Distribution Agreement Page

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I agree that the Library of the University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to copy from, or to publish, this dissertation may be granted by the professor under whose direction it was written when such copying or publication is solely for scholarly purposes and does not involve potential financial gain. In the absence of the professor, the dean of the Graduate School may grant permission. It is understood that any copying from, or publication of, this dissertation which involves potential financial gain will not be allowed without written permission.

Student's signature

_____

E. Andrew Bennett

Genetic variation caused by active retrotransposons in the human genome

By

E. Andrew Bennett
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program of Genetics and Molecular Biology

_____
Scott E. Devine, Ph.D.
Advisor

_____ _____
Ichiro Matsumura, Ph.D. James Thomas, Ph.D.
Committee Member Committee Member

_____ _____
Paula Vertino, Ph.D. Stephen Warren, Ph.D.
Committee Member Committee Member

Accepted:

_____
Lisa A. Tedesco, Ph.D.
Dean of the Graduate School

_____
Date

Genetic variation caused by active retrotransposons in the human genome

By

E. Andrew Bennett
M.S., University of Leeds, 2001
B.S., Georgia State University, 1998
B.A., Georgia State University, 1998

Advisor: Scott E. Devine, Ph.D.

An Abstract of
a dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program of Genetics and Molecular Biology

December 2, 2008

Abstract

Genetic variation caused by active retrotransposons in the human genome

By E. Andrew Bennett

Nearly one third of our genome belongs to three families of active retrotransposons: L1, Alu, and SVA, that continue to create genetic variation and cause disease in humans. We developed several novel methods to detect the contribution transposable elements have made to human genetic diversity, and identified 25-35% of transposon insertion polymorphisms commonly found in human populations. Our method improved on previous assays by identifying different families of recently mobile elements equally. Using this same approach, we identified nearly 11,000 species-specific transposon insertions that have mobilized in the past 6 million years in humans and chimps. We found that humans possess a more diverse and active collection of retrotransposon subfamilies, and have sustained almost twice as many new insertion events since our last common ancestor. The majority of recent insertions in both humans and chimps were caused by Alu elements. There are over 1 million Alus in humans and they collectively occupy 10% of our genome. It was unclear however, how many of these remained active, and what constituted an active Alu sequence. In order to define the requirements for Alu activity, we performed a comprehensive analysis using conservation data and retrotransposition assays. We show that active Alu elements display a high degree of sequence variation, but must conserve nucleotides that enable them to bind to SRP9/14 proteins. Furthermore, we show the affinity for SRP9/14 binding has decreased since the earliest Alus evolved from 7SL RNA. We estimate that at least 10,000 of the Alus in our genome are capable of causing new genetic variation through future retrotransposition.

Genetic variation caused by active retrotransposons in the human genome

By

E. Andrew Bennett
M.S., University of Leeds, 2001
B.S., Georgia State University, 1998
B.A., Georgia State University, 1998

Advisor: Scott E. Devine, Ph.D.

A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Graduate Division of Biological and Biomedical Sciences
Program of Genetics and Molecular Biology

December 2, 2008

Acknowledgements

My thanks to Scott Devine, Rebecca Iskow, Jackie Griffith, Chris Luttig, Laura Coleman, Laura McLane, and my committee members.

Special thanks for the support of my family and Christin Gray.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Introduction to retrotransposons

Both eukaryotes and prokaryotes contain discrete stretches of DNA capable of inserting themselves into new locations in their host genome (Campbell, *et al*. 1979). First recognized as repeated sequences occurring throughout genomic sequence, these 'transposons' are represented by a broad array of elements that belong to some 848 families in humans (reviewed in Mills, *et al*. 2007). That nearly half of the human genome can be attributed to transposon sequences is testament to the role these elements have played shaping our species, and indeed that of many organisms. This role is both historic and contemporary, as a fraction of these elements remain active and continue to cause heritable changes in our genomes (reviewed in Prak and Kazazian 2001; Deininger, *et al*. 2003).

Transposons were first identified as the causative agents of the mosaic coloration in certain types of corn (McClintok, *et al*. 1950). Since then they have been characterized in all kingdoms of life (reviewed in Hickey 1992; Zillig, *et al*. 1996; Kempken and Kück 1998) . Transposons are divided into two major classes based on the basic mechanisms by which they transpose or 'jump' into genomic DNA. DNA transposons encode an enzyme, transposase, which is used to mobilize itself using a 'cut and paste' mechanism, whereby the transposon is excised from its surrounding genomic sequence and reinserted into a different genomic location. This mode of transposition does not increase the absolute number of transposons in the genome. While the relics of past DNA transposon copies are evident in our genome, the last DNA transposon jumped in our ancestors about 40 million years ago, and active copies of DNA transposons in modern primates and rodents are unlikely (Pace and Feschotte 2007; Xing, *et al*. 2007).

Retrotransposons on the other hand are mobilized through a 'copy and paste' mechanism. They are transcribed by cellular RNA polymerase into an RNA intermediate, which is then reverse transcribed into a cDNA copy elsewhere in the genome. This method of transposition, or 'retrotransposition', allows for the amplification of the total number of copies in the genome, and in humans this expansion has resulted in 88% of total transposon sequence belonging to retrotransposons (reviewed in Goodier and Kazazian 2008).

Retrotransposons are divided into two groups, the retrovirus-like LTR (Long Terminal Repeat) elements (such as HERV-K, Figure 1-1), and non-LTR elements such as the mammalian LINEs (Long Interspersed Nuclear Elements) and SINEs (Small Interspersed Nuclear Elements) (Figure 1-1). The only retrotransposons with evidence of substantial ongoing activity in humans belong to these last two groups of non-LTR retrotransposons, and they are the concentration of this dissertation.

L1

The L1 (LINE-1) retroelement is the only active autonomous transposon in humans (Dombroski *et al*. 1991). This ~6000 nucleotide element encodes for two proteins, ORF1p and ORF2p, both of which are required for successful transposition (Moran, *et al*., 1996). ORF2p has endonuclease and reverse transcriptase activities (Feng, *et al*. 1996; Alisch, *et al*. 2006). Less is known of the function of ORF1p, although it possesses a RNA binding domain (Hohjoh and Singer 1997a; Martin, *et al*. 2005). At the 5' end of

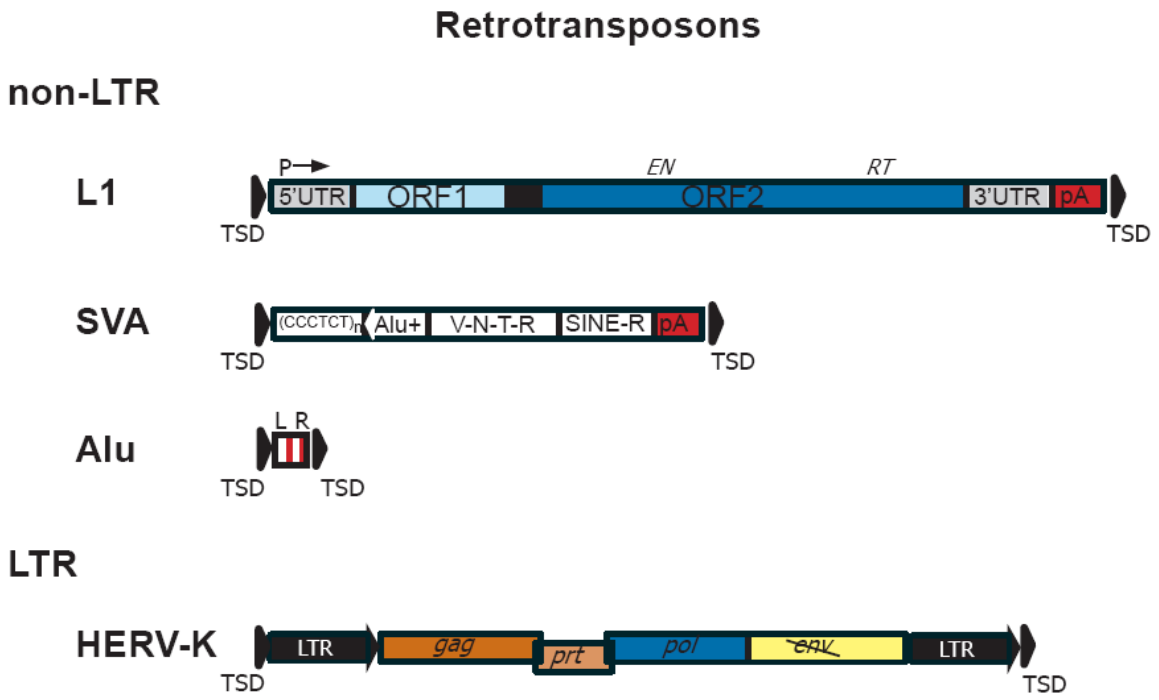


Figure 1-1. Structures of retrotransposons. The structures of the three recently active human non-LTR retrotransposons are depicted to scale. HERV-K, an LTR retrotransposon is also shown (not to scale). Abbreviations: TSD, target site duplication; P, promoter; UTR, untranslated regions; ORF, open reading frame; pA, poly(A); VNTR, variable number of tandem repeats; L, left Alu monomer; R, right Alu monomer, (red bars denote adenosine-rich regions in Alu); LTR, long terminal repeats; HERV-K genes: group-specific antigen (gag), protease (prt), polymerase (pol). The envelope gene (env) is inactive in all HERV-K copies.

L1 is a untranslated region (UTR) containing a polII promoter that directs transcription beginning upstream of itself, allowing L1 to transcribe and mobilize its own promoter (Minakami, *et al*. 1992). The 3' UTR contains a polyadenylation signal followed by an intact poly(A) tail. Of the 500,000 copies of L1 in the human genome, nearly all have accumulated mutations over time that have disrupted the reading frames or promoter

regions so that an individual typically has only 100 L1s still capable of retrotransposition (Brouha, *et al*. 2003).

Alu

Alu elements are SINEs derived from genomic 7SL RNA sequence (reviewed in Batzer and Deininger 2002). Normally, 7SL is an RNA component that binds to the six proteins of the signal recognition particle (SRP) (Gundelfinger, *et al*. 1983). The SRP complex binds to the ribosome, interfacing with the two proteins SRP9/14 which are bound together with the 7SL Alu domain (Strub and Walter 1990; Halic, *et al*. 2004). Once at the ribosome, the SRP complex arrests translation upon encounter of a specific peptide signal from a nascent protein that needs to be translated in the cell membrane or the endoplasmic reticulum (Sun *et al*., 2007; reviewed in Luirink and Sinning 2004; Halic and Beckman 2005).

Both primates and rodents have SINEs derived from 7SL (Kriegs, *et al*. 2007). In fact many SINEs are derived from RNAs that play a role in translation (Oshima, *et al*. 1996, Kapitonov and Jurka, 2003). The ancestral sequence that would become the B1 SINE in mice dimerized early in the primate lineage to give rise to the Alu element (reviewed in Kriegs, *et al*., 2007). Alu elements are 281 nucleotide sequences containing a loosely conserved polIII transcription promoter (A-Box and B-Box) at the 5' end. An adenosine-rich linker region separates the two dimers into a left and right side, and the 3' end is marked by a poly(A) tail.

Alus are non-autonomous retrotransposons. They code for no protein and are reliant on the L1 ORF2p protein for retrotransposition (Dewannieux, *et al*. 2003;

reviewed in Mills, *et al*. 2007). Unlike L1, Alu does not require L1 ORF1p to transpose (Dewannieux, *et al*. 2003). Nearly 900,000 full-length Alu sequences are present in the human genome, and over 200,000 partial and truncated sequences. Given the current lack in our knowledge of the specific details of the Alu retrotransposition mechanism, little is known about the sequence requirements of these Alu elements to maintain retrotransposition competence, and thus the subset of these Alus that have retained their retrotransposition capabilities is largely unknown.

SVA

The youngest retrotransposon showing evidence of activity in humans is the SVA element (Ostertag, *et al*. 2003; Bennett, *et al*. 2004; Mills, *et al*. 2006). SVA (SINE-R, VNTR (Variable Number Tandem Repeat), Alu-like) is a composite retrotransposon specific to the great apes, having arisen approximately 15-20 million years ago, and is represented by ~2600 copies in humans (Wang *et al*., 2005). Originally derived from a fragment of the endogenous retrovirus HERV-K10 and frequently 5' truncated, the 3' SINE-R region was first thought to be an independent retroelement (Zhu *et al*. 1994) until a dimorphic insertion revealed the structure of SVA in its entirety (Shen *et al*. 1994). 5' to 3', SVA is comprised of a CCCTCT hexamer repeat of variable length, a partially duplicated sequence derived from Alu, which is in reverse orientation and 379 nucleotides in length, a Variable Number Tandem Repeat (VNTR) of a 35-50 nucleotide CpG rich sequence, the SINE-R sequence, and a polyadenylation signal followed by a polyA tail (Figure 1-1). Due to the hexamer repeat and VNTR, the size of SVA is highly variable, from ~1500-3000 nt. SVA is not known to code for any protein, and the

mechanisms of its transcription and retrotransposition are yet to be determined. However, inserted copies share identical features with other L1-driven retrotransposition events such as target site duplications and 5' truncation (Shen, *et al*. 1994, Strichman-Almashanu, *et al*. 2001, see below).

Retrotransposon subfamilies

Retrotransposon subfamilies are classified using unique sequence changes that serve as markers to indicate a common phylogeny among subfamily members. These diagnostic changes occur in a progenitor element and are inherited in all new retrotransposon insertions that are transcribed from this founder or its progeny. For example, in the AluY family, new subfamilies are named using the number of nucleotide changes relative to the AluY consensus sequence. AluYa4 contains four specific changes from the Y consensus. AluYa5 contains those same four changes plus one additional change. Letters following Y specify a particular cluster of additive changes that track the evolution of the active elements; both Ya4 and Yj4 contain four changes from the Y consensus sequence but they occur at different positions.

Similar methods have been used to define L1, and SVA subfamilies. The youngest L1 elements in humans are designated L1-Ta (Transcribed subset A, Boissinot, *et al*. 2000). This group is further divided into Ta-0, Ta-1nd and Ta-1d subfamilies on the basis of additional diagnostic sequence changes in these elements. Six SVA subfamily

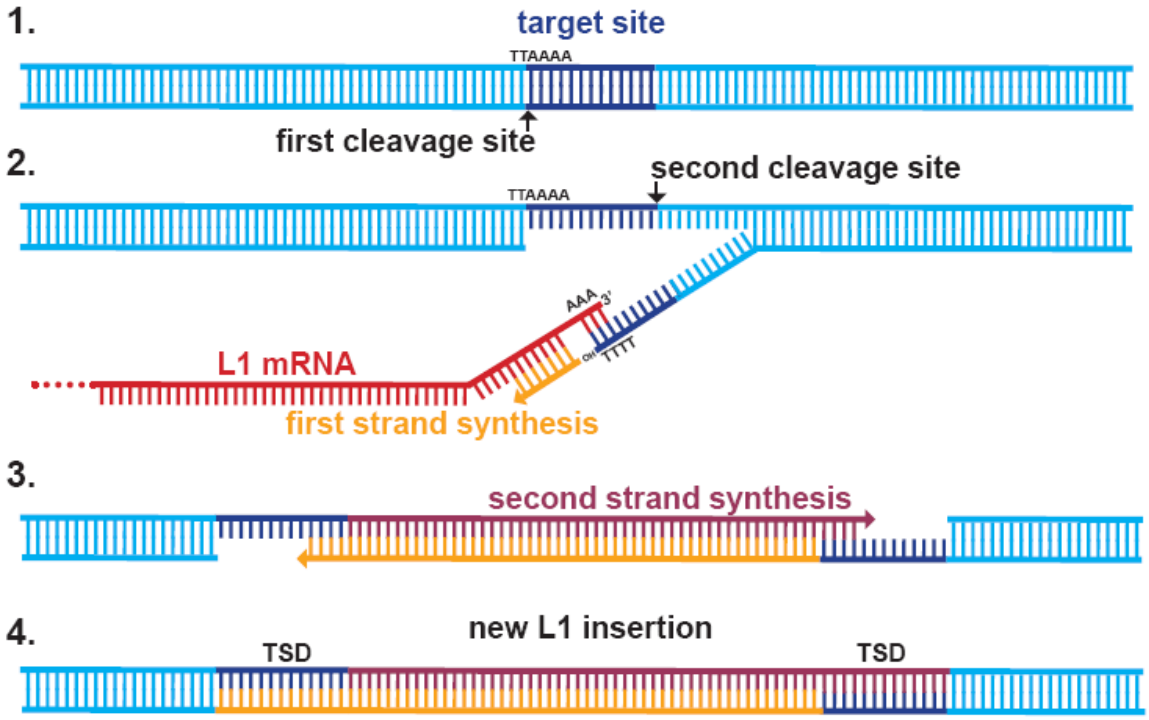


Figure 1-2. Target Primed Reverse Transcription. (1.) First strand of target DNA is cleaved by L1 ORF2p endonuclease. (2.) L1 ORF2p reverse transcriptase begins first strand synthesis primed by exposed 3' OH and using L1 RNA as a template. Cleavage of the second strand of DNA occurs several bases downstream of target site. (3.) DNA is repaired by second strand synthesis. This step creates duplicate copies of the target site. (4.) Newly inserted L1 copy.

consensus sequences, A-F, have been proposed (Wang, *et al*. 2005). SVA-A, -B, -C and -D predate the divergence of humans, chimps and gorillas. SVA-E and -F elements are only found in humans (these are the youngest SVA subfamilies). SVA-D is the largest subfamily, representing nearly half of all human SVAs.

Mechanisms of retrotransposition

L1 has been shown to transpose through a mechanism known as Target Primed Reverse Transcription (TPRT) (Luan, *et al*. 1993; Cost *et al*. 2002). In this multi-step process, L1

ORF2p binds the polyA sequence of the RNA to be reverse transcribed, while the endonuclease domain of the protein creates a nick in double stranded genomic DNA at a loosely conserved target sequence approximating 5'TTTTT/AA 3' (Repanas *et al*. 2007). The exposed 3' hydroxyl is then used to prime reverse transcription along the RNA template by the reverse transcriptase domain (Figure 1-2). Details of the second strand nick and resolution of the integrated retrotransposon are currently unknown, but staggered nicks result in a duplication of the target site flanking the newly inserted retrotransposon known as a Target Site Duplication (TSD) (Figure 1-2). TSDs are found flanking insertions of L1, Alu and SVA and are thought to be a hallmark of TPRT retrotransposition.

In cells, the L1 proteins are known to favor selection of their own RNA transcripts for retrotransposition, known as a *cis*-preference (Wei *et al*. 2001; Kulpa *et al*. 2006). On occasion however, different RNA transcripts successfully compete with the L1 transcript for retrotransposition by L1 ORF2p, allowing *trans* retrotransposition to occur. Rarely, processed cellular mRNA is used as a template by ORF2p, which generates pseudogenes, but more often the L1 machinery is successfully hijacked by non-autonomous SINEs, Alu and SVA.

The success of Alu in co-opting L1 proteins for its own retrotransposition may depend upon its homology to 7SL RNA in the ribosome-associated complex SRP. In this model, an Alu transcript displaces 7SL RNA to bind to SRP9/14 proteins. As these proteins are known to dock at the ribosome (Terzi *et al*. 2004; Halic *et al*. 2004), Alu RNA would be in close proximity in the event an L1 transcript is translated. In these cases, the newly formed ORF2p may capture the Alu RNA in *trans*, rather than its own

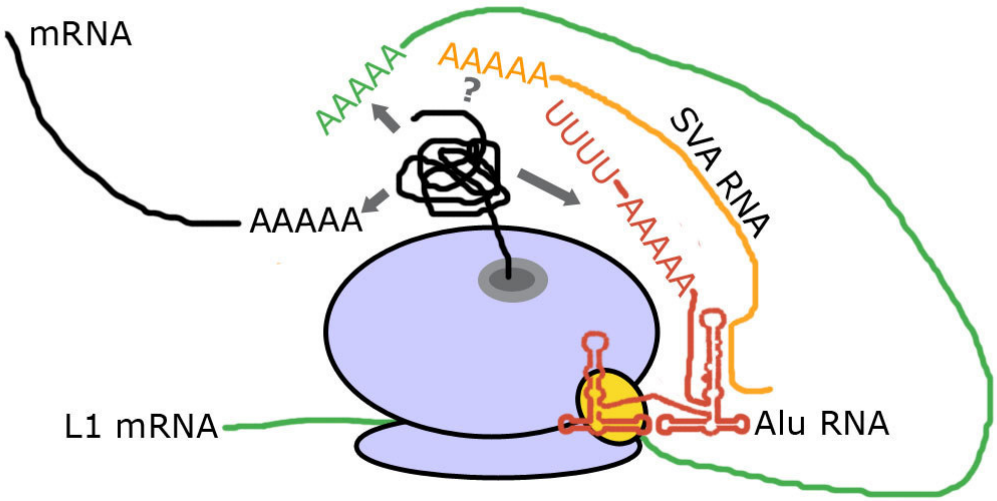


Figure 1-3. Model for L1, Alu and SVA retrotransposition. The model depicts possible scenarios for the *cis* and *trans* mechanisms of L1 retrotransposition. In the original model for *cis* preference, the L1 ORF proteins normally bind to their coding L1 mRNA (green), as they are translated. In the *trans* mechanism, docking of Alu (red) on ribosomes enables it to hijack ORF2p (black line) as it is translated. This docking is proposed to be mediated by interactions with SRP9p and SRP14p in a manner that is analogous to SRP9p/SRP14p binding on 7SL (these binding sites are similar because Alu was derived from 7SL. SVA (orange) might use a similar mechanism, perhaps by first hybridizing to Alu RNA. The black mRNA represents all other cellular mRNAs that in rare cases appear to be transposed by this mechanism, creating pseudogenes. Due to the infrequency of these events, cellular mRNA is not thought to serve as a good substrate for the L1 machinery. The gray arrows indicate possible fates of ORF2p. Figure adapted with permission from Dewannieux, *et al*. 2003.

transcript, and bring the Alu into the nucleus for retrotransposition into the chromosomal DNA by the TPRT mechanism (Boeke 1997). The role of SRP 9/14 proteins in Alu retrotransposition has not been formally demonstrated. Alu RNA does bind relatively well to SRP 9/14 proteins, although with less affinity than 7SL RNA (Birse *et al*. 1997; Sarrowa *et al*.1997), and sequences required for SRP binding are conserved across Alu subfamilies.

Although the bulk of *trans* retrotransposition events involve Alu elements, SVA insertions also share features concordant with L1-mediated retrotransposition, and SVA is also thought to use L1 proteins to retrotranspose (Ostertag, *et al*. 2003; Bennett, *et al*. 2004; Wang, *et al*. 2005). How an SVA transcript competes successfully with L1 RNA for ORF2p is currently unclear. One possibility is that the derived Alu component of SVA sequence can interact with SRP 9/14 proteins in a similar way as Alu RNA, and thus position itself near the newly formed ORF2p. As with Alu, this proximity may enhance the chances of the SVA RNA being selected for retrotransposition. A problem with this model is that the SVA-Alu sequence is the reverse complement of a standard Alu element, and thus is unlikely to fold and interact with SRP proteins in a similar way. One possible model for SVA retrotransposition that is consistent with the SRP-mediated mechanism would be that these reverse-oriented SVA-Alu sequences may hybridize with complementary sequences on active Alu transcripts (Mills, *et al*. 2007). This stably hybridized Alu RNA then, in turn, would escort the SVA RNA to the ribosome where SVA could efficiently participate in the *trans* mechanism of L1 retrotransposition (Figure 1-3). Supporting this model, the most conserved regions of the SVA-Alu correspond to

the stem structures of the Alu element (Figure 1-3), which are not crucial for SRP binding (Mills, *et al*. 2007).

Impact of retrotransposons in humans

Variation in genomic structure and gene expression

Active transposons behave as endogenous mutagens that can change the structure of chromosomes and the expression and function of genes (reviewed in Kazazian and Moran 1998; Callinan and Batzer 2006, Mills *et al*., 2007). Retrotransposition events that occur in germline cells, where much of retrotransposition is thought to take place (Brouha, *et al*. 2002; Jurka, *et al*. 2002) can permanently alter the genetic landscape of the next generation and act as a significant source of genetic variation among individuals. In time this renewing source of active elements can change the genomes they inhabit. There are several mechanisms by which retrotransposons can cause genetic or structural variation. L1 contains an inefficient RNA cleavage and polyadenylation signal, and transcription frequently continues into the 3' flanking genomic sequence until an alternate transcription terminator is encountered. The downstream flanking sequence then becomes incorporated into the L1 transcript and is transposed (and thus duplicated) to a new genomic location along with the L1, a mechanism termed 3' transduction (Moran, *et al*. 1999). 23% of L1 insertions contain 3' transduced sequence averaging 207 nucleotides in length, and together constituting 0.6% of our genome (Goodier, *et al*. 2000). 3' transduction is likewise common in Alu and SVA insertions. Low processivity of the ORF2 reverse transcriptase of L1 often results in truncations at the 5' ends of L1, Alu and

SVA elements. This can also occur during reverse transcription of the 3' transduction, which results in a transposed fragment of genomic DNA flanked by two complete TSDs, but without any retrotransposon sequence, effectively duplicating short fragments of DNA to new locations. Although rarer, 5' transductions that duplicate flanking sequence upstream of the retrotransposon can also occur. In these instances it is presumed that transcription of the retroelement was initiated upstream of the element's endogenous promoter and the upstream sequence became a part of the active transcript.

Genetic changes introduced by retrotransposition are not confined to duplications and insertional mutations. Aberrant nicking of DNA by the L1 endonuclease during reverse transcription has been described as a cause of large interstitional deletions (>3kb) near the target site, sometimes removing entire genes (Gilbert, *et al*. 2002). The increase of repetitive sequence that results from retrotransposition events continues to burden the genome long after the initial DNA strand breaks are resolved. Genomic rearrangements, inversions, duplications and deletions due to unequal crossover events between repeats with high degrees of homology are routinely documented (reviewed in Kazazian and Goodier 2002). In fact, an estimated 8.4 Mb of sequence has been removed from the human genome due to these retrotransposon-mediated deletion events during the past 6 million years since speciation (Xing, *et al*. 2007).

Apart from causing structural variation that may delete or duplicate entire genes or exons, other features of retrotransposons contribute to variable gene expression and transcript integrity. The most drastic of these is a retrotransposon insertion directly into a coding or regulatory region of a gene, reducing or eliminating its function (see McClintok 1950). However, several additional mechanisms exist beyond insertional mutagenesis for

retroelements to disturb gene expression or function. The sequence of the sense strand of the L1 itself is an inhibitor of transcriptional elongation, and can repress RNA production of genes containing L1 insertions in non-coding regions (Han, *et al*. 2004; Ustyugova, *et al*. 2006). Throughout the anti-sense strand are several cryptic polyadenylation signals that can affect the quality of transcripts containing these elements (Perepelitsa-Belancio and Deininger 2003). 80% of human genes contain L1 sequences in either orientation (Han, *et al*. 2004). Together, the number, size, and orientation of L1 elements in genes produce an added layer of regulation of transcript quantity and viability. In addition to its native promoter, the antisense strand of L1 also contains a promoter found to generate transcripts of the flanking upstream DNA (Mätlik, *et al*. 2006). Instances of Alu effecting gene expression have also been documented (Rubin, *et al*. 2002, reviewed in Schmid 1998). In its reverse orientation, the Alu sequence contains a splice acceptor site. When inserted into an intron in the reverse orientation to the gene, splice variants incorporating partial Alu sequence are generated, a process known as 'Alu exonization' (Sorek, *et al*. 2002, Lev-Maor, *et al*. 2003, Lin, *et al*. 2008). Transcripts containing inverted Alus may be subject to RNA editing by the adenosine deaminase ADAR, leading to nuclear retention of the edited transcript ((Möller-Krull, *et al*. 2008). It has been proposed that SVA may act on gene expression as a mobile CpG island (Bennett, *et al*. 2004; Wang *et al*. 2005). The vast CpG-rich VNTR region of SVA has been shown to be methylated (Strichman-Almashanu, *et al*. 2001), and the insertion of these elements near promoter or regulatory regions may inhibit transcription.

It is logical to assume that some of the effects that retrotransposons have on their host genomes have, over time, been exapted into host functions. The constant

broadcasting of new splice sites, polyadenylation signals, promoters and gene fragments into a functioning genome creates a rich environment for experimentation with novel gene structures, expression patterns or transcripts. Several instances have been found of biologically important genes and regulatory elements that are derived from retrotransposons or retrotransposition events. *BC200*, a small neuronal regulatory RNA found in primates, originated from Alu sequence (Volff and Brosius 2007). A fusion gene conferring HIV-1 resistance in owl monkeys was created by the retrotransposition of one gene into the intron of another (Sayah, *et al*. 2004). L1s have been implicated in assisting in X-inactivation in placental mammals and repairing double stranded breaks in DNA (Morrish, *et al*. 2002; Bailey, *et al*. 2000). Many other retrotransposition events are likely to be involved with cellular functions, perhaps some currently masked by redundancy. Time and future opportunities for selection should bring more of these functions to light.

Retrotransposons and disease

Given the above mechanisms of altering genomic structure and expression, the role retrotransposons play in human disease should not be surprising. The first *de novo* retrotransposon insertion documented was an L1 into an exon of the factor VIII gene causing hemophilia A (Kazazian, *et al*. 1988). Since then several dozen disease-causing retrotransposon insertions have been identified (Table 1-1), and retrotransposon-mediated deletions or recombinations have been implicated in several cancers and other disorders (reviewed in Callinan and Batzer 2006). As personal genome sequences become more

Retrotransposon Family	Subfamily	Differentially present in humans and chimps	Dimorphic among humans	Disease-causing	Active in cell culture
Alu	Sc	31	5	None Found	NT
	Sg	56	8	None Found	NT
	Sp	25	5	None Found	NT
	Sq	46	3	1	NT
	Sx	46	7	None Found	NT
	Sz	58	1	1	NT
	Y	475	66	None Found	NT
	Ya1	67	12	1	NT
	Ya4	170	54	None Found	NT
	Ya5	1676	587	11	Yes
	Ya5a2	38	9	None Found	NT
	Ya8	36	9	None Found	NT
	Yb3a1	17	10	1	NT
	Yb3a2	87	8	None Found	NT
	Yb8	1290	409	4	NT
	Yb9	137	24	4	NT
	Yc1	356	113	4	NT
	Yc2	68	13	None Found	NT
	Yd2	35	5	None Found	NT
	Yd3	40	3	None Found	NT
	Yd8	102	12	None Found	NT
	Ye2	31	2	None Found	NT
	Ye5	144	35	None Found	NT
	Yf1	19	4	None Found	NT
	Yg6	261	42	None Found	NT
	Yh9	10	4	None Found	NT
	Yi6	116	17	None Found	NT
	Yj	10	6	None Found	NT
LINE	L1-PA2	490	21	1	No
	Pre-Ta	252	4	1	Yes
	Ta*	270	101	9	Yes
	Ta-0	43	2	1	Yes
	Ta-1d	91	7	3	Yes
	Ta-1nd	20	2	None Found	Yes
SVA	A	No	No	None Found	NT
	B	5	No	None Found	NT
	C	15	No	None Found	NT
	D	259	5	None Found	NT
	E	55	32	3	NT
	F	23	26	1	NT

Table 1-1. Summary of recently mobilized non-LTR retrotransposons. Recently mobilized retrotransposons, by subfamily, as determined by comparative genomic alignment between humans and chimps (Differentially present in humans and chimps) and between individual humans (Dimorphic among humans). Disease-causing insertions and subfamily activity verified by cell culture assays are also shown. NT, not tested. *These L1 Ta elements are 5' truncated and lack the necessary 5' diagnostic sequences to classify them further into Ta subfamilies.

commonplace, the scope of retrotransposon-caused pathologies in humans is likely to extend beyond these initial cases. Active human retrotransposons have been estimated to generate about one new insertion per 10-100 live births (Cordaux, *et al*. 2006; Kazazian 1999; Li *et al*., 2001). These 'Private' *de novo* insertions occur only once in the human population (in just a single individual) and are expected to represent a particularly abundant class of insertions. Thus, in addition to the 1.6 million fixed copies, the human genome collectively receives an average of one new insertion every 6-60 nucleotides. This amounts to a huge human mutagenesis experiment generating a substantial degree of genetic variation. The full impact of these private insertions on human diversity and disease is only just beginning to be studied, and these insertions are likely to influence a range of human phenotypes. Therefore, it is crucial to determine which endogenous human retrotransposons remain active and continue to produce new insertions.

Strategies for identifying recently active retrotransposons

The moment a new retrotransposition event occurs both the source element and its copy should be 100% identical in sequence (barring any errors that may be introduced during retrotransposition due to the low fidelity of L1 ORF2p reverse transcriptase). Due to mutation over time, both copies diverge independently of each other. This random and equal divergence, given molecular context from the consensus sequence of the founding element, allows a rough dating of the age of each retrotransposition event. The degree by which various subfamilies of retrotransposons diverge from the consensus of that subfamily is indicative of the time-frame in which that subfamily was active. Building the consensus for a given subfamily helps determine the sequence of the active element that

initially transposed. Therefore subfamilies whose members diverge little from their consensus sequences are assumed to be younger and more recently active than subfamilies whose members are more divergent from their subfamily consensus sequence. Given that a consensus sequence has produced offspring elements, and thus has proven activity, younger elements more similar or identical to their consensus sequences are more likely to retain the sequence characteristics required for activity.

Another approach to identify recently active retrotransposon elements is to look for actual incidents of recent insertions. Several *de novo* retrotranspositions that disrupt functional integrity and cause pathologies have been documented (Table 1-1), but the majority of new insertions appear to be neutral, without an obvious phenotype, and are thus more difficult to detect. Comparative genomics, or pair-wise alignment studies between individuals has proven to be a useful tool in discovering these events (Bennett, *et al*. 2004; Wang, *et al*. 2006; Konkel, *et al*. 2007; Akagi, *et al*. 2008). In this analysis, a newly inserted retrotransposon appears as a gap in the alignment when compared to individuals without the insertion (care must be taken that the DNA sequence of these gaps corresponds to *de novo* retrotransposition events, and not deletions or insertions of genomic DNA by other mechanisms, by finding evidence of L1-driven TPRT retrotransposition events such as the presence of a precise TSD). Likewise, retrotransposition events that are fixed within a given population or species, can be identified by genomic comparisons with closely related populations or species that diverged prior to those retrotransposition events. These direct genomic comparisons between individuals or species can also be performed by polymerase chain reaction (PCR) assays. With PCR assays, differentially present retrotransposon insertions appear

as different sized bands on a gel. As with pair-wise alignments, DNA size alone is not a definitive indication of retrotransposon presence and sequence must be obtained and analyzed to determine actual retrotransposition events.

The most conclusive evidence of an active retrotransposon is to test retrotransposition in living cells. Moran *et al*. (1996) reported a cell culture assay to demonstrate that L1 could actively transpose in human cells and required both L1 ORF1p and L1 ORF2p to do so. This study used two L1 sequences retrieved from disease-causing insertions into the factor VIII and dystrophin genes. Later work by Brouha *et al*. (2003) expanded this finding by assaying 82 L1s with intact open reading frames in cell culture. They showed that, not only did very few of the human L1s retain activity, but that each active element had varying sequence-related degrees of activity. In fact, only six 'hot' L1s account for 84% of all L1 retrotransposition in humans, and all of these elements belong to the younger pre-Ta or Ta L1 subfamilies. Dewannieaux *et al*. (2003) adapted this retrotransposition assay to demonstrate the retrotransposition activity of a disease causing Alu insertion in cell culture, and determined the requirement of L1 ORF2p, but not L1 ORF1p for Alu activity.

Scope of the Dissertation

The goals of this dissertation were to identify which retrotransposon families are currently active in humans and the extent of that activity. We also planned to identify which retrotransposon families have been active in recent evolutionary history in humans and in their nearest relatives, chimpanzees. Once these families were identified, our aim was then to learn about the mechanisms of Alu retrotransposition and begin to predict

which of the 900,000 full length Alu copies in the human genome are likely to be active. Our approach was to define the specific sequence requirements of active Alu retrotransposons through sequence conservation among mobile copies and direct verification of activity in cell culture assays.

This study of active retrotransposons arose from a broader study of human genetic variation. By mapping DNA traces to the human genome reference sequence (Chapter 2), we searched for variation due to Single Nucleotide Polymorphisms (SNPs) and insertions/deletions (indels), including dimorphic transposons. A bioinformatics pipeline was developed to specifically recognize transposon insertions, and was applied to the indels generated in the initial study. The resulting dimorphic transposon copies in human populations were identified and validated (Chapter 3). To identify species-specific transposon insertions during the last ~6 million years in humans and chimps, human and chimp genomes were aligned and transposition events were identified from the resulting insertions and deletions (Chapter 4). From our comparative genomic studies (Chapter 3 and 4, and summed up in Table 1-1) we found evidence of the ongoing retrotransposition activity of a large number of Alu elements in the past 6 million years and continuing into modern times. Since little was known about the activity of Alu elements, which make up the greatest proportion of recently active elements in both humans and chimps, we used conservation studies, retrotransposition assays of cloned and constructed elements and biochemical binding assays to estimate how many Alus are still retrotranspositionally competent in humans (Chapter 5). Several extinct Alu subfamilies were also reconstructed, and their activity in human cells was compared with the evidence of their ancestral behavior. In addition, specific sequence changes that effect retrotransposition

activity were engineered in order to better understand the effects of selection on Alu retrotransposition in humans. Together, this work contributes to our knowledge of the evolutionary history, current status and future activity of the three active retrotransposons in humans, L1, Alu and SVA.

Chapter 2

Single Nucleotide Polymorphisms (SNPs) that map to gaps in the human SNP Map

Circe Tsui[1,2], Laura E. Coleman[1], Jacqulyn L. Griffith[1], E. Andrew Bennett[3], Summer G. Goodson[1], Jason D. Scott[4], W. Stephen Pittard[2,5] and Scott E. Devine[1,2,3,*]

[1]Department of Biochemistry, [2]Center for Bioinformatics, [3]Genetics and Molecular Biology Graduate Program, [4]DNA Sequencing Core Facility and [5]Bimcore, Emory University School of Medicine, Atlanta, GA. The authors wish it to be known that, in their opinion, the first two authors should be considered as joint First Authors.

2003. ***Nucleic Acids Research*****, 31, 4910-4916**

 DOI: 10.1093/nar/gkg664

Introduction

An international effort is underway to generate a comprehensive haplotype map (HapMap) of the human genome represented by an estimated 300,000 to 1 million `tag' single nucleotide polymorphisms (SNPs). Our analysis indicates that the current human SNP map is not sufficiently dense to support the HapMap project. For example, 24.6% of the genome currently lacks SNPs at the minimal density and spacing that would be required to construct even a conservative tag SNP map containing 300,000 SNPs. In an effort to improve the human SNP map, we identified 140,696 additional SNP candidates using a new bioinformatics pipeline. Over 51,000 of these SNPs mapped to the largest gaps in the human SNP map, leading to significant improvements in these regions. Our SNPs will be immediately useful for the HapMap project, and will allow for the inclusion of many additional genomic intervals in the final HapMap. Nevertheless, our results also indicate that additional SNP discovery projects will be required both to define the haplotype architecture of the human genome and to construct comprehensive tag SNP maps that will be useful for genetic linkage studies in humans.

With the nearing completion of the human genome sequence, a number of studies have been initiated to identify natural genetic variation in human populations (Altshuler, *et al*. 2000; The International SNP Map Working Group 2001). Several large scale projects have been conducted to identify single nucleotide polymorphisms (SNPs), including studies involving specific genes (Nickerson, *et al*. 1998; Rieder, *et al*. 1999; Taillon-Miller, *et al*. 1999; Taillon-Miller, *et al*. 2000), chromosomes (Mullikin, *et al*. 2000; Dawson, *et al*. 2001) and the whole genome (Altshuler, *et al*. 2000; The

International SNP Map Working Group 2001). Two basic experimental strategies have been used to identify SNPs on a genome-wide scale. In the first approach, DNA from 24 diverse humans was pooled together and shotgun sequenced. The traces generated were compared with each other, or to finished genomic sequence, to identify SNPs. In the second approach, base substitutions were identified within the overlap regions of adjacent bacterial artificial chromosomes (BACs) that were used to sequence the human genome. Using a combination of these two methods, The SNP Consortium (TSC) developed an initial map of human genome variation containing 1.4 million SNPs (The International SNP Map Working Group 2001). The human SNP map now has grown to ~2.2 million nonredundant polymorphisms due to contributions from a number of laboratories (www.ncbi.nlm.nih.gov/SNP).

In addition to serving as a repository for human genetic variation, this collection of human SNPs will be useful for genetic linkage studies in humans. In fact, an international effort currently is underway to develop a comprehensive haplotype map (HapMap) of the human genome using these SNPs (Daly, *et al*. 2001; Reich, *et al*. 2001; Gabriel, *et al*. 2002). The HapMap will greatly facilitate genetic linkage studies in humans by providing a genome-wide map of SNPs that are commonly inherited together as `haplotype blocks'. Tag SNPs representing these haplotype blocks then will be used to map traits to specific genomic intervals.

Haplotype architecture varies greatly across the human genome, with haplotype blocks ranging in size from <1 kb in length to >100 kb (9-14). Thus, it will be necessary

to have very high SNP densities in some regions of the genome, but lower densities in others, in order to identify all of the common haplotype blocks in humans. Since we cannot determine in advance which regions of the genome will require high or low densities, it will be necessary to have uniformly high distributions across the genome at the onset of the project in order to identify all of the haplotype blocks. Conservative estimates indicate that 300,000 tag SNPs will be necessary to represent all of the common haplotype blocks in the human genome, corresponding to an average density of one tag SNP per 10 kb. Higher estimates indicate that as many as 1 million tag SNPs (or one tag SNP per 3 kb) may be required (Gabriel, *et al*. 2002; Patil, *et al*. 2001; Stephens, *et al*. 2001; Judson, *et al*. 2002). These tag SNPs will be selected from a larger set of SNPs that define the haplotype architecture of the genome.

Although the current human SNP map (build 110) contains 2.2 million non-redundant SNPs, it is presently unclear as to whether the density and spacing of these SNPs across the human genome is sufficient to define the haplotype architecture of the genome. It is also unclear whether a sufficient number of tag SNPs could be selected from this collection to adequately represent all of the underlying haplotypes. We have studied the distribution of SNPs in the human SNP map (build 110), and have determined that there are 38,497 inter-SNP intervals that are >10 kb in length and 221,511 inter-SNP intervals that are >3 kb in length. Since even the most minimal tag SNP map will require an average spacing of one SNP per 10 kb, we conclude that the current SNP map is not sufficiently dense to support the HapMap project. In an effort to generate higher SNP densities in the largest gaps of the current SNP map, we conducted a new SNP discovery

project, which has led to significant improvements in these regions. Nevertheless, our results indicate that additional SNP discovery projects will be required both to define the haplotype architecture of the human genome and to construct comprehensive tag SNP maps that will be useful for genetic linkage studies in humans.

Materials and Methods

Computational pipeline for SNP identification

The 7.1 million trace files generated by TSC were obtained from Cold Spring Harbor Laboratory (The International SNP Map Working Group 2001). The traces we retrimmed using the VecScreen system developed by the National Center for Biotechnology Information (NCBI) together with a custom Perl trimming program that used the Phred quality scores (Ewing, *et al*. 1998) to trim traces upon encountering five bases in a row with quality scores below 25. The single longest, high quality interval from each trace then was selected for further analysis and all other data were set aside. We required that the average Phred score for the trimmed trace exceed a minimum of 25. The minimum trace length after trimming was 100 bases (due to our imposed limit), and the maximum was 905 bases, with an average of 346 bases. After the trimming step, the number of useful traces was reduced from 7,120,020 to 4,293,807 (60.3% of the original traces).

