAGTTTGAGACCAGCCTGGCCAACAGGGTG
gnl|ti|555960713/344-767 GGCCGAGGCAGGCGGATAACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACATGGTG
case52c/1-154           ------------------------------------------------------------

case52h/1-458           AAACCCCGTCTCTACTAAAAAATACAAAAATTAGCCAGGCGTGGTGGCAGGCGCCTTTAAT
gnl|ti|555960713/344-767 AAACCCTGTCTCTACTAAAAAATAGAAAAATTAGCTGGGCGTGGTGGCAGGAGCTTTAAT
case52c/1-154           ------------------------------------------------------------

case52h/1-458           CCTAGCTACTTGGGAGGCAGAGGCAGGAGAATCGTTTGAACCCAGGAGGCAGAGGTTGCA
gnl|ti|555960713/344-767 CCCAGCTACTTGGGAGGC--------------------------GGAGGCAGAGGTTGCA
case52c/1-154           ------------------------------------------------------------

case52h/1-458           GTGGGCTGAGATCGAGCCACTGCACTCAAGCCTGGGGGACAAGGGCGAGACTTCTCTGAA
gnl|ti|555960713/344-767 GTGAGGCAAGATCGAGCCATTGCACTCAAGCCTGGGGGACAAGGGTGAGACTTCTGTCAA
case52c/1-154           ------------------------------------------------------------

case52h/1-458           AAAAGGAAATGCCCCCTCTCACAAAACTGCTGGCTGCAGGGCAAACCAACTCAGTGGGCC
gnl|ti|555960713/344-767 AAAAG-AAATGTCCCCTCTCACAAAATTGCTGGCTGCCCGGCAAACCAACTCAGTGGGCC
case52c/1-154           --------------------ACAAAATTGCTGGCTGCAGGGCAAACCAACTCAGTGGGCC

case52h/1-458           CCAGGGTCACTTGGCTGTGGCCACCAAGTTCCCCAAAC
gnl|ti|555960713/344-767 CCAGGGTCACTTGGCCGTGTGCACCAAGTTCCACAAAC
case52c/1-154           CCAGGGTCACTTGGCTGTGGCCACCAAGTTCCTCAAAC
```

chr22:46857451:46857769:scaffold_37534:338287:-

...AluY partially deleted and reversed in Rhesus (slightly complicated example of NHEJ), gene conversion to AluYa5 in human, precise deletion in chimpanzee

CLUSTAL

```
case53h/1-418             GGCACTGGACCAAGCCTTCCTGCTGGGCAGAGATGGGACTGGCTTTTCATAAGATTGCGC
gnl|ti|556293551/420-813  GGCAGTGGACCAAGCCTTCCTGCCGGGCAGAGACGGGACTGGC----------------
case53c/1-100             GGCACTGGACCAAGCCTTCCTGCTGGGCAGAGACGGGACTGGCTTTTCATAAGATTGAGC

case53h/1-418             CTTGGGCCGGGCACGGTGGCTCACTCCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGG
gnl|ti|556293551/420-813  ----------------XXATGAAAAGCCAGXXTCCCAGCACTTTGGGAGGCCGAGACGGG
case53c/1-100             CTT-------------------------------------------------------

case53h/1-418             CGGATCACGAGGTCAGGAGATCGAGACCATCCCGGCTATAACGGTGAATCCCCGTCTCTA
gnl|ti|556293551/420-813  CGGATCACGAGGTCAGGAGATCGAGACCATCCTGGCTAACACGGTGAAACCCCGTCTCTA
case53c/1-100             -------------------------------------------------------

case53h/1-418             CTA-AAAATACAAAAAA-TTAGCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTT
gnl|ti|556293551/420-813  CTACAAAATACAAAAAAACTAGCCGGGCGAGGTGGCGGGCACCTGTAGTCCCAGCTACTC
case53c/1-100             -------------------------------------------------------