These trimmed traces were masked for known repeats using RepeatMasker and MaskerAid (Bedell, *et al*. 2000). The single longest unmasked interval within the trace then was used to map the trace with Mega BLAST (NCBI) and the Golden Path (June 2002 release) (Kent, *et al*. 2002). We required a minimum of a 50 base match at 100%

identity to a single location in the genome for a successful trace mapping. Traces that matched to more than one location, or that lacked at least a 50 base match at 100%identity, were eliminated from the analysis due to potential segmental duplications (Bailey, *et al*. 2002). Using this approach, 2,759,010traces were successfully mapped to unique genomic locations (64.3% of the trimmed traces, equivalent to 980,655,789 bases, or ~33% of the genome). The traces then were unmasked and aligned with their assigned genomic locations by pair-wise alignment, allowing for insertions and deletions of up to 16 bases in length using the program BL2Seq (NCBI).Sixty-five percent of the traces were completely identical to their assigned genomic locations within the Golden Path after unmasking and alignment, whereas another 35% of the traces contained sequence variations. Ninety-nine percent of these `variant' traces had <10% variation from their assigned genomic locations. We have found that this collection of variant traces is an excellent resource for identifying a variety of DNA sequence polymorphisms, including SNPs, insertion/ deletion polymorphisms and transposon polymorphisms (Figure 2-2 and unpublished data).

A total of 709,492 SNPs were identified from the 7.1 million traces analyzed with this pipeline. We used a neighborhood quality assessment to call the final base differences. The SNP itself was required to have a minimum Phred score of 25, and all five bases on each side flanking the SNP were required to have minimum quality scores of 20. Of the 709,492 SNPs identified in our analysis, 568,796 (80.2%) were present in dbSNP (build 110) and, therefore, had been identified previously (The International SNP Map Working Group 2001). An additional 140,696 SNPs were identified only by our

analysis. These new SNPs were deposited into build 111 of dbSNP under accession numbers ss7844264 through ss7984919. TSC identified ~1 million SNPs using the same 7.1 million traces (compared with 709,492 here). We did not expect to find all of the TSC SNPs because we aggressively trimmed the traces and discarded over half of the sequence information from those traces.

The distribution of our SNPs also was examined relative to genes in the human genome using the classification system established by dbSNP (www.ncbi.nlm.nih.gov/SNP). Build113 of dbSNP, which included our SNPs, was used for this analysis. Overall, 70,046 of our 140,696 SNPs (49.7%) fell within `locus regions' according to the dbSNP definition (i.e. fell within 3 kb of a gene or predicted gene in the upstream direction, or within 500 bp of a gene or predicted gene in the downstream direction). In comparison, 45.2% of all SNPs in the same build of dbSNP (113) fell within such regions. Our 70,046 `locus SNPs' could be broken down further according to the locations of these SNPs within the genes. A total of 29,316 of our SNPs (or 40.5% of our locus SNPs) were located within 3 kb (upstream) or 500 bp (downstream) of a gene but were not located within the transcribed portion of the gene. In comparison, 36.9% of all locus SNPs in build 113 of dbSNP were found within this category. A total of 343 of our locus SNPs (0.47%) were predicted to cause synonymous changes within coding regions (compared with 1.0% of all locus SNPs in build 113 of dbSNP). A total of 513 of our locus SNPs (0.71%) were predicted to cause non-synonymous changes in coding regions (compared with 1.2% for all locus SNPs in build 113 of dbSNP). A total of 6676 (9.3%) of our locus SNPs were found within predicted untranslated regions (compared

with 16.5% of all locus SNPs in build 113 of dbSNP). A total of 34,440 of our locus SNPs (47.7%) were located within predicted introns (compared with 47.7% for all locus SNPs in build 113 of dbSNP). Finally, a total of 13 of our locus SNPs (0.02%) fell within splice sites (compared with 0.02% for all locus SNPs within build 113 of dbSNP). Thus, the overall distribution of our SNPs relative to genes and other genomic features was remarkably similar to the distribution of all SNPs in the same build of dbSNP.

Table 2-1. Distribution of SNP intervals (intervals between adjacent SNPs) based on dbSNP (build 110)

Class	*Interval size (Kb)*	*No. of intervals*	*Total length (bp)*	*%Genome covered by this class*
0	≤ 0.1	540,376	22,982,827	0.76%
1	> 0.1 and ≤ 0.5	728,703	187,697,440	6.17%
2	> 0.5 and ≤ 1	330,788	237,817,929	7.82%
3	> 1 and ≤ 3	399,275	689,502,930	22.67%
4	> 3 and ≤ 5	105,666	405,154,007	13.32%
5	> 5 and ≤ 10	77,348	533,016,204	17.52%
6	> 10 and ≤ 50	37,046	618,629,083	20.34%
7	> 50 and ≤ 500	1,451	113,191,946	3.72%
	Total	2,220,653	2,807,992,366	92.32%

Analysis of the SNP map

We examined the 2.2 million SNPs of dbSNP (build 110) that could be mapped to unique locations within the Golden Path (June 2002 release) (Kent, *et al*. 2002). The SNP TSC (random read) and SNP NIH (overlap) tables were downloaded from the UC Santa Cruz website (www.genome.ucsc.edu) and combined into a single table. These tables then were merged with a third table containing the locations of all gaps in the genome (also obtained from the UC Santa Cruz website). A custom Perl program then was used to

measure the intervals between all adjacent SNPs, excluding intervals caused by genomic gaps. The results were deposited into an Oracle database to assign intervals to size classes and then count each class (Table 2-1).Similar results were obtained using equivalent SNP tables obtained from dbSNP (www.ncbi.nlm.nih.gov/SNP/). Finally, this process was repeated after adding our new SNPs to assess the impact of our SNPs on the intervals (Table 2-2).

Analysis of SNPs by PCR and DNA sequencing

Primer pairs were designed to amplify each SNP using the flanking DNA sequences. In each case, primers were designed to have melting temperatures above 64°C. The M13 Reverse (M13R) primer sequence (5'-CAGGAAACAGCTATGACC-3') was added to the 5'-end of the downstream primer such that this sequence was incorporated into the PCR product to generate a recombinant template for sequencing. PCR products were amplified from the first 12 human DNAs from the Coriell diversity panel (Collins, *et al*. 1999), and these 12 PCR products were sequenced using M13R BigDye Primer kits from Applied Biosystems (version 1.0 and version 3.0). The PCR products were diluted in water 1:4 before sequencing, and 1.0 ml of DNA was used as a template. Upon completion of the cycling protocol recommended by Applied Biosystems, the four dye primer sequencing reactions were pooled together and precipitated with ethanol. The dried precipitates were resuspended in formamide and run on an ABI 3100 GeneticAnalyzer. If the SNP was not verified in the first 12 samples, then the remaining 12 samples were sequenced.

Table 2-2. Distribution of 10-500 Kb SNP intervals (between adjacent SNPs) by chromosome

	Current distribution		*With our newly discovered SNPs*		
Chr	*Number*	*Total Length (bp)*	*Number*	*Total Length (bp)*	*% Change*
SNP Intervals of length > 50Kb and ≤ 500Kb					
1	78	5,585,535	12	844,883	-84.87%
2	106	7,505,978	15	1,125,383	-85.01%
3	76	5,537,160	13	1,161,678	-79.02%
4	146	11,323,424	16	1,217,843	-89.24%
5	115	8,779,105	17	1,272,405	-85.51%
6	110	9,053,122	13	911,330	-89.93%
7	63	4,910,285	12	1,343,660	-72.64%
8	136	10,621,236	31	2,306,466	-78.28%
9	30	2,223,391	10	900,842	-59.48%
10	37	2,606,920	8	632,838	-75.72%
11	42	3,025,765	9	706,986	-76.63%
12	48	3,804,568	11	979,539	-74.25%
13	22	1,601,924	3	186,953	-88.33%
14	8	516,358	1	50,299	-90.26%
15	27	1,866,030	7	467,171	-74.96%
16	39	3,143,940	13	1,067,630	-66.04%
17	25	1,634,730	8	494,328	-69.76%
18	30	2,421,414	4	375,489	-84.49%
19	9	743,099	8	665,629	-10.43%
20	0	0	0	0	0
21	0	0	0	0	0
22	0	0	0	0	0
X	202	15,497,366	111	7,670,345	-50.51%
Y	102	10,790,596	100	10,457,071	-3.09%
Total	1451	113,191,946	422	34,838,768	-69.22%
SNP Intervals of length > 10Kb and ≤ 50Kb					
1	2,477	40,672,204	1,676	25,252,780	-37.91%
2	3,453	56,757,325	2,514	37,698,436	-33.58%
3	2,519	42,062,241	1,692	25,473,302	-39.44%
4	2,822	47,874,122	1,828	27,109,984	-43.37%
5	1,886	32,513,067	1,162	17,455,100	-46.31%
6	2,008	33,560,739	1,374	20,408,992	-39.19%
7	2,041	32,910,970	1,552	23,218,090	-29.45%
8	1,835	32,010,697	1,093	16,755,825	-47.66%
9	1,569	25,556,884	1,164	18,030,068	-29.45%
10	1,688	27,772,540	1,158	17,701,341	-36.26%
11	1,332	21,844,811	871	13,007,496	-40.45%
12	1,634	26,718,713	1,233	18,639,999	-30.24%
13	1,152	18,687,944	783	11,638,775	-37.72%
14	1,148	17,969,446	950	14,359,234	-20.09%
15	1,079	17,676,577	727	10,902,358	-38.32%
16	1,018	17,378,709	683	10,903,977	-37.26%
17	1,064	16,989,104	765	11,554,814	-31.99%
18	696	11,233,681	416	5,808,116	-48.30%
19	807	12,880,609	667	10,406,368	-19.21%
20	766	11,268,333	653	9,430,374	-16.31%
21	133	1,767,854	118	1,571,726	-11.09%
22	332	5,040,570	276	4,201,143	-16.65%
X	3,216	60,143,734	2,758	49,583,696	-17.56%
Y	371	7,338,209	367	7,196,772	-1.93%
Total	37,046	618,629,083	26,480	408,308,766	-34.00%

Results

In order to examine the distribution and spacing of human SNPs across the genome, we measured the distances between all adjacent SNPs in the human SNP map (build 110), and then classified the observed intervals by size (Table 2-1). We found a wide degree of variation in the distances between adjacent SNPs, ranging from 1 to 427,780 bases. A large number of intervals were greater than the 10 kb average spacing that will be necessary to construct even a minimal tag SNP map containing 300,000 SNPs (Table 2-1). For example, 1451 SNP intervals were 50-500 kb in length, and 37,046 SNP intervals were 10-50 kb in length (Table 2-1). In fact, we found a total of 38,497 intervals (occupying 24.6% of the human genome) that currently lack SNPs at the minimal 10 kb average spacing required for a tag SNP map of 300,000 SNPs. Moreover, 221,511 intervals (occupying 54.9% of the genome) currently fall below a density of one SNP per 3 kb (the density required for a tag SNP map of 1 million SNPs; Table 2-1). Therefore, between 24.6 and 54.9% of the human genome currently lacks SNPs at the minimal densities that are estimated to be required to construct tag SNP maps containing 300,000 and 1 million SNPs, respectively.

In an effort to generate higher SNP densities in the largest gaps of the current SNP map, we re-mined the original 7million traces that were generated by The SNP Consortium(TSC) using a new computational pipeline (see Materials and Methods), and identified 140,696 additional SNP candidates that had not been identified previously. These SNP candidates were distributed on all 24 chromosomes (Figure 1A), and also fell within genes at a frequency that was very similar to the frequency that all SNPs in dbSNP

fell within genes (Figure 1A and Materials and Methods). A significant fraction of these SNPs were located within the largest SNP gaps of the genome (Figure 1B). For example, 14,253 of our SNP candidates mapped to the 1451 largest SNP `deserts' of the human genome (ranging from 50 to 500 kb in length; Figure 1B, class 7). As a result of adding our SNP candidates to the map, 1029 (69.2%)of these largest gaps were reduced to intervals that were smaller than 50 kb in length, and this group now has an average size that is <10 kb (Figure 1 and Table 2-2). In fact, 80-90% of the largest gaps were eliminated on chromosomes 1, 2, 4, 5, 6, 13, 14 and 18 (Table 2-2). Another 37,517 of our SNP candidates (20.3%) mapped to the second-largest class of intervals (10-50 kb in length; Figure 1B, class 6). The addition of these SNPs to the map resulted in a ~48% reduction of the 10-50 kb intervals on chromosomes 8 and 18, and led to a 34%genome-wide reduction in 10-50 kb intervals overall (Table 2-2). Therefore, over 51,000 of our SNPs mapped to the two largest interval groups (50-500 and 10-50 kb), and yielded significant reductions in the amount of DNA contained within these groups.

It is noteworthy that chromosomes 20, 21 and 22 lacked 50-500 kb gaps altogether, presumably due to SNP discovery projects that were focused on these chromosomes (Mullikin, *et al*. 2000; Dawson, *et al*. 2001; Patil, *et al*. 2001). In contrast, both the X and Y chromosomes had many more gaps than the average chromosome, and also had a significant number of gaps that were refractory to closure by our SNPs (Table 2-2). This also was the trend with SNPs identified by TSC (The International SNP Map Working Group 2001) and others (Gabriel, *et al*. 2002), presumably due to the smaller effective population sizes and different mutation rates of these chromosomes. Natural

selection (both positive and negative), and the lower recombination rates of these chromosomes, also may have contributed to the creation of SNP deserts on these chromosomes.

Although the two largest interval groups discussed above (50-500 and 10-50 kb) are the most important with respect to closing gaps in the SNP map, we also identified SNPs that will be useful for identifying smaller haplotype blocks in the genome (<10 kb). Indeed, up to half of the haplotype blocks identified in previous studies were found to be <3 kb in length, and a large number of blocks were in the 3-10 kb range as well (Patil, *et al*. 2001). Many of our SNPs mapped to these smaller intervals. For example, 18,760 of our SNPs mapped to 5-10 kb intervals (Figure 1B, class 5), and another 15,197 mapped to 3-5 kb intervals (Figure 1B, class 4). A total of 85,757 (61.0%) of our 140,696 SNP candidates mapped to intervals that were >3 kb in length and thus are likely to be immediately useful for the construction of the HapMap. The remaining SNP candidates (54,969 or 39.1%) mapped to genomic regions containing SNP densities more than one SNP per 3 kb. Although these more densely populated regions already contain a large number of SNPs, our SNPs may be useful in cases where other SNPs in the region have low allelic frequencies.

In order to measure the accuracy of our SNP predictions, thirty SNP candidates were chosen randomly from the collection of 140,696 and examined further. If our predictions were accurate for a given SNP, then we expected to be able to identify that SNP in at least one of the original 24 people that were used to generate the 7.1 million

TSC traces (The International SNP Map Working Group 2001). If, on the other hand, the SNP was not confirmed in at least one of the 24individuals, then we would know with certainty that our bioinformatics prediction was incorrect. To perform these validation studies, each SNP candidate was amplified by PCR in 12-24 of the original 24 humans and individually sequenced. Twenty-nine of the thirty SNP candidates examined were confirmed by this analysis for a confirmation rate of 96.7% (Figure 2-2 and Table 2-3). This is comparable with the overall verification rate of 95% obtained by TSC (The International SNP Map Working Group 2001), indicating that our pipeline is highly accurate.

A range of allelic frequencies was observed among the SNPs sequenced in our validation study (Table 2-3).Interestingly, four of the SNPs in the study were present in all individuals examined except for the person(s) represented by the Golden Path sequence (Table 2-3). The most likely explanation for these results is that the person(s) represented by the Golden Path sequence had rare `private' alleles at those positions that were absent from the majority of humans. Overall, 83% of the SNPs examined in our validation study had minor allelic frequencies that were >5% (Table 2-3). This fraction is similar to estimates obtained previously by TSC and others (The International SNP Map Working Group 2001), indicating that our SNPs will be equally useful for mapping studies.

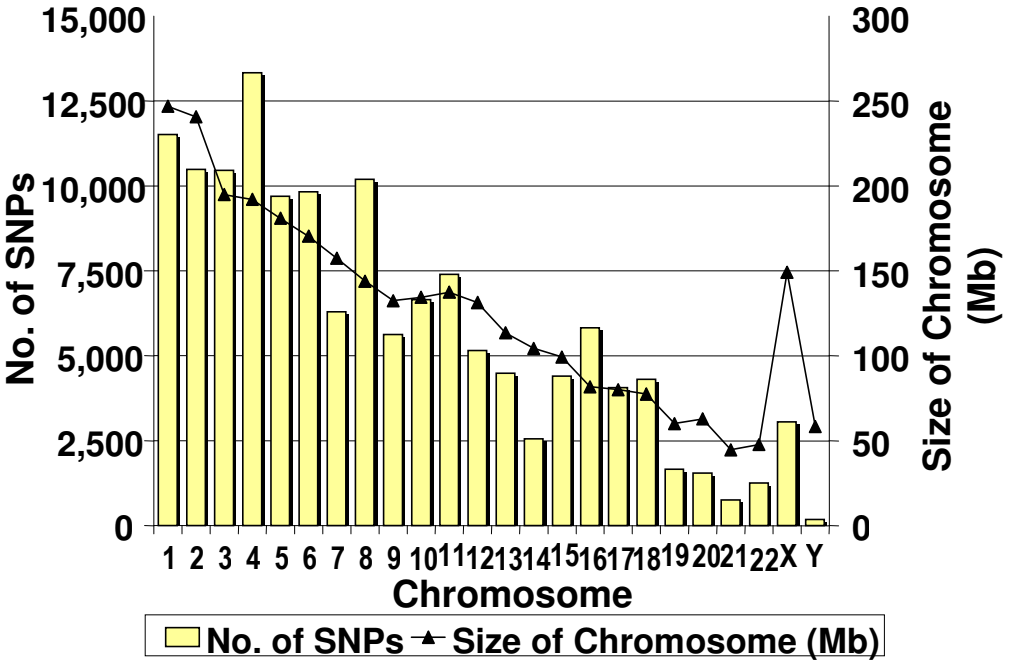
Discussion

Since the HapMap project is rapidly ramping up to meet the goal of completion within 3 years, it is desirable at this stage to identify all of the SNPs that will be necessary to meet

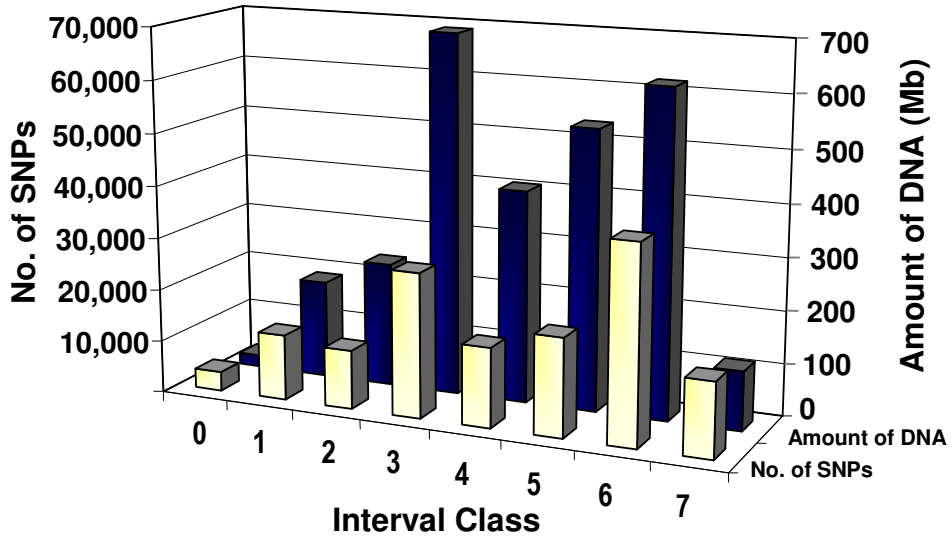
the demands of the project. Our SNP discovery pipeline, which has several notable differences from previous pipelines (see below), has allowed us to discover a novel collection of human SNPs that will be immediately useful for the HapMap project. The addition of our SNP candidates to the current SNP map has improved the most problematic areas of the map, and has been particularly useful in the 24% of the genome that (prior to our study) contained SNP intervals of 10-500 kb in length. We have contributed over 51,000 SNPs to these largest intervals and, as a consequence, have reduced the amount of DNA contained in these gaps by 34 (10-50 kb gaps) to 69%

Figure 2-1. (following page) Analysis of 140,696 new SNP candidates. (A) The bar graph depicts the genomic distribution of our 140,696 SNP candidates by chromosome. Note that the SNPs are distributed on all 24 chromosomes proportional to the amount of DNA (indicated by a line above the bar graph). The possible exceptions are the X and Y chromosomes, which have fewer SNPs than average per kb of DNA. The distribution of our SNPs also was examined relative to genes (see Materials and Methods). Overall, 49.7% of our 140 696 SNPs fell within or near genes (defined as being within 3 kb of a gene or predicted gene in the upstream direction, or within 500 bp of agene or predicted gene in the downstream direction). In comparison, 45.2%of all SNPs in the equivalent build of dbSNP (build 113) fell within such regions using the same criteria. Thus, the overall distribution of our SNPs relative to genes was highly similar to the overall distribution of all SNPs in the same build of dbSNP. (B) The row of white bars (front) shows the number of SNPs from our collection in interval classes 0 through to 7 (defined in Table 2-1). The row of black bars (back) depicts the amount of DNA (in Mb) contained within each class before adding our SNPs to the map (listed in Table 2-1 under `Total length' column). Note that our SNPs occur in all classes, and are generally proportional to the amount of DNA in each class. The 51,770 SNPs in classes 6 and 7 close many of the largest gaps in the SNP map, and the 85,757 SNPs in classes 4-7 are likely to be immediately useful for construction of the HapMap. The SNPs in classes 0-3 may be useful in cases where the allelic frequencies of existing SNPs are unfavorable.

A.


15,000
12,500
10,000
7,500
5,000
2,500
0
No. of SNPs
300
250
200
150
100
50
0
Size of Chromosome (Mb)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Chromosome
No. of SNPs
Size of Chromosome (Mb)

B.


70,000
60,000
50,000
40,000
30,000
20,000
10,000
No. of SNPs
700
600
500
400
300
200
100
0
Amount of DNA (Mb)
0 1 2 3 4 5 6 7
Interval Class
Amount of DNA
No. of SNPs

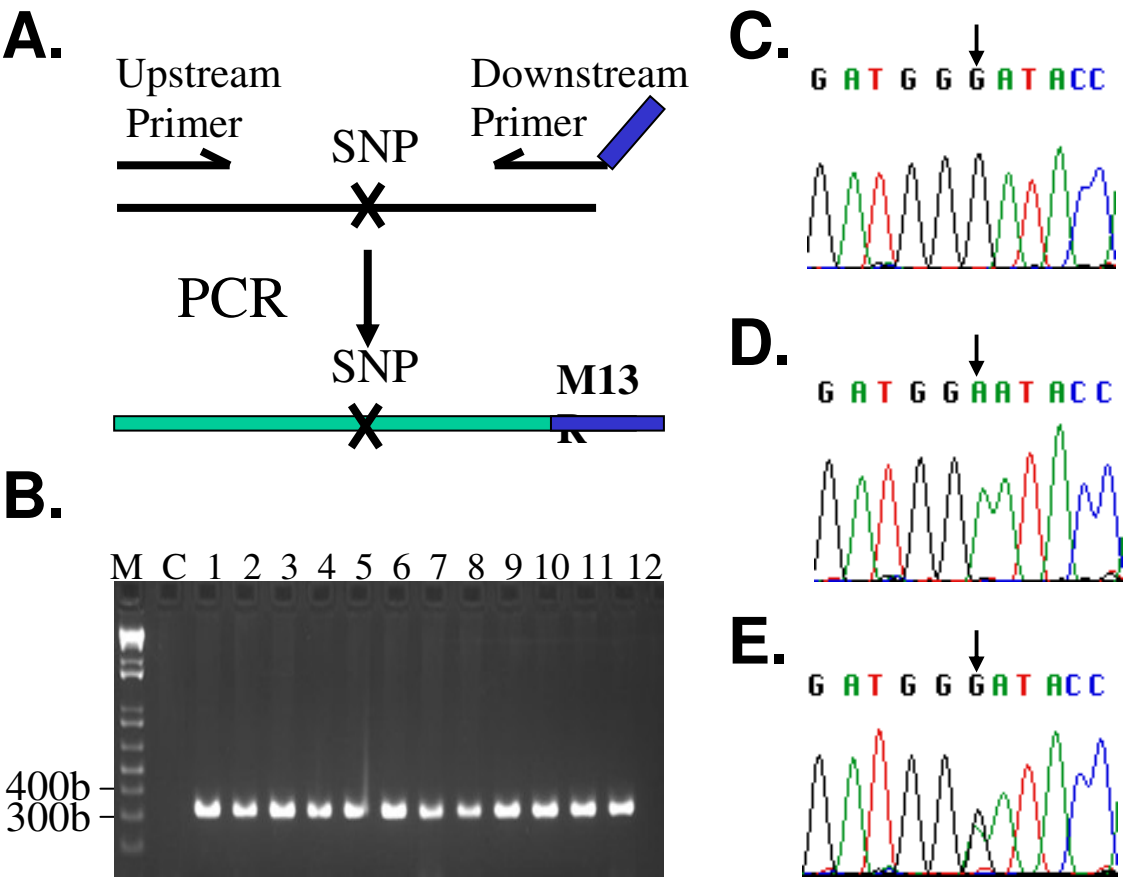


Figure 2-2. Strategy for SNP validation. (**A**) The diagram shows the strategy used to amplify each SNP candidate by PCR. A recombinant PCR product is generated by introducing the M13R sequence into the downstream primer.(**B**) A gel showing PCR products generated for a typical SNP (SNP 9 is shown, generating a PCR product of 319 bp in length). Note the robust PCR products in lanes 1-12, and the absence of a band in the negative control lane (C). A 1 kb ladder marker is shown on the left (M). (C-E) Examples of chromatograms from SNP 9 (a `C' is present at position 67,671,151 of chromosome 8 according to our predictions, whereas a `T' is present in the Golden Path. The sequences shown are of the complementary strand. Arrows show the single nucleotide of interest. (**C**) Homozygous for our prediction based on the TSC trace sequence (human sample 3). (**D**) Homozygous for the Golden Path (human sample 6). (**E**) Heterozygous for our prediction and the Golden Path sequence (human sample 5).

Table 2-3. SNP verification by PCR and sequencing

SNP	*Genomic location*	*Trace base*	*Golden Path base*	*Alleles sequenced*	*% Trace*	*% Golden Path*
1	chr8: 138543564	t	G	22	95.0	5.0
2	chr11: 2786593	g	A	24	83.0	17.0
3	chr4: 173806221	c	G	24	37.5	62.5
4	chr4: 140987424	a	G	18	94.0	6.0
5	chr8: 117177669	t	C	22	77.0	23.0
6	chr18: 39493166	g	T	44	100.0	0.0
7	chr6: 85928202	a	G	22	9.0	91.0
8	chr3: 51223303	t	C	22	91.0	9.0
9	chr8: 67671151	c	T	20	50.0	50.0
10	chr8: 27015545	t	C	24	67.0	33.0
11	chr4: 52910312	a	G	24	75.0	25.0
12	chr4: 44538473	c	T	42	100.0	0.0
13	chr5: 177158144	a	G	24	21.0	79.0
14	chr15: 84691055	c	T	22	23.0	77.0
15	chr3: 186142005	g	A	22	59.0	41.0
16	chr3: 99873851	g	A	22	32.0	68.0
17	chr4: 3649237	a	G	24	54.0	46.0
18	chr15: 55487474	t	C	24	96.0	4.0
19	chr1: 178793436	g	A	10	40.0	60.0
20	chr11: 124193407	t	G	24	37.5	62.5
21	chr16: 16356684	g	A	22	0.0	100.0
22	chr20: 11220172	g	T	22	27.0	73.0
23	chr2: 19825504	a	T	18	28.0	72.0
24	chr18: 38060888	g	A	18	83.0	17.0
25	chr19: 36651599	g	A	22	32.0	68.0
26	chr1: 12919864	c	T	14	21.0	79.0
27	chr5: 58480456	g	A	24	17.0	83.0
28	chr1: 186077564	g	A	22	100.0	0.0
29	chr2: 29123555	g	A	22	100.0	0.0
30	chr10: 56311213	t	C	24	58.0	42.0

(50-500 kb gaps) genome-wide. Therefore, these SNPs will allow for the inclusion of many additional genomic intervals in the final HapMap. Nevertheless, our efforts did not completely close all of the gaps in the SNP map, and our analysis indicates that additional SNP discovery projects should be launched now to meet the demands of the HapMap project (Table 2-2). Ideally, a uniformly dense SNP map would be generated now in order to ensure an adequate supply of SNPs as each region of the HapMap is developed.

For example, global SNP discovery methods such as shotgun re-sequencing could be extended such that every region of the genome exceeds a minimum SNP density (one SNP per 1-3 kb?). Such densities may not be required throughout the genome; however, up to half of the haplotype blocks in the genome will be missed at lower SNP densities (Patil, *et al*. 2001). Having more SNPs at the beginning of the project also will ensure that the final HapMap contains SNPs that have more desirable allelic frequencies and are technically easier to genotype. An alternative approach would be to construct an initial draft of the HapMap using all currently available SNPs, followed by a gap closure phase in which additional SNPs are identified as necessary. Irrespective of the specific strategy that is chosen, however, additional SNPs will be required to complete the goals of the project.

There are several key differences between our pipeline and previous pipelines that allowed us to discover new SNPs in the TSC traces. First, 13 different methods were used previously to identify SNPs in the TSC traces, and no single method was applied to all of the traces uniformly (The International SNP Map Working Group 2001). In contrast, we processed all 7.1 million traces using only a single pipeline (see Materials and Methods). Since some of the 13 methods used previously compared only the traces to each other, our analysis represents the first time that some of these traces were compared with the draft sequence for the purpose of cataloging SNPs. Secondly, we used only the largest, high quality segment from each trace, which allowed us to include insertions and deletions (Indels) in our analysis. The inclusion of Indels facilitated SNP discovery in traces that otherwise would have been set aside (TSC discarded traces with <99% match). Thirdly, our procedure for mapping traces was different from methods employed

previously and allowed accurate mapping while still retaining flexibility in the subsequent pair-wise alignment step. We required a minimum of a 50 base match at 100% identity to a single location in the genome for a successful trace mapping. In contrast, many of the previous methods used the entire trace for mapping and required a match of >99%. Some of the methods also set aside traces that contained >50% repetitive sequences; however, our mapping methods allowed us to utilize many of these traces successfully. Fourthly, because a number of gaps in the human genome sequence have been closed recently, we were able to map traces for the first time to these segments of the genome. Since our collection of SNPs mapped to all SNP interval sizes, however, only a fraction of the SNPs discovered in our study can be attributed to such gaps (Table 2-2). Therefore, the success of our pipeline is due to a combination of the factors listed above.

In conclusion, the current human SNP map will serve as an excellent resource to fuel the construction of an initial `draft' of the human HapMap. Although the addition of our new SNPs to the human SNP map led to significant improvements in the largest gaps of the map, additional SNP discovery projects will be required to fully support the HapMap project. Higher SNP densities will be necessary to fully discover the haplotype architecture of the genome and to construct comprehensive tag SNP maps that will be useful for genetic linkage studies in humans.

Acknowlegements

This work was supported by Research Scholar Grant RSG-01-173-01-MBC from the American Cancer Society (S.E.D.), the Woodruff Foundation (S.E.D.), and the Emory University Research Council (S.E.D.). We thank Shari Corin for helpful discussions and critical review of the manuscript. We thank the people at TSC, dbSNP and UC Santa Cruz for the use of their data and for helpful advice.

Chapter 3

Natural genetic variation caused by transposable elements in humans

E. Andrew Bennett,*,†,1 Laura E. Coleman,*,1 Circe Tsui,*,‡,1 W. Stephen Pittard‡,§ and Scott E. Devine*,†,‡,2

*Department of Biochemistry, ‡Center for Bioinformatics, †Genetics and Molecular Biology Graduate Program and §Bimcore, Emory University School of Medicine, Atlanta, Georgia. 1 These authors contributed equally to this work.

2004. *Genetics* 168, 933-951

 DOI: 10.1534/genetics.104.031757

Introduction

Transposons and transposon-like repetitive elements collectively occupy 44% of the human genome sequence. In an effort to measure the levels of genetic variation that are caused by human transposons, we have developed a new method to broadly detect transposon insertion polymorphisms of all kinds in humans. We began by identifying 606,093 insertion and deletion (indel) polymorphisms in the genomes of diverse humans. We then screened these polymorphisms to detect indels that were caused by de novo transposon insertions. Our method was highly efficient and led to the identification of 605 nonredundant transposon insertion polymorphisms in 36 diverse humans. We estimate that this represents 25-35% of ~2075 common transposon polymorphisms in human populations. Because we identified all transposon insertion polymorphisms with a single method, we could evaluate the relative levels of variation that were caused by each transposon class. The average human in our study was estimated to harbor 1283 Alu insertion polymorphisms, 180 L1 polymorphisms, 56 SVA polymorphisms, and 17 polymorphisms related to other forms of mobilized DNA. Overall, our study provides significant steps toward (i) measuring the genetic variation that is caused by transposon insertions in humans and (ii) identifying the transposon copies that produce this variation.

Transposons and transposon-like repetitive elements collectively occupy an impressive 44% of the human genome sequence (Lander, *et al.* 2001). Alu and LINE (L1) elements alone account for ~30% of the genome sequence and are the most abundant transposable elements in humans (Lander, *et al.* 2001). Both Alu and L1 also

are actively mobile in the genome today and serve as ongoing sources of human genetic variation (Moran, *et al.* 1996; Ostertag and Kazazian 2001; Batzer and Deininger 2002; Brouha *et al.* 2003; Dewannieux, *et al.* 2003). The remaining transposon-like elements in the genome have some or all of the hallmark features of transposons, such as target site duplications (TSDs), terminal repeats, and/or poly(A) tails, but are not known to remain functional (Smit and Riggs 1996; Smit 1999; Lander, *et al.* 2001).

Alu elements have been actively mobile in primate genomes during the past 65 million years and consequently have expanded to ~1 million copies in the human genome today (Batzer and Deininger 2002 and references therein). The earliest Alu elements appear to have been monomeric derivatives of 7SL RNA, and these monomers later gave rise to dimeric Alu elements (Ullu and Tschudi 1984; Slagel, *et al.* 1987; Britten, *et al.* 1988; Jurka and Zuckerkandl 1991). Alu J elements are the oldest dimeric elements in the human genome (Jurka and Smith 1988; Batzer and Deininger 2002). Although these elements were highly active ~55-65 million years ago, they are thought to have lost the ability to transpose long ago (Jurka and Smith 1988; Batzer and Deininger 2002). Likewise, Alu S elements, which are intermediate in age, are thought to have become inactive at least 35 million years ago (Jurka and Smith 1988; Batzer and Deininger 2002; Johanning, *et al.* 2003). Alu Y elements, in contrast, are the youngest Alu elements in the genome and these elements remain actively mobile today (Batzer and Deininger 2002; Dewannieux, *et al.* 2003). The Alu J, S, and Y families (and their subfamilies) contain a series of hierarchical DNA sequence changes that arose during Alu evolution (Slagel, *et al.* 1987; Jurka and Smith 1988; Batzer and Deininger 2002; Jurka, *et al.* 2002). Each Alu

family contains a unique set of diagnostic base changes that can be used to identify copies belonging to that family.

The second most abundant class of transposons in humans, the LINE (L1) elements, are autonomous poly(A) retrotransposons (Ostertag and Kazazian 2001 and references therein). These elements also have reached high copy numbers in the human genome (~500,000) and collectively occupy ~17% of the genome sequence (Lander, *et al.* 2001). Like Alu, L1 elements have been actively mobile over a long period of time and have been classified according to their respective ages using specific base changes (Boissinot, *et al.* 2000; Ovchinnikov, *et al.* 2002; Brouha, *et al.* 2003). The oldest L1 elements in the genome have accumulated deleterious mutations that render them inactive. However, younger L1 elements have been identified that remain actively mobile today (Moran, *et al.* 1996; Brouha, *et al.* 2003). These active copies contain two intact open reading frames, ORF1 and ORF2, which encode proteins that are necessary for L1 retrotransposition (Feng, *et al.* 1996; Moran, *et al.* 1996). ORF1 encodes a 40-kD protein with RNA-binding activity (Hohjoh and Singer 1996, 1997a, 1997b; Kolosha and Martin 1997; Martin, *et al.* 2000, 2003; Martin and Bushman 2001), whereas ORF2 encodes a protein with both endonuclease (EN) and reverse transcriptase (RT) activities (Mathias, *et al.* 1991; Feng, *et al.* 1996; Moran, *et al.* 1996; Cost, *et al.* 2002). EN and RT work together in a process known as target-primed reverse transcription (TPRT; Luan, *et al.* 1993) that integrates a newly synthesized L1 cDNA into a DNA target site (Cost, *et al.* 2002). Alu RNA (and other cellular RNAs) can compete for the L1 machinery during the TPRT process, leading to the retrotransposition of these alternative RNAs instead of the

normal L1 mRNA (Esnault, *et al.* 2000; Wei, *et al.* 2001; Dewannieux, *et al.* 2003). This "*trans*" replication mechanism is thought to account for the massive expansion of Alu elements in the human genome and for the existence of processed pseudogenes.

Because Alu and L1 remain actively mobile in the human genome today, they serve as ongoing sources of genetic variation by generating new transposon insertions (reviewed in Ostertag and Kazazian 2001 and Batzer and Deininger 2002). For example, estimates suggest that a new Alu insertion occurs approximately once every 200 live births (Deininger and Batzer 1999). As a consequence, a large number of polymorphic Alu and L1 insertions have accumulated in human populations. Many of these insertions are expected to be genetically neutral and, therefore, would have little or no impact on human phenotypes. However, other insertions (primarily those within genes) have been found to cause altered human phenotypes, including diseases. For example, disease-causing Alu insertions have been observed in the BRCA2 gene (Miki, *et al.* 1996), the glycerol kinase gene (Zhang, *et al.* 2000), and others (Deininger and Batzer 1999). Disease-causing L1 insertions likewise have been observed in at least 14 different genes, causing cancers (Morse, *et al.* 1988; Miki, *et al.* 1992; Liu, *et al.* 1997), hemophilia (Kazazian, *et al.* 1988), muscular dystrophy (Narita, *et al.* 1993), and other diseases. It is likely that additional transposon insertions will be found to affect human phenotypes as well.

As an initial step toward studying the potential phenotypic variation that is caused by Alu and L1 elements, it is necessary to identify all of the polymorphic insertions that exist in human populations. Only a fraction of such insertions have been identified to

date, largely because the methods for detecting transposon insertion polymorphisms are labor intensive. Most of the known Alu and L1 insertion polymorphisms have been identified by systematically screening individual element copies in human populations using PCR assays (Carroll, *et al.* 2001; Roy-Engel, *et al.* 2001; Myers, *et al.* 2002; Abdel-Halim, *et al.* 2003; reviewed in Ostertag and Kazazian 2001 and Batzer and Deininger 2002). Transposon display assays also have been used to identify transposon insertion polymorphisms (Sheen, *et al.* 2000; Badge, *et al.* 2003). Although these methods have been useful for identifying polymorphisms, they are not likely to be sufficient on a genome-wide scale to identify all of the polymorphic Alu and L1 copies that exist in human populations. Thus, new and more efficient methods are necessary to identify transposon insertion polymorphisms.

In addition to Alu and L1 elements, some of the remaining transposons and transposon-like elements in the genome also might be polymorphic and, therefore, would contribute to human genetic diversity. Despite the fact that there are many families of such elements in humans (Smit and Riggs 1996; Smit 1999; Lander, *et al.* 2001), no comprehensive studies have been conducted to examine whether these elements are polymorphic or remain actively mobile. As is the case for Alu and L1, such elements would be of interest because they represent sources of human genetic variation and might also cause mutations that lead to human diseases.

In an effort to measure the levels of genetic variation that are caused by human transposons, we have developed an efficient method to broadly detect transposon

insertion polymorphisms of all kinds in humans. The method exploits DNA sequencing traces that originally were generated from diverse humans for single-nucleotide polymorphism (SNP) discovery projects (Sachidanandam, *et al.* 2001; International HapMap Consortium 2003). We have developed a computational pipeline that now analyzes these traces to identify transposon insertion polymorphisms. Our study provides significant steps toward (i) measuring the genetic variation that is caused by transposon insertions in humans and (ii) identifying the transposon copies that produce this variation.

Materials and Methods

Identifying insertion and deletion candidates using DNA sequencing traces from diverse humans: DNA sequencing traces and accompanying quality files were obtained from Cold Spring Harbor Laboratory [traces generated by the SNP Consortium (TSC)] or from the Trace DB archive at the National Center for Biotechnology Information (NCBI). Insertion and deletion (indel) and transposon insertion polymorphisms were identified from these traces using a sequential series of computer programs and databases as outlined in Figure 3-1. Many of these programs were obtained from NCBI or from other sources as indicated below. Other custom Perl programs were developed for indel and transposon polymorphism discovery as necessary and are available upon request. Most of these programs and databases were installed locally on Dell workstations running Microsoft 2000, XP, or Red Hat Linux operating systems. A 12-CPU Linux cluster also was constructed and utilized for the RepeatMasker and Mega-BLAST steps of the pipeline (Figure 3-1).

A total of 16.4 million DNA sequencing traces were processed using the pipeline depicted in Figure 3-1. The traces first were screened for vector contamination using the VecScreen system developed by NCBI and were trimmed as necessary. Low-quality regions of the traces then were identified and trimmed with a custom Perl program that uses the Phred quality scores in the accompanying quality files to identify such regions (Ewing and Green 1998; Ewing, *et al.* 1998). Our method identified the longest high-quality region of each trace and then trimmed the flanking data upon encountering 5 bases in a row with Phred scores ~25. The longest high-quality interval from each trace was chosen for further analysis and the remaining data were set aside. Trimmed traces also were required to have average Phred scores of at least 25 and minimum lengths of 100 bases.

After trimming, each trace then was mapped to a unique location in the human genome sequence (build 33 for the TSC traces and build 34 for the remaining traces). Builds 33 and 34 of the human genome sequence database were obtained from the University of California (Santa Cruz) and installed locally to perform this step (Kent, *et al.* 2002). All known repeats (including all transposons and transposon-like repetitive elements defined in Repbase Volume 7, Issue7; Jurka 2000) first were temporarily masked in the traces using RepeatMasker (version 2001/07/07; A. Smit, unpublished data) and MaskerAid (Bedell, *et al.* 2000). The single longest unmasked "anchor sequence" of the trace then was used to assign each trace to a unique genomic location using Mega-BLAST (NCBI). The anchor sequence was required to have a minimum of a 50-base match at 100% identity for a trace to be mapped successfully. Traces with anchor

sequences that matched to more than one genomic location with 100% identity, or that did not have a minimum of a 50-base match at 100% identity, were set aside to avoid traces that mapped to duplicated regions of the genome (Bailey, *et al.* 2002). After the traces were successfully mapped to unique genomic locations, they were unmasked and aligned to their assigned genomic locations using the Bl2Seq program (NCBI). The Bl2Seq program allowed for as much as a 16-base gap in the alignments and led to identification of indels as large as 16 bases in length.

A new algorithm also was developed to identify indels that were >16 bp in length. Our strategy was designed to split trace data into two blocks upon encountering a region in the pairwise alignment that no longer matched the query. The first block of sequence that matched was maintained in the correct position, and the nonmatching sequence was moved over as a block, 1 base at a time, until a match was obtained. The Perl program that was developed to accomplish this task moved the nonmatching block until it detected either a perfect alignment or a distance of 10,000 bases (the maximum distance allowed by the program). The 5 bases on each side of an indel candidate were required to have Phred scores of ≥20 to ensure that high-quality bases were being used to locate the indel junctions. Indel candidates were deposited into dbSNP under accession nos. ss8029278-ss8176133, ss8475737- ss8484870, ss14926095-ss15354938, and ss15357378-ss15378640.

Identifying transposon insertion polymorphisms by screening human indels

Transposon insertion polymorphisms were identified among indels using a custom computer algorithm. First, indels were identified for which at least 80% of the indel sequence was occupied by a known transposon as defined by the definitions of all human transposons and repeats in Rep-base (Vol. 7, Issue 7; Jurka 2000). This step was accomplished by querying an Oracle database that stored RepeatMasker output data (and other information) for each indel. Next, selected candidates were examined with a custom Perl program to determine whether potential TSDs were present. Such duplications generally flank transposon insertions and are hallmarks of most transposons (Berg and Howe 1989). Therefore, if an indel was caused by a transposon insertion, it generally would be expected to be flanked by a TSD (one copy of the duplicated sequence actually is contained within the indel itself, since the duplication is created during the insertion of the transposon). Candidate transposon insertions also were screened with a custom Perl program to identify potential poly(A) tails, which are associated with certain retrotransposons. Finally, the genomic contexts of all transposon indel candidates were examined to identify true de novo insertions vs. indels that were caused by deletions or duplications within existing transposon copies. All indels that met at least the first test were inspected and curated manually (see supplemental Table 3-1 at http://www.genetics.org/supplemental/ for the final curated set). Six hundred and five nonredundant polymorphisms were identified that were caused by de novo transposon insertions (these are listed in the “Alu,” “L1,” “SVA,” and “Other” sections of supplemental Table 3-1). Another 50 nonredundant polymorphisms were caused by

deletions or duplications within existing transposons (these are listed in the "Deletions and duplications" section of supplemental Table 3-1).

Analysis of transposon subfamilies

Alu transposon insertions were classified initially using RepeatMasker (A. Smit, unpublished data) and Repbase (Vol. 7, Issue 7; Jurka 2000). Each polymorphic copy also was compared independently to the consensus sequences of all known Alu subfamilies (Repbase Vol. 7, Issue 7; Jurka 2000). To accomplish this goal, all Alu insertions identified were coaligned with the consensus sequences of all Alu subfamilies using the ClustalW program. Key diagnostic bases then were analyzed to further assist with the assignments of these elements to specific subfamilies. Each copy then was compared to the assigned subfamily consensus using Bl2seq (NCBI). In some cases, element copies also were compared to the consensus sequences of several neighboring families. A final assignment was made on the basis of the best match obtained. L1-Hs and L1-P elements were classified initially using RepeatMasker (A. Smit, unpublished data). The L1-Hs elements then were assigned to a given subfamily using the classification system described by Brouha, *et al.* (2003). All other transposons were classified using the RepeatMasker system (A. Smit, unpublished data) and Repbase (Vol. 7, Issue 7; Jurka 2000).

Validation of the computational pipeline by PCR

Sixty-one transposon insertions were chosen arbitrarily from the TSC data set and examined by PCR to evaluate the accuracy of our computational predictions (Table 3-6).

PCR assays were designed for each of the 61 polymorphic transposon copies using primers that either flanked (A and D primers) or were located within (B and C primers) a given transposon as depicted in Figure 3-3. All primers used in these studies are listed in supplemental Table 3-2 at http://www.genetics.org/supplemental/. A total of 68 PCR assays were designed initially. Seven (10%) of these assays failed due to technical reasons and these assays were abandoned. The remaining 61 assays (90%) yielded band(s) of the expected size(s) and were used to assay 12-24 DNA samples from the Coriell diversity panel (Figure 3-3 and Table 3-6). The Coriell diversity panel of 24 DNA samples was obtained from the Coriell Repository, Camden, New Jersey (Collins, *et al.* 1999). Lymphocyte cultures of this panel also were obtained from Coriell and, in some cases, DNA was prepared from these cells. PCR reactions were carried out in 50-µl volumes as described previously (Kimmel, *et al.* 1997). PCR products were run on 1.5% agarose gels and sized using a 1-kb ladder marker (Invitrogen, San Diego).

Analysis of additional genomic SVA elements

In addition to the SVA copies identified in the trace experiments, 28 other genomic SVA copies were selected from the human genome sequence using SVA element query sequences and the BLAT program (Kent 2002). These SVA copies were examined by PCR to assess whether they were polymorphic in at least one individual of the Coriell panel. PCR primers were developed to examine the status of each SVA copy as described in Figure 3-3 and supplemental Table 3-2. PCR reactions were carried out as outlined above and in Figure 3-3. An SVA copy was considered to be polymorphic if both alleles (one with and one without the transposon insertion) could be identified at least once.

Fifty-nine additional SVA element copies were identified by manual inspection of the first 50 Mb of human chromosome 1 using the University of California, Santa Cruz genome browser (Kent, *et al.* 2002). The genomic regions surrounding all of these SVA copies were compared to the equivalent chimp genomic sequences to determine whether the chimp contained an SVA element at the equivalent position (supplemental Table 3-1).

Results

A strategy for detecting genetic variation caused by transposon insertions in humans: Our strategy for detecting transposon insertion polymorphisms in humans involved identifying a large number of indel polymorphisms in human populations and then screening these polymorphisms to identify de novo transposon insertions. We reasoned that this strategy should be successful since transposon insertion polymorphisms are equivalent to insertions and deletions in genomes. Relatively few indels had been identified in human populations prior to our study, despite the fact that indels are abundant in the genomes of model organisms such as *Drosophila melanogaster* (Berger, *et al.* 2001) and *Caenorhabditis elegans* (Wicks, *et al.* 2001) and were likely to be abundant in humans as well. Therefore, we began our study by developing new computational methods to discover in-del polymorphisms in the genomes of diverse humans (materials and methods).

Our strategy involved mining indels from DNA sequencing traces that previously had been generated for SNP discovery projects. All of the traces used in our study originally were generated at genome centers by resequencing pools of genomic DNA

from diverse humans. For example, a set of 7.1 million traces, which originally had been generated by shotgun sequencing the DNA of 24 diverse humans (Sachidanandam, *et al.* 2001), was obtained from TSC. A second set of 8.2 million whole-genome shotgun (WGS) traces, which originally had been generated by shotgun sequencing the DNA of eight unrelated African-American adults (four males and four females from the Baylor Polymorphism Resource; International HapMap Consortium 2003), was obtained from the Baylor and Whitehead Genome Centers. Finally, a much smaller set of 0.9 million whole-chromosome shotgun (WCS) traces, which had been generated by shotgun sequencing chromosome 20-specific libraries from four diverse humans (International HapMap Consortium 2003), was obtained from the Sanger Center. Because these DNA sequencing traces were derived from diverse humans, we expected them to harbor various forms of genetic variation, including indels. We developed a computational pipeline to identify indels within these traces by comparing them to the human genome reference sequence (builds 33 and 34; Figure 3-1).

A total of 606,093 indel candidates were identified by analyzing 16.4 million traces with our computational pipeline (Figure 3-1 and materials and methods). The majority of these indels (428,838 or 70.8%) were identified from the WGS traces. An additional 155,992 indels (25.7%) were identified from the TSC traces, and 21,263 indels (3.5%) were identified from the WCS traces. Overall, these indel candidates were distributed throughout the human genome and were found on all 24 chromosomes (data not shown). They ranged in size from 1 to 9969 bp in length and contained a wide array of different DNA sequences. All indel candidates were deposited into dbSNP under the "Devine_lab" handle (http://www.nih.nlm.gov/SNP).

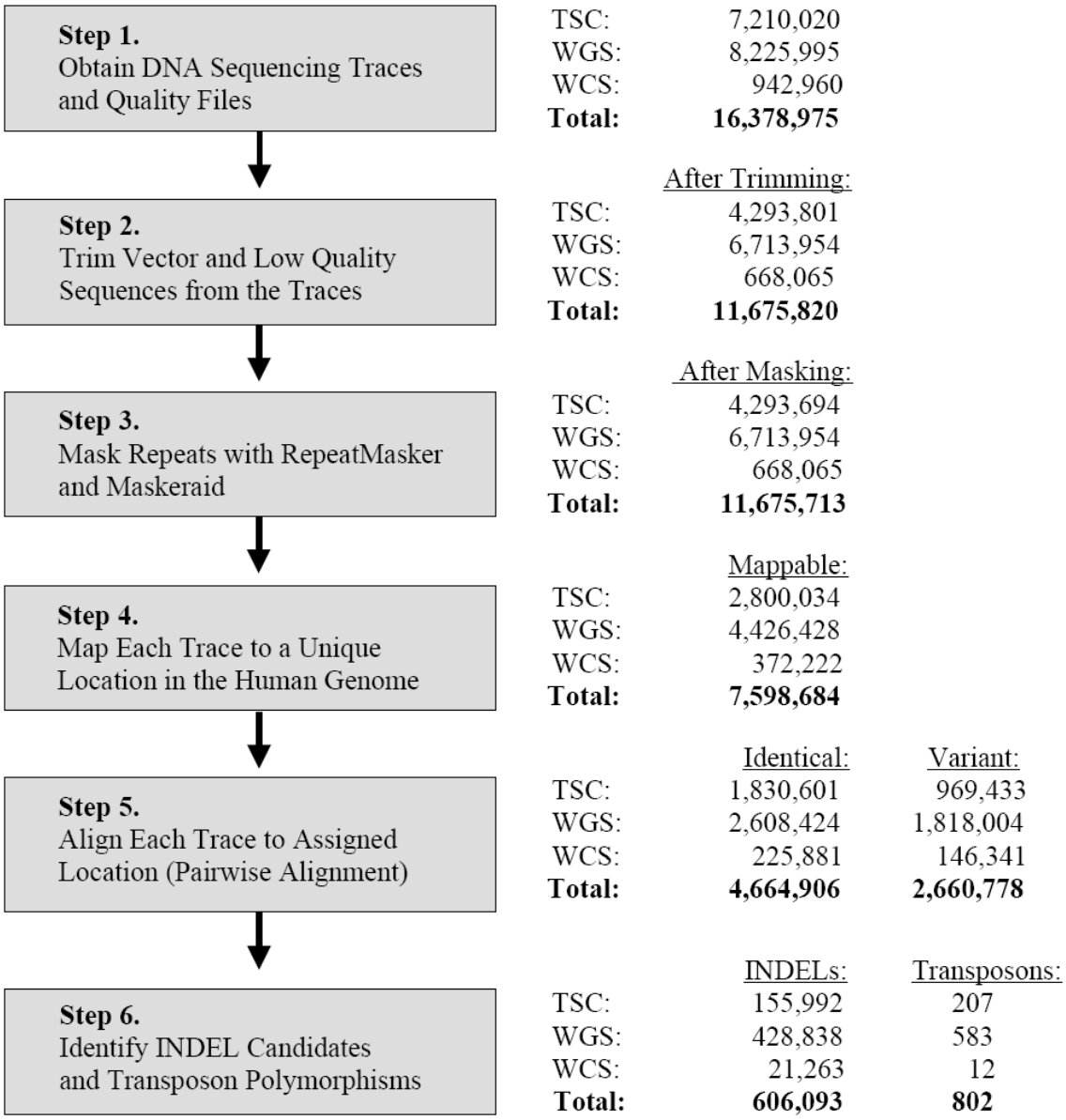


Figure 3-1. Computational pipeline for indel and transposon polymorphism discovery. A flow chart of the computational steps that were followed for the discovery of indel and transposon polymorphisms is shown on the left (boxes). A breakdown of the number of traces present at the end of each step is shown on the right. The number of indel and transposon polymorphisms identified is listed at the bottom. Note that the numbers are broken down for each of the three populations examined. TSC, the SNP Consortium traces; WGS, whole-genome shotgun traces; WCS, whole-chromosome shotgun traces.

We next developed a computer algorithm to identify indels that were caused by de novo transposon insertions. The method was designed to identify indels for which a single transposon copy and its associated sequences (e.g., its target site duplication) accounted for the indel (see materials and methods). Eight hundred and two transposon insertion polymorphisms were detected with these methods in the three populations examined (Table 3-1). Four major classes of transposon insertions were identified in these experiments: (i) Alu insertions, (ii) L1 insertions, (iii) SVA insertions, and (iv) insertions of "other" elements.

Alu insertion polymorphisms were by far the most abundant polymorphisms identified in the three experiments (Table 3-1). A total of 173 of 207 (83.6%) of the polymorphisms in the TSC data set were Alu insertions. Likewise, 487 of 583 (83.5%) of the polymorphisms in the WGS set were Alu insertions, and 10 of 12 (83.3%) of the polymorphisms in the WCS set were Alu insertions. L1 insertions were the next most abundant polymorphisms identified, representing 12.6 and 11.0% of the TSC and WGS data sets, respectively (Table 3-1). Although the L1-Ta class was the most abundant subfamily of L1, other non-Ta L1 elements were identified as well (see below). SVA element insertions were the third most abundant class of transposon polymorphisms identified, representing 2.9 and 4.3% of the TSC and WGS data sets, respectively. Finally, the remaining transposon insertion polymorphisms were caused by a miscellaneous collection of low-frequency insertions. These elements were pooled into a single group of other polymorphisms (Table 3-1).

	TSC		WGS		WCS	
Bases analyzed	989,283,997		2,271,983,242		110,411,692	
Fraction of genome	0.30		0.69		0.033	
No. polymorphisms observed/(% total):						
Alu-total	173	(0.836)	487	(0.835)	10	(0.833)
Alu Ya5	67	(0.324)	148	(0.254)	7	(0.583)
Alu Yb8	30	(0.145)	132	(0.226)	1	(0.083)
L1-total	26	(0.126)	64	(0.110)	2	(0.167)
L1-Ta	25	(0.121)	58	(0.099)	2	(0.167)
L1-other	1	(0.005)	6	(0.010)	0	(0)
SVA	6	(0.029)	25	(0.043)	0	(0)
Other	2	(0.010)	7	(0.012)	0	(0)
Total	207		583		12	

Table 3-1. Transposon insertion polymorphisms identified in humans

It is important to note that our measurements were remarkably consistent between the data sets. This was particularly true for the TSC and WGS data sets, which were significantly larger than the remaining WCS set. For example, as noted above, Alu polymorphisms represented ~83% of the transposon insertions in all three of the populations examined. The percentages of L1 insertions likewise were very similar in these experiments (Table 3-1). Overall, the TSC and WGS experiments were remarkably similar given the differences in the populations that were used to generate these trace sets (Table 3-1). Nevertheless, the results were not completely identical between the populations. For example, Alu Ya5 polymorphisms represented 32.4% of the insertions in the TSC population and 25.4% of the insertions in the WGS population (Table 3-1). Therefore, at least some of these element families might have amplified at slightly different rates in the populations examined.

Alu (505 total)

Alu S	
Alu Sc-derived	1
Alu Sp-derived	1
Alu Sq-derived	1
Alu Sx-derived	1
Alu Ya	
Alu Ya1	4
Alu Ya2	1
Alu Ya4	27
Alu Ya4-derived	3
Alu Ya5	170
Alu Ya5a2	11
Alu Ya8	2
Alu Yb	
Alu Yb3a1	1
Alu Yb3a1-derived	1
Alu Yb3a2	1
Alu Yb3a2-derived	5
Alu Yb8	125
Alu Yb8-derived	4
Alu Yb9	15
Alu Yc	
Alu Yc1	38
Alu Yc2	2
Alu Yd	
Alu Yd8	4
Alu Ye	
Alu Ye2	1
Alu Ye2-derived	1
Alu Ye5	15
Alu Ye5-derived	3
Alu Yf	
Alu Yf1	1
Alu Yg	
Alu Yg6	12
Alu Yi	
Alu Yi6-derived	8
Alu Y	
Alu Y	22
Alu Y-derived	10
Too short to classify	
Alu Ya4/5	1
Alu Ya5/8	5
Alu Yc	1
Alu Y	6
Alu	1

L1 (65 total)

L1 Hs	
Ta-0	8
Ta-1	12
Ta-1nd	3
Ta-1d	12
Pre-TA(acg/a)	4
Pre-TA(acg/g)	9
L1 Hs cont.	
Hs (unclassifiable)	1
Ta (unclassifiable)	7
Ta-1d/Ta-0	3
L1-P	
L1 PA2	5
L1 PA3	1

SVA (39 total)

From traces	28
Other genomic	11

Other (7 total)

DNA/Mariner	1
LTR/ERVK	2
U2 RNA	1
U5 RNA	1
5S rDNA	2

Table 3-2. Nonredundent transposon insertion polymorphisms.

We next inspected all of the polymorphic transposon insertions from the three populations to determine whether any of the copies were redundant in the three data sets. In fact, 149 polymorphisms were identified in which the same transposon allele was detected more than once in our trace experiments. In most of these cases (115 of 149 or 77.2%), the alleles were detected independently twice. Another 25 of 149 alleles (16.8%) were detected three times and the remaining 9 of 149 alleles (6%) were detected four to six times. These results provide confidence in our method and suggest that at least some of our transposon insertion polymorphisms are present at high frequencies in human populations. To perform additional analyses of these transposons, we developed a nonredundant data set of 605 transposon insertions.

Alu Y insertion polymorphisms

A total of 505 nonredundant Alu insertion polymorphisms were identified in the three populations of our study, including both full-length and partial Alu insertions (Table 3-2). These elements were compared to all known Alu families and were classified to determine which Alu elements were detected in our experiments (materials and methods). The vast majority of our Alu insertions were Alu Y elements, with 500 of 505 (99%) of the insertions falling in this category (Table 3-2). Alu Ya5 elements were the most abundant subfamily in our study, representing 33.7% of the insertions (Table 3-2). Alu Yb8 polymorphisms also were abundant, representing 25.5% of the insertions (Table 3-2). Alu Y, Alu Yc1, and Alu Ya4 elements were present at intermediate levels (between 5.9 and 7.5%), and these three families together represented 19.7% of all nonredundant Alu insertions in our study. Most of the remaining Alu Y-related insertions were present

at relatively low levels and were distributed among 15 different Alu Y subfamilies (Table 3-2). Notably, Alu polymorphisms were detected from most of the known Alu Y subfamilies, including Alu Ya, Yb, Yc, Yd, Ye, Yf, Yg, and Yi (Table 3-2). Moreover, although we did detect several new small groups of Alu Y insertions that might be considered novel subfamilies (see below and Figure 3-2), no new extended Alu Y families of significant size were detected in our study.

As outlined above, Alu Ya5 and Alu Yb8 insertions were the most abundant Alu elements in our data sets. Carroll, *et al.* (2001) previously demonstrated that these two Alu subfamilies were highly polymorphic in human populations. In fact, they estimated that 25% of Alu Ya5 elements and 20% of Alu Yb8 elements were polymorphic in at least one individual of a panel of 80 diverse humans (Carroll, *et al.* 2001). On the basis of their copy number estimates for these two elements, we can predict that at least 660 Alu Ya5 insertion polymorphisms and 370 Alu Yb8 insertion polymorphisms should exist in human populations. We found a total of 170 nonredundant Alu Ya5 insertions and 129 nonredundant Alu Yb8 insertions in our study (Table 3-2). Only 8 of these polymorphic insertions (4 Alu Ya5 and 4 Alu Yb8) were identified by Carroll, *et al.* (2001). Therefore, 291 of 299 (98.6%) of our Alu Ya5 and Alu Yb8 polymorphisms had not been detected previously. Similar results were obtained with the remaining Alu classes, indicating that our method efficiently identified a large number of novel Alu insertion polymorphisms in human populations.

Polymorphic ancient Alu elements

In addition to Alu Y elements, we also identified four polymorphic copies of older Alu S elements (Table 3-2). Two of these examples (Alu ss14941867 and Alu ss8480425) were intact, full-length Alu S insertions with all of the expected features of Alu retrotransposition events, including poly(A) tails and target site duplications. Two additional examples of 5'-truncated or otherwise fragmented copies of Alu S also were identified (Table 3-2). One of these insertions (ss14931773) was a 5'-truncated Alu Sc element with a perfect target site duplication. The second insertion (ss15143442) was an Alu Sq element that was truncated at both the 5' and the 3' ends and lacked a target site duplication altogether. It is not clear how this second Alu polymorphism was formed. One possibility is that it was caused by an endonuclease-independent mechanism of retrotransposition involving partial Alu RNA templates. Both Alu and L1 elements are known to use an endonuclease-independent mechanism that does not generate target site duplications surrounding the newly transposed copy (Morrish, *et al.* 2002; Abdel-Halim, *et al.* 2003). Perhaps this older Alu Sq element was mobilized by the L1 machinery using this alternative mechanism.

The fact that we identified four ancient Alu S insertion polymorphisms indicates that at least some of the Alu S copies are likely to have retained the ability to transpose long after the majority of Alu S elements became transpositionally inactive. This is most probable for the intact Alu copies discussed above (Alu ss14941867 and Alu ss8480425). These copies do not appear to have been caused by gene conversion events and have estimated ages of 7-23 million years, suggesting that they are younger than most of the

Alu S elements. Prior to our study, only the Alu Y elements were thought to be polymorphic in humans, whereas the older Alu S, Alu J, and Alu monomers were thought to have only fixed alleles in human populations. Recent evidence from Johanning, *et al.* (2003) showed that at least some Alu Sx elements appear to have transposed later than previously estimated (~35 million years ago); however, Alu S insertion polymorphisms were not detected in humans prior to our study.

Sequence variation within polymorphic copies of Alu indicates patterns of Alu evolution

Significant DNA sequence variation was noted within the polymorphic Alu insertions identified in our study. In most cases, a given element copy could be placed unambiguously within a known Alu subfamily using key diagnostic base changes (materials and methods). Nevertheless, a large number of additional single-and multiple-base changes were noted in these elements relative to their respective consensus sequences. Of particular interest were small groups of Alu elements that clearly belonged to a given element family, but differed from the consensus by one or more shared base changes. Since CpG changes occur independently at a high frequency, it was possible that some of these groups were caused by independent changes at CpG hotspots. However, at least 10 of these groups possessed shared base changes at non-CpG sites (or had unusually high frequencies of a given CpG change along with additional shared changes). Most of these groups also showed evidence for the progressive accumulation of shared mutations. In these cases, a single base change was shared by an initial subset of the elements and additional shared changes appear to have been acquired later. We

propose that these groups represent novel evolutionary lineages of Alu elements that are defined by these new base changes (Figure 3-2). These data suggest that a significant number of Alu insertions go on to serve as new source genes for small numbers of additional retrotransposition events (Deininger, *et al.* 1992).

L1 insertion polymorphisms

Although most of the ~500,000 L1 copies in the haploid human genome have accumulated deleterious mutations that render them inactive, some of the younger L1 copies remain actively mobile today (Moran, *et al.* 1996; Brouha, *et al.* 2003). These younger copies belong to the L1-Hs (Human-specific) family of elements (Brouha, *et al.* 2003). The Hs family has been subdivided further into the Ta-0, Ta-1, Ta-nd, and Ta-d subfamilies on the basis of the presence or absence of specific nucleotide changes within the L1 sequence (Boissinot, *et al.* 2000; Ovchinnikov, *et al.* 2002; Brouha, *et al.* 2003). Reflective of their younger ages, L1-Hs elements are highly polymorphic in human populations (Sheen, *et al.* 2000; Ovchinnikov, *et al.* 2001; Myers, *et al.* 2002; Badge, *et al.* 2003; Brouha, *et al.* 2003).

We identified 65 nonredundant L1 insertion polymorphisms in our study (Table 3-2). Each of these L1 elements ended in a poly(A) sequence and was flanked by a typical L1 target site duplication. We classified these elements using the system described by Brouha, *et al.* (2003) and found that most of the copies belonged to Ta subfamilies of L1 elements. In fact, elements belonging to the L1 Ta-0, Ta-1, Ta-

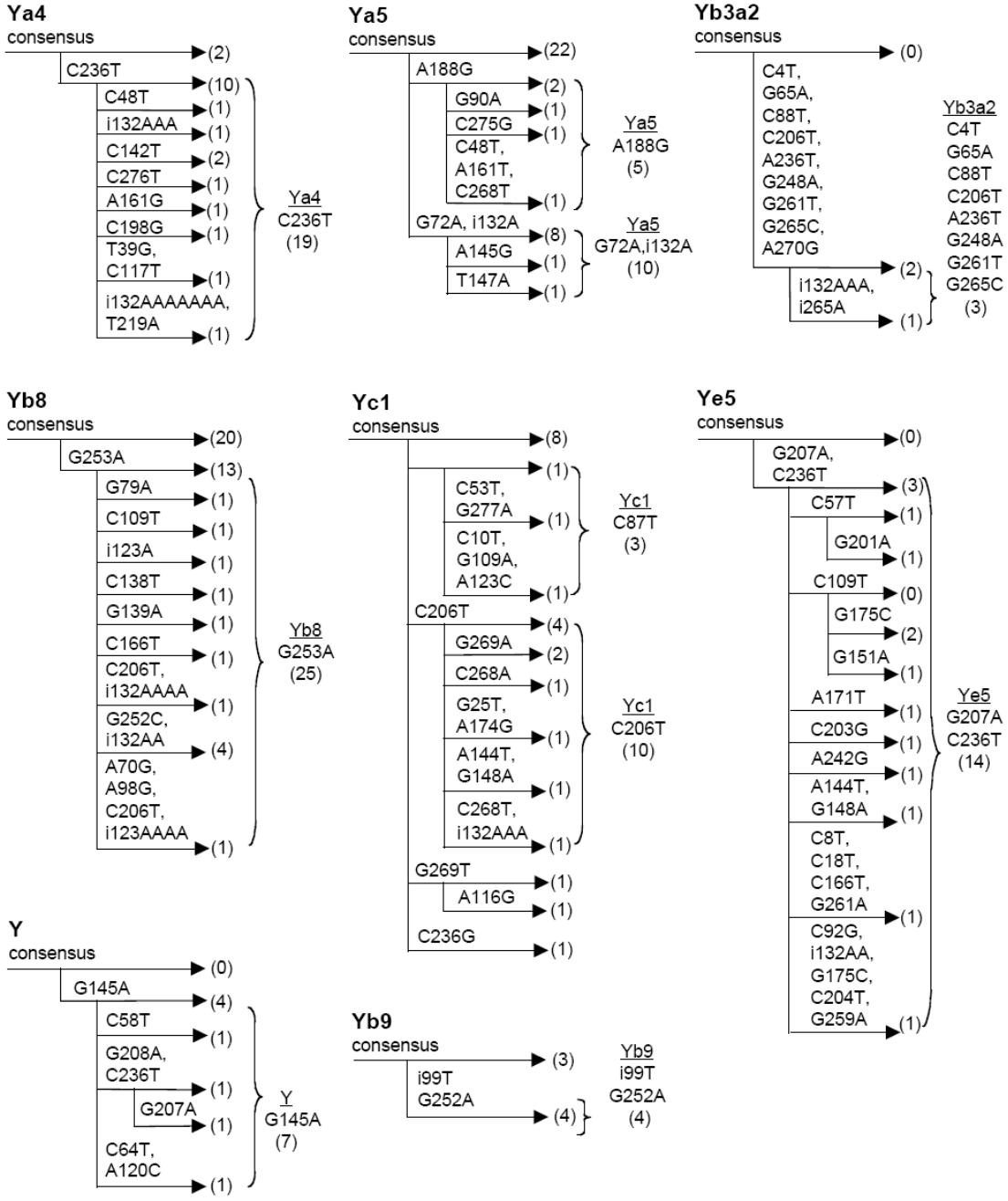

Ya4
consensus
(2)
C236T
(10)
C48T
(1)
i132AAA
(1)
C142T
(2)
C276T
(1)
A161G
(1)
C198G
(1)
T39G,
C117T
(1)
i132AAAAAAA,
T219A
(1)
Ya4
C236T
(19)
Ya5
consensus
(22)
A188G
(2)
G90A
(1)
C275G
(1)
C48T,
A161T,
C268T
(1)
Ya5
A188G
(5)
G72A, i132A
(8)
A145G
(1)
T147A
(1)
Ya5
G72A,i132A
(10)
Yb3a2
consensus
(0)
C4T,
G65A,
C88T,
C206T,
A236T,
G248A,
G261T,
G265C,
A270G
(2)
i132AAA,
i265A
(1)
Yb3a2
C4T
G65A
C88T
C206T
A236T
G248A
G261T
G265C
(3)
Yb8
consensus
(20)
G253A
(13)
G79A
(1)
C109T
(1)
i123A
(1)
C138T
(1)
G139A
(1)
C166T
(1)
C206T,
i132AAAA
(1)
G252C,
i132AA
(4)
A70G,
A98G,
C206T,
i123AAAA
(1)
Yb8
G253A
(25)
Yc1
consensus
(8)
(1)
C53T,
G277A
(1)
C10T,
G109A,
A123C
(1)
Yc1
C87T
(3)
C206T
(4)
G269A
(2)
C268A
(1)
G25T,
A174G
(1)
A144T,
G148A
(1)
C268T,
i132AAA
(1)
Yc1
C206T
(10)
G269T
(1)
A116G
(1)
C236G
(1)
Ye5
consensus
(0)
G207A,
C236T
(3)
C57T
(1)
G201A
(1)
C109T
(0)
G175C
(2)
G151A
(1)
A171T
(1)
C203G
(1)
A242G
(1)
A144T,
G148A
(1)
C8T,
C18T,
C166T,
G261A
(1)
C92G,
i132AA,
G175C,
C204T,
G259A
(1)
Ye5
G207A
C236T
(14)
Y
consensus
(0)
G145A
(4)
C58T
(1)
G208A,
C236T
(1)
G207A
(1)
C64T,
A120C
(1)
Y
G145A
(7)
Yb9
consensus
(3)
i99T
G252A
(4)
Yb9
i99T
G252A
(4)

Figure 3-2. Proposed new evolutionary lineages of Alu. For each Alu subfamily, the number of polymorphic copies retaining the subfamily consensus sequence is compared to groups sharing one or more base pair changes (in parentheses). In several cases, the majority of polymorphic copies of a given subfamily diverge from the subfamily consensus by a few shared changes. An evolutionary progression can be inferred (from left to right) in which new shared base changes appear to have been acquired. The 10 (Figure 3-2 cont.) proposed novel evolutionary groups are indicated by braces to the right of the groups. The total number of elements within the group is indicated in the parentheses. These data suggest that a significant number of Alu insertions in the genome can serve as new source genes to produce offspring elements. Only elements that are at least 80% full length are represented.

nd, and Ta-d subfamilies were identified along with some older pre-Ta elements (Table 3-2). These results are consistent with the observation that 13 of the 14 L1 insertions that have been found to cause human diseases were L1-Ta elements and the remaining element was a pre-Ta element (reviewed in Moran 1999). Our results also are consistent with previous studies demonstrating that L1-Ta elements are highly polymorphic (Sheen, *et al.* 2000; Ovchinnikov, *et al.* 2001; Myers, *et al.* 2002; Brouha, *et al.* 2003). Although some of our L1 insertion polymorphisms were identified previously, most were unique to our study.

Interestingly, we also identified six polymorphic copies of older L1-P insertions, including five polymorphic L1PA2 insertions and a single L1PA3 insertion (Table 3-2). Therefore, in addition to L1-Hs elements, older L1-P elements also are polymorphic in humans. In fact, these elements collectively accounted for 9.2% of the L1 insertion polymorphisms in our study (Table 3-2). Thus, the spectrum of L1 elements that cause human genetic variation, and perhaps human disease, is broader than previously

established (Ovchinnikov, *et al.* 2002). Moreover, since a high level of polymorphism is associated with active transposons, our results suggest that at least some of the L1-P elements in humans and chimps might also remain actively mobile today.

SVA insertion polymorphisms are abundant in humans

The human SVA element is a transposon-like repetitive element that was first identified within the RP gene on human chromosome 6 (Shen, *et al.* 1994). The authors of this original report proposed that SVA represented a composite retrotransposon that contains two previously identified elements (SINE-R and Alu) as well as a variable nucleotide tandem repeat (VNTR) region. Although the authors of this study had no evidence that their proposed element was actively mobile, they suggested that SVA is a retrotransposon because it ended in a poly(A) tail and was flanked by an apparent target site duplication (Shen, *et al.* 1994). Strichman-Almashanu, *et al.* (2001) later estimated that the haploid human genome contains approximately ~5000 copies of the SVA element.

We identified 28 nonredundant SVA insertion polymorphisms in our trace experiments (Tables 2 and 3). These insertion polymorphisms have all of the hallmark features of retrotransposon insertions. In each case: (i) both empty and SVA-occupied sites were identified in different humans, (ii) the newly inserted SVA copy ended in a poly(A) tail, and (iii) each inserted copy was precisely flanked by a new target site duplication (Table 3-3). These copies ranged in size from 396 to 2806 bp in length, with the shorter elements lacking 5' ends due to truncation, or lacking internal VNTR repeats (Table 3-3). The target site duplications of all insertions closely resembled (in both length

and sequence) the target site duplications of Alu and L1 (Table 3-3). To further confirm that SVA insertion polymorphisms were indeed abundant in human populations, we developed PCR assays to individually examine 28 additional genomic copies of SVA (materials and methods). These copies were identified arbitrarily from the ~5000 copies in the human genome by searching the human genome database with SVA element query sequences. A total of 11 of 28 (39%) of the copies tested were found to be polymorphic for insertion in at least one individual of the Coriell panel (Collins, *et al.* 1999; Table 3-4). Thus, together with the 28 SVA polymorphic copies identified in the trace experiments (Table 3-3), we have identified a total of 39 independent SVA insertion polymorphisms in human populations. These insertion polymorphisms have all of the sequence features of bona fide SVA retrotransposition events (Tables 3 and 4). These results indicate that not only Alu and L1, but also a third transposon, SVA, is highly polymorphic in human populations (Tables 1-4). Collectively, our results indicate that Alu, L1, and SVA provide the bulk of genetic variation that is caused by transposon insertion polymorphisms in humans.

The recent completion of a draft sequence for the chimpanzee genome allowed us to determine whether the SVA element also is present in the chimp genome. We manually inspected the SVA copies listed in Table 3-3 and found that 27 of 28 (96.4%) of these copies were absent from the equivalent positions of the chimp genome. The remaining copy appeared to be present, but had only partial sequence coverage in the chimp genome sequence. Therefore, most of these 28 polymorphic SVA copies appear to have been generated relatively recently in humans, at a point in time following the

divergence of chimps and humans. However, since these 28 copies were selected on the basis of the fact that they were polymorphic in humans, it was possible that these copies were not representative of all SVAs in the human genome. Therefore, we arbitrarily selected 87 additional copies from the human genome to determine whether they were present in the chimp genome. Twenty-eight of these copies are listed in Table 3-4, and 59 additional copies were identified in the first 50 Mb interval of human chromosome 1, for a total of 87 copies. Our analysis revealed that only 11 of these 87 SVA copies (12.6%) were precisely present at the equivalent positions of the chimp genome. Seven additional copies had partial sequence coverage in the chimp genome and thus also appeared to be present at equivalent positions. Therefore, the available evidence indicates that □18 of the 87 SVA copies (20.7%) are likely to be present at equivalent positions of the human and chimp genomes. The remaining 69 SVA copies (79.3%) were completely absent from the chimp genome sequence. In a few cases, the chimp genome completely lacked sequence coverage in the area of the element, so it is unclear whether the SVA is truly absent in these cases. However, in most of these 69 cases, the SVA element and 1 of the 2 copies of the target site duplication were precisely absent from the chimp genome. Taken together, these data indicate that ~20.7% of SVA insertions in the human genome were generated prior to the evolutionary divergence of chimps and humans, and up to ~79.3% of the remaining SVA insertions were generated after the divergence of these species. Thus, SVA is a relatively young transposon that has expanded in the human during the past several million years. At least some of the SVA copies appear to have been mobilized very recently, suggesting that SVA might also remain actively mobile in humans and chimps today.

SVA name (Nearest gene)	Chromosomal location*	dbSNP#	INDEL size	Element size	Target site duplication sequence (5' to 3')
CLIC4-A	chr1: 24,509,856- 24,512,677	ss15143842	2,822	2,806	AAAAATAAAAAATCAA
LRIG2	chr1: 112,910,067-112,912,521	ss15142943	2,455	2,438	AAAAACACATATTTGCG
PPP2R5A	chr1: 209,527,634-209,529,114	ss14942498	1,481	1,463	AACAATTCACTTCATCTT
SSB	chr2: 170,844,052-170,846,495	ss15142807	2,444	2,433	GAAAATAATGA
XRCC5	chr2: 217,292,473-217,295,040	ss15143464	2,568	2,555	AAGAACACATGGC
AFURS1	chr3: 195,440,681-195,441,437	ss8481522	757	749	AAGACTTC
KLHL3	chr5: 137,098,795-137,100,118	ss15143086	1,324	1,308	AAAATATTACCTCCCT
HLA-F	chr6: 29,793,437-29,796,056	ss8483556	2,620	2,601	AAAAGAAAGACCCAAGCCT
HLA-G	chr6: 30,005,583-30,007,340	ss15143277	1,758	1,744	AAGAATTGAGGAGC
SLC17A5	chr6: 74,365,117-74,367,305	ss14142874	2,189	2,181	AAAAATGG
SERAC1	chr6: 158,457,569-158,458,368	ss8483321	800	784	GAAAAATGAACATATC
LOC90637	chr7: 929,403-931,992	ss15143254	2,590	2,576	AAAACTTAAGAGTG
BC002644	chr7: 10,248,637-10,250,470	ss8483421	1,834	1,817	AAAGAAAAGAGGTTTAA
EGFR	chr7: 55,028,493-55,030,810	ss8483579	2,318	2,303	TAAAAGCACATTGCA
AQP3	chr9: 33,413,362-33,414,654	ss15142745	1,293	1,055	AAGAATCTAGTTTTT
BC036431	chr9: 79,781,468-79,784,045	ss15143735	2,578	2,563	TAAAATGGCTCTAGC
RAD23B	chr9: 105,413,293-105,415,295	ss15143222	2,003	1,990	AATTATTATTATT
PRMT3	chr11: 20,533,487-20,535,487	ss15143629	2,001	1,988	AATACAGAAATGT
CNOT2	chr12: 68,881,132-68,883,862	ss15143419	2,731	2,713	AAAAAAGTATGACACTTC
EPSTI1	chr13: 41,337,810-41,339,462	ss14938190	1,653	1,638	GAAAATCAGCTGGAG
FLJ12577	chr13: 47,749,416-47,751,876	ss15143296	2,461	2,325	AAACAAAAACAGT
C14orf24	chr14: 33,497,755-33,499,561	ss8478175	1,807	1,792	AAGACTTACGAATAG
PRPSAP2	chr17: 18,991,720-18,992,744	ss15143022	1,025	1,010	AAGAAAGAACAAGTT
PRKCA	chr17: 64,815,159-64,816,230	ss15143644	1,072	1,057	AAAAAATGTTTTAAG
ZNF137	chr19: 57,787,819-57,789420	ss15143783	1,602	1,586	AAAAATACAAAATTAG
AK09261	chr19: 58,380,999-58,383,780	ss15143831	2,782	2,774	AAAAAAAA
C20orf100	chr20: 43,361,555-43,361,953	ss14942665	399	396	TAA
AK026502	chr22: 25,492,631-25,494,356	ss15143832	1,726	1,716	AGAGGTTAAG

Table 3-3. SVA insertion polymorphisms identified in trace experiments. [a] Chromosomal coordinates are given for the July 2003 build (build 34) of the human genome sequence (University of California, Santa Cruz; Kent, *et al*. 2002). All of the elements shown have poly(A) tails at their 3' ends. Additional data are provided in supplemental Table 3-1 and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/). Chr, chromosome.

HERV-K and other examples of mobilized DNA

In addition to the three most abundant groups of insertion polymorphisms in this study (Alu, L1, and SVA), we also identified several classes of less abundant insertion polymorphisms that were caused by mobilized DNA. For example, two human endogenous retrovirus K (HERV-K) insertion polymorphisms were identified in our trace experiments (Tables 2 and 5). One of these HERV-K copies was a 969-bp solo LTR

element that was flanked by a perfect 5-bp target site duplication (Table 3-5). The other HERV-K element was a full-length 9462-bp copy that was flanked by a perfect 6-bp target site (Table 3-5). This full-length copy had intact LTRs at its termini and four intact open reading frames capable of encoding homologs of the retroviral Gag, protease, Pol, and Env proteins (data not shown). We also identified four examples of mobilized genome small cellular RNAs, including two polymorphic copies of 5S rDNA and single examples of mobilized U2 and U5 RNA (Table 3-5). All four of these polymorphic insertions were flanked by L1-like target site duplications, strongly suggesting that these RNAs were mobilized by the L1 machinery. However, none of these mobilized elements contained a poly(A) tail, perhaps suggesting that the poly(A) sequences on template RNAs are not strictly required for the TPRT process (Boeke 1997; Roy-Engel, *et al.* 2003). Finally, we identified a single example of a Mariner dependent-1 (Made1) insertion with an inverted repeat structure that was flanked by a perfect 5 bp target site duplication (Table 3-5).

Validation studies

Several lines of evidence suggested that our computational methods were highly accurate. For example, we detected 149 redundant transposon duplication polymorphisms in the three data sets. Therefore, our methods independently detected identical transposon insertion polymorphisms using totally different traces from different populations. Moreover, we also detected a number of Alu and L1 insertion polymorphisms that had been detected in previous studies with totally different methods. Nevertheless, as outlined

SVA name (nearest gene)	Chromosomal location*	Element size	Target site duplication sequence (5' to 3')	Polymorphic in Coriell panel?
EIF4G3-A	chr1: 20,610,706- 20,613,260	2,544	TAAAAATTCAT	No
SPATA6	chr1: 48,217,230- 48,220,014	2,773	AAAAGAAAAACC	No
CASP8	chr2: 202,349,192-202,351,983	2,782	AAGAATTTGA	Yes
PLOD2	chr3: 147,134,530-147,136,524	1,976	AAAGAAAATGTGGCATATA	Yes
CD38	chr4: 15,542,595-15,544,984	2,374	GAAAAGCAGCAAGCC	No
RAP1B	chr5: 75,550,653- 75,552,974	2,305	AAAAATTAAAAAAACT	Yes
Neurestin	chr5: 167,266,427-167,268,968	2,528	GAAAACAACGTCAA	No
POLH	chr6: 43,594,478-43,596,505	2,015	AAGATTCTTTCAC	No
PCMT1	chr6: 150,137,472-150,139,466	1,930	GAAAACAGCCA	No
LOC90693	chr7: 23,417,744- 23,420,774	2,865	AAGACTGTCCCCTGC	No
FLJ14117	chr7: 100,561,431-100,563,976	2,529	AAAAATACAAAAATTGG	Yes
TNKS	chr8: 9,551,502- 9,553,345	1,829	GAAAATTCTTTTCTT	Yes
FLJ10871	chr8: 28,759,637-28,761,987	2,260	AGAAAAATGTAGACATA	No
MELK	chr9: 36,604,406-36,606,089	1,669	CAAAAATAATTTTTT	No
PBX3	chr9: 123,916,920- 123,919,948	2,361	GAAAAGATCA	No
HHEX	chr10: 94,098,024- 94,100,358	2,320	GAGAGATGGGATGTG	No
ITPR2	chr12: 26,828,444-26,830,665	2,209	AAAAATGGAGAAT	No
SNRPF	chr12: 94,736,050-94,738,795	2,735	AAAACTGTGGA	No
BRMS1	chr14: 34,435,361-34,437,930	2,553	TAAATACCTACGAGTAG	No
SPTB	chr14: 63,350,428-63,353,581	2,686	GAAAATTCTT	Yes
PRPSAP2	chr17: 18,991,705-18,992,729	1,010	AAGAAAGAACAAGTT	Yes
RNF135	chr17: 29,451,155-29,453,977	2,808	GAAATAATTAAAAATC	Yes
CPX-1	chr20: 2,798,148-2,801,409	3,247	AAAAGAACTTGATTT	Yes
REM	chr20: 30,802,228-30,805,279	3,036	AAGATTTGTTTCTTTT	No
SLC2A11	chr22: 22,520,488-22,523,059	2,556	GAAAAAAAATTAACCT	Yes
GPR24	chr22: 39,383,157-39,385,673	2,503	AAAACAAACAAACA	Yes
UTX	chrX: 43,944,720-43,947,538	2,808	TTATCAAATGA	No
OPHN1	chrX: 66,412,207-66,414,450	2,232	TTTTAAACTTTT	No

Table 3-4. Analysis of additional genomic SVA elements. [a] Chromosomal coordinates are given for the July 2003 build (build 34) of the human genome sequence (University of California, Santa Cruz; Kent et al. 2002). All of the elements shown have poly(A) tails at their 3' ends. Additional data are provided in supplemental Table 3-1.