case53h/1-418             GGGAGGCTGAGGCAGGAGAATGGCGTGAACCCGGGAGGCGGAGCTTGCAGTGAGCCGAGA
gnl|ti|556293551/420-813  GGGAGGCTGAGGCAGGAGAATGGCGTGAACCTGGGAGGCGGAGCTTGCAGTGAGCTGAGA
case53c/1-100             -------------------------------------------------------

case53h/1-418             TCCCGCCACTGCACTCCAGCCTGGGCGACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAA
gnl|ti|556293551/420-813  TCCGGCCACTGTACTCCAGCCTGGGCCACAGAGCGAGACTCCGTCTCAAAAAAAAAAAAA
case53c/1-100             -------------------------------------------------------

case53h/1-418             ACAAAAAAAA-------AAGATTGCGCCTTTCAGCAACATCAACTCTTCCAGGAAATGTG
gnl|ti|556293551/420-813  AAAAAAAAAAGAAAAXXAAGATTGTGCCTTTCAGCAACATCGAGTCTTCCAGGAAATGTG
case53c/1-100             -------------------------------TCAGCAACATCAACTCTTCCAGGAAATGTG

case53h/1-418             CTTATAT
gnl|ti|556293551/420-813  CTTACAT
case53c/1-100             CTTATAT
```

chrX:5202006:5202292:scaffold_25582:6996:-
...precise deletion of AluSp in chimpanzee

CLUSTAL

```
case54h/1-431             TGGAAGCCTGACCAGAAAATATCATGGCATCAGTTCAACGCATCCCAAAAACTCACTTAA
gnl|ti|485777938/470-934  TGGCAGCCTGACCAGAAAATATGATGGCATCAGTTCAACGCATCCCAAAAACTCACTTAA
case54c/1-140             TGGTAATCTGATCAGAAGATATCA------CAGTACAACGCATCCCAAAAACTCACTTAA

case54h/1-431             AAGCCAAAGGCAGGCCGGGCGTGGTGGCTCACGCCTATAATCCCAGCACTTTGGGAGGCC
gnl|ti|485777938/470-934  AAGCCAAAGGCAGGCCGGGAATGGTGGCTCACGCCTATAATCCCAGCACTCTGGGAGGCC
case54c/1-140             AAGCTAAAGGCA----------------------------------------------

case54h/1-431             GAGGCAGGTGGATCACCTGAGGTCGGGAGTTCAAGACCAGCCTGACCAACATGGAGAAAT
gnl|ti|485777938/470-934  GAGGCAGGTGGATCACCTGAGGTCGGGAGTTCAAGACCAGCCTGACCAACATGGTGAAAC
case54c/1-140             -------------------------------------------------------

case54h/1-431             CCCATCTCTACTAAAAATACAAAATTAGCCAGGTGTGGTGGCACATGCCTGTAATCCCAG
gnl|ti|485777938/470-934  CCCATCTTTACTAAAATTACAAAATTAGCTGGGTGTGGGGGCACATGCCTGTAATCCCAG
case54c/1-140             -------------------------------------------------------

case54h/1-431             CTACTCGGGAGG---CTGAGGCAGGAGAATGGCTTGAACCTGGGAGGGGGAGGCTGCAGT
gnl|ti|485777938/470-934  CTACTCGGGAGGAGGCTGAGGCAGGAGAATGGCTTGAACCTGGGAGGCAGAGGCTACGGG
case54c/1-140             -------------------------------------------------------

case54h/1-431             GAGCGAAACTC-----------------------CATCAAAAAAAAAAA--------AAA
gnl|ti|485777938/470-934  GGGCCAAGATCGCGCCATTGCACTCCAGACTGGGCAACAAGAGGGAAACTCCGTTTCAAA
case54c/1-140             -------------------------------------------------------

case54h/1-431             AAAAAGGAAAGAAAAAAAAAAGCCAAAGACAAACAAATCATCTGACAGCTGCAAAGAAAA
gnl|ti|485777938/470-934  AAAAAAAAAAAAAAAAAAAATACCAAAGGCAAACAAATCATCTGACATCTGCAAAGAAAA
case54c/1-140             -------------------------------AACAAATCAGCTGACGTCTGCAAATAAAC
```