Element type	Chromosomal location*	dbSNP#	INDEL size	Element size	Target site duplication sequence (5' to 3')
5S rDNA	chr3: 12,171,534-12,171,588	ss14942367	55	40	GAAAGGTGAAAAGGA
HERV-K (LTR)	chr8: 7,342,808-7,352,275	ss15143090	9,468	9,463	AAAGGT
U5 RNA	chr9: 195,440,681-195,441,437	ss14936914	76	60	GAGAATCCTGGGTTCT
5S rDNA	chr12: 13,090,355-13,090,458	ss8477420	104	90	GCAAGTGAACATTT
HERV-K (LTR)	chr12: 54,013,482-54,014,455	ss14933585	974	969	GCTAT
U2 RNA	chr13: 108,834,029-108,834,078	ss14938045	50	35	GAAACTGCGAATCCA
MADE1 (Mariner)	chr20: 1,037,602- 1,037,680	ss14935893	79	74	GCAAA

Table 3-5. Insertion polymorphisms generated by other forms of mobilized DNA. [a] Chromosomal coordinates are given for the July 2003 build (build 34) of the human genome sequence (University of California, Santa Cruz; Kent, *et al*. 2002). All of the above elements lack poly(A) tails. Additional data are provided in supplemental Table 1 and dbSNP (http://www.ncbi.nlm.nih.gov/SNP/).

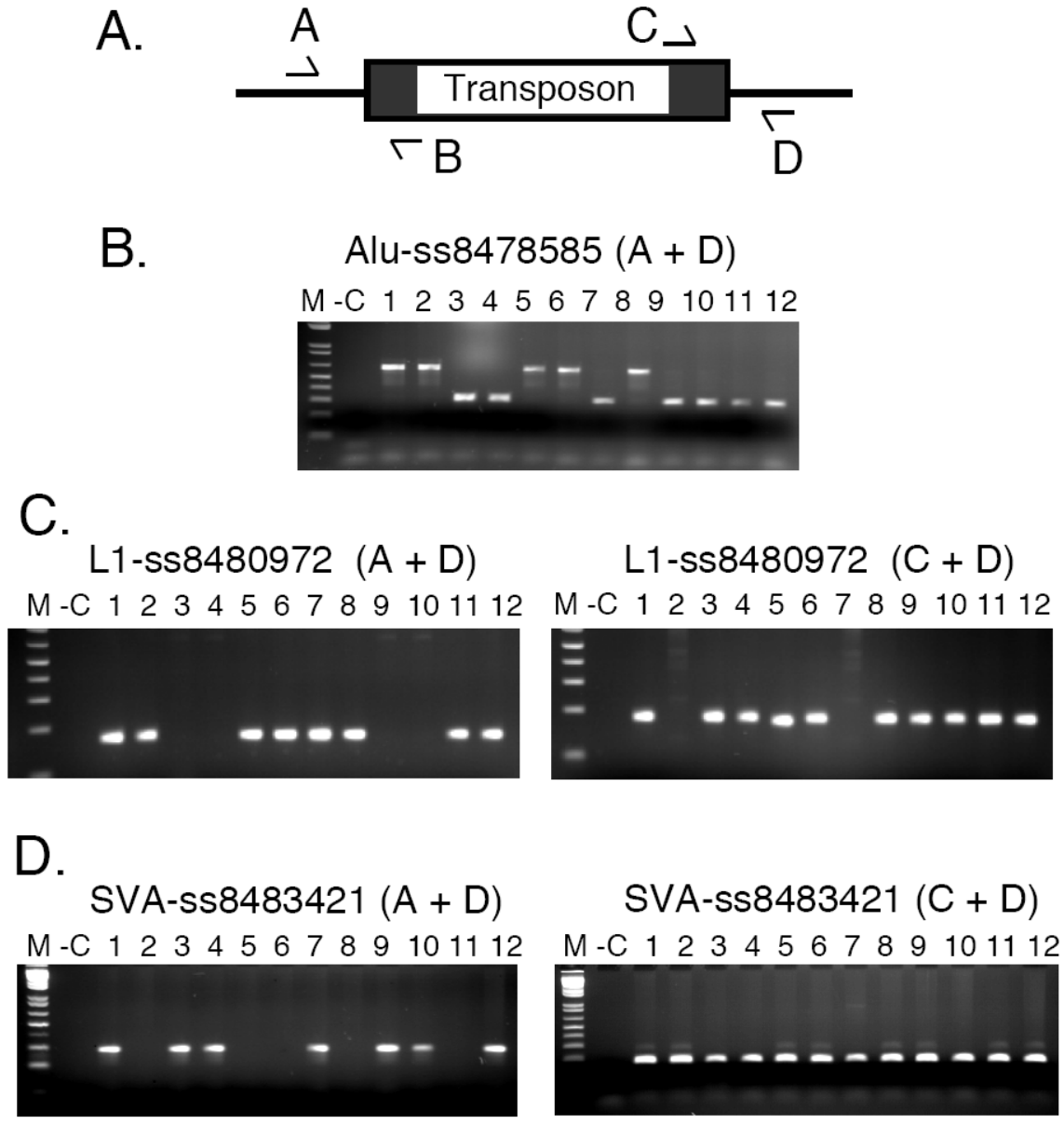


Figure 3-3. PCR validation studies. The strategy for the PCR validation assays is shown along with some examples of these assays. (A) The locations of the four primers (A, B, C, and D) that were used to evaluate transposon polymorphisms by PCR are shown. (B) A typical Alu PCR assay is shown in which primers flanking the transposon (A and D) are used to determine whether a given Alu copy is present in the

(Figure 3-3 cont.) genome of an individual. The larger band is produced when the element is present, whereas the smaller band is produced when the element is absent. Lane M, 1-kb marker; -C, negative control lacking template DNA; lanes 1-12, 12 PCR reactions evaluating DNA samples from the Coriell panel. (C) A typical L1 PCR assay is shown in which two PCRs are performed to identify all of the alleles present in the Coriell panel. The assay on the left uses the A and D primers to identify alleles that lack an L1 insertion, whereas the assay on the right uses primers C and D (or A and B) to identify alleles containing the L1 insertion. These two assays are used together to evaluate whether a given allele is homozygous or heterozygous in a given individual. The lanes are the same as for B. For cases in which the L1 element is relatively short (<2 kb), the allele containing the L1 insertion often was detected in the A plus D assay as well. (D) A typical SVA assay is depicted. These assays are performed the same way as the L1 assays (in C above), using two assays to evaluate all SVA alleles present.

below, we also conducted a systematic validation study to further evaluate the accuracy of our computational pipeline (Figure 3-3; Table 3-6; materials and methods).

Sixty-one transposon insertion polymorphisms were selected from the TSC data set to conduct this validation study (Table 3-6). We focused on the TSC data set because DNA was available for the entire panel of 24 diverse humans used in that study (Collins *et al.* 1999). Since all DNA traces in that experiment were derived from only these 24 individuals (Sachidanandam, *et al.* 2001), any transposon polymorphism that was predicted from the TSC data set also should be found by PCR in at least 1 of these 24 individuals. If the polymorphism was not found in any of the 24, then we would know with certainty that our bioinformatics prediction was incorrect. PCR assays were developed for all 61 of the selected transposon insertion polymorphisms and 12-24 members of the Coriell panel were evaluated to determine whether the polymorphism

could be verified. In all 61 cases, the allele predicted by the trace was confirmed in at least 1 of the 24 individuals of the panel (Figure 3-3 and Table 3-6). Therefore, our methods produced a 100% success rate for this arbitrarily selected sample of 61 TSC polymorphisms, indicating that our methods are highly accurate.

In five cases, we detected only the allele predicted by the trace in the panel of 24 individuals, and the allele predicted by the reference human genome sequence was not detected by PCR (Table 3-6). The most likely explanation for these results is that the person(s) represented by the reference human genome sequence had rare "private" transposon insertions at these positions that were absent from the majority of humans. We (and others) have observed similar results with SNPs identified from the TSC traces (Sachidanandam, *et al.* 2001; Tsui, *et al.* 2003), and Myers, *et al.* (2002) have reported similar results with private alleles of transposon insertions. In cases where they have been examined, these private alleles have been verifiable in the DNA clones that were used to sequence the human genome (Myers, *et al.* 2002).

Estimating the number of transposon insertion polymorphisms in humans

Our study provided a unique opportunity to measure the levels of variation that are caused by transposon insertions in humans. Because our methods utilized DNA sequencing traces, it was possible to determine exactly how many bases of the human

Table 3-6 (following page). Verification of TSC trace predictions by PCR. [a] Chromosomal coordinates are given for the July 2003 build (build 34) of the human genome sequence (University of California, Santa Cruz; Kent et al. 2002). [b] In cases where both chromosomal coordinates are the same, the insertion occurred in the trace. [c] Too short to classify.

dbSNP#	Chromosomal Location[a]	Type of Transposon	Alleles Examined	%Golden Path Allele	%Trace Allele
ss8475858	chr1: 35,902,357-35,902,357[b]	Alu Ya5	48	4	96
ss8475874	chr1: 43,270,765-43,271,100	Alu Y	46	11	89
ss8476214	chr1: 180,191,288-180,191,688	Alu Yg6	46	46	54
ss8480173	chr2: 184,452,525-184,452,867	Alu Ya5	46	4	96
ss8481265	chr3: 117,596,173-117,596,496	Alu Ye5	44	0	100
ss8481475	chr3: 182,282,816-182,283,143	Alu Y	16	63	38
ss8481677	chr4: 36,367,085-36,367,417	Alu Ya4	20	30	70
ss8481874	chr4: 98,569,039-98,569,364	Alu Ya5	44	43	57
ss8481886	chr4: 100,528,683-100,528,998	Alu Ya5	48	0	100
ss8481943	chr4: 127,436,762-127,437,090	Alu Ya4	44	11	89
ss8482744	chr5: 174,741,730-174,742,046	Alu Y	14	21	79
ss8482858	chr6: 20,873,358-20,873,677	Alu Ya5	32	44	56
ss8482986	chr6: 52,830,475-52,830,475	Alu Y	48	96	4
ss8483054	chr6: 74,416,861-74,417,166	Alu Ya5	36	14	86
ss8483191	chr6: 116,796,384-116,796,702	Alu Yi6	14	14	86
ss8483704	chr7: 94,998,283-94,998,613	Alu Ya4	32	22	78
ss8484187	chr8: 76,176,865-76,177,188	Alu Ya5	48	40	60
ss8484397	chr8: 140,446,315-140,446,646	Alu Yg6	48	4	96
ss8484534	chr9: 30,481,284-30,481,594	Alu Ya5	44	20	80
ss8484574	chr9: 41,460,809-41,461,130	Alu Ya5	46	2	98
ss8484611	chr9: 74,581,632-74,581,947	Alu Yb8	48	33	67
ss8476640	chr10: 53,785,800-53,786,116	Alu Y	10	50	50
ss8477096	chr11: 42,428,408-42,428,718	Alu Yb9	20	5	95
ss8477278	chr11: 102,948,539-102,948,851	Alu Yd8	48	0	100
ss8477519	chr12: 45,361,618-45,361,965	Alu Yf1	48	0	100
ss8477629	chr12: 89,348,858-89,349,061	Alu Ya4/Ya5	48	4	96
ss8478585	chr15: 79,108,405-79,108,719	Alu Ya5	48	29	71
ss8478989	chr17: 13,905,127-13,905,415	Alu Y	44	9	91
ss8480426	chr20: 11,512,266-11,512,588	Alu Sx	26	31	69
ss8480530	chr20: 53,978,074-53,978,387	Alu Ya5	48	4	96
ss8480534	chr20: 55,148,643-55,148,959	Alu Y	20	10	90
ss8480629	chr21: 29,522,273-29,522,612	Alu Yb8	46	17	83
ss8484940	chrX: 115,208,464-115,208,782	Alu*	20	60	40
ss8480093	chr2: 160,351,662-160,354,192	L1 Ta-0	20	10	90
ss8480896	chr3: 7,322,025-7,322,322	L1 pre TA	20	40	60
ss8480972	chr3: 33,524,976-33,525,861	L1 Ta-1nd	20	60	40
ss8481820	chr4: 78,694,670-78,694,935	L1 pre TA	12	33	67
ss8482149	chr4: 182,572,693-182,574,194	L1 Ta-1	48	0	100
ss8482285	chr5: 24,416,029-24,418,937	L1 Ta-1	18	11	89
ss8482213	chr5: 2,021,152-2,024,339	L1 Ta-1d	20	50	50
ss8483298	chr6: 153,060,963-153,064,780	L1 Ta-1d	20	30	70
ss8483034	chr6: 66,255,638-66,257,631	L1 Ta-1	20	25	75
ss8483013	chr6: 57,469,905-57,475,956	L1 PA3	48	0	100
ss8484339	chr8: 126,551,711-126,557,723	L1 Ta-1d/Ta-0	20	75	25
ss8484350	chr8: 129,421,743-129,427,855	L1 Ta-1d	16	6	94
ss8484672	chr9: 96,187,490-96,188,291	L1 pre TA	18	11	89
ss8477608	chr12: 82,346,513-82,347,076	L1 Ta-1	16	12.5	87.5
ss8477509	chr12: 40,524,321-40,525,144	L1 Ta-1	18	22	78
ss8478229	chr14: 49,510,268-49,510,585	L1 pre TA	20	90	10
ss8478193	chr14: 38,088,431-38,090,252	L1 pre TA	20	60	40
ss8478558	chr15: 68,737,658-68,743,326	L1 Ta-1d/Ta-0	20	75	25
ss8479344	chr18: 50,014,733-50,014,733	L1 Ta	18	39	61
ss8479308	chr18: 39,283,540-39,284,523	L1 Ta-1	20	20	80
ss8480503	chr20: 42,710,042-42,711,239	L1 Ta-1	20	70	30
ss8481522	chr3: 195,440,681-195,441,437	SVA	48	46	54
ss8483321	chr6: 158,457,569-158,458,368	SVA	48	17	83
ss8483556	chr6: 29,793,437-29,796,056	SVA	48	4	96
ss8483421	chr7: 10,248,637-10,250,470	SVA	48	67	33
ss8483579	chr7: 55,028,493-55,030,810	SVA	44	32	68
ss8478175	chr14: 33,497,755-33,499,561	SVA	48	75	25
ss8477420	chr12: 13,090,355-13,090,458	5S	48	67	33

genome were sampled for a given trace experiment. In the case of the TSC experiment, 989,283,997 bp were sampled (equivalent to 30% of the haploid human genome). Similarly, 2,271,983,242 bp were sampled in the WGS experiment, equivalent to ~69% of the human genome. Therefore, it was possible to normalize the data from these experiments to a genome size of 3.3 billion base pairs (100%) to estimate the total number of transposon insertion polymorphisms that were present in the average haploid genome in our study. By doubling these estimates, we determined that the average (diploid) human in our study harbored ~1283 Alu insertion polymorphisms, 180 L1 polymorphisms, 56 SVA polymorphisms, and 17 other polymorphisms (Table 3-7). The TSC and WGS populations gave estimates that generally differed by less than twofold with the WGS giving a higher estimate. Given that the WGS data were generated from eight African-Americans, these results are consistent with the observation of higher levels of genetic diversity in African populations (International HapMap Consortium 2003).

We also used our data to estimate the total number of common transposon insertion polymorphisms that are present in human populations. We detected 26% of the 660 polymorphic insertions of Alu Ya5 estimated to exist in human populations by Carroll, *et al.* (2001). Similarly, we identified 35% of the 370 polymorphic insertions of Alu Yb8 estimated to exist in humans by Carroll, *et al.* (2001). We also identified 25% of the 234 polymorphic L1-Ta insertions predicted by Myers, *et al.* (2002). Thus, using the Caroll, *et al.* and Myers, *et al.* studies to calibrate our study, we conclude that we are detecting between 25 and 35% of a given class of transposon insertion polymorphisms.

Therefore, given that we identified 605 nonredundant transposon polymorphisms, we estimate that human populations harbor a total of 1730-2420 common transposon insertion polymorphisms (for an average estimate of 2075).

Polymorphism frequencies for Alu, L1, and SVA

Since we recovered data on Alu, L1, and SVA insertion polymorphisms using a single method, we could calculate the relative polymorphism frequencies for these elements (Table 3-7). To perform these calculations, we compared the number of polymorphisms that were identified for each transposon to the genomic copy numbers for each element (Table 3-7). We found that the average polymorphism frequency for all copies of L1 was the lowest, at 0.00018 (one polymorphic L1 insertion per 5556 copies in the genome; Table 3-7). The average polymorphism frequency for all genomic Alu copies likewise was relatively low at 0.00058 (one polymorphic insertion per 1724 copies in the genome). The average polymorphism frequency for SVA, in contrast, was an order of magnitude higher, at 0.0057 (one polymorphic insertion per 175 copies). Therefore, SVA is the most polymorphic element in humans and is likely to be one of the youngest elements to expand in the human genome. Nevertheless, when the L1-Ta, Alu Ya5, and Alu Yb8 subfamilies were examined separately from all L1 and Alu copies, these younger L1 and Alu subfamilies had even higher levels of polymorphism frequencies than SVA (Table 3-7). The L1-Ta subfamily, for example, with ~1040 copies in the diploid genome, has the highest polymorphism frequency at 0.161 (one insertion per 6.2 copies; Table 3-7). The Alu Ya5 and Alu Yb8 subfamilies likewise have much higher

No. polymorphisms per average diploid human in group (calculated from Table 1):

Element	TSC	WGS	Average
Alu-total	1,153.9	1,411.3	1,282.6
Alu Ya5	446.9	428.9	437.9
Alu Yb8	200.1	382.5	291.3
L1-total	173.4	185.5	179.5
L1-Ta	166.8	168.1	167.5
L1-other	6.7	17.4	12.1
SVA-total	40.0	72.5	56.3
Other	13.3	20.3	16.8

No. element copies in the human genome:

Element	Haploid	Diploid	Reference
Alu-total	1,100,000	2,200,000	Lander et al. 2001
Alu Ya5	2,640	5,280	Caroll et al. 2001
Alu Yb8	1,852	3,704	Caroll et al. 2001
L1-total	500,000	1,000,000	Lander et al. 2001
L1 (Ta)	520	1,040	Myers et al. 2002
SVA-total	5,000	10,000	Strichman-Almashanu et al. 2001

Average polymorphism frequencies per average diploid human in group:
(No. polymorphisms listed above/ total copies in genome listed above)

Element	TSC		WGS		Average	
Alu-total	0.00052	(1 per 1,923)	0.00064	(1 per 1,563)	0.00048	(1 per 1,724)
Alu Ya5 only	0.085	(1 per 11.8)	0.081	(1 per 12.3)	0.068	(1 per 12.0)
Alu Yb8 only	0.054	(1 per 18.5)	0.103	(1 per 9.7)	0.064	(1 per 12.7)
L1-total	0.00017	(1 per 5,882)	0.00019	(1 per 5263)	0.00015	(1 per 5,556)
L1-Ta only	0.160	(1 per 6.3)	0.162	(1 per 6.2)	0.136	(1 per 6.2)
SVA-total	0.0040	(1 per 250)	0.0073	(1 per 137)	0.0057	(1 per 175)

Table 3-7. Average polymorphism frequencies in humans (diploid).

polymorphism frequencies than all genomic Alu elements (Table 3-7). Therefore, the most active subfamilies of L1 and Alu have the highest polymorphism frequencies, followed by SVA. Like these other elements, SVA might also harbor extremely polymorphic subfamilies that remain to be discovered but are as polymorphic for insertion as the L1-Ta, Alu Ya5, and Alu Yb8 subfamilies. Alternatively, SVA might have a uniformly lower rate of polymorphism but collectively produces relatively high

levels of genetic variation through its higher copy number (SVA has almost 10 times the number of L1-Ta copies). Either of these models would account for the relatively high levels of genetic variation that are caused by SVA insertion polymorphism (Tables 1 and 7).

Discussion

The spectrum of mobile DNA in humans: In an effort to measure the overall levels of genetic variation that are caused by human transposons, we have developed a new method to broadly detect transposon insertion polymorphisms of all kinds in humans. Our strategy was highly efficient and led to the identification of 605 nonredundant transposon insertion polymorphisms in 36 diverse humans. Since the majority of these insertion polymorphisms had not been identified previously, our method was highly successful at discovering novel transposon polymorphisms. In fact, we estimate that our collection of 605 polymorphisms represents ~25-35% of all common transposon insertion polymorphisms in human populations (see below). Our strategy, in principle, now could be used to identify all of the common transposon insertion polymorphisms that exist in human populations. Together with all previously identified Alu and L1 insertion polymorphisms, our 605 insertions provide significant progress toward this goal. Approximately 20 million additional human traces (beyond the 16.4 million used here) currently are available from SNP discovery projects, and more traces are being generated by ongoing SNP discovery projects daily. Another ~17 million chimp traces are available that could be used to identify transposon insertion polymorphisms in the chimp genome

relative to the human genome. Thus, our method is likely to be useful in humans as well as other organisms.

Unlike previous strategies, our polymorphism discovery strategy yielded data regarding the relative levels of genetic variation from all classes of transposons. The three most abundant classes of insertion polymorphisms in our study were Alu, L1, and SVA insertions. Although Alu and L1 insertion polymorphisms were expected to occur at high frequencies, no studies had been conducted previously to measure the polymorphism frequency of the SVA element or the remaining elements in the human genome. Therefore, our method has revealed that three transposons, Alu, L1, and SVA, are highly polymorphic in humans, and that these three elements together provide the bulk of genetic variation that is caused by transposon insertions in humans (Tables 1 and 7). Our data also indicate that few, if any, insertion polymorphisms exist for the remaining classes of elements in the human genome.

Most human transposon families are not highly polymorphic

As mentioned above, an interesting finding of our study is that many transposon families in humans have not generated insertion polymorphisms to any great extent in recent history. Although Alu, L1, and SVA account for a little more than 30% of the human genome sequence, a total of 44% of the genome sequence is occupied by transposons and transposon-like repetitive elements. Therefore, ~14% of the human genome is occupied by essentially extinct transposon-like families that contain mostly (or totally) inactive, fixed transposon alleles. Our study does not necessarily indicate that these elements are

completely inactive, since we have sampled only 36 human genomes. Therefore, it is likely that we have identified only the most polymorphic classes of elements in the genome, and we may have missed polymorphic copies that occur at lower frequencies within smaller families. Such elements would be of great interest, and we do not rule out the existence of these elements, particularly since the heterochromatic regions of the human genome remain unsequenced. For example, our data indicate that elements such as HERV-K have been mobile recently enough to generate polymorphic insertions in human populations (Table 3-5). Although such elements do not generate a great deal of genetic variation, they would be of great interest if at least some of the polymorphic copies have retained the ability to function as autonomous retrotransposons. Additional studies will be required to determine whether the full-length HERV-K copy discovered in this study (ss15143090, Table 3-5) remains actively mobile today.

Alu and L1 insertion polymorphisms

Our results regarding Alu element polymorphisms generally are in good agreement with a large number of previous studies examining the young Alu Y subfamilies of the human genome (reviewed in Batzer and Deininger 2002). Because we detected all Alu subfamilies with a single method, we also were able to measure the relative levels of variation that are caused by each subfamily (Tables 1, 2, and 7). Consistent with previous studies, we found that Alu Ya5 and Alu Yb8 insertion polymorphisms are highly abundant in humans. We also found that Alu Y, Alu Yc1, and Alu Ya4 insertion polymorphisms are moderately abundant in human populations (Table 3-2). The remaining Alu Y insertions were less abundant and were distributed among 15 different

Alu Y subfamilies (Table 3-2). Our results further indicate that additional polymorphic Alu Y families of any significant size are not likely to exist in humans. Finally, we unexpectedly found a small number of ancient Alu S insertion polymorphisms in humans (Table 3-2).

Our results regarding L1 element polymorphisms likewise are in good agreement with previous studies examining L1-Hs insertion polymorphisms in humans (Sheen *et al.* 2000; Ovchinnikov, *et al.* 2001; Myers, *et al.* 2002; Badge, *et al.* 2003; Brouha, *et al.* 2003). However, we also found that older L1-P (primate) insertions represent a significant source of human genetic variation (Ovchinnikov, *et al.* 2002). In fact, L1-P insertions represented close to 10% of all L1 insertion polymorphisms in our study (Table 3-2). Because we detected all L1 subfamilies with a single method, it was possible to assess the relative levels of variation that were caused by each subfamily and integrate these values with all other elements in the genome (Tables 1, 2, and 7).

SVA is highly polymorphic in humans

Since the initial discovery of the SVA element, a number of retrotransposon-like insertions have been reported that might have been caused by SVA retrotransposition events (Hassoun, *et al.* 1994; Kobayashi, *et al.* 1998; Rohrer, *et al.* 1999). One of the best candidates in this regard is a 3-kb retrotransposon insertion in the Fukutin gene that was reportedly responsible for 70% of the Fukuyama type muscular dystrophy in Japan (Kobayashi, *et al.* 1998). The element described in that report has some of the features of a de novo SVA insertion; however, it was not referred to as an SVA element in that

article, and the sequence of the insertion was not provided (Kobayashi, *et al.* 1998). Two additional retrotransposon insertions also have been referred to as SVA elements by Ostertag and Kazazian (2001) and Ostertag, *et al.* (2003). In one of these cases, the element was reported in the original study to be a SINE-R element rather than an SVA element (Rohrer, *et al.* 1999). Since SINE-R is itself a retrotransposon and also a component of the SVA element, the insertion could be a SINE-R element or a truncated SVA. In the final case, no SVA sequences were actually present within the DNA insertion that was identified (Hassoun, *et al.* 1994), and the inserted DNA segment was proposed to have been mobilized by a 3'-transduction event sponsored by an adjacent SVA element (Ostertag, *et al.* 2003).

We now provide clear evidence for the existence of at least 39 de novo SVA element insertions in humans (Tables 3 and 4). For these 39 insertions: (i) both empty and SVA-occupied sites were identified in the human genome in different individuals, (ii) the SVA copies ended in poly(A) tails, and (iii) the inserted copies were precisely flanked by new target site duplications. Thus, our study indicates that SVA insertion polymorphisms are highly abundant in humans. In fact, SVA insertion polymorphisms provide about one-third the level of genetic variation that is caused by L1 insertions (Tables 1 and 7). Finally, our data also indicate that SVA has amplified independently and perhaps at different rates in the genomes of humans and chimps. Approximately 79% of the SVA insertions in the human genome are absent from the equivalent positions of the chimp genome, indicating that these insertions occurred relatively recently in human history (within the past ~6 million years).

Since a high rate of polymorphism is a hallmark feature of an active transposon, our data suggest that SVA might be actively mobile in the human genome today. Because SVA does not encode any obvious proteins of its own (it lacks substantial open reading frames), it is likely to be a nonautonomous element that relies upon another transposon for its own transposition. Several aspects of our SVA insertions suggest that they might be mobilized by L1 elements *in trans* by the same mechanism that mobilizes Alu elements (Dewannieux, *et al.* 2003). For example, the target site duplications of our SVA insertions closely resemble those of Alu and L1 elements in length and sequence (Tables 3 and 4). Our SVA insertions also have poly(A) tails, indicating that they are poly(A) retrotransposons. Finally, many of our polymorphic SVA insertions had 5'-truncations that were similar to the 5'-truncations of L1 elements. Given these similarities to Alu and L1 elements, SVA is likely to be mobilized in trans by L1encoded proteins (Esnault, *et al.* 2000; Ostertag and Kazazian 2001; Wei, *et al.* 2001; Dewannieux, *et al.* 2003; Ostertag, *et al.* 2003). Therefore, it appears that all three classes of highly polymorphic elements in our study (Alu, L1, and SVA) were generated by the L1 retrotransposition machinery in *cis* or in *trans*.

Estimating the levels of variation caused by transposon insertions in human populations

Our method provided a unique opportunity to measure the levels of genetic variation that are caused by transposon polymorphisms in humans. The measure of variation that is most commonly reported for transposons is the percentage of copies that are polymorphic

in at least one individual of a population (Sheen, *et al.* 2000; Carroll, *et al.* 2001; Ovchinnikov, *et al.* 2001, 2002; Roy-Engel, *et al.* 2001; Myers, *et al.* 2002; Abdel-Halim, *et al.* 2003; reviewed in Ostertag and Kazazian 2001 and Batzer and Deininger 2002). This approach is useful from the viewpoint of assessing whether a given transposon family is polymorphic in populations; however, it tends to overestimate the levels of genetic variation that are caused by transposons. This is because high-frequency and low-frequency alleles are counted equally with this type of measurement. In contrast, we measured the polymorphism rates in a manner that included the allelic frequencies of the transposon alleles. High-frequency alleles are encountered more often than rare alleles in our trace experiments, and therefore our method naturally takes into account the allelic frequencies of the transposon insertions. Thus, with this approach, we have been able to estimate the levels of genetic variation that are caused by transposon insertions in populations. As a consequence of factoring in the allelic frequencies, our estimates for polymorphism rates are four-to fivefold lower than those reported previously. For example, 25% of the Alu Ya5 copies previously were reported to be polymorphic in at least one individual of a population of 80 humans (Carroll, *et al.* 2001). We now estimate that ~6.8% of the Alu Ya5 copies are polymorphic in the average human of our study (Table 3-7). Likewise, 45% of the L1-Ta element copies previously were reported to be polymorphic in at least one member of a large population (Myers, *et al.* 2002), whereas we now calculate that ~13.6% of the L1-Ta copies are polymorphic in the average human of our study (Table 3-7).

We also generated an estimate of the total number of common transposon insertion polymorphisms that exist in human populations. We generated this estimate by comparing the number of insertion polymorphisms discovered in our study for a given element such as Alu Ya5 to the total number expected. For example, Carroll, *et al.* previously had predicted that ~25% of the 2640 genomic Alu Ya5 copies were polymorphic for insertion in at least one member of a population of 80 diverse humans (Carroll, *et al.* 2001). Therefore, since we identified 170 Alu Ya5 insertion polymorphisms, we determined that we had identified 26% of all expected Alu Ya5 insertion polymorphisms in humans. By calibrating our study with several of these previous studies, we estimate that our 605 transposon insertions represent between 25 and 35% of all common transposon insertion polymorphisms in human populations. Therefore, on the basis of these comparisons, human populations are estimated to harbor between 1730 and 2420 common transposon insertion polymorphisms (for an average of 2075). Together with our 605 polymorphisms, less than half of these polymorphisms have been identified to date, indicating that additional efforts will be required to identify the full set of polymorphic transposon insertions (i.e., the "transposon insertion polymorphome") in humans.

Together with previous studies, our analysis indicates that SNPs, indels, and transposon insertion polymorphisms represent significant sources of genetic variation in humans. Human populations are estimated to harbor ~10 million common SNPs (Judson, *et al.* 2002), ~2 million common indels (our unpublished data), and ~2000 common transposon insertion polymorphisms (this study). Therefore, with 10 million bases of

variation, SNPs account for the majority of common human genetic variation, followed by indels and then transposon insertion polymorphisms. On the other hand, if we assume that the average transposon polymorphism in humans is ~500-1000 bp in length, then the total amount of variation caused by common transposon insertions is 1-2 million base pairs (equivalent to 10-20% of the base pair variation caused by SNPs). Thus, in terms of the number of base pairs, common transposon insertions cause significant levels of human genetic variation. Moreover, humans also are likely to harbor >10 million rare private transposon insertions (cases in which only one or a few individuals have the insertion). Therefore, transposon insertion polymorphisms cause significant levels of human variation. A number of studies now have shown that SNPs, indels, and transposon insertions all may cause serious phenotypic changes when positioned at critical sites within genes (Collins, *et al.* 1987; Kazazian, *et al.* 1988; Sachidanandam, *et al.* 2001). Nevertheless, a comprehensive map of genetic variation that integrates SNPs, indels, and transposon insertions currently is lacking. A fully integrated map that includes all forms of genetic variation will be necessary to efficiently identify genetic polymorphisms that influence human phenotypes and diseases.

Acknowlegements

We thank the people at the SNP Consortium, the Sanger Centre, the Baylor Genome Center, and the Whitehead Genome Center for the use of their trace data, and University of California, Santa Cruz for its human genome database. We also thank Shari Corin for helpful advice on this project and for critical review of the manuscript. Finally, we thank Karen Ventii and Summer Goodson for help with some PCR experiments. This work was supported by training grant 2T32GM008490-11 from the National Institutes of Health (E.A.B.), grant 2-80302 from the Emory University Research Council (S.E.D.), grant RSG-01-173-01-MBC from the American Cancer Society (S.E.D.), and grant 1R01HG02898-01A1 from the National Institutes of Health (S.E.D.).

Chapter 4

Recently mobilized transposons in the human and chimpanzee genomes

Ryan E. Mills,[1,2] E. Andrew Bennett,[1,2,3] Rebecca C. Iskow,[1,2,3] Christopher T. Luttig,[1] Circe Tsui,[1,2] W. Stephen Pittard,[1,2,4] and Scott E. Devine[1,2,3]

[1]Department of Biochemistry, [2]Emory Center for Bioinformatics, [3]Graduate Program in Genetics and Molecular Biology, and [4]BimCore, Emory University School of Medicine, Atlanta, Georgia.

2006. ***Am. J. Hum. Genet.*** **78, 671-679**

 DOI:10.1086/501028

Introduction

Transposable genetic elements are abundant in the genomes of most organisms, including humans. These endogenous mutagens can alter genes, promote genomic rearrangements, and may help to drive the speciation of organisms. In this study, we identified almost 11,000 transposon copies that are differentially present in the human and chimpanzee genomes. Most of these transposon copies were mobilized after the existence of a common ancestor of humans and chimpanzees, ~6 million years ago. Alu, L1, and SVA insertions accounted for >95% of the insertions in both species. Our data indicate that humans have supported higher levels of transposition than have chimpanzees during the past several million years and have amplified different transposon subfamilies. In both species, ~34% of the insertions were located within known genes. These insertions represent a form of species-specific genetic variation that may have contributed to the differential evolution of humans and chimpanzees. In addition to providing an initial overview of recently mobilized elements, our collections will be useful for assessing the impact of these insertions on their hosts and for studying the transposition mechanisms of these elements.

Transposable genetic elements collectively occupy ~44%of the human genome (International Human Genome Sequencing Consortium 2001). Although most of these transposons lost the ability to transpose long ago, some copies have transposed in relatively recent human history (Kazazian, *et al.* 1988; Wallace, *et al.* 1991; Moran, *et al.*1996; reviewed by Ostertag and Kazazian 2001; reviewed by Batzer and Deininger 2002; Bennett, *et al.* 2004).These recently mobilized transposons are of great interest for

a number of reasons. First, recent insertions within or near genes may cause phenotypic changes in humans, including diseases (Kazazian, *et al.* 1988; Wallace, *et al.* 1991, reviewed by Ostertag and Kazazian 2001; reviewed by Batzer and Deininger 2002). Several dozen transposon insertions have been identified to date that cause human diseases, and human populations are likely to harbor additional transposon insertions that influence phenotypes as well. Some of these recently mobilized transposons also remain actively mobile today and continue to generate new transposition events elsewhere in the genome (Moran, *et al.* 1996, reviewed by Ostertag and Kazazian 2001; Dewannieux, *et al.* 2003). Active retrotransposons in particular have been observed to be the most potent endogenous mutagens in humans, and these elements continue to generate mutations and genetic variation in human populations (reviewed by Ostertag and Kazazian 2001). In some cases, transposon insertions also may go on to create genomic rearrangements by recombining with other transposon copies (reviewed by Ostertag and Kazazian 2001). Thus, recently mobilized transposons continue to restructure the human genome through a variety of mechanisms.

The completion of a draft chimpanzee genome sequence provided an opportunity to identify these recently mobilized transposons in both humans and chimpanzees (Chimpanzee Sequencing and Analysis Consortium2005). Transposons that inserted into either of these genomes during the past ~6 million years (i.e., since the existence of the most recent common ancestor of humans and chimpanzees) would be expected to be present in only one of the two genomes. We used a comparative genomics approach to identify these recently inserted transposon copies (Figure 4-1). We began by aligning the

sequences of the human and chimpanzee genomes to identify all insertions and deletions (indels).

Results

We screened indels for the presence of transposable elements by comparing each indel to a library of known transposons (Repbase v.10.02) (Jurka 2000). Using this approach, we initially identified a total of 14,783 transposon copies that were differentially present in the two genomes. Many of these copies appeared to be recently mobilized transposon insertions, whereas others were simply transposon copies that happened to be located within larger genomic duplications or deletions in the two genomes.

To identify all of the insertions that were caused by actual transposition events, we next screened our collections for insertions that (1) were precisely flanked by target-site duplications (TSDs) and (2) precisely accounted for a gap in one of the two genomes. Using these criteria, we identified 10,719 insertions of single transposon copies that appeared to have been caused by transposition events. The remaining 4,064 examples lacked TSDs or, in general, did not precisely account for the indels, which suggests that they were caused by alternative mechanisms. Of the 10,719 transposon insertions, 7,786 (72.6%) were found in humans and only 2,933 (27.4%) were found in chimpanzees. Therefore, it appears that transposons have been significantly more active in the human genome during the parallel evolution of these organisms. The different population

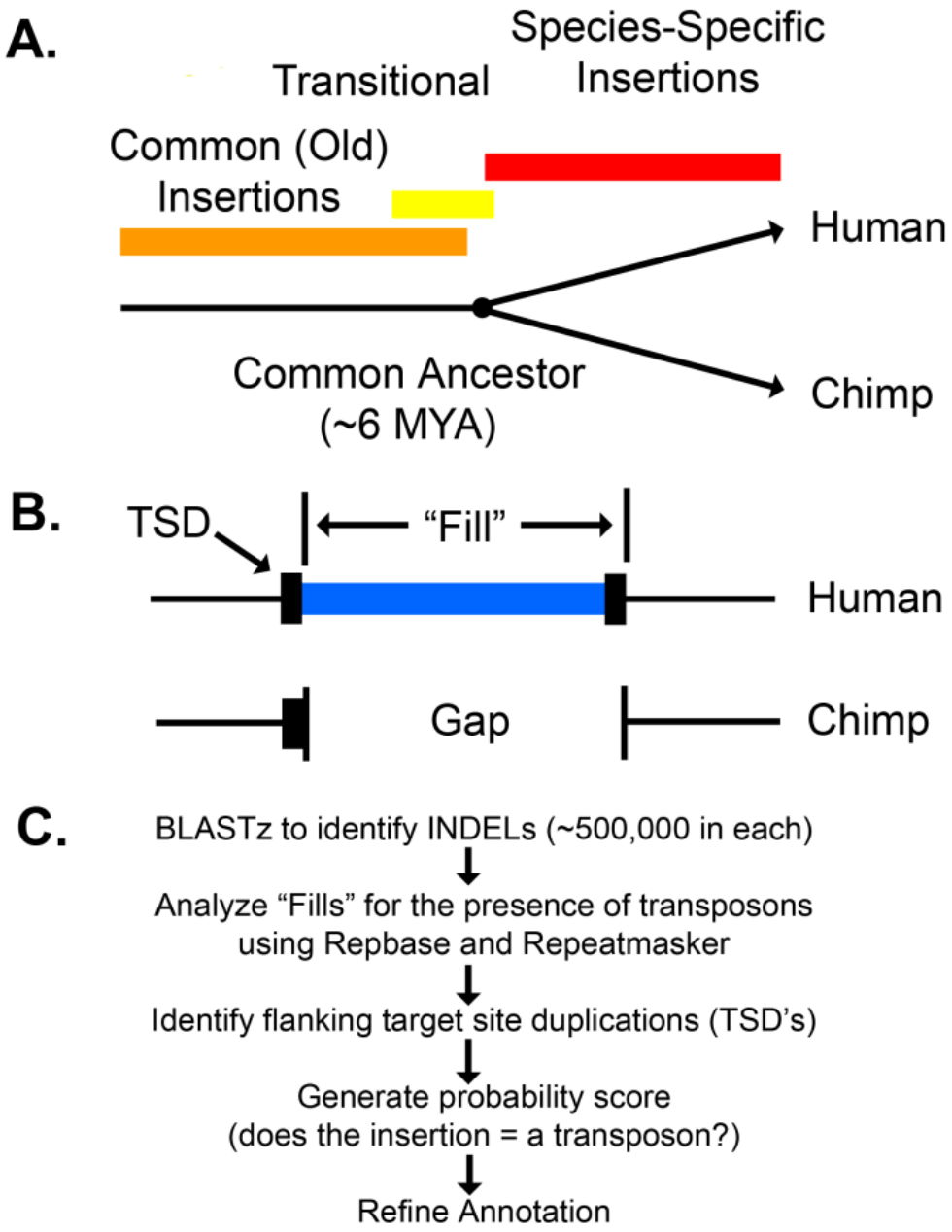


Figure 4-1. Overview of our transposon insertion-discovery pipeline. (A) The time line for speciation of humans and chimpanzees is compared with the time line for the generation of transposon insertions. Common insertions occurred a very long time ago and are fixed in both species. "Species-specific" insertions are differentially present in the two species and occurred mostly during the past ~6 million years. MYA = million years ago. (B) Our strategy for identifying new transposon insertions in humans and chimpanzees. Recently mobilized transposons are flanked by TSDs and are precisely absent from one of the two genomes. One of the two copies of the TSD is actually found within the indel. Thus, the transposon

(Figure 4-1 cont.) plus one TSD copy equals the "fill." (C) Our computational pipeline. The five sequential steps of our computational pipeline for discovering species-specific transposon insertions in humans and chimpanzees are depicted.

dynamics of these organisms during the past several million years also may have helped to shape the final patterns of transposons observed.

The most abundant classes of new transposon insertions in both chimpanzees and humans were Alu, L1, and SVA element insertions, and these three classes collectively accounted for >95% of the recently mobilized transposons in both species (Table 4-1 and Figure 4-2). However, the relative abundance of these elements and their subfamilies differed between the two species (Figure 4-2). Other, less-abundant classes of transposon insertions also were identified in our study. For example, long terminal repeat (LTR) retroelement insertions were observed in both species, including insertions of human endogenous retroviruses (HERVs) and solo LTRs of these elements. Solo LTR insertions have been shown to influence the expression of nearby genes, which makes these insertions of particular interest (Landry and Mager 2003). Also identified were five full-length HERV-K insertions with relatively long ORFs (up to several thousand amino acids in length) that could remain capable of retrotransposition. Insertions of chimpanzee endogenous retroviruses (CERVs) also were identified (Yohn, *et al.* 2005). Finally, mammalian interspersed repetitive elements, copies of satellite DNA flanked by unusual TSDs, and small numbers of other interesting transposable elements were identified in the two species.

humans (5,530) was 3.4 fold higher than the number observed in chimpanzees (1,642). The distributions of these elements among various Alu subfamilies also differed between the two organisms (Table 4-1, Figure 4-2). For example, Alu Ya5, Alu Yb8, Alu Y, and Alu Yc1 were highly abundant in humans, whereas only Alu Yc1 and Alu Y were highly abundant in chimpanzees. Our data indicate that Alu S elements, which have been presumed to have been inactive for the past 35 million years (Johanning, *et al.* 2003); apparently have been active in humans and less active in chimpanzees during the past ~6 million years (Table 4-1). It is possible that some of these older Alu S "insertions" were caused by the precise deletion of Alu S elements from one of the two genomes (van de Lagemaat, *et al.* 2005) or by gene-conversion events (reviewed by Batzer and Deininger 2002). However, these results also are in agreement with recent data from our laboratory, which indicates that a small number of younger Alu S elements are polymorphic in humans and appear to have transposed more recently than the bulk of Alu S elements (Bennett, *et al.* 2004). Overall, our results indicate that the human genome has supported higher levels of Alu retrotransposition and has amplified a different set of Alu elements than has the chimpanzee genome (Table 4-1 and Figure 4-2). These results confirm and extend previous classifications of Alu elements of chimpanzee chromosome 22 (Hedges, *et al.* 2004; Watanabe, *et al.* 2004).

Transposon Class	Human (n = 7,786)	Chimp (n = 2,933)
Alu (All)	**5,530 (71.0%)**	**1,642 (56.1%)**
Alu S	263 (3.3%)	50 (1.7%)
Alu Ya5	1,709 (21.9%)	10 (0.3%)
Alu Yb8	1,290 (16.6%)	9 (0.3%)
Alu Y	484 (6.2%)	360 (12.3%)
Alu Yc1	356 (4.6%)	979 (33.4%)
Alu Yg6	261 (3.4%)	1 (0.1%)
L1 (All)	**1,174 (15.1%)**	**758 (25.9%)**
L1 Hs (Ta)	271 (3.5%)	0 (0.0%)
L1 Hs (Non Ta)	252 (3.2%)	210 (7.2%)
L1 PA2	490 (6.3%)	476 (16.2%)
SVA (All)	**864 (11.1%)**	**396 (13.6%)**
Other	**219 (2.8%)**	**127 (4.4%)**

Table 4-1. Summary of Transposon Insertions

L1 insertions also were abundant in both organisms. In humans, almost 1,200 recently mobilized L1 insertions with TSDs were identified that precisely accounted for gaps in the chimpanzee genome (Table 4-1). These human L1 elements predominantly included members of the L1-Hs and L1-PA2 families (Table 4-1, Figure 4-2) (Boissinot, *et al.* 2000; Brouha, *et al.* 2003). The human L1-Hs elements included members of the pre-Ta, Ta0, and Ta1 subfamilies (grouped together as "L1Hs Ta" in Table 4-1 and Figure 4-2), which are known to be highly active in humans (Brouha, *et al.* 2003). Also identified in humans were additional L1-Hs and L1-PA2 subfamilies that had unique base combinations at the nine key positions described elsewhere (grouped together as "L1-Hs non-Ta" or "L1-PA2" in Table 4-1 and Figure 4-2) (Boissinot, *et al.* 2000; Brouha, *et al.*

2003). These novel subfamilies contained 3-65 copies (The remaining L1 insertions in humans belonged to older L1-PA2, L1-PA3, and L1PA4 groups, Figure 4-2).

The L1 insertions identified in chimpanzees, in contrast, were notably different from those outlined above for humans (Table 4-1, Figure 4-2). For example, fewer recently mobilized L1 insertions were identified in chimpanzees than in humans (758 in chimpanzees vs. 1,174 in humans). Only 4 of the chimpanzee L1 insertions were full-length (compared with ≥200 new full-length insertions in humans), and none of the chimpanzee L1 insertions had intact ORFs. The initial draft sequence of the chimpanzee genome is likely to contain assembly errors that may account for at least some of these observed differences. However, we also observed differences in the L1 subfamilies of these organisms that are unrelated to genome assembly issues. For example, proportionally more L1PA2 insertions and fewer L1-Hs insertions were observed in chimpanzees than in humans (Figure 4-2). Initially, we were surprised to find L1-Hs elements in chimpanzees at all, since these elements were expected to be found only in humans. However, further analysis revealed that most of the L1-Hs elements in chimpanzees actually were “intermediate” elements that matched L1-Hs overall but had ORF1 sequences that were more similar to L1-PA2 elements. Therefore, the L1-Hs family of elements includes subfamilies that are truly human specific as well as other L1-Hs–like elements that are not human specific. We also aligned and analyzed all of our chimpanzee L1 insertions, using ClustalW and PAUP, to determine whether any new L1 subfamilies (equivalent to L1-Ta elements in humans) were present in chimpanzees. In addition, we classified all of our chimpanzee L1 insertions, using the nine key positions that have been used elsewhere to classify human L1 elements (Boissinot, *et al.* 2000;

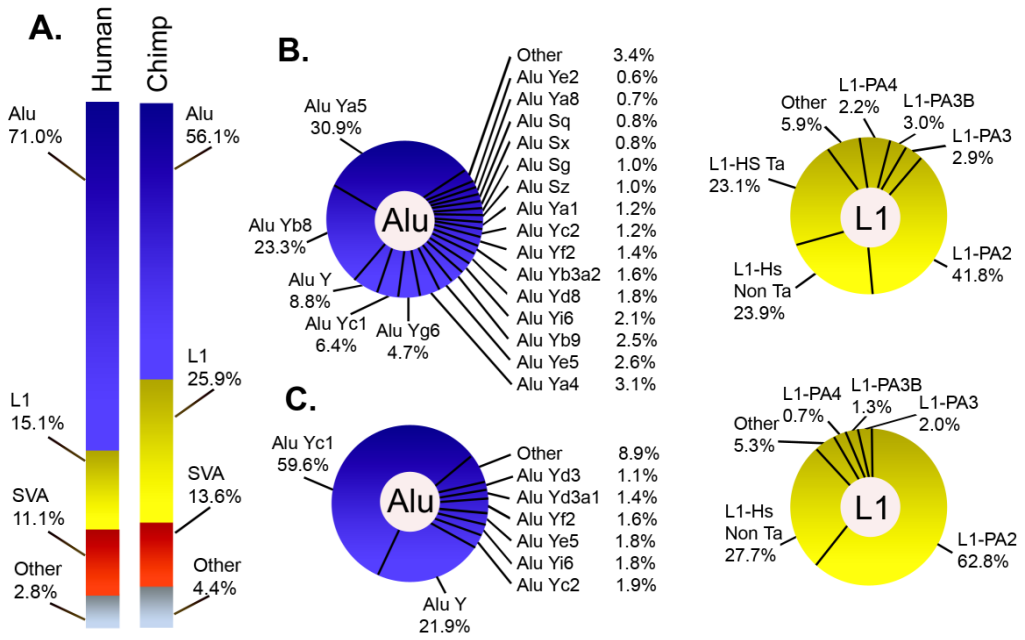


Figure 4-2. Classes of species-specific transposons in humans and chimpanzees. (A) The overall composition of species-specific insertions in humans and chimpanzees. Note that 97.2% of all insertions in humans and 95.6% of all insertions in chimpanzees are Alu, L1, and SVA insertions. (B) The distributions of Alu and L1 subfamilies for humans. (C) The distributions of Alu and L1 subfamilies for chimpanzees. Note that different Alu and L1 subfamilies were amplified in humans (B) and chimpanzees (C).

Brouha, *et al.* 2003). In both cases, we failed to identify any new extended subfamilies of L1 elements within our collection of chimpanzee insertions. Therefore, a dominant class of new offspring elements analogous to L1-Ta elements in humans does not appear to have been produced in recent chimpanzee history.

We next examined all of the existing L1 ORFs in the human and chimpanzee genomes to further characterize possible differences between the L1 elements of these species. We screened the human and chimpanzee genomes for ORFs in all nearly full-

	Human	**Chimp**	
	Full Genome (version hg17)	BACs (260 Mb)	Full Genome (extrapolation)
L1 larger than 5,500	**8,483**	**702**	**8,100**
Intact ORF1 (1,017 bp)	**633**	**20**	**230**
Intact ORF2 (3,828 bp)	**205**	**4**	**46**
Intact ORF1and ORF2	**126**	**2**	**23**

Table 4-2. Analysis of L1 ORFs

length elements (>5,500 bp) and identified 633 L1 elements with intact ORF1 sequences in the human genome but only 39 elements with intact ORF1 sequences in the draft chimpanzee sequence. Moreover, we identified 205 L1 elements with intact ORF2 sequences in the human genome (Table 4-2) but failed to detect intact ORF2 sequences in the draft chimpanzee sequence. These results suggested that functional L1 elements were likely to be rare in chimpanzees. As outlined above, however, it also was possible that the quality of the chimpanzee draft sequence affected our ability to detect ORFs accurately. We determined that the sequence quality of the >5,500 bp L1 elements in chimpanzees had average scores that generally were high (>40 Phred scores) (Ewing, *et al.* 1998; Kent, *et al.* 2002). However, single bases of low quality (<10 Phred scores) (Ewing, *et al.* 1998) also were distributed throughout the draft sequence at sporadic intervals (Kent, *et al.* 2002). These single bases, although rare, often resulted in frame-shifts. Therefore, it was possible that these sporadic low-quality bases were interfering with our ability to detect ORFs accurately.

To independently examine the frequency of intact L1 ORFs in chimpanzees, we analyzed all L1 elements that were present in finished BAC sequences that had been generated for the chimpanzee genome project. Approximately 260 Mb of finished sequence was available in GenBank from chimpanzee BACs, and the quality of these sequences was identical to that of the finished human genome sequence. We identified a total of two L1 elements in these BACs that were >5,500 bp in length and also had intact ORFs (Table 4-2). Neither of these two elements was present in the draft sequence, so it is unclear whether the quality of the draft affected our ability to detect these ORFs. Nevertheless, extrapolation of these results to the whole genome (3,000 Mb) predicts that chimpanzees harbor ~23 full-length L1s with intact ORFs (compared with 126 in humans, Table 4-2). Thus, the chimpanzee draft sequence indicates that L1 elements with intact ORFs are up to 20-fold less abundant in chimpanzees than in humans, whereas the BACs indicate that such elements are ~5.5 fold less abundant (Table 4-2). Since the draft chimpanzee sequence contains some sporadic low-quality bases, the BAC estimate is likely to be more accurate. Both of these estimates indicate that functional L1 elements are less abundant in chimpanzees than in humans.

We next examined the L1 ORF1 and ORF2 coding regions from chimpanzees to determine whether the encoded proteins are likely to remain active today. Brouha, *et al.* (2003) showed elsewhere that human L1 elements that differ by an average of only 21 nucleotide changes from an active human L1 consensus were inactive. Thus, even

elements that were >99% identical to this consensus could be inactive, and no human elements that were <99% identical were found to be active. We determined that the human genome contains at least 119 elements with >99% nucleotide identity to the active human L1 consensus (Brouha, *et al.* 2003) within the regions encoding ORF1 and ORF2. In contrast, no L1 ORFs in the chimpanzee genome or BACs had >99% identity to this active human L1 consensus. We cannot rule out the possibility that some of the chimpanzee L1 elements that are <99% identical to the human L1 consensus could remain active. Other "hot L1" elements might have evolved separately in chimpanzees that are >1% variant from the most-active human elements. However, since we failed to detect extended subfamilies of L1-PA2 or L1-Hs elements within our collection of chimpanzee insertions (analogous to the L1-Ta subfamily in humans), such elements generally would be present at low copy numbers in the chimpanzee genome. Thus, the landscape of potentially functional L1 elements in chimpanzees appears to be quite different from the landscape of active L1 elements in humans.

To verify these ORF results, we used an ORF1 trapping method to recover full-length ORF1 sequences from L1 elements in humans and chimpanzees (Ivics and Izsvak 1997). We recovered and sequenced 41 intact ORF1 sequences from humans and 51 intact ORF1 sequences from chimpanzees and observed results that were very similar to those obtained through the computational methods described above. None of the intact chimpanzee ORF1 sequences recovered through ORF1 trapping were >99% identical to the active human L1 consensus, whereas 17 (41%) of the 41 intact human ORF1 sequences were >99% identical to the active human L1 consensus. Almost all of the

intact ORF1 sequences trapped from chimpanzees (48/51, 94%) were L1-PA2 ORF1 sequences. In contrast, only 21 (51%) of the 41 intact ORF1 sequences recovered from humans were L1-PA2 ORF1 sequences and 18 (44%) were L1-Hs sequences. Thus, our ORF1-trapping experiments confirmed that the most recently active elements in chimpanzees (i.e., those with intact ORFs) contained ORF1 sequences that were divergent from the active L1 consensus in humans (Brouha, *et al.* 2003).

Our method for trapping ORF1 in humans and chimpanzees employed the pβFUS plasmid (Ivics and Izsvak 1997). Briefly, full-length ORF1 sequences were amplified from human and chimpanzee genomic DNA (NA1MR91 and NA03448A, respectively [Coriell Cell Repository]) using PCR. The PCR primers were identical for humans and chimpanzees, and had the following sequences 5'-CCTGATCTGCGGCCGCATGGGGAAAAAACAGAACAGAAAAACTGG-3' and 5'-CGTCCGAACGATATCCATTTTGGCATGATTTTGCAGCGGCTGG-3'.We used a combination of human and chimpanzee ORF1 sequences to design these primers. The ORF1 sequences identified in our chimpanzee BAC experiments were aligned to generate a consensus sequence using ClustalW. This sequence was compared to the human L1 consensus, and we determined that the primers chosen were conserved in human L1 sequences. Finally, we compared the candidate primer regions with L1-PA2, L1-PA3, and L1-PA4 elements and determined that the primer sequences also were completely conserved in these elements. Thus, the primers chosen were capable of amplifying a wide spectrum of ORF1 sequences in both humans and chimpanzees (including L1-Hs, L1-PA2, L1PA3, and L1-PA4 elements). The *Not*I and *Eco*RV restriction sites that were

introduced for cloning purposes are underlined. PCR products were cut with these enzymes and ligated to *Not*I/*Sma*I-digested pβFUS, such that the complete ORF1 sequence would be in frame with the AUG-less *lacZ* in the plasmid. Recombinants were identified on LB medium containing X-gal (recombinants with ORFs were blue, whereas those without ORFs and the empty vector alone were white). DNA was prepared and sequenced at Agencourt Biosciences using the primers 5'-CCAGTCACGTTGTAAAACGAC3' and 5'-CTAGGCCTGTACGGAAGTGTTAC-3'. High-quality sequences were analyzed and assembled using Sequencher version 4.1.2.

In addition to Alu and L1 insertions, we also found that SVA elements have been highly active in humans and chimpanzees (Table 4-1). In fact, SVA insertions were almost as abundant as L1 insertions in humans during the past ~6 million years (Table 4-1). SVA is an unusual composite element that contains four components: (1) a tandem repeat of TCTCCC(n), (2) an unusual Alu element in reverse orientation, (3) a central variable-number-of tandem-repeat (VNTR) region that is rich in CpG sequences, and (4) a SINE-R sequence that was derived from an LTR element (Shen, *et al.* 1994). SVA ends with a poly (A) tail and is flanked by TSDs that closely resemble the TSDs of Alu and L1 elements (Ostertag, *et al.* 2003; Bennett, *et al.* 2004). SVA recently was found to be highly polymorphic among humans (Bennett, *et al.* 2004), and a few instances have been reported of SVA insertions causing diseases (Kobayashi, *et al.* 1998; Ostertag, *et al.* 2003). Our study now provides further evidence that SVA has been actively mobile in relatively recent primate history and may remain active today.

The ORF1 and ORF2 proteins of L1 elements perform a specific retrotransposition mechanism known as "target-primed reverse transcription" (TPRT) (Luan *et al.* 1993), in which L1 mRNAs are copied into cDNAs and integrated into the genome (reviewed by Ostertag and Kazazian 2001). Alu RNAs (and other cellular RNAs) can compete for the L1 machinery during the TPRT process, which leads to the retrotransposition of these alternative RNAs instead of the normal L1 mRNAs (Esnault, *et al.* 2000; Wei, *et al.* 2001; Dewannieux, *et al.* 2003). This "trans" mechanism of retrotransposition is thought to have led to the massive expansion of Alu (Dewannieux, *et al.* 2003) and SVA (Ostertag, *et al.* 2003; Bennett, *et al.* 2004) elements in the human genome. Therefore, if L1 elements are indeed less functional in chimpanzees, as predicted above (Table 4-2), we likewise might expect to see fewer Alu and SVA insertions in the chimpanzee genome. Table 4-1 and Figure 4-3 show that this is, in fact, the case. Since other factors also influence the amplification rates of Alu (and probably SVA) elements, these differences may not be totally caused by lower levels of L1 activity in chimpanzees. It is possible, for example, that humans had a larger number of potentially active Alu and SVA source elements than did chimpanzees in recent history. However, when combined with our other data demonstrating that chimpanzees (1) have fewer full-length L1 insertions than humans, (2) have fewer L1 elements with intact ORFs than humans, (3) have ORF sequences that are divergent from active human L1 elements, and (4) lack extended subfamilies of new insertions, these data collectively indicate that chimpanzees are likely to have supported lower levels of L1 activity in recent history compared with humans.

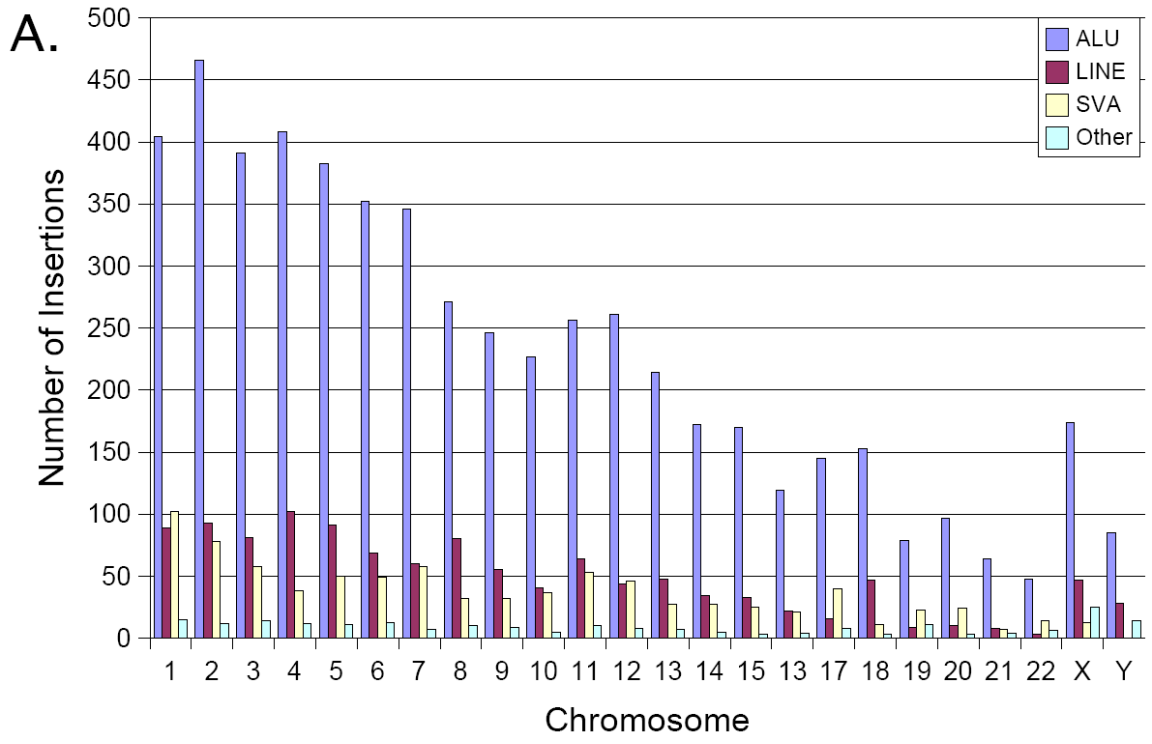
We next examined the genomic distributions of our recently mobilized transposon insertions. In both humans and chimpanzees, these insertions generally were distributed according to the amount of DNA that was present on each chromosome (Figure 4-3). We also examined the distributions of new insertions relative to genes (Table 4-3). Approximately 34% of the new insertions in both genomes were located within known genes (defined as 3 kb upstream to 0.5 kb downstream of a RefSeq gene) (Table 4-3). Using the same criteria, we determined that genes occupy ~34% of the human and chimpanzee genomes (33.5% and 34.8%, respectively). Therefore, the fraction of insertions in genes was very close to that expected if integration (and mechanisms that subsequently remove insertions) had occurred randomly during the past ~6 million years.

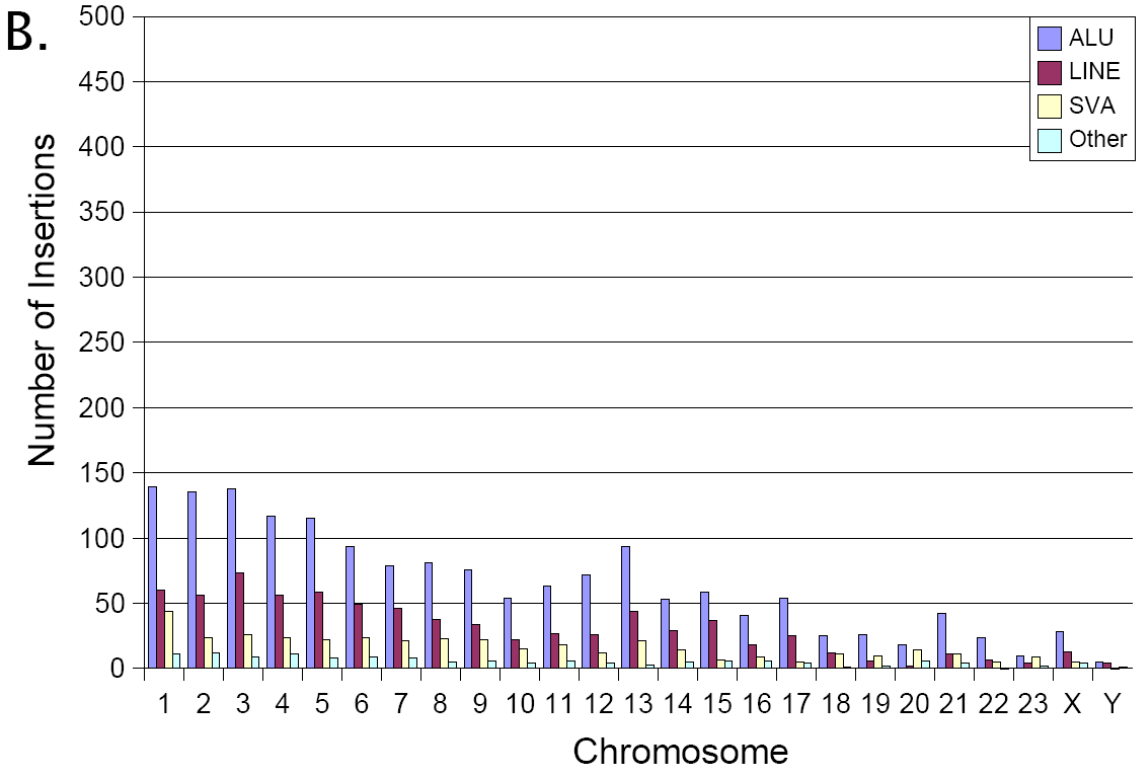
However, further analysis of these patterns revealed that they were not, in fact, random. Although we identified insertions in only ~14% of all human genes, many of these genes had more than one insertion (Table 4-3). Overall, about a third of the human genes with insertions contained multiple insertions. Similar results were observed in the chimpanzee genome (16.5% of the genes with insertions had multiple insertions). We performed one-sample Z tests with our insertions and determined that the observed patterns were not consistent with a random integration model. For example, we observed 16,901 human genes that lacked new transposon insertions from our collections (Table 4-3). The chance of observing this many human genes without such insertions is zero ($P = 0$) with a random integration model. Similar results were observed with Z tests for the remaining integration classes listed in Table 4-3 (data not shown). Therefore, our statistical tests allowed us to reject the hypothesis of random integration with a very high degree of confidence. On the basis of this analysis, it appears that a large fraction of the

new transposon insertions in humans and chimpanzees (the majority of which were Alu, L1, and SVA elements) were targeted preferentially to specific genes. It is also possible that negative selection eliminated insertions from a larger initial collection over time, and this led to the appearance of nonrandom integration. Although targeted integration of L1 has not been observed previously in biochemical or cell culture experiments, previous studies indicate that transposons are eliminated through negative selection (Boissinot, *et al.* 2001). Thus, negative selection is likely to have played a role in dictating the final patterns of transposons observed. Our data also may reflect an integration targeting mechanism that is not functional in cell-culture systems but is active in the germ line of whole organisms, where all of our insertions occurred.

Our study indicates that a relatively large number of insertions occurred within genes during the evolution of humans and chimpanzees (2,642 in humans and 990 in chimpanzees) (Table 4-3). It is likely that at least some of these insertions altered the expression of the target genes, perhaps to the extent that mutant phenotypes emerged. Thus, at least some of the insertions might have had an impact on the differential speciation of humans and chimpanzees by influencing the expression of nearby genes. Since humans received at least 4,853 additional transposon insertions compared with

Figure 4-3 (following page). Genomic distributions of transposon insertions. (A) Genomic distribution of Alu, L1, SVA, and other elements in the human genome. (B) Genomic distribution of Alu, L1, SVA and other elements in the chimpanzee genome. For both genomes, the number of insertions in each chromosome is generally proportional to the amount of DNA present. Note that the Y-axis is the same for both charts. Thus, many more transposon insertions are present throughout the human genome than the chimpanzee genome (compare the number of insertions depicted in panels A and B).


A.
Number of Insertions
0
50
100
150
200
250
300
350
400
450
500
ALU
LINE
SVA
Other
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 13 17 18 19 20 21 22 X Y
Chromosome


B.
Number of Insertions
0
50
100
150
200
250
300
350
400
450
500
ALU
LINE
SVA
Other
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 X Y
Chromosome

Insertions in genes	Human	Chimp
Total Insertions in Genes	2,642	990
Number of Unique Genes Hit	1,891	828
Promoter	50	13
Exon	7	4
Intron	2,478	973
Terminator	17	4
Unclassified	90	0
Number of Insertions per Gene		
0	16,901	19,328
1	1,457	704
2	265	97
3	99	22
4	37	3
5	13	1
6	9	0
7	4	0
8	1	0
9	5	1
10	1	0

Table 4-3. Transposon insertions within genes

chimpanzees, the impact of transposon mutagenesis was likely to be greatest in humans during the past several million years.

In conclusion, we have determined that the original set of transposons in the common ancestor of humans and chimpanzees behaved differently during the subsequent evolution of these organisms. More than 95% of the new transposon insertions in both organisms were Alu, L1, and SVA insertions. However, our data indicate that humans and chimpanzees have amplified very different subfamilies of these elements. Our

combined data also indicate that chimpanzees have supported lower levels of L1 activity than have humans during the past several million years, and this has led to decreased levels of Alu, L1, and SVA transposition in chimpanzees. Other factors, such as differences in population sizes and differences in population bottlenecks, also are likely to have influenced the final patterns of transposon insertions observed in these organisms. In some cases, apparent "insertions" may have been caused by the precise deletion of transposon copies through homologous recombination at the TSDs flanking these elements (van de Lagemaat, *et al.* 2005). A fraction of our insertions also may have been older polymorphisms that were subject to lineage sorting. Thus, the final patterns of transposons in these genomes are likely to have been shaped not only by integration and excision mechanisms but also by the population dynamics of these organisms during the past several million years.

Materials and Methods

The draft chimpanzee-genome (build panTro1) and human-genome (build hg17) sequences were obtained from the University of California Santa Cruz browser (Kent *et al.* 2002). BAC clone sequences for the chimpanzee genome were obtained from GenBank (National Center for Biotechnology Information [NCBI]). BLAST programs also were obtained from NCBI. RepeatMasker was obtained from Arian Smit (Institute for Systems Biology). Repbase version 10.02 and the consensus sequence for the L1-Hs element were obtained from Jurzy Jurka (Jurka 2000). Full-length consensus sequences for L1-PA2, L1-PA3, L1-PA4, and L1-PA5 were obtained from GenBank (Boissinot, *et al.* 2000). Custom MySQL databases and PERL scripts were generated as necessary. All

analysis was performed locally on SUN SunFire v40z or Dell Power Edge 2500 servers running Linux operating systems. Our computational pipeline began with identification of all indels in humans versus chimpanzees using genomic alignments that were generated with BLASTz. Next, indels containing transposons were identified using RepeatMasker (A. Smit, unpublished material) and Repbase version 10.02 (Jurka 2000). Repbase libraries for humans and chimpanzees were modified to include full-length L1-PA2, L1-PA3, L1-PA4, L1-PA5 consensus sequences (Boissinot, *et al.* 2000). TSDs were identified using a Smith-Waterman local alignment algorithm on the regions flanking each indel junction. The algorithm was restricted to require the optimum alignment to be located within 5 bp of the indel junction. Aligned sequences smaller than 4 bp or having an identity <90% were not scored as TSDs. A probability scoring system was developed to determine the likelihood that a given indel was caused by a single transposon insertion plus its TSD. This score was obtained by adding together the fraction of the indel that was accounted for by the transposon, its TSD, and a poly (A) tail (if present). A score of 1.0 indicated that the gap was fully accounted for by the transposon and associated sequences. We empirically determined that a lower cutoff of 0.85 provided accurate results while eliminating few, if any, true positives. SVA elements initially were annotated poorly by RepeatMasker. This program often split SVA elements into 2-3 segments (and thus counted most elements more than once). We developed a new method to reassemble these segments into a single element, where appropriate.

Acknowledgments

We thank the Washington University Genome Sequencing Center and the Whitehead Genome Center for their chimpanzee draft sequence data. We thank Zoltan Ivics for the pβFUS plasmid. We thank Shari Corin for critical review of the manuscript and for helpful advice. We thank Haiyan Wu for help with statistical analysis. This work was supported by National Institutes of Health training grant 2T32GM00849 (to E.A.B. and R.C.I.), a grant from SUN Microsystems (to W.S.P. and S.E.D.), and National Institutes of Health grant 1R01HG002898 (to S.E.D.).

Chapter 5

Active Alu retrotransposons in the human genome

E. Andrew Bennett[1,2,3], Heiko Keller[4,6], Ryan E. Mills[2,3,6], Steffen Schmidt[4], John V. Moran[5], Oliver Weichenrieder[4], and Scott E. Devine[1,2,3,7]

[1]Genetics and Molecular Biology Graduate Program, Emory University School of Medicine, Atlanta, Georgia 30322, USA; [2]Department of Biochemistry, Emory University School of Medicine, Atlanta, Georgia 30322, USA; [3]Center for Bioinformatics, Emory University School of Medicine, Atlanta, Georgia 30322, USA; [4]Department of Biochemistry, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany; [5]Howard Hughes Medical Institute, Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA; [6]These authors contributed equally to this work.

2008. *Genome Research.* 18, 1875-1883

 DOI:10.1101/gr.081737.108

Introduction

Alu retrotransposons evolved from 7SL RNA ~65 million years ago and underwent several rounds of massive expansion in primate genomes. Consequently, the human genome currently harbors 1.1 million Alu copies. Some of these copies remain actively mobile and continue to produce both genetic variation and diseases by "jumping" to new genomic locations. However, it is unclear how many active Alu copies exist in the human genome and which Alu subfamilies harbor such copies. Here, we present a comprehensive functional analysis of Alu copies across the human genome. We cloned Alu copies from a variety of genomic locations and tested these copies in a plasmid-based mobilization assay. We show that functionally intact core Alu elements are highly abundant and far outnumber all other active transposons in humans. A range of Alu lineages were found to harbor such copies, including all modern AluY subfamilies and most AluS subfamilies. We also identified two major determinants of Alu activity: (1) The primary sequence of a given Alu copy, and (2) the ability of the encoded RNA to interact with SRP9/14 to form RNA/protein (RNP) complexes. We conclude that Alu elements pose the largest transposon-based mutagenic threat to the human genome. On the basis of our data, we have begun to identify Alu copies that are likely to produce genetic variation and diseases in humans.

Several lines of evidence indicate that the human genome harbors active Alu retrotransposons (Mills, *et al*. 2007). In fact, one new Alu insertion is estimated to occur for every 20 live human births (Cordaux, *et al*. 2006). An extrapolation of these data to a global population of 6 billion people suggests a total of ~300 million recent Alu

insertions in human populations. This is an impressive mutagenesis of the human genome and is equivalent to an average density of one insertion per 10 bp of DNA. Therefore, Alu retrotransposition events are expected to have a major impact on human biology and diseases (Batzer and Deininger 2002; Mills, *et al*. 2007; Belancio, *et al*. 2008). Forty-three disease-causing Alu insertions have been identified already (Belancio, *et al*. 2008), and such insertions are expected to be discovered routinely as we enter the age of personalized genomics (Mills, *et al*. 2007). However, to understand which Alu elements will continue to produce these new insertions, it is necessary to first define the active Alu copies that reside in the human genome. Only two Alu copies have been tested for mobilization in mammalian cells (Roy, *et al*. 2000; Dewannieux, *et al*. 2003; Hagan, *et al*. 2003) and the number of functional Alu copies in the human genome is unknown.

To fill this gap in our knowledge, we systematically examined the mobilization capacity of Alu copies across the human genome. In particular, we examined the retrotransposition capacity of the ~280-bp central "core" regions of Alu copies using a plasmid-based mobilization assay (Dewannieux, *et al*. 2003). A plasmid-based system is ideal for comparing the relative mobilization efficiencies of diverse core elements, because it keeps all other factors constant and eliminates possible variation due to flanking sequences. We first developed an annotated database of 850,044 full-length human Alu copies that was based upon the reference genome sequence (Lander, *et al*. 2001). We then strategically identified specific Alu copies from this database to test in mobilization assays. We also tested several synthetic Alu elements, including some older consensus elements that are no longer present in the modern human genome. By systematically testing 89 representatives from many Alu families and subfamilies, we

developed the first comprehensive view of functional Alu core elements in the human genome.

Results

Functional analysis of the AluJ, S, and Y lineages

We began by examining the most ancient AluJ lineage for possible retrotransposition activity. Given that this lineage is ~65 million years old and is thought to be functionally extinct (Batzer and Deininger 2002; Mills, *et al*. 2007; Belancio, *et al*. 2008), we were unlikely to find any functional AluJ copies in the genome. Accordingly, our database contains 163,368 full-length AluJ elements but completely lacks intact AluJ copies with consensus AluJ sequences (Figure 5-1). In fact, the AluJ lineage has degraded to the point where the average copy has ~52 changes relative to the 280-bp AluJo and AluJb consensus sequences (equivalent to 18.6% sequence variation) (Figure 5-1). We cloned and tested representatives of the most highly conserved AluJo and AluJb elements that remain in the human genome; however, none of these elements was active in the mobilization assay (Figures 5-1, 5-2E; data not shown). Thus, our combined data indicate that the AluJ lineage is likely to be completely inactive in humans. In further support of this conclusion, no species-specific AluJ copies have been observed in comparisons of the human and chimpanzee genomes (Hedges, *et al*. 2004; The Chimpanzee Sequencing and Analysis Consortium 2005; Mills *et al*. 2006), and no polymorphic or disease-causing alleles of AluJ have been reported (Batzer and Deininger 2002; Bennett, *et al*. 2004; Chen, *et al*. 2005; Wang, *et al*. 2006; Mills, *et al*. 2007; Belancio, *et al*. 2008).

In contrast, the second oldest Alu lineage, AluS, clearly contains functional Alu core elements. This lineage is ~30 million years old and contains 551,383 full-length copies (Figure 5-1). Overall, four of the 16 AluS elements that were selected from the genome and tested in mobilization assays were active (Sg_h11.1, Sp_h12.1, Sc_h1.1, Sx_425) (Figure 5-2B, 5-2D). Functionally intact Alu core elements were identified from four of the six AluS subfamilies. Moreover, some AluS elements were at least as active as consensus AluSx and AluYa5 elements (Figure 5-2B, 5-2E; see below). Our results are consistent with the fact that species-specific AluS copies have been identified in comparisons of the human and chimpanzee genomes (Mills, *et al*. 2006), and that both polymorphic and disease-causing AluS copies have been reported (Bennett, *et al*. 2004; Wang, *et al*. 2006; Mills, *et al*. 2007). Finally, we found that the youngest Alu lineage, AluY, harbors the largest number of functionally intact Alu core elements (Figure 5-1, 5-2C, 5-2D). In fact, AluY and all of its major subfamilies were active in mobilization assays (Figure 5-2C). Consensus AluYa5, AluYb8, and AluYd8 elements had the highest levels of mobilization, followed by the remaining AluY subfamilies. The higher mobilization efficiencies of AluYa5 and AluYb8 might account for the fact that 58.3% of all polymorphic Alus in humans belong to these two subfamilies (Wang, *et al*. 2006). Given the range of activity levels observed among AluY subfamilies, the diagnostic base changes that define these subfamilies appear to have affected the mobilization efficiencies of these elements. Overall, our data indicate that the subfamily status of a given Alu copy largely dictates its mobilization capacity (Figures 5-1, 5-2), though other factors influence mobilization as well (see below).

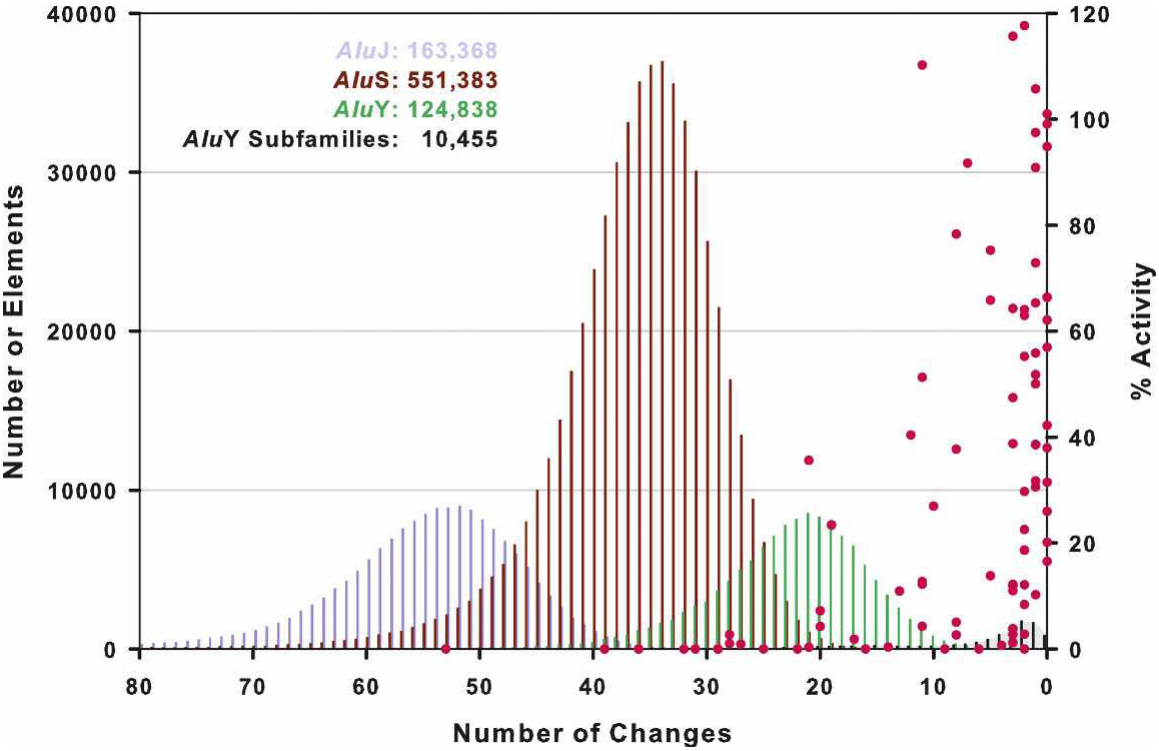


Figure 5-1. A genome-wide view of human Alu activity. A total of 850,044 full-length (>268 bp) genomic Alus were identified in hg18 of the reference human genome sequence and assigned to known Alu subfamilies. Alu elements frequently have sequence changes relative to consensus sequences. The number of changes for each full-length copy is indicated on the x-axis; the copy number for a given level of sequence variation is indicated on the left y-axis. Pink data points mark the mobilization activities of the 89 Alu copies that were examined in this study (labeled on the right y-axis). In sum, 8 AluJ, 27 AluS, and 54 AluY copies were tested, spanning a range of subfamilies and variation levels. Note that elements with fewer changes relative to consensus sequences (zero changes) generally had the highest levels of activity; no elements below 10% variation (28 changes) were active.

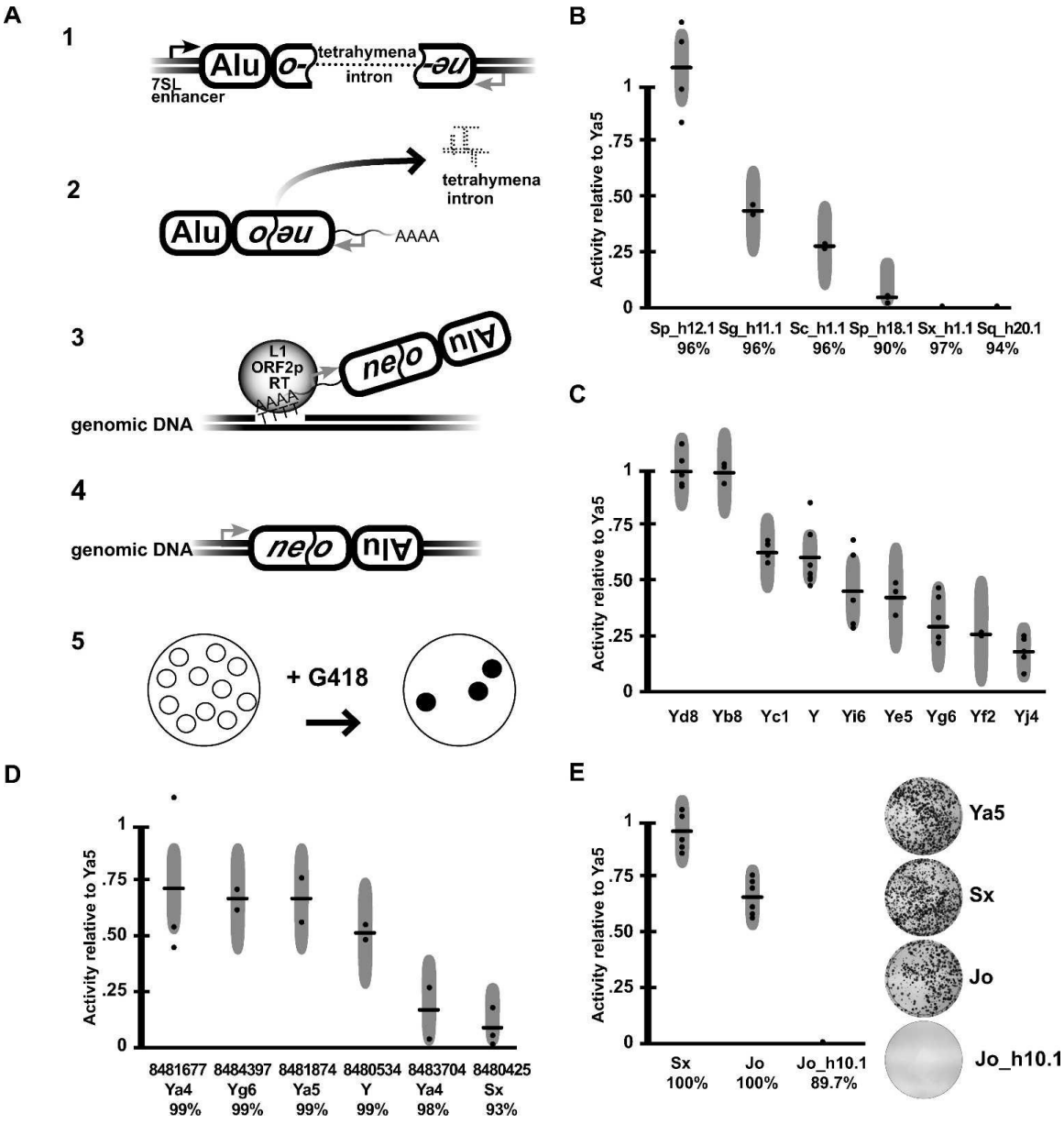


Figure 5-2. Alu mobilization assays. (A) Alu retrotransposition assay (Dewannieux, *et al*. 2003). (1) Alus were cloned into a test plasmid containing the 7SL polIII enhancer and a neo retrotransposition selection cassette. The cassette contains a neo G418 resistance gene that is interrupted by the self-splicing tetrahymena intron. (2) Upon polIII transcription, the tetrahymena intron is spliced out. (3) When cotransfected with L1 ORF2p, Alu RNAs are reverse transcribed along with the neo gene, and (4) integrated into the genome, conferring G418 resistance. (5) After a 2-wk treatment with G418, resistant colonies are stained, photographed, and counted. (B) Assay results for a sample of genomic AluS elements. Activities are given relative to AluYa5 activity (100%) within each assay. Each horizontal bar indicates the

(Figure 5-2 cont.) mean of multiple independent assays (dots). Each dot represents the average of a single (triplicate) experiment. Gray vertical bars represent 95% confidence intervals. The percent consensus identity is indicated below each element. (C) AluY subfamily results. (D) Mobilization activities of known polymorphic AluY elements and a polymorphic AluSx. The dbSNP ss numbers are listed for each. (E) Resurrected AluJo and Sx elements. The mobilization results for artificially constructed 100% consensus AluSx and AluJo elements are compared with a highly conserved (but inactive) genomic AluJo (Jo_h10.1).