```
case54h/1-431              GTGCAAGTCCCTATGTTTTGTTTTGTTTTTCATTCTATTTCCAGA
gnl|ti|485777938/470-934   GTGCAAGTCCCTATGTTTTGTTTTGTTTTTCATTCTATTTCCAGA
case54c/1-140              ATGCAAGTCTCTATGTTTTGTCTTGGTTTTCACCCTATCTCCAGA


chrX:86865679:86865830:scaffold_37382:835478:+
...imprecise deletion of low complexity region in chimpanzee, no flanking identity

CLUSTAL

case55h/1-251              GTAATAGA-------ATAGGAAAAGTTTATTTCTTATTCTTAAAGATGAATCATTTAGAA
gnl|ti|495823394/144-417   GTAATATACAGTGTAATAGGAAAAGTTTATTTCTTACTCTTAAAGATGAATCATTTGGAA
case55c/1-100              GTAATAGA-------ATAGGAAAAGTTTATTTCTTATTCTTACAGATGAATCATTTA---

case55h/1-251              CAAAAATTTTTGCTTTTTCTTTTAGAATATATATATGTGTGTATATATATGT--------
gnl|ti|495823394/144-417   CAA--ATTTTTGCTTTTTCTTTTAGAATATATATGTGTGTGTATATATATGTGTGTGTAA
case55c/1-100              ------------------------------------------------------------

case55h/1-251              ----------ATATATGTGTGTGTATATATATGTATATATATGTGTGTGTATATATATAT
gnl|ti|495823394/144-417   ATGTGTGTATATGTGTATGTGTGTGTATATATGTGTATGTATATATGTGTGTGTGTATGT
case55c/1-100              ------------------------------------------------------------

case55h/1-251              GTGTATATATATGTACTGAAGCATATTCTCAAAATGTGCAAAGAGGCTGCAGTAATATTA
gnl|ti|495823394/144-417   GTGTATATATAAGCACTAAAGCATATTCTCAAAATGTGCAAAGAGGCTGCAGTAATATTA
case55c/1-100              ----------------------------------------------CTGCAGTAATATTA

case55h/1-251              TAGATAATTAAAATGAGTCAAACTCTGATTTTGAGG
gnl|ti|495823394/144-417   TGGATAAGTAAAATGAGTCAAACTCTGATTTTGAGG
case55c/1-100              TTGATAATTAAAATGAGTCAAACTCTGATTTTGAGG

Primers and Sequences

>caseC1_3
GTACAGTTGAGGCATTGCTAC
>caseC1_5
TCAGTCTCCAGGGAAGCAATG
>caseC2_3
AGGCAATAAAAGAGGCCGGCT
>caseC2_5
CAGAGCTCTTTCCTTCCACTC
>caseC3_3
TGGGTTATAGGCTTACAGATG
>caseC3_5
GAGATAGGCCAAGAACTATAG
>caseC4_3
AGAGTACCACCAAGGTATTAG
>caseC4_5
GAACTGATGTCTGCAACTTTG
>case14_3
CATACACATATAAGACCCTTC
>case14_5
GTCTCAGTGATAACTTGATGA
>case33_3
TTGTAGGGTTGAGAGAGCCTC
>case33_5
TGGCCACTTACCTTCTGCTTC
>case42_3
CTTTCTGTCTTATGGGAACTC
>case42_5
GAACATCTCTATTCACCTTCG
>case43_3
TTAGTGCAGGATGAAGTTGGC
>case43_5
TTCTCCCATCTGGTCATGTGA
>case52_3
CAGAAAGACACCATGGGTGAA
>case52_5
GCCTGTGGATAGATCATAGTC
```