Resurrection of ancient AluJ and AluS elements

We next determined that sequence variation is ultimately responsible for the functional extinction of older Alu elements. As outlined above, AluJ elements appear to have accumulated deleterious sequence changes to the point where no intact, functional AluJ copies exist in the modern human genome. To evaluate this hypothesis further, we resurrected an ancient AluJ element carrying the consensus AluJo sequence and tested it using modern L1 ORF2 proteins to drive retrotransposition. Remarkably, this ancient AluJo element was highly active in the mobilization assay (Figure 5-2E). We also resurrected an old AluSx consensus element, which was highly active as well (Figure 5-2E; see also Hagan, *et al*. 2003). These data support a model of sequence decay for the extinction of older AluJ and AluS elements, in which deleterious sequence changes accumulated more rapidly than the pool of active elements was replenished by retrotransposition.

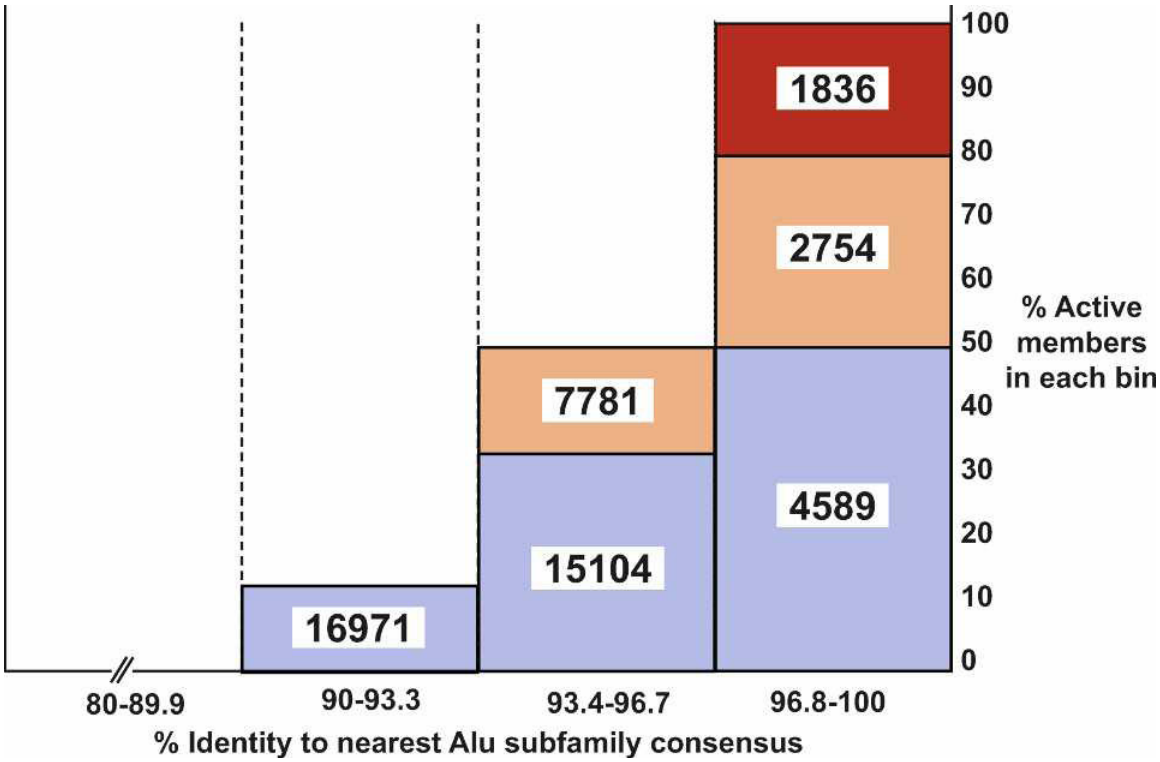


Figure 5-3. How many potentially active Alu core elements in the human genome? A model was developed for estimating the number of potentially active Alu core elements in the reference human genome. A set of 33 unbiased Alu copies (see Methods) were placed into four bins according to their level of sequence variation (96.8%–100%, 93.4%–96.7%, 90%–93.3%, and <89.9%). The percentage of "active copies" was calculated for each bin, where "active" was defined as >5% of AluYa5 activity level in the mobilization assay. The percentages of active copies within each bin were then used to estimate how many genomic copies with the same levels of variation are present in the human genome (the results are depicted as numbers). The levels of activity were broken down further into "hot" (red; 100%–66.6% of AluYa5 activity level), "moderate" (yellow; 66.5%–40% of AluYa5), and "cool" (blue; 39.9%–5% of AluYa5). No elements below 90% conservation were active in the mobilization assay. This method provides a liberal estimate of the number of Alu core elements that would be active if cloned and tested in our mobilization assay. The actual number of elements expressed and mobilized from their natural genomic locations is likely to be lower than the numbers presented due to the impact of flanking genomic regions on Alu expression (see Discussion).

Sequence variation also affects AluY mobilization

Sequence variation also is very common among modern AluY elements, and 134,441/135,293 (99.4%) of the AluY copies in our database had sequence changes compared with consensus sequences. To assess the potential impact of this sequence variation on activity, we next examined the mobilization efficiencies of 22 polymorphic and nine randomly chosen AluY copies that carried core sequence changes. Polymorphic AluY copies (i.e., copies that were differentially present in humans and thus had moved recently) (see Batzer and Deininger 2002) generally had robust levels of mobilization in retrotransposition assays, indicating that sequence variation did not appreciably affect the mobilization of these elements (Figure 5-2D). In contrast, randomly chosen AluY copies that contained sequence variation often had low levels of activity or were completely inactive, presumably because of the mutations in these elements (e.g., elements Y_h5.1, Y_h13.1, and Y_h16.1). On the other hand, some AluY copies with up to 7.4% sequence variation were active in mobilization assays, indicating that Alu can, in some cases, carry a large burden of mutations while still retaining function (e.g., Y_h14.1). These results indicate that there is a general relationship between the amount of sequence variation in a given Alu copy and its level of activity. However, it also appears that some sequence changes are more effective than others at altering activity.

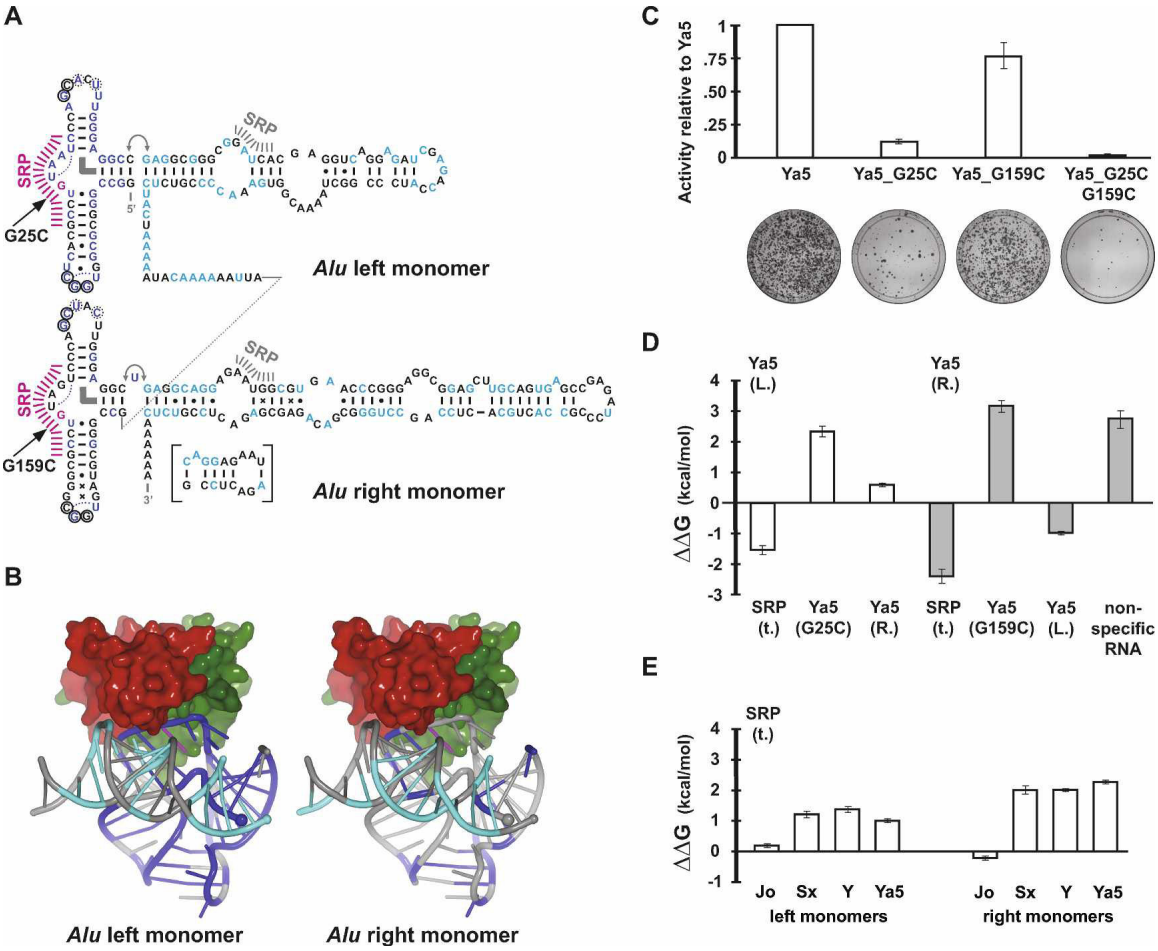


Figure 5-4. SRP9/14 host proteins are necessary for efficient Alu retrotransposition. (A) Secondary structure representation of an AluYa5 RNA. The 124 positions (Figure 5-7) that are conserved among 70 consensus sequences and 45 experimentally tested, active Alu elements are highlighted in blue (Alu 5' domain) and cyan (Alu 3' domain). Positions (G25 and G159) that were mutated to prevent SRP9/14 binding are in magenta. Hash marks indicate major (magenta) and minor (gray) SRP9/14 contact sites. The inset shows an alternative base-pairing that is possible since the emergence of the AluS family and that may be responsible for the drop in SRP9/14 affinity at the transition from the AluJ to S families. Thin curves indicate U-turns, the thick, curved bar indicates a stacking interaction, and the double arrow a flexible linkage. Circles indicate tertiary base pairs between the loops. The dotted circles symbolize an alternative base pair of the respective nucleotides with G14 (left monomer) and G148 (right monomer). Additional symbols: (|) Watson-Crick base pairs; (.) wobble base pairs; (x) other (potential) base pairs. (B) Three-dimensional representation of Alu RNA and SRP9/14 binding. Positions highlighted in A have

(Figure 5-4 cont.) been mapped onto the crystal structure of the SRP Alu RNP (Weichenrieder, *et al*. 2000) using the same colors. SRP9 is in red and SRP14 is in green. The 5' and 3' RNA ends are indicated by a large and small sphere, respectively. (C) Retrotransposition activity of consensus AluYa5 and SRP9/14 binding mutants. The mobilization activities relative to AluYa5 are shown along with representative assay plates below. (D) Relative affinities of Alu RNA mutants for SRP9/14. Left and right mutant Alu RNA monomers competed against the wild-type AluYa5 left monomer RNA (Ya5 (L.), white bars) or AluYa5 right monomer RNA (Ya5 (R.), gray bars) as labeled references in an *in vitro* assay based on nitrocellulose filter binding of SRP9/14. Positive values of $\Delta\Delta G$ reflect loss of affinity with respect to wild-type RNA. The full data set and a representative binding experiment are presented in Table 5-1. A representative binding experiment is shown in Figure 5-6. (E) Relative affinities of Alu consensus sequences for SRP9/14. In contrast to D, truncated SRP RNA (SRP(t.)) was used as reference. These binding results agree with previous gel-shift assays examining SRP9/14 binding to AluS and Y monomers (Sarrowa, *et al*. 1997).

We next developed a model to examine the impact of core sequence variation on Alu activity by plotting the level of sequence variation vs. the mobilization efficiencies for a random selection of unbiased Alu copies in our study (Figure 5-3; Methods). Our model predicts the following: All copies with intact consensus sequences are active in the mobilization assay (852 copies in the genome). However, as changes are introduced, the likelihood that critical sites are mutated increases to the point where all elements are inactive below ~90% conservation. By applying this model to the human genome, we estimate that there are up to 1836 "hot" Alu copies that would be highly active in mobilization assays, 10,535 elements that would be moderately active, and 36,664 copies that would have low levels of activity (Figure 5-3). It should be noted that this approach provides a liberal estimate for the number of functional copies (see Discussion section below). We conclude that the pool of potentially active Alu copies in the reference

genome includes at least 852 consensus copies and is likely to include thousands of copies. For comparison, ~80–100 copies of the human L1 retrotransposon are active in similar assays (Brouha, *et al*. 2003). Thus, the number of potentially active Alu elements in the reference human genome is unexpectedly large and exceeds that of all other human transposons.

In a parallel approach, we compiled a list of 124 positions that are conserved in all known active Alu elements. We first identified 190 sites that are conserved among 70 Alu subfamily consensus sequences (Figure 5-7). Since all consensus elements tested thus far have been active in the mobilization assay, this alignment begins to identify internal sites that must be conserved for function. The preservation of these sites might be required for proper Alu RNA folding and/or to form interactions with essential host factors (see below). We then added to this alignment the 45 elements that were found to be active in this study. This led to the identification of 124 positions that are conserved in all known active Alu elements (Figure 5-4A; Figure 5-7). Since our data set is not exhaustive, it is likely that additional sites can sustain changes within these 124 positions. We identified 3437 elements in our database that conserved all 124 of these positions, and a total of 12,431 elements with up to two changes at these positions (Table 5-2). Importantly, no AluJ elements in our database conserved these 124 positions (Table 5-2). Thus, this independent (and more conservative) approach also predicts that there are thousands of potentially active Alu elements in the human genome.

Alu RNAs must interact productively with SRP9/14 host proteins for successful mobilization

We also identified a second major determinant of Alu activity: SRP9/14 host proteins. Alu RNA originally was derived from a region of 7SL RNA that includes SRP9/14 contact sites (Figure 5-4A, B; Weichenrieder, *et al*. 2000). The first 50 nucleotides of 7SL RNA (the Alu RNA 5' domain) (Weichenrieder, *et al*. 2000) adopt a complicated three-dimensional fold that is recognized by the SRP9/14 heterodimer and is clamped against the helical Alu RNA 3' domain (Figure 5-4A,B). Alu retrotransposons encode two 7SLderived domains in tandem (the Alu left and Alu right monomers) (Sinnett, *et al*. 1991). Surprisingly, each of these domains has conserved this three-dimensional fold, and hence, the ability to bind SRP9/14 (Walter and Blobel 1983; Weichenrieder, *et al*. 2000).

But what impact, if any, does SRP9/14 protein binding have on Alu mobilization? A popular model suggests that SRP9/14 binding facilitates the docking of Alu RNAs on ribosomes, which in turn allows these RNAs to capture L1 ORF2 proteins as they are translated from active L1 mRNAs (Fig 5) (Sinnett, *et al*. 1991; Boeke 1997; Dewannieux, *et al*. 2003; Mills, *et al*. 2007). By hijacking the L1 reverse transcriptase, Alu ensures that its own RNA is copied into the genome instead of L1's mRNA. This model predicts that SRP9/14 binding is necessary for efficient Alu mobilization. We tested this model by constructing a G25C mutation within a predicted SRP9/14 binding site on AluYa5 RNA (Figure 5-4A, B). In the closely related 7SL RNA, this mutation changes a key nucleotide in the SRP binding site and lowers SRP9/14 binding affinity ~50-fold (Chang, *et al*. 1997). We confirmed that our AluYa5_G25C mutation had a similar effect on SRP9/14 binding (Figure 5-4D) and found that mobilization also was decreased to 12% of wild-

type AluYa5 levels (Figure 5-4C). The corresponding mutation in the right monomer (G159C) resulted in a similar decrease in SRP9/14 binding (Figure 5-4D), but led to only a modest decrease in retrotransposition (Figure 5-4C). The combination of both mutations led to severely diminished levels of retrotransposition, indicating that SRP9/14 binding is essential for Alu retrotransposition (Figure 5-4C). These data provide strong experimental support for the SRP9/14 docking model, and indicate that left Alu monomer binding to SRP9/14 is more important for mobilization than the right Alu monomer binding.

Finally, we found that primary sequence changes within Alu have led to diminished SRP9/14 binding during the course of evolution (Figure 5-4E). Our binding assays indicate that 7SL RNA and AluJo RNA have the strongest affinities for SRP9/14, followed by AluSx and AluY RNAs (Figure 5-4E). Remarkably, a major drop in SRP9/14 binding affinity appears to have occurred at the evolutionary transition between AluJ and AluS (Figure 5-4E), and modern AluY elements have preserved this lower affinity. However, our results with the AluYa5 G25C and G159C mutants clearly show that some level of SRP9/14 binding must be maintained for efficient mobilization (Figure 5-4C). Therefore, AluS and Y elements appear to have evolved the lowest possible affinities for SRP9/14 that are still compatible with efficient mobilization.

One possible explanation for these data is that modern Alu RNAs have evolved the ability to disengage from SRP9/14 (Figure 5-5). The ability to disengage from SRP9/14 would not necessarily be required by 7SL RNA, because 7SL RNA serves as a structural scaffold within the signal-recognition particle (Walter and Blobel 1983; Weichenrieder, *et al*. 2000). However, efficient release from SRP9/14 could be

envisioned to improve Alu retrotransposition. According to this model, SRP9/14 would still facilitate the initial docking of Alu RNAs on ribosomes. But at some downstream step of retrotransposition, such as reverse transcription, SRP9/14 would be more efficiently displaced from modern Alu RNA templates. This could have improved the efficiency of reverse transcription and could have led to a competitive advantage over the older Alu RNA templates.

Discussion

In this study, we have identified two major determinants of Alu activity in humans: (1) The primary sequence of the ~280-bp core region, and (2) the ability of the encoded RNA to interact with SRP9/14 to form RNA/protein (RNP) complexes. The closer an element's core sequence is to an active consensus sequence, the more likely it is to remain functional in mobilization assays (Figure 5-1). Likewise, SRP9/14 binding is essential for Alu retrotransposition, and a given Alu RNA sequence must retain the ability to interact productively with SRP9/14 (Figures 5-4, 5-5). Another finding of our study is that the number of functionally intact core sequences in the reference human genome is unexpectedly large. The pool of functionally intact cores includes at least 852 intact consensus elements and is likely to include thousands of copies (Figures 5-3, 5-4A; Table 5-2). Thus, the number of potentially active Alu copies in the human genome greatly exceeds that of all other active human transposons.

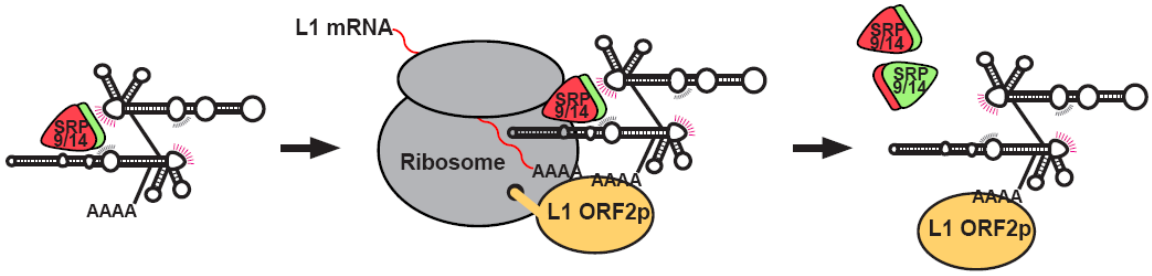


Figure 5-5. Model for Alu retrotransposition. Alu RNA competes with 7SL RNA for SRP9/14 binding and RNP formation. It appears that at least one SRP9/14 heterdimer is necessary for Alu mobilization, although the binding of two heterodimers provides more efficient mobilization. Alu RNPs, once formed, can dock on ribosomes. As L1 mRNA is translated, the poly(A) tail of an SRP9/14-bound Alu competes for nascent L1 ORF2 reverse transcriptase (Sinnett, *et al*. 1991; Boeke, 1997; Dewannieux, *et al*. 2003; Mills, *et al*. 2007). Finally, a new Alu sequence is inserted into the genome by target-primed reverse transcription (Luan, *et al*. 1993). Modern Alu RNAs have evolved weaker SRP9/14 binding affinities, perhaps to disengage from SRP9/14 more readily during reverse transcription.

Additional factors that influence Alu mobilization

Although our mobilization assays measure the ability of Alu RNAs to fold, interact with SRP9/14, and carry out downstream steps of retrotransposition, they do not evaluate all parameters that are likely to be critical for Alu retrotransposition. For example, because we launch Alu mobilization from plasmids, our assays do not take into account natural differences in Alu expression that occur within the context of the genome. Both methylation (Liu and Schmid 1993) and flanking genomic sequences (Ullu and Weiner 1985; Chu, *et al*. 1995; Goodier and Maraia 1998; Roy, *et al*. 2000; Li and Schmid 2001) have been shown to affect Alu element transcription. Likewise, poly(A) tail length has been shown to influence Alu retrotransposition efficiency (Roy-Engel, *et al*. 2002; Dewannieux and Heidmann 2005), and our assay does not evaluate differences in

poly(A) tail length (a constant poly(A) tail length was used). Our analysis was focused solely on the contribution of the ~280-bp core sequence toward mobilization, and our assay did not measure the impact of flanking genomic sequences. Therefore, we currently do no know how many of our elements would be expressed and mobilized from their normal chromosomal positions in biologically relevant cells. One way to estimate how many genomic elements are actually expressed and mobilized would be to examine the number of "source" genes that exist for a typical Alu subfamily. Source genes are differentiated from other copies in that, once integrated, they remain functional and can give rise to new offspring insertions elsewhere in the genome. Clearly, such copies must be getting expressed and mobilized in biologically relevant cell types, and these data allow us to estimate the fraction of Alu copies that are located at favorable (permissive) genomic sites. Batzer and colleagues reported that between 6% and 20% of a given AluY subfamily's copies are capable of serving as source genes, and thus, of producing new retrotransposition events (Cordaux, *et al*. 2004). On the basis of the Batzer study, we expect that ~6%–20% of the functional Alu cores in our study (Figures 5-3, 5-4A) likewise would be located within favorable genomic contexts, and thus, would be able to produce new insertions in the human genome. Therefore, even when adjusted in this manner, we still conclude that the number of active Alu copies in the human genome far exceeds that of all other human transposons. Additional studies will be necessary to identify the exact copies that are being expressed and mobilized within our collections. One way to tackle this problem would be to examine the expression of our elements in a variety of cell types, particularly in germ cells where Alu mobilization is likely to occur. Li and Schmid (2001), for example, studied the expression of six Alu copies under

baseline conditions in several cell lines and in response to stress induction. Their studies revealed diverse expression profiles for each of the six Alus. This approach now could be applied on a much larger scale to a range of embryonic (and possibly somatic) cell types that are likely to derepress Alu expression. Up to several thousand Alu cDNAs could be cloned and sequenced to gain an understanding of which elements are actually expressed from their natural chromosomal sites. The expressed elements then could be compared with those predicted to have active core sequences from our study, ultimately providing a better picture of which elements are most likely to produce new offspring insertions in humans. Finally, such data could be combined with parallel L1 studies to identify Alu copies that would be coexpressed with L1 ORF2p (a condition that also is essential for Alu mobilization). Collectively, these studies would allow us to make better predictions of which Alu copies are likely to produce genetic variation and diseases in humans.

Why are there so many potentially active Alu copies in the genome?

There might be evolutionary advantages to maintaining large pools of potentially active Alu copies in the genome. Given that some of the factors that inhibit Alu activity such as methylation and poly(A) tail length can be reversed, these pools are likely to be dynamic. Dormant Alu copies could be envisioned to become reactivated provided that they had "active" core sequences that could support mobilization (Han, *et al*. 2005). Large pools of diverse Alu sequences could help Alu to modify its interactions with host factors such as SRP9/14 and might be useful in overcoming host suppression. Indeed, our SRP9/14 binding data suggest that AluS evolved a competitive advantage over AluJ by changing its interaction with SRP9/14. Moreover, this advantage could explain the extinction of

AluJ and the subsequent expansion of AluS. Thus, the number of active Alu elements in the genome might change during specific developmental stages or in the face of selective pressure.

Materials and Methods

Database of full-length Alu elements

Alu locations were obtained from the RepeatMasker track on the UCSC genome browser (Kent, *et al*. 2002). Alu elements with core sequences of >268 bp were considered to be full length and were included in the database. Full-length Alus were reclassified using an in-house Alu identification program entitled CAlu, ("call you") which aligns an Alu sequence to an alignment profile consisting of known Alu subfamilies (obtained from RepBase version 21) (Jurka 2000) using ClustalW. Positional changes were identified compared with an ancestral Alu sequence (AluY, AluSz, AluJo); these changes were then compared with a library of positional changes and the Alu was classified accordingly. The newly classified Alus were organized into a database using genomic coordinates, nearest subfamily, and nucleotide changes beyond the diagnostic subfamily positions, if present.

Plasmids

pCEP 5'UTR ORF2 No Neo, containing ORF2 of the L1.3 retrotransposon was described previously (Alisch, *et al*. 2006). Marked Alu plasmids were created using pAlu-eab2 (a modified version of the pAluNF1-neoIII plasmid) (Dewannieux, *et al*. 2003), which contains the 7SL polIII enhancer upstream of the NF1-Alu10, and a downstream neo

retrotransposition selection cassette consisting of a neo G418 resistance gene interrupted by the self-splicing tetrahymena intron (Esnault, *et al*. 2002) cloned into pUC19. A SpeI restriction site was introduced immediately following the 3' end of the NF1-Alu, and an AflIII site was introduced into the NF1-Alu to facilitate clone selection. Alus were amplified by PCR and cloned using sequence-specific primers to preserve the individual 5' and 3' sequences of the target Alu. Typical primer sequences included an upstream primer containing a PstI restriction site (underlined): 5'-TGCCCTGCAGCTTCTAGTAGCTTTTCGCAGCGTCTCCGACCGGCCGGGCGCGGTGGCT-' and a downstream primer containing a SpeI restriction site (underlined): 5'-TTCTGAACTAGTATTTGAGACGGAGTCTCGCT-3'. Alu consensus sequences were either amplified and cloned from the genome by PCR or synthesized by annealing short, overlapping oligos, ligating cohesive ends, and then performing PCR amplification. Specific genomic copies were first PCR amplified using primers in flanking sequences, followed by a second PCR amplification using the primers described above or similar primers. Random genomic copies were PCR amplified directly from BAC DNA or genomic DNA from the SNP Discovery Resource Panel of 24 diverse humans (Coriell) (Collins, *et al*. 1999). Site-directed mutagenesis was performed on Alus by PCR using primers containing the desired mutations, and amplified fragments were cloned into the appropriate plasmid. For each genomic Alu, the genomic source is indicated in the plasmid named using the convention: "(subfamily)h(chromosome).(ID number)." For runoff *in vitro* transcription by T7 RNA polymerase, left and right Alu monomers were amplified from the respective source plasmid using primers containing the T7 promotor. Products were inserted into a derivative of plasmid pSP64 (Promega) that provided a 3'

terminal HDV ribozyme to the transcript, allowing the generation of precise 3' ends (Walker, *et al*. 2003). Primer sequences are available upon request. All plasmids were sequenced by QuickLane DNA sequencing (Agencourt Bioscience Corp.) using the M13-rev or SP6 primers.

DNA preparation

Plasmids were purified using midi- and maxiprep columns from QIAGEN according to the manufacturers' protocols. Plasmid DNA purity and concentrations were determined by spectrophotometer.

Cell culture

Hela cells were grown in 100-mm plates at 37°C with 5% CO2 in Dulbecco's modified Eagle medium (DMEM) with 4.5 g/L glucose, L-glutamine, sodium pyruvate (Cellgro), supplemented with 10% Fetal Calf Serum, and passaged using standard protocols.

Alu retrotransposition assay

Retrotransposition assays were carried out essentially as described (Dewannieux, *et al*. 2003) except that G418 was added directly to cells 72 h after transfection. Twenty-four hours before transfection, cells were pooled in a 50-mL conical tube in 50 mL of DMEM, and counted using a hemocytometer. Then, 6 x 10^5 cells were plated from an agitated 50-mL conical tube onto 100-mm plates. Sample plates were trypsinized the next day and counted to confirm uniformity of cell number across plates. DNA concentrations were measured prior to each assay. Transfections were performed in triplicate using FuGene6

transfection reagent (Roche). A total of 2 μg of pCEP 5'UTR ORF2 No Neo was cotransfected with 6 μg of pAlu-eab2 (varying Alu plasmid concentration +/- 25% showed no difference in final colony count). For each triplicate, 2 μg of EGFP-N1 (Clontech) or a modified version with ampicillin resistance (EGFP-ampR) was cotransfected on a fourth plate to measure transfection efficiencies. A total of 8 μg of pYa5-neo without the L1 ORF2 driver was used as a negative control, and 6 μg of pYa5-neo with 2 ug of pCEP 5'UTR ORF2 No Neo was transfected as a positive control for all assays. After 24 h, transfection efficiencies were determined. After 72 h, cells were given DMEM containing 600 μg/mL G418 and 100 μg/mL Penicillin-Streptomycin (Cellgro). Fourteen days later, plates were washed with methanol, Giemsa stained, and photographed. Colonies were counted manually using ImageJ (W.S. Rasband, ImageJ, U.S. National Institutes of Health, Bethesda, Maryland, [1997–2007]; http://rsb.info.nih.gov/ij/) and normalized to the pYa5-neo with pCEP 5'UTR ORF2 No Neo result within each assay. Data for featured Alus combine results from two to seven independent assays. All Alus were expressed using the 7SL promoter enhancer sequence immediately upstream of the Alu. To examine the possible variation in construct expression, RT–PCR was performed using primers for (1) sequences in the Alu, (2) a non-expressed plasmid backbone sequence to control for DNA contamination, and (3) a cotransfected expressed protein as a loading control. RT–PCR was performed from cells transfected with AluYa5, and AluJo, having high and medium levels of activity, respectively, and Sz_hX.1, being completely inactive and containing multiple disruptions of its A-Box and B-Box sequences. PCR product was removed at cycles 23, 30, and 35

and run on a 1% agarose gel. Equal levels of Alu RNA were detected across all Alus, indicating that varying the Alu sequence had no effect on expression.

Models for estimating the number of potentially active Alu copies

We identified 33 unbiased Alu copies from our data set of 89 copies (Figure 5-3). Elements that were known to be polymorphic in humans were excluded from the analysis, and only naturally occurring elements were used. The following elements were included: 1_Sc_h5.1, 2_Y_h5.1, 3_Sp_h18.1, 9_Sx_h19.1, 11_Y_h10.1, 12_Y_h13.1, 21_Yb8, 26_Yc1, 46_Yg6, 47_Yi6, 48_Yd8, 58_Sx_h14.1, 59_Sc_h5.2, 60_Y_h14 0.1, 65_Ya4_15, 66_Sgxz_20, 69_Yj4, 83_Yf2, 86_Yf2_38, 88_Jo_h10.1, 89_Jo_h11.1, 90_Jb_h8.1, 93_Jb_h9.1, 95_Sc_h15.1, 96_S_5, 97_Y_h16.1, 100_Y_h6.1, 101_Y, 109_Sz_hX.1, 110_Sc_57, 111_Yc1_h20.1, 114_Ya5a2. These elements were placed in bins as described in Figure 5-3, and the percentage of active copies was calculated ("active" being defined as having >5% AluYa5 levels of activity). These percentages then were used to count the number of Alu copies in the genome with the same percentages of conservation. For the 124-position analysis, we developed the list of critical positions using the alignment depicted in Figure 5-7, including 70 consensus sequences and 45 active elements (elements with >5% AluYa5 levels of activity). We then examined our database for full-length elements that conserved either all 124 positions or that had up to two changes in these positions. The data are presented in Table 5-2.

SRP9/14 binding assays

Large-scale *in vitro* transcription from linearized plasmid templates and gel purification of Alu RNA was done as described previously (Weichenrieder, *et al*. 1997, 2001). RNA was quantified spectroscopically using a value of 40 mg/mL per OD_{260}. Reference RNA was labeled cotranscriptionally in the presence of [α^{32}P]UTP (20 μCi/20 μL reaction) and was gel purified. The human SRP9/14ΔR protein heterodimer was expressed, purified, and quantified as described previously (Weichenrieder, *et al*. 2000). Protein was prepared in reaction buffer (20 mM Tris-HCl at pH 7.5, 10 mM $MgCl_2$, 200 mM Na-acetate) supplemented with 30 mM DTT and 0.3 mg/mL BSA. Unlabeled reference RNA, together with traces of radioactive material, was annealed in reaction buffer separately from unlabeled competitor RNA by incubating 10 min at 65°C and slow cooling to 37°C. Finally, SRP9/14 (10 μL) was added to a mixture of reference RNA (5 μL) with different concentrations of competitor RNA (15 μL). Final samples contained SRP9/14 (~90 nM), reference RNA (100 nM), and competitor RNA (23–17,000 nM) in 20 mM Tris-HCl (pH 7.5), 10 mM $MgCl_2$, 200 mM Na-acetate, 10 mM DTT, and 0.1 mg/mL BSA. After 15 min at room temperature, allowing for full equilibration, samples (25 μL) were filtered through a nitrocellulose membrane (PROTRAN, Schleicher & Schuell) and washed with 100 μL of reaction buffer using an S&S Minifold Slot Blot System (Schleicher & Schuell) according to the instructions of the manufacturer. Depending on the relative affinities, competitor RNA replaces labeled reference RNA retained on the filter by SRP9/14. Filters were exposed to PhosphorImager screens (Molecular Dynamics), scanned with a Storm 820 (GE Healthcare) and quantified with the associated software (Image Quant TL).

After appropriate pilot experiments and controls, we determined the fraction saturation, ν, of SRP9/14 as a function of the ratio, ρ, of competitor to reference RNA. As a parameter for curve fitting we used κ, the ratio of dissociation constants of reference to competitor RNA. For convenience of calculation and graphical representation we chose to replace κ by (e^ln(κ)) in Equation 1 and fit ln(κ) directly. Finally, Equation 2 was used to calculate differences in binding energy (ΔΔG). Three independent measurements were done for each Alu RNA construct, using cold reference RNA for normalization and as a positive control on each filter. An RNA aptamer for tetracycline (Müller, *et al*. 2006) served as a negative control for nonspecific competition.

(1)

$$\nu = f_\kappa(\rho) = \frac{S - S_\infty}{S_0 - S_\infty} = \frac{\kappa(1-\rho) - 2 + \sqrt{(\kappa(1-\rho)-2)^2 + 4(\kappa-1)}}{2(\kappa-1)}$$

Equation 1 relates the fraction saturation, ν, to the ratio, ρ, of competitor to reference RNA. The fraction saturation is calculated as $((S-S_\infty)/(S_o-S_\infty))$, where S and S_o correspond to Phosphor-Imager counts in the presence and absence of competitor RNA and where S_∞ accounts for background counts.

(2)

$$\Delta\Delta G = \Delta G_{comp} - \Delta G_{ref} = - RT\ln\kappa$$

Equation 2 relates κ, the ratio of dissociation constants of reference to competitor RNA to ΔΔG, the difference in affinity, where R corresponds to the gas constant (1.986 cal * K^{-1} *

mol^{-1}) and T corresponds to the temperature (in Kelvin). A positive value of $\Delta\Delta G$ indicates that competitor RNA has less affinity for SRP9/14 than reference RNA.

Statistics

Activity fractions and their 95% confidence intervals were calculated with maximum likelihood using SAS PROC MIXED (SAS Institute, Inc.). All active Alus were included and were treated as fixed effects, while a random assay term accommodated the repetition of each Alu within each replicate and across assays. The randomly selected Alus were categorized by their average activity fraction into four groups: 0–<0.05 (inactive), 0.05–<0.40 (low activity), 0.40–<0.66.6 (moderate activity), and 66.6%–100% (high activity). These randomly selected Alus were further categorized by their consensus levels: <90%, 90%–<93.3%, 93.3%–<96.6%, and >96.6%. Within each consensus group, the proportion of Alus in each activity levels was calculated using Wilson's method (Altman, *et al*. 2000).

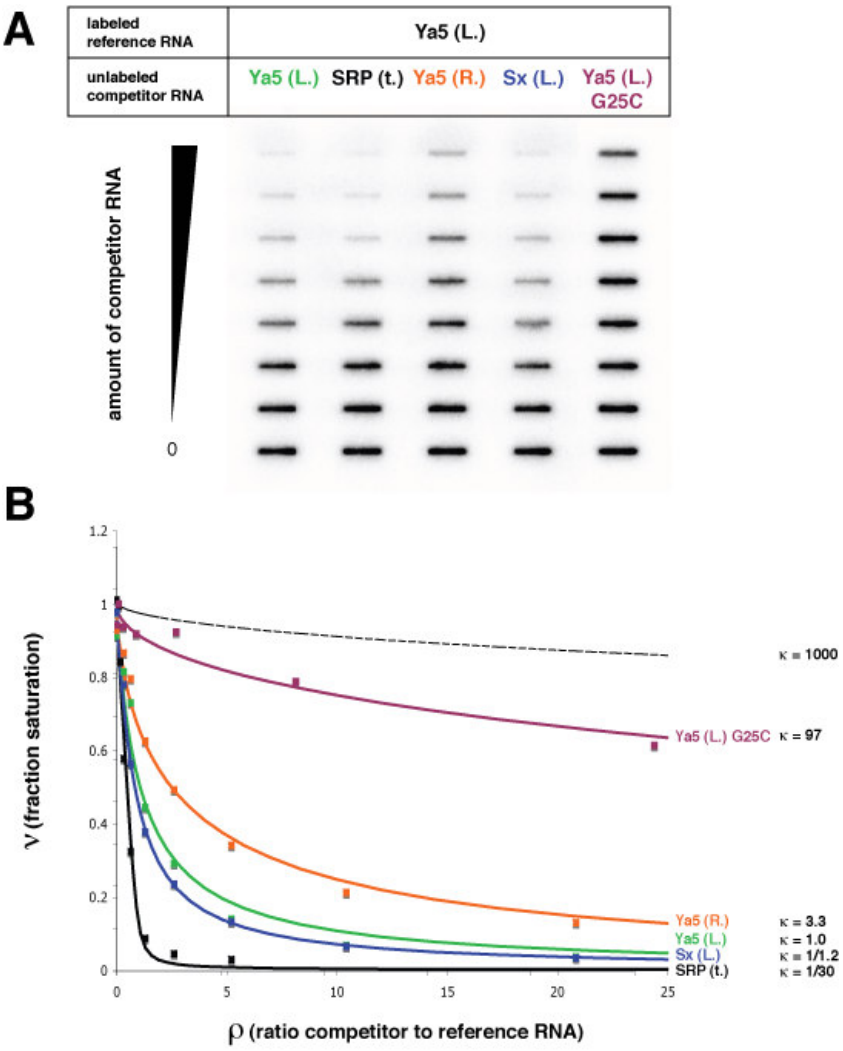


Figure 5-6. **Relative SRP9/14 affinity of various Alu RNA constructs.** Protein affinities were compared in an RNA competition assay based on nitrocellulose filter binding of SRP9/14. We determined the fraction saturation, ν, of SRP9/14 as a function of the ratio, ρ, of competitor to reference RNA. As a parameter for curve-fitting we used κ, the ratio of dissociation constants of reference to competitor RNA. (A) Representative phosphoimager scan of a nitrocellulose filter from a slot-blot device. Labeled reference RNA (Ya5 (L), the left arm monomer of a Ya5 consensus sequence) was challenged with various unlabeled competitor RNAs for binding to a limiting amount of SRP9/14. (B) Quantification of (a). The fraction saturation, ν, is plotted as a function of the ratio, ρ, of competitor to reference RNA. Theoretical curves are drawn, corresponding to the indicated values of κ, the ratio of dissociation constants of reference to competitor RNA.

Figures	reference RNA	competitor RNA	ΔΔG (kcal/mol)	error ΔΔG (kcal/mol)	p-value	affinity loss (fold)	affinity gain (fold)
Fig 4 e	SRP (t.)	Jo (L.)	0.19	0.08	1.4E-01	1.4	
		Sx (L.)	1.19	0.11	8.4E-03	7.5	
		Y (L.)	1.33	0.11	6.2E-03	9.5	
		Ya5 (L.)	1.14	0.01	7.7E-05	6.9	
		Jo (R.)	-0.26	0.07	6.1E-02		1.6
		Sx (R.)	2.12	0.14	4.3E-03	36	
		Y (R.)	2.10	0.03	2.3E-04	35	
		Ya5 (R.)	2.23	0.06	6.6E-04	43	
Fig 4 d	Ya5 (L.)	SRP (t.)	-1.61	0.12	5.7E-03		15
		Ya5 (L.) G25C	2.43	0.14	3.4E-03	61	
		Ya5 (R.)	0.67	0.03	2.4E-03	3	
	Ya5 (R.)	SRP (t.)	-2.44	0.24	9.3E-03		61
		Ya5 (R.) G159C	3.30	0.14	1.9E-03	265	
		Ya5 (L.)	-0.85	0.06	4.4E-03		4.2
		non-specific RNA	3.21	0.25	6.0E-03	226	

Table 5-1. SRP9/14 binding data. Relative affinities of Alu RNA for SRP9/14. Left and right Alu RNA monomers competed against the wild-type AluYa5 left or right monomer RNA as labeled references in an *in vitro* assay based on nitrocellulose filter binding of SRP9/14. Positive values of ΔΔG reflect loss of affinity with respect to wild-type RNA.

Number of Alu elements with exactly n mutation			
Number of mutations	**0**	**1**	**2**
All	3437	3818	5176
AluJ	0	0	0
AluS	7	110	606
AluY	3430	3708	4570
Number of Alu Elements with up to n mutations			
Number of mutations	**0**	**1**	**2**
All	3437	7255	12431
AluJ	0	0	0
AluS	7	117	723
AluY	3430	7138	11708
Proportion of each Alu family having up to n mutations (%)			
Number of mutations	**0**	**1**	**2**
All	100	100	100
AluJ	0	0	0
AluS	0.2	1.6	5.8
AluY	99.8	98.4	94.2

Table 5-2. Human Alu elements that conserve 124 key sites listed by family. Full length Alu elements that preserve the 124 conserved positions were identified in the human genome. These elements are listed by major family (Alu J, S, Y). Note that no Alu J elements, which are presumed to be extinct, preserve these positions. Because our model of 124 positions was based upon limited data, it is likely that some of the 124 positions do not need to be conserved. Thus, we also calculated the number of elements in the genome that had one or two changes within the 124 positions. This approach is likely to provide a conservative estimate of potentially active Alu copies. Additional refinements of this model should be possible as more elements are tested. Alu elements with mutations in positions (n=124) that are 100% conserved in active elements.

Figure 5-7 (following pages). Clustal alignment of 70 consensus and 45 active Alu element sequences. The 70 consensus sequences and 45 active sequences were aligned and used to identify 124 positions that were conserved among these elements. CLUSTAL X (1.83) multiple sequence alignment, manually curated.

```
AluYa1_1/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
10_Y_534/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
11_Y_h10.1/1-281            GGCCGGGCGCGGTGGCTCATG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYd6/1-269                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYd8/1-269                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYc5/1-269                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb10/1-288               GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb11/1-289               GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb8/1-288                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb9/1-288                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb7_2/1-288              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb7_1/1-288              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb7_3/1-288              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb6_1/1-288              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb7_4/1-288              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb5/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb6_2/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb3a1/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYb3a2/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYbc3a/1-282              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYi6/1-282                GGCCGGGCGCGGTGGCTCACG-CTTGTAATCCCAGCACTTTGGGAGGCCG
111_Yc1_h20.1/1-281         GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYg6a2/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYg6/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
22_Yg6_397/1-280            GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
28_Yc1_65/1-280             GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYg5b3/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
100_Y_h6.1/1-281            GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCTG
AluYj4/1-281                GGCCGGGCGCGGTGGCTCGCG-CCTGTAATCCCAGCACTTTGGGAGGCCG
50_Yj4_11/1-281             GGCCGGGCGCGGTGGCTCGCG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYj3/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
60_Y_h14/1-282              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCTAGCACTTTGGGAGGCCG
AluYh9/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCAG
71_Yh9_20/1-282             GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCAG
AluYh7/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCAG
AluYh3/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCAG
AluYe5/1-280                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
128_Ye5_d196/1-279          GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYe6/1-280                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYe4/1-282                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYe2/1-279                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
102_Y_T144C/1-281           GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYf2/1-283                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
86_Yf2_38/1-283             AGCCGGGGGCGGTGGCTCACG-CCTGTAATCCCAGTATTTTGGGAGGCCG
AluYf1/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluY/1-281                  GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYc1/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
25_Yc1_73/1-279             GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYc2/1-281                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
27_Yc1_66/1-280             GGCCGG-CGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
13_pAlu_A/1-286             GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa1_2/1-281              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYd3a1/1-270              GGCCGGGCGCGGTGGCTCACG-CTTGTAATCCCAGCACTTTGGGAGGCCG
AluYd3/1-269                GGCCGGGCGCGGTGGCTCACG-CTTGTAATCCCAGCACTTTGGGAGGCCG
AluYd2/1-269                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYd/1-269                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
19_Ya5_173/1-281            GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
125_Ya5_C138T/1-281         GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
64_Ya5_G25C/1-281           GGCCGGGCGCGGTGGCTCACG-CCTCTAATCCCAGCACTTTGGGAGGCCG
92_Ya5#-Ya5#_95/1-278       GGCCGGGCGCGGTGGCTCACG-CCTCTAATCCCAGCACTTTGGGAGGCCG
117_Ya5_G159C/1-281         GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
116_Ya5_G25C/1-281          GGCCGGGCGCGGTGGCTCACG-CCTCTAATCCCAGCACTTTGGGAGGCCG
119_Ya5_G25C/1-281          GGCCGGGCGCGGTGGCTCACG-CCTCTAATCCCAGCACTTTGGGAGGCCG
118_Ya5_G159C/1-281         GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
```

```
120_Ya5_AC275TG/1-281          GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa5a1/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa5b1/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa5a2/1-282                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
20_Ya5_054/1-285               GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa5/1-281                   GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa8/1-280                   GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
81_Ya5_C150T/1-281             GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa5c1/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
124_Ya5_G21A_C23T/1-281        GGCCGGGCGCGGTGGCTCACA-CTTGTAATCCCAGCACTTTGGGAGGCCG
112_Ya5_rtSRPko/1-281          GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
18_Ya5_874/1-281               GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa5b2/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
99_Ya5_A131T_C206T/1-281       GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa3_2/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa4_1/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa3_3/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa4_3/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa4_2/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa4b/1-281                  GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa3_1/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
15_Ya4_677/1-281               GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
16_Ya4_704/1-283               GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa4/1-281                   GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa4_5/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa4_4/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa3_5/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa3_4/1-281                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa3/1-281                   GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluYa2/1-281                   GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
62_Sx-Ya5/1-282                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluSc/1-279                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
108_Sc_h1.1/1-279              GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluSx/1-282                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
5_Sx_5/1-280                   GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
42_Sx_10/1-279                 GGCCGGGCGCG-TGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
73_Sx_i22T/1-283               GGCCGGGCGCGGTGGCTCACGTCCTGTAATCCCAGCACTTTGGGAGGCCG
4_Sx_13/1-280                  GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluSz/1-282                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluSg/1-280                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
104_Sg_h11.1/1-280             GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluSp/1-283                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
105_Sp_h12.1/1-283             GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
8_Sx_425/1-280                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCTCAGCACTTTGGGAGGCTG
AluSq/1-283                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
79_Sx_11/1-282                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluJo/1-282                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
32_Jo_18/1-280                 GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
30_Jo_5/1-282                  GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
AluJb/1-282                    GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
63_Jo-Ya5/1-282                GGCCGGGCGCGGTGGCTCACG-CCTGTAATCCCAGCACTTTGGGAGGCCG
                                ***** *** ******    * * *****  ** * ********** *
```

```
AluYa1_1/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
10_Y_534/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
11_Y_h10.1/1-281            AGGTGGGCGGATCAC--GGGGTCAGGAGATCGAGACTATCCTGGCTAACT
AluYd6/1-269                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCA------------
AluYd8/1-269                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCA------------
AluYc5/1-269                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCA------------
AluYb10/1-288               AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb11/1-289               AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb8/1-288                AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb9/1-288                AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb7_2/1-288              AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb7_1/1-288              AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb7_3/1-288              AGGCGGGTGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb6_1/1-288              AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb7_4/1-288              AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb5/1-281                AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb6_2/1-281              AGGCGGGTGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb3a1/1-281              AGGCGGGTGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYb3a2/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYbc3a/1-282              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYi6/1-282                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
111_Yc1_h20.1/1-281         AGGCGGGCGGATCAC--AAGGTCAGGAGATCAAGACCATCCTGGCTAACA
AluYg6a2/1-281              AGACGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYg6/1-281                AGACGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
22_Yg6_397/1-280            AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
28_Yc1_65/1-280             AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYg5b3/1-281              AGACGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTGACA
100_Y_h6.1/1-281            AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATGCTGGCTAACA
AluYj4/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
50_Yj4_11/1-281             AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYj3/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
60_Y_h14/1-282              AGGCAGGTGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYh9/1-281                AGGCGGGCGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACG
71_Yh9_20/1-282             AGGCGGGCGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACG
AluYh7/1-281                AGGCGGGCGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACG
AluYh3/1-281                AGGCGGGCGGATCAT--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYe5/1-280                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
128_Ye5_d196/1-279          AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYe6/1-280                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYe4/1-282                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYe2/1-279                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
102_Y_T144C/1-281           AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYf2/1-283                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
86_Yf2_38/1-283             AGGCGGGCGGATCAC--GATGTCAAGAGATCGAGACCATCCTGGCTAACA
AluYf1/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluY/1-281                  AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYc1/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
25_Yc1_73/1-279             AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYc2/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
27_Yc1_66/1-280             AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
13_pAlu_A/1-286             AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYa1_2/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYd3a1/1-270              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCA------------
AluYd3/1-269                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCA------------
AluYd2/1-269                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCA------------
AluYd/1-269                 AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCA------------
19_Ya5_173/1-281            AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
125_Ya5_C138T/1-281         AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
64_Ya5_G25C/1-281           AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
92_Ya5#-Ya5#_95/1-278       AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
117_Ya5_G159C/1-281         AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
116_Ya5_G25C/1-281          AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
119_Ya5_G25C/1-281          AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
118_Ya5_G159C/1-281         AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
120_Ya5_AC275TG/1-281       AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa5a1/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGGCCATCCCGGCTAAAA
AluYa5b1/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa5a2/1-282              AGGCGGGCGGATCAC--GAGGTCAAGAGATCGAGACCATCCCGGCTAAAA
20_Ya5_054/1-285            AGGCGGGCGGATCAC--GAGGTCAAGAGATCGAGACCATCCCGGCTAAAA
AluYa5/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa8/1-280                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
```

```
81_Ya5_C150T/1-281          AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa5c1/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
124_Ya5_G21A_C23T/1-281     AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
112_Ya5_rtSRPko/1-281       AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
18_Ya5_874/1-281            AGGCGGGCGGATCAT--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa5b2/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCAACCCCGCTAAAA
99_Ya5_A131T_C206T/1-281    AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa3_2/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAACA
AluYa4_1/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAACA
AluYa3_3/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa4_3/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa4_2/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa4b/1-281               AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa3_1/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
15_Ya4_677/1-281            AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
16_Ya4_704/1-283            AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa4/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa4_5/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCCGGCTAAAA
AluYa4_4/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAAAA
AluYa3_5/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAAAA
AluYa3_4/1-281              AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYa3/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
AluYa2/1-281                AGGCGGGCGGATCAC--GAGGTCAGGAGATCGAGACCATCCTGGCTAACA
62_Sx-Ya5/1-282             AGGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
AluSc/1-279                 AGGCGGGCGGATCAC--GAGGTCAAGAGATCGAGACCATCCTGGCCAACA
108_Sc_h1.1/1-279           AGGCGGGCGGATCAC--GAGGTCAAAAGATCGAGACCATCCTAGCCAACA
AluSx/1-282                 AGGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
5_Sx_5/1-280                AGGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
42_Sx_10/1-279              AGGC--GCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
73_Sx_i22T/1-283            AGGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
4_Sx_13/1-280               AGGCGGGCG-ATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
AluSz/1-282                 AGGCGGGCGGATCACTTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
AluSg/1-280                 AGGCGGGCGGATCAC--GAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
104_Sg_h11.1/1-280          AGGCAGGCGGATCAC--GAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
AluSp/1-283                 AGGCGGGCGGATCACCTGAGGTCGGGAGTTCGAGACCAGCCTGACCAACA
105_Sp_h12.1/1-283          AGGCGGGCGGATCACCTGAGGTCGGGAGTTCAAGACCAGCCTGACCAACA
8_Sx_425/1-280              AGGTGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGACCAACA
AluSq/1-283                 AGGCGGGTGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
79_Sx_11/1-282              AGGCGGGCGGATCACCTGAGGTCAGGAGTTCGAGACCAGCCTGGCCAACA
AluJo/1-282                 AGGCGGGAGGATTGCTTGAGCCCAGGAGTTCGAGACCAGCCTGGGCAACA
32_Jo_18/1-280              AGGCGGGAGGATTGCTTGAGCCCAGGAGTTCGAGACCAGCCTGGGCAACA
30_Jo_5/1-282               AGGCGGGAGGATTGCTTGAGCCCAGGAGTTCGAGACCAGCCTGGGCAACA
AluJb/1-282                 AGGCGGGAGGATCACTTGAGCCCAGGAGTTCGAGACCAGCCTGGGCAACA
63_Jo-Ya5/1-282             AGGCGGGAGGATTGCTTGAGCCCAGGAGTTCGAGACCAGCCTGGGCAACA
                            **    * * **          *   ** ** ** * *
```

```
AluYa1_1/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
10_Y_534/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
11_Y_h10.1/1-281            ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYd6/1-269                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYd8/1-269                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYc5/1-269                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb10/1-288               ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb11/1-289               ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb8/1-288                ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb9/1-288                ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb7_2/1-288              ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb7_1/1-288              ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb7_3/1-288              ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb6_1/1-288              ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb7_4/1-288              ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb5/1-281                ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb6_2/1-281              ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb3a1/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYb3a2/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYbc3a/1-282              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAAA---TTA
AluYi6/1-282                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAAA---TTA
111_Yc1_h20.1/1-281         ---CAGTGAAACCCCATCTCTACTAAAAA-----TACAAAAAA----TTA
AluYg6a2/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYg6/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
22_Yg6_397/1-280            ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
28_Yc1_65/1-280             ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
AluYg5b3/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
100_Y_h6.1/1-281            ---CGGTGAAACCCCATCTCTACTAAAAA-----TACAAAAAA----TTA
AluYj4/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TGCAAAAAA----TTA
50_Yj4_11/1-281             ---CGGTGAAACCCCGTCTCTACTAAAAA-----TGCAAAAAA----TTA
AluYj3/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TGCAAAAAA----TTA
60_Y_h14/1-282              ---AGGCGAAACCCTGTCTCTACTAAAAA-----TACAAAAAAA---TTA
AluYh9/1-281                ---CGGTGAAACCCCGCCTCTACTAAAAA-----TACAAAAAA----TTA
71_Yh9_20/1-282             ---CGGTGAAACCCCGCCTCTACTAAAAA-----TACAAAAAA----TTA
AluYh7/1-281                ---CGGTGAAACCCCGCCTCTACTAAAAA-----TACAAAAAA----TTA
AluYh3/1-281                ---CGGTGAAACCCCGCCTCTACTAAAAA-----TACAAAAAA----TTA
AluYe5/1-280                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
128_Ye5_d196/1-279          ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYe6/1-280                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYe4/1-282                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYe2/1-279                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
102_Y_T144C/1-281           ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYf2/1-283                ---CAGTGAAACCCCGTCTCTACTAAAAAA----CACAAAAAAA---TTA
86_Yf2_38/1-283             ---CAGTGAAACCCCGTCTCTACTAAAAAA----CACAAAAAAA---TTA
AluYf1/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluY/1-281                  ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYc1/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
25_Yc1_73/1-279             ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYc2/1-281                ---AGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
27_Yc1_66/1-280             ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
13_pAlu_A/1-286             ---CGGTGAAACCCCGTCTCTACTAAAAAAAAAATACAAAAAA----TTA
AluYa1_2/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYd3a1/1-270              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYd3/1-269                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYd2/1-269                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYd/1-269                 ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
19_Ya5_173/1-281            ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
125_Ya5_C138T/1-281         ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
64_Ya5_G25C/1-281           ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
92_Ya5#-Ya5#_95/1-278       ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
117_Ya5_G159C/1-281         ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
116_Ya5_G25C/1-281          ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
119_Ya5_G25C/1-281          ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
118_Ya5_G159C/1-281         ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
120_Ya5_AC275TG/1-281       ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa5a1/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa5b1/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa5a2/1-282              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAAA---TTA
20_Ya5_054/1-285            ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAAAAAATTA
AluYa5/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa8/1-280                ---CGGTGAAACCCCGTCTCTACTAAAAC-----TACAAAAAA----T-A
```

```
81_Ya5_C150T/1-281          ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa5c1/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
124_Ya5_G21A_C23T/1-281     ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
112_Ya5_rtSRPko/1-281       ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
18_Ya5_874/1-281            ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa5b2/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
99_Ya5_A131T_C206T/1-281    ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAATA----TTA
AluYa3_2/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa4_1/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa3_3/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa4_3/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa4_2/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa4b/1-281               ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa3_1/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
15_Ya4_677/1-281            ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
16_Ya4_704/1-283            AAACGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
AluYa4/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa4_5/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa4_4/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa3_5/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa3_4/1-281              ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa3/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
AluYa2/1-281                ---CGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAAA----TTA
62_Sx-Ya5/1-282             ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
AluSc/1-279                 ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
108_Sc_h1.1/1-279           ---TGATGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
AluSx/1-282                 ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
5_Sx_5/1-280                ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
42_Sx_10/1-279              ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
73_Sx_i22T/1-283            ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
4_Sx_13/1-280               ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
AluSz/1-282                 ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
AluSg/1-280                 ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
104_Sg_h11.1/1-280          ---TGGCGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
AluSp/1-283                 ---TGGAGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
105_Sp_h12.1/1-283          ---TGGAGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
8_Sx_425/1-280              ---TAGTGAAACCCCGTCTCTACTAAAA------TACAAAAA-----TTA
AluSq/1-283                 ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTA
79_Sx_11/1-282              ---TGGTGAAACCCCGTCTCTACTAAAAA-----TACAAAAA-----TTT
AluJo/1-282                 ---TAGCGAGACCCCGTCTCTACAAAAAA-----TACAAAAA-----TTA
32_Jo_18/1-280              ---TAGCGAGACCCCGT-TCTACAAAAAA-----TACAAAAA-----TTA
30_Jo_5/1-282               ---TAGCGAGACCCCGTCTCTACAAAAAA-----TACAAAAA-----TTA
AluJb/1-282                 ---TGGTGAAACCCCGTCTCTACAAAAAA-----TACAAAAA-----TTA
63_Jo-Ya5/1-282             ---TAGCGAGACCCCGTCTCTACAAAAAA-----TACAAAAA-----TTA
                                  ** ****    ***** ****        *****      *
```

```
AluYa1_1/1-281              GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
10_Y_534/1-281              GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
11_Y_h10.1/1-281            GCCGAGCGTAGTGGCAGGTGCCTGTAGTCCCAGCTACTCAG-GAGG-CTG
AluYd6/1-269                GCCGGGCGCAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYd8/1-269                GCCGGGCGCAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYc5/1-269                GCCGGGCGCAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb10/1-288               GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGG-GAGG-CTG
AluYb11/1-289               GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGG-GAGG-CTG
AluYb8/1-288                GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb9/1-288                GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTGGG-GAGG-CTG
AluYb7_2/1-288              GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb7_1/1-288              GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb7_3/1-288              GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb6_1/1-288              GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb7_4/1-288              GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb5/1-281                GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb6_2/1-281              GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb3a1/1-281              GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYb3a2/1-281              GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYbc3a/1-282              GCCGGGCGTGGTAGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYi6/1-282                GCCGGGCGTGATGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
111_Yc1_h20.1/1-281         GCAGGGCGTGGTAGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYg6a2/1-281              GCCGGGCATGGTGGCGCGTGCCTGTAGTCCCAGCTACACAG-GAGG-CTG
AluYg6/1-281                GCCGGGCATGGTGGCGCGCGCCTGTAGTCCCAGCTACACGG-GAGG-CTG
22_Yg6_397/1-280            GCCGGGCATGGTGGCGCGCGCCTGTAGTCCCAGCTACACGG-GAGG-CTG
28_Yc1_65/1-280             GCCGGGCATGGTGGCGCGCGCCTGTAGTCCCAGCTACACGG-GAGG-CTG
AluYg5b3/1-281              GCCGGGCATGGTGGCGCGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
100_Y_h6.1/1-281            GCCGGGCATGGTGGCGGGCGCCTGTAGTGCCAGCTACTCCG-GAGG-CTG
AluYj4/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTGCTCGG-GAGG-CTG
50_Yj4_11/1-281             GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTGCTCGG-GACG-CTG
AluYj3/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTGCTCGG-GAGG-CTG
60_Y_h14/1-282              GCTGGGCATGGTGGTGGGCACCTGCAGTCCGAGCTACTCGG-GAGG-CTG
AluYh9/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTGGTCCCGGCTACTCGG-GAGG-CTG
71_Yh9_20/1-282             GCCGGGCGTGGTGGCGGGCGCCTGTGGTCCCGGCTACTCGG-GAGGTCTG
AluYh7/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTGGTCCCGGCTACTCGG-GAGG-CTG
AluYh3/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYe5/1-280                GCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
128_Ye5_d196/1-279          GCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYe6/1-280                GCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYe4/1-282                GCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYe2/1-279                GCCGGGCGAGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
102_Y_T144C/1-281           GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYf2/1-283                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACGCGG-GAGG-CTG
86_Yf2_38/1-283             TCCGGGCGTGGTGGCGGGCGCCTGTAAACCCAGCTACGCGG-GAGG-CTG
AluYf1/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACGCGG-GAGG-CTG
AluY/1-281                  GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYc1/1-281                GCCGGGCGTGGTAGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
25_Yc1_73/1-279             GCCGGGCGTGGTAGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYc2/1-281                GCCGGGCGTGGTAGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
27_Yc1_66/1-280             GCCGGGCGTGGTAGCGGGCGCCTGTAGTCCCAGCTACTCGG-CAGG-CTG
13_pAlu_A/1-286             GCCGGGCGTGGTAGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYa1_2/1-281              GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYd3a1/1-270              GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGGAGAGG-CTG
AluYd3/1-269                GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYd2/1-269                GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYd/1-269                 GCCGGGCGCGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
19_Ya5_173/1-281            GCTGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
125_Ya5_C138T/1-281         GCTGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
64_Ya5_G25C/1-281           GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
92_Ya5#-Ya5#_95/1-278       GCCGGGCGTAGTGGCGGCCACAT-CAG--CCAGCTACTTGG-GAGG-CTG
117_Ya5_G159C/1-281         GCCGGGCGTAGTGGCGGGCGCCTCTAGTCCCAGCTACTTGG-GAGG-CTG
116_Ya5_G25C/1-281          GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
119_Ya5_G25C/1-281          GCCGGGCGTAGTGGCGGGCGCCTCTAGTCCCAGCTACTTGG-GAGG-CTG
118_Ya5_G159C/1-281         GCCGGGCGTAGTGGCGGGCGCCTCTAGTCCCAGCTACTTGG-GAGG-CTG
120_Ya5_AC275TG/1-281       GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa5a1/1-281              GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa5b1/1-281              GCCGGGCGTAGTGGCGGGCGCCTGTGGTCCCAGCTACTTGG-GAGG-CTG
AluYa5a2/1-282              GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
20_Ya5_054/1-285            GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa5/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa8/1-280                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCTAGCTACTTGG-GAGG-CTG
```

```
81_Ya5_C150T/1-281            GCCGGGCGTAGTGGTGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa5c1/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
124_Ya5_G21A_C23T/1-281       GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
112_Ya5_rtSRPko/1-281         GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
18_Ya5_874/1-281              GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa5b2/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
99_Ya5_A131T_C206T/1-281      GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa3_2/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa4_1/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa3_3/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYa4_3/1-281                GCCGGGCGTGGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa4_2/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYa4b/1-281                 GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluYa3_1/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
15_Ya4_677/1-281              GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
16_Ya4_704/1-283              GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa4/1-281                  GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa4_5/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa4_4/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa3_5/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa3_4/1-281                GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa3/1-281                  GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluYa2/1-281                  GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
62_Sx-Ya5/1-282               GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
AluSc/1-279                   GCTGGGCGTGGTGGCGCGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
108_Sc_h1.1/1-279             GCTAGGCGTGGTGGCGCGAGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluSx/1-282                   GCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
5_Sx_5/1-280                  GCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGG--AGG--TG
42_Sx_10/1-279                GCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
73_Sx_i22T/1-283              GCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
4_Sx_13/1-280                 GCCGGGCGTGGTGGCGCGCGCCT-TAATCCCAGCTACTCGG-GAGG-CTG
AluSz/1-282                   GCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
AluSg/1-280                   GCCGGGCGTGGTGGCGCGCGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
104_Sg_h11.1/1-280            GCTGGGCGTGGTGGCACGCGCCTGTAACCCCAGCTACTCGG-GAGG-CTG
AluSp/1-283                   GCCGGGCGTGGTGGCGCATGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
105_Sp_h12.1/1-283            GCTAGGTGTGGTGGCGCGGGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
8_Sx_425/1-280                GCCGGGCGCAGTGGCGGGCGCCTGTAATCTCAGCTACTTGG-GAGC-CTG
AluSq/1-283                   GCCGGGCGTGGTGGCGGGCGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
79_Sx_11/1-282                GCCGGGCGTAGTGGCGGGCGCCTGTAATCCCAGCTACTCGG-GAGG-CTG
AluJo/1-282                   GCCGGGCGTGGTGGCGCGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
32_Jo_18/1-280                GCCGGGCGTGGTGGCGCGCGC-TGTAGTCCCAGCTACTCGG-GAGG-CTG
30_Jo_5/1-282                 GCCGGGCGTGGTGGCGCGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
AluJb/1-282                   GCCGGGCGTGGTGGCGCGCGCCTGTAGTCCCAGCTACTCGG-GAGG-CTG
63_Jo-Ya5/1-282               GCCGGGCGTAGTGGCGGGCGCCTGTAGTCCCAGCTACTTGG-GAGG-CTG
                               *   *     * *      * *         *** *   *  *    **
```

```
AluYa1_1/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
10_Y_534/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
11_Y_h10.1/1-281            AGGCAGGAGAATGGCGT-GAACCC-AGGAGGTGGAGCTTGCAGTGAGCCG
AluYd6/1-269                AGGCAGGAGAATGGCGT-GAACCC-GGAAGGCGGAGCTTGCAGTGAGCGG
AluYd8/1-269                AGGCAGGAGAATGGCGT-GAACCC-GGAAGGCGGAGCTTGCAGTGAGCGG
AluYc5/1-269                AGGCAGGAGAATGGCGT-GAACCC-GGAAGGCGGAGCTTGCAGTGAGCGG
AluYb10/1-288               AGGCAGGAGAATGGCGT-GAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb11/1-289               AGGCAGGAGAATGGCGTTGAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb8/1-288                AGGCAGGAGAATGGCGT-GAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb9/1-288                AGGCAGGAGAATGGCGT-GAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb7_2/1-288              AGGCAGGAGAATGGCGT-GAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb7_1/1-288              AGGCAGGAGAATGGCGT-GAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb7_3/1-288              AGGCAGGAGAATGGCGT-GAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb6_1/1-288              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYb7_4/1-288              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYb5/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYb6_2/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAAGCGGAGCTTGCAGTGAGCCG
AluYb3a1/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYb3a2/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYbc3a/1-282              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCA
AluYi6/1-282                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
111_Yc1_h20.1/1-281         AGGCAGGAGAATGGCGT-GAACCT-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYg6a2/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGTCG
AluYg6/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGTCG
22_Yg6_397/1-280            AGGCAGGAGAATGGCGT-AAACCC-GGGAGGCGGAGCTTGCAGTGAGTCG
28_Yc1_65/1-280             AGGCAGGAGAATGGCGT-AAACCC-GGGAGGCGGAGCTTGCAGTGAGTCG
AluYg5b3/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGTCG
100_Y_h6.1/1-281            AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGTTTGCAGTGAGCTG
AluYj4/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
50_Yj4_11/1-281             AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYj3/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
60_Y_h14/1-282              AGGCAGGAGAATGGCAT-GAACCA-GTGAGGCGGAGCTTGCAGTGAGCCG
AluYh9/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCTG
71_Yh9_20/1-282             AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYh7/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYh3/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYe5/1-280                AGGCAGGAGAATGGCGT-GAACCCCGGGGGGCGGAGCCTGCAGTGAGCCG
128_Ye5_d196/1-279          AGGCAGGAGAATG-CGT-GAACCCCGGGGGGCGGAGCCTGCAGTGAGCCG
AluYe6/1-280                AGGCAGGAGAATGGCGT-GAACCCCGGGGGGCGGAGCCTGCAGTGAGCCG
AluYe4/1-282                AGGCAGGAGAATGGCGT-GAACCCCGGGGGGCGGAGCCTGCAGTGAGCCG
AluYe2/1-279                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
102_Y_T144C/1-281           AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYf2/1-283                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
86_Yf2_38/1-283             AGGCAGGAAAATGGCAT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYf1/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluY/1-281                  AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYc1/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
25_Yc1_73/1-279             AGGCAGGAGAATGGCGT-GA-CCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYc2/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
27_Yc1_66/1-280             AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
13_pAlu_A/1-286             AGGCAGGGGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa1_2/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYd3a1/1-270              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYd3/1-269                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYd2/1-269                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYd/1-269                 AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
19_Ya5_173/1-281            AGGCAGGAGAATGGCGT-GAACCC-AGGAGGCGGAGCTTGCAGTGAGCCG
125_Ya5_C138T/1-281         AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
64_Ya5_G25C/1-281           AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
92_Ya5#-Ya5#_95/1-278       AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
117_Ya5_G159C/1-281         AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
116_Ya5_G25C/1-281          AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
119_Ya5_G25C/1-281          AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
118_Ya5_G159C/1-281         AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
120_Ya5_AC275TG/1-281       AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa5a1/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa5b1/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa5a2/1-282              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
20_Ya5_054/1-285            AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa5/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa8/1-280                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
```

```
81_Ya5_C150T/1-281            AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa5c1/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
124_Ya5_G21A_C23T/1-281       AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
112_Ya5_rtSRPko/1-281         AGGCAGGAGTTACCCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
18_Ya5_874/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCTG
AluYa5b2/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
99_Ya5_A131T_C206T/1-281      AGGCAGGAGAATGGCGT-GAACCT-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa3_2/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa4_1/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa3_3/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa4_3/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa4_2/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa4b/1-281                 AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa3_1/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
15_Ya4_677/1-281              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
16_Ya4_704/1-283              AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAATGAGCCG
AluYa4/1-281                  AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa4_5/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa4_4/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa3_5/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa3_4/1-281                AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa3/1-281                  AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluYa2/1-281                  AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
62_Sx-Ya5/1-282               AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
AluSc/1-279                   AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
108_Sc_h1.1/1-279             AAGCAGGAGAATCGCTT-GAACCC-GGGAGGCAGAGGTTGCAGTGAGCCG
AluSx/1-282                   AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
5_Sx_5/1-280                  AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
42_Sx_10/1-279                AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
73_Sx_i22T/1-283              AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
4_Sx_13/1-280                 AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
AluSz/1-282                   AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
AluSg/1-280                   AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
104_Sg_h11.1/1-280            AGGCAGGAGAATTGCTT-GAATCC-GGGAGGTGGAGGTTGCAGTGAGCTG
AluSp/1-283                   AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCGGTGAGCCG
105_Sp_h12.1/1-283            AGGCAGGAGAATCTCTT-GAACCC-AGGAGGCGGAGGTTGCAGTGAGCCG
8_Sx_425/1-280                AGGCAGGAGAATCGCTT-GAACCC-AGGAGGCGGAGGTTGCAGTGACCCA
AluSq/1-283                   AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
79_Sx_11/1-282                AGGCAGGAGAATCGCTT-GAACCC-GGGAGGCGGAGGTTGCAGTGAGCCG
AluJo/1-282                   AGGCAGGAGGATCGCTT-GAGCCC-AGGAGTTCGAGGCTGCAGTGAGCTA
32_Jo_18/1-280                AGGCAGGAGGATCGCTT-GAGCCC-AGGAGTTCGAGGCTGCAGTGAGCTA
30_Jo_5/1-282                 AGGCAGGAGGATCGCTT-GAGCCC-AGGAGTTCGGGGCTGCAGTGAGCTA
AluJb/1-282                   AGGCAGGAGGATCGCTT-GAGCCC-GGGAGGTCGAGGCTGCAGTGAGCCG
63_Jo-Ya5/1-282               AGGCAGGAGAATGGCGT-GAACCC-GGGAGGCGGAGCTTGCAGTGAGCCG
                              * *****       * *  *  *          * *  ***  ***
```

```
AluYa1_1/1-281              AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
10_Y_534/1-281              AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
11_Y_h10.1/1-281            AGATTGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYd6/1-269                AGATCGCGCCACAGCACTCC-------CGCCTGGGCGACA-GAACGAGAC
AluYd8/1-269                AGATCGCGCCACAGCACTCC-------CGCCTGGGCGACA-GAACGAGAC
AluYc5/1-269                AGATCGCGCCACAGCACTCC-------CGCCTGGGCGACA-GAGCGAGAC
AluYb10/1-288               AGATTGCGCCACTGCAGTCCGCAGTCCAGCCTGGGCGACA-GAGCGAGAC
AluYb11/1-289               AGATTGCGCCACTGCAGTCCGCAGTCCAGCCTGGGCGACA-GAGCGAGAC
AluYb8/1-288                AGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGAC
AluYb9/1-288                AGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGAC
AluYb7_2/1-288              AGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGAC
AluYb7_1/1-288              AGATCGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGAC
AluYb7_3/1-288              AGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGAC
AluYb6_1/1-288              AGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGAC
AluYb7_4/1-288              AGATTGCGCCACTGCAGTCCGCAGTCCGGCCTGGGCGACA-GAGCGAGAC
AluYb5/1-281                AGATTGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYb6_2/1-281              AGATTGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYb3a1/1-281              AGATTGCGCCACTGCAGTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYb3a2/1-281              AGATAGCGCCACTGCAGTCC-------GGCCTGGGCGAAA-GAGCGAGAC
AluYbc3a/1-282              AGATAGCGCCACTGCAGTCC-------AGCCTGGGCGAAA-GAGCGAGAC
AluYi6/1-282                AGATTGCGCCACTGCACTCC-------CGCCTGGGCCACA-GAGCGAGAC
111_Yc1_h20.1/1-281         AGATTGCGCCACTGCACTCC-------CACCTGGGCCACA-GAGCGAGAC
AluYg6a2/1-281              AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAAAC
AluYg6/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAAAC
22_Yg6_397/1-280            AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAAAC
28_Yc1_65/1-280             AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAAAC
AluYg5b3/1-281              AGATCGCGCCACTGCGCTCC-------AGCCTGGGCGACA-GAGCGAAAC
100_Y_h6.1/1-281            ACATCGCACCACTGCACTCC-------AGCCTGGTCGACA-GAGCGAGAC
AluYj4/1-281                AGATCGTGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
50_Yj4_11/1-281             AGATCGTGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYj3/1-281                AGATCGTGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
60_Y_h14/1-282              AGATCATGCCACTGCACTCC-------AACCTGGGCGACA-GAGCGAGAC
AluYh9/1-281                AGGTCGCGCCACTGCACCCC-------AGCCTGGGCGACA-GAGCGAGAC
71_Yh9_20/1-282             AGGTCGCGCCACTGCACCCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYh7/1-281                AGGTCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYh3/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYe5/1-280                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-G--CGAGAC
128_Ye5_d196/1-279          AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-G--CGAGAC
AluYe6/1-280                AGATCGCGCCACTGCACTCC-------AACCTGGGCGACA-G--CGAGAC
AluYe4/1-282                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYe2/1-279                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-G--CGAGAC
102_Y_T144C/1-281           AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYf2/1-283                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
86_Yf2_38/1-283             AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYf1/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluY/1-281                  AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYc1/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
25_Yc1_73/1-279             AGATCGCGCCA-TGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYc2/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
27_Yc1_66/1-280             AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
13_pAlu_A/1-286             AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa1_2/1-281              AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYd3a1/1-270              AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYd3/1-269                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYd2/1-269                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYd/1-269                 AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
19_Ya5_173/1-281            AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
125_Ya5_C138T/1-281         AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
64_Ya5_G25C/1-281           AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
92_Ya5#-Ya5#_95/1-278       AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
117_Ya5_G159C/1-281         AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
116_Ya5_G25C/1-281          AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGTG
119_Ya5_G25C/1-281          AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGTG
118_Ya5_G159C/1-281         AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGTG
120_Ya5_AC275TG/1-281       AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGTG
AluYa5a1/1-281              AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa5b1/1-281              AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa5a2/1-282              AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
20_Ya5_054/1-285            AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa5/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa8/1-280                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
```

```
81_Ya5_C150T/1-281            AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa5c1/1-281                AGGTCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
124_Ya5_G21A_C23T/1-281       AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
112_Ya5_rtSRPko/1-281         AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
18_Ya5_874/1-281              AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa5b2/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
99_Ya5_A131T_C206T/1-281      AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa3_2/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa4_1/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa3_3/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa4_3/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa4_2/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa4b/1-281                 AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa3_1/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
15_Ya4_677/1-281              AGATCGTGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
16_Ya4_704/1-283              AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa4/1-281                  AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa4_5/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa4_4/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa3_5/1-281                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa3_4/1-281                AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa3/1-281                  AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluYa2/1-281                  AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
62_Sx-Ya5/1-282               AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluSc/1-279                   AGATCGCGCCACTGCACTCC-------AGCCTGG-CGACA-GAGCGAGAC
108_Sc_h1.1/1-279             ACATCGCGCCACTGCACTCC-------AGCCTGG-CGACA-GAGCAAGAC
AluSx/1-282                   AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
5_Sx_5/1-280                  AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
42_Sx_10/1-279                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
73_Sx_i22T/1-283              AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
4_Sx_13/1-280                 AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluSz/1-282                   AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluSg/1-280                   AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
104_Sg_h11.1/1-280            AGATCGCACCACTGCACTCC-------AGCCTGGGTGACA-GAGCGAGAC
AluSp/1-283                   AGATCGCGCCATTGCACTCC-------AGCCTGGGCAACAAGAGCGAAAC
105_Sp_h12.1/1-283            AGATCGCGCCATTGCACTCC-------AGCCTGGGCAACAAAAGTGAAAC
8_Sx_425/1-280                AGATCGCGCCATTGCACTCC-------AGCCTGG-CAACA-GAGCGAGAC
AluSq/1-283                   AGATCGCGCCACTGCACTCC-------AGCCTGGGCAACAAGAGCGAAAC
79_Sx_11/1-282                AGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluJo/1-282                   TGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
32_Jo_18/1-280                TGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
30_Jo_5/1-282                 TGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
AluJb/1-282                   TGATCGCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
63_Jo-Ya5/1-282               AGATCCCGCCACTGCACTCC-------AGCCTGGGCGACA-GAGCGAGAC
                                 *    ***  **   **         *****   * *      *
```

AluYa1_1/1-281	TCCGTCTC
10_Y_534/1-281	TCCGTCTC
11_Y_h10.1/1-281	TCCGTCTC
AluYd6/1-269	TCCGTCTC
AluYd8/1-269	TCCGTCTC
AluYc5/1-269	TCCGTCTC
AluYb10/1-288	TCCGTCTC
AluYb11/1-289	TCCGTCTC
AluYb8/1-288	TCCGTCTC
AluYb9/1-288	TCCGTCTC
AluYb7_2/1-288	TCCGTCTC
AluYb7_1/1-288	TCCGTCTC
AluYb7_3/1-288	TCCGTCTC
AluYb6_1/1-288	TCCGTCTC
AluYb7_4/1-288	TCCGTCTC
AluYb5/1-281	TCCGTCTC
AluYb6_2/1-281	TCCGTCTC
AluYb3a1/1-281	TCCGTCTC
AluYb3a2/1-281	TCCGTCTC
AluYbc3a/1-282	TCCGTCTC
AluYi6/1-282	TCCGTCTC
111_Yc1_h20.1/1-281	TCCGTCTC
AluYg6a2/1-281	TCCGTCTC
AluYg6/1-281	TCCGTCTC
22_Yg6_397/1-280	TCCGTCTC
28_Yc1_65/1-280	TCCGTCTC
AluYg5b3/1-281	TCCGTCTC
100_Y_h6.1/1-281	TCCGTCTC
AluYj4/1-281	TCCGTCTC
50_Yj4_11/1-281	TCCGTCTC
AluYj3/1-281	TCCGTCTC
60_Y_h14/1-282	CCTGTCTC
AluYh9/1-281	TCCGTCTC
71_Yh9_20/1-282	TCCGTCTC
AluYh7/1-281	TCCGTCTC
AluYh3/1-281	TCCGTCTC
AluYe5/1-280	TCCGTCTC
128_Ye5_d196/1-279	TCCGTCTC
AluYe6/1-280	TCCGTCTC
AluYe4/1-282	TCCGTCTC
AluYe2/1-279	TCCGTCTC
102_Y_T144C/1-281	TCCGTCTC
AluYf2/1-283	TCCGTCTC
86_Yf2_38/1-283	TCCGTCTC
AluYf1/1-281	TCCGTCTC
AluY/1-281	TCCGTCTC
AluYc1/1-281	TCCGTCTC
25_Yc1_73/1-279	TCCGTCTC
AluYc2/1-281	TCCGTCTC
27_Yc1_66/1-280	TCCGTCTC
13_pAlu_A/1-286	TCCGTCTC
AluYa1_2/1-281	TCCGTCTC
AluYd3a1/1-270	TCCGTCTC
AluYd3/1-269	TCCGTCTC
AluYd2/1-269	TCCGTCTC
AluYd/1-269	TCCGTCTC
19_Ya5_173/1-281	TCCGTCTC
125_Ya5_C138T/1-281	TCCGTCTC
64_Ya5_G25C/1-281	TCCGTCTC
92_Ya5#-Ya5#_95/1-278	TCCGTCTC
117_Ya5_G159C/1-281	TCCGTCTC
116_Ya5_G25C/1-281	TCCGTCTC
119_Ya5_G25C/1-281	TCCGTCTC
118_Ya5_G159C/1-281	TCCGTCTC
120_Ya5_AC275TG/1-281	TCCGTCTC
AluYa5a1/1-281	TCCGTCTC
AluYa5b1/1-281	TCCGTCTC
AluYa5a2/1-282	TCCGTCTC
20_Ya5_054/1-285	TCCGTCTC
AluYa5/1-281	TCCGTCTC
AluYa8/1-280	TCCGTCTC

```
81_Ya5_C150T/1-281            TCCGTCTC
AluYa5c1/1-281                TCCGTCTC
124_Ya5_G21A_C23T/1-281       TCCGTCTC
112_Ya5_rtSRPko/1-281         TCCGTCTC
18_Ya5_874/1-281              TCCGTCTC
AluYa5b2/1-281                TCCGTCTC
99_Ya5_A131T_C206T/1-281      TCCGTCTC
AluYa3_2/1-281                TCCGTCTC
AluYa4_1/1-281                TCCGTCTC
AluYa3_3/1-281                TCCGTCTC
AluYa4_3/1-281                TCCGTCTC
AluYa4_2/1-281                TCCGTCTC
AluYa4b/1-281                 TCCGTCTC
AluYa3_1/1-281                TCCGTCTC
15_Ya4_677/1-281              TCCGTCTC
16_Ya4_704/1-283              TCCGTCTC
AluYa4/1-281                  TCCGTCTC
AluYa4_5/1-281                TCCGTCTC
AluYa4_4/1-281                TCCGTCTC
AluYa3_5/1-281                TCCGTCTC
AluYa3_4/1-281                TCCGTCTC
AluYa3/1-281                  TCCGTCTC
AluYa2/1-281                  TCCGTCTC
62_Sx-Ya5/1-282               TCCGTCTC
AluSc/1-279                   TCCGTCTC
108_Sc_h1.1/1-279             TCTGTCTC
AluSx/1-282                   TCCGTCTC
5_Sx_5/1-280                  TCCGTCTC
42_Sx_10/1-279                TCCGTCTC
73_Sx_i22T/1-283              TCCGTCTC
4_Sx_13/1-280                 TCCGTCTC
AluSz/1-282                   TCCGTCTC
AluSg/1-280                   TCCGTCTC
104_Sg_h11.1/1-280            TCCATCTC
AluSp/1-283                   TCCGTCTC
105_Sp_h12.1/1-283            TCCGTCTC
8_Sx_425/1-280                TCCGTCTC
AluSq/1-283                   TCCGTCTC
79_Sx_11/1-282                TCCGTCTC
AluJo/1-282                   CCTGTCTC
32_Jo_18/1-280                CCTGTCTC
30_Jo_5/1-282                 CCTGTCTC
AluJb/1-282                   CCTGTCTC
63_Jo-Ya5/1-282               TCCGTCTC
                               *  ****
```


Acknowledgments

We thank Shari Corin, Paul Doetsch, Natasha Degtyareva, Rebecca Iskow, and James Schroeder for critical discussions and reading of the manuscript. We also thank Jackie Griffith for technical assistance and Lisa Elon for statistical analysis. This work was funded by grants from Sun Microsystems (S.E.D.) and grants F32HG004207 (R.E.M.), R01GM060518 (J.V.M.), and R01HG002898 (S.E.D.) from the National Institutes of Health. H.K. and O.W. are supported by a VIDI grant from the Dutch National Science Organization (NWO-CW [Nr. 700.54.427]) awarded to O.W. S.E.D. dedicates this study to the memory of Jeffrey Devine.

Chapter 6

Discussion

Discussion

Seven years after the first publication of the human genome sequence (Lander, *et al*., 2001), there is still much to be learned about the natural variation that distinguishes individuals from each other. Human genetic variation underlies our differences and commonalities, and defining it will help us learn how our genome came to be and what it does. Genome sequences of several individuals have now been published (Venter, *et al*. 2007; Bentley, *et al*. 2008; Wang, *et al*., 2008; Wheeler, *et al*. 2008), but due to the considerable resources these projects still require, large-scale comparative analysis remains an efficient way to learn about human diversity. Only very recently were we finally able to observe the burden of transposable elements that our genomes have always maintained. In this short time we have made considerable progress, but we are far from fully understanding the role these genetic parasites have played in our history and will continue to play in our future.

While genetic variation can be discovered through automated comparative genomics methods, understanding the underlying causes of that variation require insight and experimentation. The work of this dissertation has both increased our knowledge of genetic variation in humans, and our understanding of how retrotransposons contribute to that variation. We have identified a substantial amount of new single nucleotide polymorphisms (Chapter 2), and we also identified as many as a third of the common transposon polymorphisms in human populations (Chapter 3). Many of these markers have been incorporated into the HapMap and will continue to inform as markers in

genome-wide association studies, where they can be used to identify the genetic basis of human disease and traits.

We have also learned which of the 848 transposon families found in our genomes are currently active (Chapter 3), and which have transposed since we last shared a common ancestor with our species' closest living relative, the chimpanzee (Chapter 4). Building on work done on human L1s (Brouha, *et al*., 2003), our analysis of the requirements for Alu activity has confirmed aspects of previous models, such as the requirement for binding to SRP9/14 (Boeke 1999; Dewannieux, *et al*. 2003). Likewise, our work has enhanced our understanding of this mechanism by demonstrating an optimal SRP9/14 binding affinity for activity (Chapter 5). Experimental and conservation studies of Alu were used to develop a model to predict the likelihood of a given Alu transcript to retrotranspose. With this model, we estimated the total number of active Alu sequences to be near 10,000, or 1% of all Alus in our genome (Chapter 5).

Human genetic variation

The findings of Chapter 2 demonstrated the power of uniform computational methods in the detection of SNP analysis and genetic variation. In 2003 a collection of public and private scientists and funding agencies announced the HapMap project (International Human Genome Sequencing Consortium 2001). Their goal was to create a simplified map of uniformly-spaced genetic markers that took advantage of existing human haplotype architecture. The HapMap would standardize markers and jumpstart genome-wide association studies in order to locate regions containing genetic factors relating to disease, susceptibility to infection, and variation in drug and environmental responses.

The human population was estimated to contain10 million common SNPs, which would be represented in the HapMap by 600,000 tag SNPs (The International HapMap Consortium 2003).

The 140, 696 SNPs discovered in Chapter 2 made up 6% of the 2.2 million already uploaded into the public database dbSNP. When phase 1 of the project was completed and the first Tag SNPs selected, the number had reached 2.8 million. By completion of phase 2 of the project, the 3.1 million SNPs that were successfully genotyped included approximately 40,000 of our SNPs. The HapMap continues to be an important source for large-scale, genome wide association studies, and in addition to its mandate, also contributes to studies of recombination, population stratification and positive selection.

Recent transposon activity

As the large-scale genotyping required for the HapMap project progressed, the commonality of genetic structural variants began to emerge. The high-quality trace sequences (Phred scores >25) generated for the SNP analysis in Chapter 2, were used in Chapter 3 to search for insertion/deletion polymorphisms. A new bioinformatics pipeline was used to scan over 600,000 indel polymorphisms for recent transposon insertions. This method involved aligning trimmed trace sequences to the reference genome through a short anchor sequence, and provided an improvement over selectivity biases inherent in PCR display assays (Sheen, *et al*. 2000), which allowed for relative comparisons of recent transposition activities among the various families and subfamilies. It also could identify recent L1-driven retrotransposition of pseudogenes and small cellular RNAs. The

ability to identify polymorphic transposon insertions without the requirement of primer design also allowed for the *a priori* discovery of recent transposition events from transposons whose activity had not yet been reported. This advantage was responsible for the discovery that SVA was the most polymorphic transposon family in humans (Chapter 3). Disease-causing insertions of SVA had been described (Hassoun, *et al*. 1994; Kobayashi, *et al*. 1998; Rohrer, *et al*. 1999; Wilund, *et al*. 2002), but the proportion of active elements of this young transposon family had previously been unknown. The status of SVA along with L1 and Alu as the only active transposons in humans is now widely accepted (Mandal and Kazazian 2008), though the precise mechanism required for the retrotransposition of this element has yet to be determined.

The unexpected finding of several polymorphic 'ancient' Alu S subfamilies in Chapter 3 raised the prospect that older Alu elements might retain some level of retrotransposition competence. The existence of active copies of Alu S, which had previously been considered extinct, is particularly compelling as older Alu S and J elements make up over 80% of the ~900,000 full length copies present in humans. Assuming these are genuine recent insertions and not due to gene conversion events or lineage sorting of ancient polymorphisms, a large number of these older elements may continue to contribute to human genetic diversity and disease. In fact, two genuine disease-causing Alu S insertions have been described since this study (Kloor, *et al*. 2004; Teugels, *et al*. 2005), and current activity of the Alu S family has been established in Chapter 5 of this dissertation.

The polymorphic transposons described in Chapter 3, and the 600,000 polymorphic insertion/deletions from which they were identified, along with the ongoing

discovery of CNVs and intermediate sized structural variants (ISVs) represents a trend beyond the cataloging of SNPs. There is increasing appreciation for the structural variation component of human genetic diversity, and examining these forms of genetic variation continues (Mills, *et al*. 2006; Newman, *et al*. 2006; Khaja, *et al*. 2006; Redon, *et al*. 2006). The eventual integration of polymorphic transposon insertions, small indels, and CNVs into the human HapMap will broaden our understanding of human genetic diversity and genomic dynamics, and increase the current marker density of the genomic landscape.

The power of active retrotransposons to create human genetic variation might also, over time, help to shape human speciation. The comprehensive identification of transposable elements that are specific to either the human or chimp allowed us to observe how the founders of our own elements might behave, if granted six million years in a slightly different organism. As in our own species, the same three retroelements, L1, Alu, and SVA, were responsible for nearly all new transposon insertions in chimpanzees since our last common ancestor. A large difference between the two datasets however, was the total number of species-specific retrotransposition events. During the past 6 million years, Alu elements have inserted in humans at three times the rate that they have in chimps, and 263 of these somewhat recent insertions belong to the ancient Alu S family. One reason for this difference may have been the evolution of 'hotter' L1s in humans, which, in turn, drove up Alu and SVA retrotransposition events. The chimp L1s in contrast, were highly truncated and nearly all incapable of making functional L1 proteins. A later study by Lee, *et al* (2007) identified two chimp-specific L1 lineages,

only one of which had L1s with intact ORFs. This group addressed the prospect that sequencing errors may have accounted for the poor representation of intact chimp L1 ORFs. By resequencing several more conserved chimp L1s they identified only 5 potentially competent elements driving retrotransposition in chimpanzees, whereas the average human has 80-100 (Brouha, *et al*. 2003).

The many mechanisms by which retrotransposons manipulate the host genomes they occupy (see Chapter 1), and the diversity they exhibit in populations, make them good candidates for generating functional variation that contributed to our evolution (reviewed in Böhne, *et al*. 2008). In certain environmental backgrounds, the slight functional effects retrotransposon insertions may have on neighboring genes, or the introduction of gene fragments, splice sites or regulatory regions via 5' or 3' transductions to new locations, could subject them to selection. An analysis of all human SVA 3' transduction events in humans by Xing, *et al* (2006) finds 143 events duplicating 53 kb of genomic sequence. An entire gene of undetermined function, AMAC, was duplicated three times by such events, two of which exhibit differential expression patterns. The human-specific Alu insertions that we identified in the coding regions of genes (Chapter 4) may account for some of the phenotypic differences between humans in chimps (Table 6-1).

Gene	Description	Function
LEP	leptin precursor (Obesity factor)	This protein has several endocrine functions
RAB21	member RAS oncogene family	Unknown function
ZNF543	zinc finger protein 543	Unknown function
APOL1	apolipoprotein-L1 precursor	May play a role in lipid exchange and transport
CP135	centrosomal protein of 135 kDa	Unknown function
DNAJA5	DNAJ homology subfamily A member 5 isoform 2	Unknown function
AK1D1	3-oxo-5-beta-steroid 4-dehydrogenas	Catalyzes the reduction of progesterone

Table 6-1. Human-specific Alu insertions into coding regions of genes.

The lack of any other species closely related to humans makes the 6 million year window available to us through chimpanzee sequences the best resolution we have to study recent retrotransposition activity. However, the announcement of nuclear sequences obtained from Neanderthal bones and the initiation of the Neanderthal genome sequencing project (Green, *et al*, 2006; Noonan, *et al*. 2006), raises the prospect of a closer look. The difficulty in obtaining insertion sites and repetitive retrotransposon sequences from the short fragments of ancient DNA, combined with issues relating to the reliable mapping of these elements to the proper genomic loci is considerable. But if this can be overcome we may one day better know the state of our genome not only 6 million, but also 500,000 years ago. A wealth of genetic variation possibly causing slight phenotypic differences would then be available.

Exploring Alu activity

Several lines of evidence from Chapters 3 and 4 revealed limitations in our understanding of Alu activity. The earlier concept of the serial evolution of a small number of ‘master’

or ‘source’ genes, which generate all new Alu insertions (Shen, *et al*. 1991; Deininger, *et al*. 1992; Shaikh, *et al*. 1996) contrasted with the radial evolution of many different active Alus described in our polymorphic studies. The apparent activity of Alu S elements in modern times also caused us to explore the sequence requirements for activity of Alu elements. By concentrating only on the core sequence, holding expression constant through a uniform 7SL enhancer element and identical flanking sequences, we separated activity due to transcript quality from activity due to transcript quantity.

Previous studies noted the conservation of 7SL SRP9/14 binding sites in the Alu retrotransposon (Hsu, *et al*. 1995; Sarrowa, *et al*. 1997). The ability of Alu RNA to bind SRP9/14 has been demonstrated (Bovia, *et al*. 1997), and models of Alu retrotransposition requiring binding to SRP9/14 proteins have been proposed (Sinnett, *et al*. 1991; Boeke, *et al*. 1997; Dewannieux, *et al*. 2003; Mills, *et al*. 2007). Two SRP9/14 binding sites are present on the Alu element, one on each monomer (left and right). However, the conservation at these two sites differs between left and right monomers (Mills, *et al*. 2007). The SRP9/14 binding site on the left monomer of Alu is much more conserved than that of the right monomer in all Alu lineages. Nevertheless, some conservation is still evident in the right monomer SRP9/14 binding site, and all Alu subfamilies show fewer mutations in both left and right SRP9/14 binding sequences than non-SRP binding sequences. It is expected from this observation that SRP9/14 binding to the left monomer is more integral for retrotransposition than right monomer binding. Alternatively, Alu retrotransposition may have at one point lost the requirement for right monomer binding and the right monomer SRP9/14 binding site has not yet lost signs of conservation.

One goal of the Alu retrotransposition assays was to resolve the roles of left vs. right monomer binding. This was approached using a previously characterized single nucleotide mutation in 7SL RNA shown to severely disrupt SRP9/14 binding (Chang, *et al*. 1997). Disrupting left monomer SRP9/14 binding in Alu resulted in a severe decrease in retrotransposition activity, whereas disruption to the corresponding right monomer resulted in only a mild decrease in activity. These data do support the importance, but not the requirement, of SRP9/14 binding to the left Alu monomer for retrotransposition. Alu does appear to be able to retrotranspose with only a single SRP9/14 heteromer bound to either monomer, but does so with less efficiency. This may be a result of the instability of the ribonucleoprotein structure when bound to only a single SRP9/14 heteromer (Oliver Weichenrieder, personal communication). The different reductions in activity levels (88% and 25%) resulting from these experiments may be due to lower affinity binding of the right monomer to SRP9/14 (Figure 5-4D and 5-4E), or may also be attributed to physical differences in the ability of L1 ORF2 protein to access the Alu poly(A) due to Alu orientation at the ribosome (Figure 6-1).

The identification of active elements beyond the expected younger subfamilies suggests an evolutionary model of Alu source elements from both young and old subfamilies. New insertions found from human-human (Chapter 3, Table 1-1) and human-chimp (Chapter 4, Table 1-1) polymorphic studies, bare this out. Contrasting Table 1-1 with Table 5-2 and data from Chapter 5 indicates that the degree to which a given Alu subfamily contributes to new (polymorphic) insertions is a function

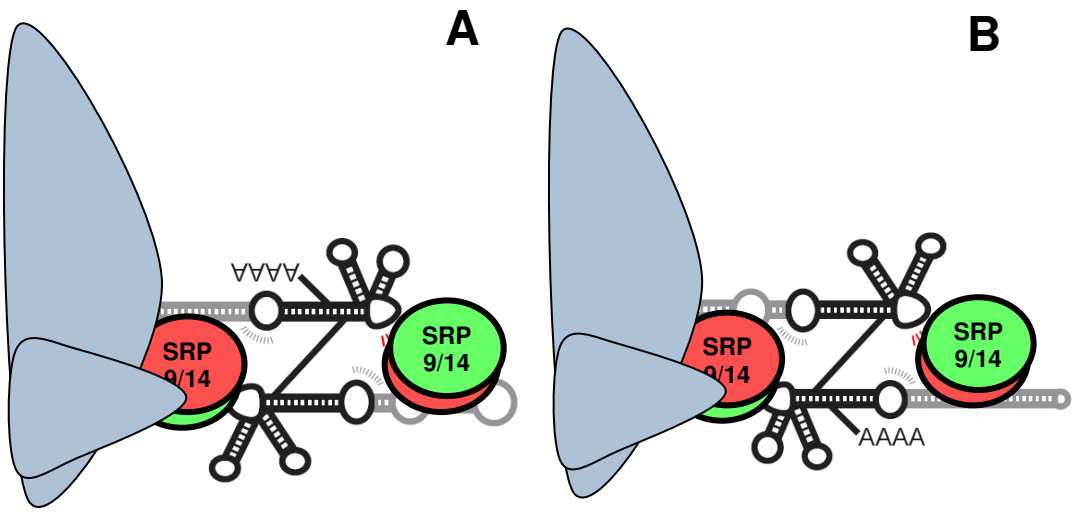


Figure 6-1. Alu retrotransposition favors SRP9/14 binding at the left monomer. Alu dimer RNA complexed with two SRP9/14 heteromers is a stable structure that can bind to the ribosome (gray) in two orientations, left monomer facing (A) or right monomer facing (B). Disrupting left monomer binding decreased activity 88%. Disrupting right monomer binding only decreased activity 25% (Chapter 5).

of the (1) activity level (as ascertained based on sequence conservation of key sites, Table 5-2) and (2) copy number of the element. This relationship is supported by conservation analysis and polymorphic insertion data (Table 6-2).

Alu activity is a result of both sequence integrity of the Alu core, and 'expressability' due to surrounding sequence and genomic context. The findings in Chapter 5 suggest that, possibly due in part to the absence of a coding region, core sequence requirements for Alu activity are more plastic than those of L1. Indeed, Alu elements are difficult to inactivate by small numbers of mutations. Given currently accepted estimations of Alu mutation rates (Carrol, *et al*. 2001), the expectation is that a

	Total number present in the genome	Total number conserving 124 'key' sites common in active elements. (allowing 1 mutation, from Table 5-2)	Proportion	Total number of polymorphic insertions in humans, by family (from Table 1-1)	Proportion
Alu S	**552,383**	**117**	**.016**	**29**	**.020**
Alu Y	**135,293**	**7,138**	**.984**	**1,444**	**.980**

Table 6-2. Contribution of Alu S and Y elements to total predicted active elements and total recent insertions in humans. The smaller proportion of recent insertions due to older Alu S elements corresponds with the proportion not of total Alu S elements, but of Alu S elements with predicted activity based on conservation of key sites (from Table 1-1 and Table 5-2).

highly active element may continue to produce active transcripts up to 17-27 million years after integration.

Future models and experiments are needed to refine our current awareness of 'key sites' to a more sophisticated understanding of the base-specific interactions these sites perform. A true predictive model should aim to recognize the 'key structure' of Alu RNA that ultimately grants it the leverage necessary to compete with L1 mRNA for L1 protein. While examination of the Alu element core sequence can determine the retrotransposition capability of a given Alu transcript, when it comes to predicting which specific Alu elements in the genome might actually retrotranspose, core sequence is only half the story. The other half awaits comprehensive analysis of the flanking sequences and genomic and cellular context that allow Alu transcription.

It is perhaps not surprising that the study of human variation and active transposons should be intertwined. As recurrent sources of new genetic variation, active retrotransposons take their place alongside more passive mechanisms of heritable genetic change such as DNA damage, replication errors or aberrant recombination events. On the surface, the potential threat of disruption to the function of our genomic architecture is great. Accumulating cases of the pathological consequences of transposon activity bare that out. Yet as a whole, our species, like nearly all other forms of life, have managed to host these elements for hundreds of millions of years.

References

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

Akagi, K., Li, J., Stephens, R.M., Volfovsky, N., Symer, D.E. (2008) Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. *Genome Res* **18**: 869-880.

Alemán, C., Roy-Engel, A.M., Shaikh, T.H., Deininger, P.L. (2000) Cis-acting influences on Alu RNA levels. *Nucleic Acids Res* **28**: 4755-61.

Alisch, R.S., Garcia-Perez, J.L., Muotri, A.R., Gage, F.H., Moran, J.V. (2006) Unconventional translation of mammalian LINE-1 retrotransposons. *Genes & Dev* **20:** 210–224.

Altman, D.G., Machin, D., Bryant, T.N., Gardner, M.J. (2000) Statistics with Confidence, 2d edition. BMJ books, Bristol, UK.

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., Lander, E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513-516.

Badge, R.M., Alisch, R.S., Moran, J.V. (2003) ATLAS, a system to selectively identify human-specific L1 insertions. *Am J Hum Genet* **72**: 823-838.

Bailey, J.A., Carrel, L., Chakravarti, A., Eichler, E.E. (2000) Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc Natl Acad Sci U S A* **97**:6634-6639.

Bailey, J.A., Gu, Z., Clark, R.A., Reinert, K., Samonte, R.V., Schwartz, S., Adams, M.D., Myers, E.W., Li, P.W., Eichler, E.E. (2002) Recent segmental duplications in the human genome. *Science* **297**: 1003-1007.

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

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

Belancio, V.P., Hedges, D.J., Deininger, P.L. (2008) Mammalian non-LTR retrotransposons: For better or worse, in sickness and in health. *Genome Res* **18**: 343–358.

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

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., Boutell, J.M., Bryant, J., Carter, R.J., Keira, Cheetham, R., Cox, A.J., Ellis, D.J., Flatbush, M.R., Gormley, N.A., Humphray, S.J., Irving, L.J., Karbelashvili, M.S., Kirk, S.M., Li, H., Liu, X., Maisinger, K.S., Murray, L.J., Obradovic, B., Ost, T., Parkinson, M.L., Pratt, M.R., Rasolonjatovo, I.M., Reed, M.T., Rigatti, R., Rodighiero, C., Ross, M.T., Sabot, A., Sankar, S.V., Scally, A., Schroth, G.P., Smith, M.E., Smith, V.P., Spiridou, A., Torrance, P.E., Tzonev, S.S., Vermaas, E.H., Walter, K., Wu, X., Zhang, L., Alam, M.D., Anastasi, C., Aniebo, I.C., Bailey, D.M., Bancarz, I.R., Banerjee, S., Barbour, S.G., Baybayan, P.A., Benoit, V.A., Benson, K.F., Bevis, C., Black, P.J., Boodhun, A., Brennan, J.S., Bridgham, J.A., Brown, R.C., Brown, A.A., Buermann, D.H., Bundu, A.A., Burrows, J.C., Carter, N.P., Castillo, N., Chiara, E., Catenazzi, M., Chang, S., Neil Cooley, R., Crake, N.R., Dada, O.O., Diakoumakos, K.D., Dominguez-Fernandez, B., Earnshaw, D.J., Egbujor, U.C., Elmore, D.W., Etchin, S.S., Ewan, M.R., Fedurco, M., Fraser, L.J., Fuentes, Fajardo, K.V., Scott, Furey, W., George, D., Gietzen, K.J., Goddard, C.P., Golda, G.S., Granieri, P.A., Green, D.E., Gustafson, D.L., Hansen, N.F., Harnish, K., Haudenschild, C.D., Heyer, N.I., Hims, M.M., Ho, J.T., Horgan, A.M., Hoschler, K., Hurwitz, S., Ivanov, D.V., Johnson, M.Q., James, T., Huw, Jones, T.A., Kang, G.D., Kerelska, T.H., Kersey, A.D., Khrebtukova, I., Kindwall, A.P., Kingsbury, Z., Kokko-Gonzales, P.I., Kumar, A., Laurent, M.A., Lawley, C.T., Lee, S.E., Lee, X., Liao, A.K., Loch, J.A., Lok, M., Luo, S., Mammen, R.M., Martin, J.W., McCauley, P.G., McNitt, P., Mehta, P., Moon, K.W., Mullens, J.W., Newington, T., Ning, Z., Ling, Ng, B., Novo, S.M., O'Neill, M.J., Osborne, M.A.,

Osnowski, A., Ostadan, O., Paraschos, L.L., Pickering, L., Pike, A.C., Pike, A.C., Chris Pinkard, D., Pliskin, D.P., Podhasky, J., Quijano, V.J., Raczy, C., Rae, V.H., Rawlings, S.R., Chiva Rodriguez, A., Roe, P.M., Rogers, J., Rogert Bacigalupo, M.C., Romanov, N., Romieu, A., Roth, R.K., Rourke, N.J., Ruediger, S.T., Rusman, E., Sanches-Kuiper, R.M., Schenker, M.R., Seoane, J.M., Shaw, R.J., Shiver, M.K., Short, S.W., Sizto, N.L., Sluis, J.P., Smith, M.A., Ernest Sohna Sohna, J., Spence, E.J., Stevens, K., Sutton, N., Szajkowski, L., Tregidgo, C.L., Turcatti, G., Vandevondele, S., Verhovsky, Y., Virk, S.M., Wakelin, S., Walcott, G.C., Wang, J., Worsley, G.J., Yan, J., Yau, L., Zuerlein, M., Rogers, J., Mullikin, J.C., Hurles, M.E., McCooke, N.J., West, J.S., Oaks, F.L., Lundberg, P.L., Klenerman, D., Durbin, R., Smith, A.J. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**: 53-59.

Berg, D.E., and Howe, M.M. (1989) Mobile DNA. American Society for Microbiology, Washington, D.C.

Berger, J., Suzuki, T., Senti, K.A., Stubbs, J., Schaffner, G., Dickson, B.J. (2001) Genetic mapping with SNP markers in Drosophila. *Nat Genet* **29**: 475-481.

Birse, D.E., Kapp, U., Strub, K., Cusack, S., Aberg, A. (1997) The crystal structure of the signal recognition particle Alu RNA binding heterodimer, SRP9/14. *EMBO J* **16**: 3757-3766.

Boeke, J.D. (1997) LINEs and Alus—The polyA connection. *Nat Genet* **16**: 6–7.

Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., Volff, J.N. (2008) Transposable elements as drivers of genomic and biological diversity in vertebrates. *Chromosome Res* **16**: 203-215.

Boissinot, S., Chevret, P., Furano, A. (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* **17**: 915-928.

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

Bovia, F., Wolff, N., Ryser, S., Strub, K. (1997) The SRP9/14 subunit of the human signal recognition particle binds to a variety of Alu-like RNAs and with higher affinity than its mouse homolog. *Nucleic Acids Res* **25**: 318-26.

Britten, R.J., Baron, W.F., Stout, D., Davidson, E.H. (1988) Sources and evolution of human Alu repeated sequences. *Proc Natl Acad Sci USA* **85**: 4770-4774.

Brouha, B., Meischl, C., Ostertag, E., de Boer, M., Zhang, Y., Neijens, H., Roos, D., Kazazian, H.H. (2002) Evidence consistent with human L1 retrotransposition in maternal meiosis I. *Am J Hum Genet* **71**: 327–336.

Brouha, B., Schstak, J., Badge, R.M., Lutz-Prigg, S., Farbey, A.H., Moran, J.V., Kazazian, H.H. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci USA* **100**: 5280-5285.

Callinan, P.A. and Batzer, M.A. (2006) Retrotransposable elements and human disease. *Genome Dyn* **1**: 104-115.

Campbell, A., Berg, D.E., Botstein, D., Lederberg, E.M., Novick, R.P., Starlinger, P., Szybalski, W. (1979) Nomenclature of transposable elements in prokaryotes. *Gene* **5**: 197–206.

Carroll, M.L., Roy-Engel, A.M., Nguyen, S.V., Salem, A.H., Vogel, E., Vincent, B., Myers, J., Ahmad, Z., Nguyen, L., Sammarco, M., Watkins, W.S., Henke, J., Makalowski, W., Jorde, L.B., Deininger, P.L., Batzer, M.A. (2001) Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity. *J Mol Biol* **311**: 17-40.

Chang, D.Y., Newitt, J.A., Hsu, K., Bernstein, H.D., Maraia, R.J. (1997) A highly conserved nucleotide in the Alu domain of SRP RNA mediates translation arrest through high affinity binding to SRP9/14. *Nucleic Acids Res* **25**: 1117–1122.

Chen, J.M., Chuzhanova, N., Stenson, P.D., Ferec, C., Cooper, D.N. (2005) Meta-analysis of gross insertions causing human genetic disease, novel mutational mechanisms and the role of replication slippage. *Hum Mutat* **25**: 207–221.

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**: 69-87.

Chu, W.M., Liu, W.M., Schimd, C.W. (1995) RNA polymerase III promoter and terminator elements affect Alu RNA expression. *Nucleic Acids Res* **23**: 1750–1757.

Collins, F.S., Brooks, L.D., Chakravarti, A., (1999) A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res* **8**: 1229-1231.

Collins, F.S., Drumm, M.L., Cole, J.L., Lockwood, W.K., Vande Woude, G.F., Iannuzzi, M.C. (1987) Construction of a general human chromosome jumping library, with application to cystic fibrosis. *Science* **235**: 1046-1049.

Cordaux, R., Hedges, D.J., Batzer, M.A. (2004) Retrotransposition of Alu elements: How many sources? *Trends Genet* **20**: 464–467.

Cordaux, R., Hedges, D.J., Herke, S.W., Batzer, M.A. (2006) Estimating the retrotransposition rate of human Alu elements. *Gene* **373**: 134–137.

Cost, G.J., Feng Q., Jacquier, A., Boeke, J.D. (2002) Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**: 5899-5910.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J., Lander,E.S. (2001) High resolution haplotype structure in the human genome. *Nature Genet* **29**: 229-232.

Dawson, E., Chen, Y., Hunt, S., Smink, L.J., Hunt, A., Rice, K., Livingston, S., Bumpstead, S., Bruskiewich, R., Sham, P. Ganske, R., Adams, M., Kawasaki, K., Shimizu, N., Minoshima, S., Roe, B. Bentley, D., Dunham, I. (2001) A SNP resource for human chromosome 22: extracting dense clusters of SNPs from the genomic sequence. *Genome Res* **11**: 170-178.

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

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

Deininger, P.L., Moran, J.V., Batzer, M.A., Kazazian, H.H. (2003) Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13**: 651-658.

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

Dewannieux, M. and Heidmann, T. (2005) Role of poly(A) tail length in Alu retrotransposition. *Genomics* **86**: 378–381.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., Kazazian, H.H. (1991) Isolation of an active human transposable element. *Science* **254**: 1805–1808.

Esnault, C., Casella, J., Heidmann, T. (2002) A Tetrahymena thermophila ribozyme-based indicator gene to detect transposition of marked retroelements in mammalian cells. *Nucleic Acids Res* **30**:e49. doi: 10/1093/nar/30.11.e49.

Esnault, C., Maestre, J., Heidmann, T. (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* **24**: 363-367.

Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using Phred II error probabilities. *Genome Res* **8**: 186-194.

Ewing, B., Hillier, L., Wendl, M.C., Green, P. (1998) Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Res* **8**: 175-185.

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

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J.,Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S.N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E.S., Daly, M.J., Altshuler, D. (2002) The structure of haplotype blocks in the human genome. *Science*, **296**: 2225-2229.

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

Goodier, J.L. and Maraia, R.J. (1998) Terminator-specific recycling of a B1-Alu transcription complex by RNA polymerase III is mediated by the RNA terminus-binding protein La. *J Biol Chem* **273**: 26110–26116.

Goodier, J.L., Ostertag, E.M., Kazazian, H.H. (2000) Transduction of 3' flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* **9**: 653-657.

Green, R.E., Krause, J., Ptak, S.E., Briggs, A.W., Ronan, M.T., Simons, J.F., Du, L., Egholm, M., Rothberg, J.M., Paunovic, M., Pääbo, S. (2006) Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**: 330-336.

Gundelfinger, E.D., Krause, E., Melli, M., Dobberstein, B. (1983) The organization of the 7SL RNA in the signal recognition particle. *Nucleic Acids Res* **11**: 7363-7374.

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

Halic, M., Becker, T., Pool, M.R., Spahn, C.M., Grassucci, R.A., Frank, J., Beckmann, R. (2004) Structure of the signal recognition particle interacting with the elongation-arrested ribosome. *Nature* **427**: 808-814.

Halic, M. and Beckmann, R. (2005) The signal recognition particle and its interactions during protein targeting. *Curr Opin Struct Biol* **15**: 116-25.

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

Han, K., Xing, J., Wang, H., Hedges, D.J., Garber, R.K., Cordaux, R., Batzer, M.A. (2005) Under the genomic radar: The stealth model of Alu amplification. *Genome Res* **15**: 655–664.

Hassoun, H., Coetzer, T.L., Vassiliadis, J.N., Sahr, K.E., Maalouf, G.J., Saad, S.T., Catanzariti, L., Palek, J. (1994) A novel mobile element inserted in the alpha spectrin gene. Spectrin Dayton. *J Clin Invest* **94**: 643-648.

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

Hickey, D.A. (1992) Evolutionary dynamics of transposable elements in prokaryotes and eukaryotes. *Genetica* **86**: 269-74.

Hohjoh, H., and Singer, M.F. (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J* **15**: 630-639.

Hohjoh, H., and Singer, M.F. (1997a) Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J* **16**: 6034-6043.

Hohjoh, H., and Singer, M.F. (1997b) Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon. *J Mol Biol* **271**: 7-12.

International HapMap Consortium. (2003) The international HapMap project. *Nature* **426**: 789-796.

Hsu, K., Dau-Yin, C., Maraia, R. (1995) Human signal recognition particle (SRP) Alu-associated protein also binds Alu interspersed repeat sequence RNAs. *J Biol Chem* **270**: 10179-10186.

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

Ivics, Z., and Izsvak, Z. (1997) Family of plasmid vectors for the expression of b-galactosidase fusion proteins in eukaryotic cells. *Biotechniques* **22**: 254-256.

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

Judson, R., Salisbury, B., Schneider, J., Windemuth, A., Stephens, J.C. (2002) How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* **3**: 379-391.

Jurka, J. (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* **16**: 418-420.

Jurka, J., and Smith, T. (1988) A fundamental division in the Alu family of repeated sequences. *Proc Natl Acad Sci USA* **85**: 4775-4778.

Jurka, J., and Zuckerkandl, E. (1991) Free left arms as precursor molecules in the evolution of Alu sequences. *J Mol Evol* **33**: 49-56.

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

Kapitonov, V.V. and Jurka, J. (2003) A novel class of SINE elements derived from 5S rRNA. *Mol Biol Evol* **20**: 694-702.

Kazazian, H.H., Wong, C., Youssoufian, H., Scott, A.F., Phillips, D.G., Antonarakis, S.E. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164-166.

Kazazian, H.H. (1998) Mobile elements and disease. *Curr Opin Genet Dev* **8**: 343-350.

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

Kazazian, H.H. (1999) An estimated frequency of endogenous insertional mutations in humans. *Nat Genet* **22**: 130.

Kazazian, H.H. and Goodier, J.L. (2002) LINE drive: retrotransposon and genome instability. *Cell* **110**: 277-280.

Kazazian, H.H. (2004) Mobile elements: drivers of genome evolution. *Science* **303**: 1626-1632.

Kempken, F. and Kück, U. (1998) Transposons in filamentous fungi--facts and perspectives. *Bioessays* **20**: 652-659.

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

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Hausler, D. (2002) The human genome browser at UCSC. *Genome Res* **12**: 996-1006.

Khaja, R., Zhang, J., MacDonald, J.R., He, Y., Joseph-George, A.M., Wei, J., Rafiq, M.A., Qian, C., Shago, M., Pantano, L., Aburatani, H., Jones, K., Redon, R., Hurles, M., Armengol, L., Estivill, X., Mural, R.J., Lee, C., Scherer, S.W., Feuk, L. (2006) Genome assembly comparison identifies structural variants in the human genome. *Nat Genet* **8**:1413-1418.

Kimmel, B., Palozzolo, M., Martin, C., Boeke, J.D., Devine, S.E. (1997) Transposon-mediated DNA sequencing, pp. 455-532 in Genome Analysis, A Laboratory Manual, Vol. 1, edited by B. Birren, E.D. Green, R.M. Myers and P. Hieter. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Kloor, M., Sutter, C., Wentzensen, N., Cremer, F.W., Buckowitz, A., Keller, M., Doeberitz, M., Gebert, J. (2004) A large MSH2 Alu insertion mutation causes HNPCC in a German kindred. *Hum Genet* **115**: 432-438.

Kobayashi, K., Nakahori, Y., Miyake, M., Matsumura, K., Kondo-lida, E., Nomura, Y., Segawa, M., Yoshioka, M., Saito, K., Osawa, M., Hamano, K., Sakakihara, Y., Nonaka, I., Nakagome, Y., Kanazawa, I., Nakamura, Y., Tokunaga, K., Toda, T. (1998) An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. *Nature* **394**: 388-392.

Kolosha, V.O., and Martin, S.L. (1997) In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc Natl Acad Sci USA* **94**: 10155-10160.

Konkel, M.K., Wang, J., Liang, P., Batzer, M.A. (2007) Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies. *Gene* **390**: 28-38.

Kriegs, J.O., Churakov, G., Jurka, J., Brosius, J., Schmitz, J. (2007) Evolutionary history of 7SL RNA-derived SINEs in Supraprimates. *Trends Genet* **23**: 158-161.

Kulpa, D.A., and Moran, J.V. (2005) Cis-preferantial LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat Struct Mol Biol* **13**: 655-660.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A.,

Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R.,

Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de, la, Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowski, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de, Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J; International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.

Landry, J.R. and Mager, D.L. (2003) Functional analysis of the endogenous retroviral promoter of the human endothelin B receptor gene. *J Virol* **77**: 7459-7466.

Lee, J., Cordaux, R., Han, K., Wang, J., Hedges, D.J., Liang, P., Batzer, M.A.. (2007) Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons. *Gene* **390**: 18-27.

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

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L., Venter, J.C. (2007) The diploid genome sequence of an individual human. *PLoS Biol* **5**: e254.

Li, T. and Schmid, C.W. (2001) Differential stress induction of individual Alu loci: Implications for transcription and retrotransposition. *Gene* **276**: 135–141.

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

Lin, L., Shen, S., Tye, A., Cai, J.J., Jiang, P., Davidson, B.L., Xing, Y. (2008) Diverse splicing patterns of exonized Alu elements in human tissues. *PLoS Genet* **4**: e1000225.

Liu, J., Nau, M.M., Zucman-Rossi, J., Powell, J.I., Allegra, C.J., Wright, J.J. (1997) LINE-1 element insertion at the t(11;22) translocation breakpoint of a desmoplastic small round cell tumor. *Genes Chromosomes Cancer* **18**: 232-239.

Liu, W. and Schmid, C.W. (1993) Proposed roles for DNA methylation in Alu transcriptional repression and mutational inactivation. *Nucleic Acids Res* **21**: 1351–1359.

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

Luirink, J., Sinning, I. (2004) SRP-mediated protein targeting: structure and function revisited. *Biochim Biophys Acta*. **1694**: 17-35.

Mandal, P.K. and Kazazian, H.H. (2008) SnapShot: Vertebrate transposons. *Cell* **135**: 192.

Martin, S.L., and Bushman, F.D. (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* **21**: 467-475.

Martin, S.L., Branciforte, D., Keller, D., Bain, D.L. (2003) Trimeric structure for an essential protein in L1 retrotransposition. *Proc Natl Acad Sci USA* **100**: 13815-13820.

Martin, S.L., Li, J., Weisz, J.A. (2000) Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. *J Mol Biol* **304**: 11-20.

Martin, S.L., Cruceanu, M., Branciforte, D., Wai-Lun Li, P., Kwok, S.C., Hodges, R.S., Williams, M.C. (2005) LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. *J Mol Biol* **348**: 549–561.

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

Mätlik K, Redik K, Speek M. (2006) L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* **71753**: 1-16.

McClintok, B. (1950) The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A* **36**: 344-355.

Miki, Y., Katagiri, T., Kasumi, F., Yoshimoto, T., Nakamura, Y. (1996) Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat Genet* **13**: 245-247.

Miki, Y., Nishisho, I., Horii, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K.W., Vogelstein, B., Nakamura, Y. (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* **52**: 643-645.

Mills, R.E., Bennett, E.A., Iskow, R.C., Luttig, C.T., Tsui, C., Pittard, W.S., Devine, S.E. (2006) Recently-mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet* **78**: 671–679.

Mills, R.E., Bennett, E.A., Iskow, R.C., Devine, S.E. (2007) Which transposable elements are active in the human genome? *Trends Genet* **23**: 183–191.

Minakami, R., Kurose, K., Etoh, K., Furuhata, Y., Hattori, M., Sakaki, Y. (1992) Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res* **20**: 3139-3145.

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

Moran, J. V. (1999) Human L1 retrotransposition, insights and peculiarities learned from a cultured cell retrotransposition assay. *Genetica* **107**: 39-51.

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

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

Morse, B., Rothberg, P.G., South, V.J., Spandorfer, J.M., Astrin, S.M. (1988) Insertional mutagenesis of the myc locus by a LINE-1 sequence in human breast carcinoma. *Nature* **333**: 87-90.

Müller, M., Weigand, J.E., Weichenrieder, O., Suess, B. (2006) Thermodynamic characterization of an engineered tetracycline-binding riboswitch. *Nucleic Acids Res* **34**: 2607–2617.

Mullikin, J.C., Hunt, S.E., Cole, C.G., Mortimore, B.J., Rice, C.M., Burton, J., Matthews, L.H., Pavitt, R., Plumb, R.W., Sims, S.K., Ainscough, R.M., Attwood, J., Bailey, J.M., Barlow, K., Bruskiewich, R.M., Butcher, P.N., Carter, N.P., Chen, Y., Clee, C.M., Coggill, P.C., Davies, J., Davies, R.M., Dawson, E., Francis, M.D., Joy, A.A., Lamble, R.G., Langford, C.F., Macarthy, J., Mall, V., Moreland, A., Overton-Larty, E.K., Ross, M.T., Smith, L.C., Steward, C.A., Sulston, J.E., Tinsley, E.J., Turney, K.J., Willey, D.L., Wilson, G.D., McMurray, A.A., Dunham, I., Rogers, J., Bentley, D.R. (2000) An SNP map of human chromosome 22. *Nature* **407**: 516-520.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V., Jorde, L.B., Batzer, M.A. (2002) A

comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71**: 312-326.

Nachman, M.W. and Crowell, S.L. (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* **156**: 297-304.

Narita, N., Nishio, H., Kitoh, Y., Ishikawa, Y., Minami, R. (1993) Insertion of a 5. truncated L1 element into the 3. end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *J Clin Invest* **91**: 1862-1867.

Newman, T.L., Rieder, M.J., Morrison, V.A., Sharp, A.J., Smith, J.D., Sprague, L.J., Kaul, R., Carlson, C.S., Olson, M.V., Nickerson, D.A., Eichler, E.E. (2006) High-throughput genotyping of intermediate-size structural variation. *Hum Mol Genet* **15**: 1159-1167.

Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Hutchinson, R.G.,Stengard, J., Salomaa, V., Vartiainen, E., Boerwinkle, E., Sing, C.F. (1998) DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene. *Nature Genet* **19**: 216-217.

Noonan, J.P., Coop, G., Kudaravalli, S., Smith, D., Krause, J., Alessi, J., Chen, F., Platt, D., Pääbo, S., Pritchard, J.K., Rubin, E.M. (2006) Sequencing and analysis of Neanderthal genomic DNA. *Science* **314**: 1113-1118.

Ohshima, K., Hamada, M., Terai, Y., Okada, N. (1996) The 39 ends of tRNA-derived short interspersed repetitive elements are derived from the 39 ends of long interspersed repetitive elements. *Mol Cell Biol* **16**: 3756–3764.

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

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

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

Ovchinnikov, I., Rubin, A., Swergold, G.D. (2002) Tracing the LINEs of human evolution. *Proc Natl Acad Sci USA* **99**: 10522-10527.

Ovchinnikov, I., Troxel, A.B., Swergold, G.D. (2001) Genomic characterization of recent human LINE-1 insertions, evidence supporting random insertion. *Genome Res* **11**: 2050-2058.

Pace, J.K. II, Feschotte, C. (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res* **17**: 422-432.

Patil, N., Berno, .J., Hinds, D.A., Barrett, W.A., Doshi, J.M., Hacker, C.R.,Kautzer, C.R., Lee, H.H., Marjoribands, C., McDonough, D.P., Nguyen, B.T., Norris, M.C., Sheehan, J.B., Shen, N., Stern, D., Stokowski, R.P., Thomas, D.J., Trulson, M.O., Vyas, K.R., Frazer, K.A., Fodor, S.P., Cox, D.R. (2001) Blocks of limited haplotype diversity revealed by high resolution scanning of human chromosome 21. *Science* **294**: 1719-1723.

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., Cho, E.K., Dallaire, S., Freeman, J.L., González, J.R., Gratacòs, M., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J.R., Marshall, C.R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K. Shen, F., Somerville, M.J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D.F., Estivill, X., Tyler-Smith, C., Carter, N.P., Aburatani, H., Lee, C., Jones, K.W., Scherer, S.W., Hurles, M.E. (2006) Global variation in copy number in the human genome. *Nature* **444**: 444-454.

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

Prak, E.T. and Kazazian, H.H. (2001) Mobile elements and the human genome. *Nat Rev Genet* **1**: 134-144.

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., Richter, D.J., Lavery, T., Kouyoumjian, R., Farhadian, S.F., Ward, R., Lander, E.S. (2001) Linkage disequilibrium in the human genome. *Nature* **411**: 199-204.

Repanas, K., Zingler, N., Layer, L.E., Schumann, G.G., Perrakis, A., Weichenrieder, O. (2007) Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res* **35**: 4914-4926.

Rieder, M.J., Taylor, S.L., Clark, A.G.,Nickerson, D.A. (1999) Sequence variation in the human angiotensin converting enzyme. *Nat Genet* **22**: 59-62.

Rohrer, J., Minegishi, Y., Richter, D., Eguiguren, J., Conley, M.E. (1999) Unusual mutations in Btk, an insertion, a duplication, and four large deletions. *Clin Immunol* **90**: 28-37.

Roy, A.M., West, N.C., Rao, A., Adhikari, P., Aleman, C., Barnes, A.P., Deininger, P.L. (2000) Upstream flanking sequences and transcription of SINEs. *J Mol Biol* **302**: 17–25.

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

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

Rubin, C.M., Kimura, R.H., Schmid, C.W. (2002) Selective stimulation of translational expression by Alu RNA. *Nucleic Acids Res* **30**: 3253-3261.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., Schaffner, S.F., Lander, E.S., and The International HapMap Consortium. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**: 913-918.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., Hunt, S.E., Cole, C.G., Coggill, P.C., Rice, C.M., Ning, Z., Rogers, J., Bentley, D.R., Kwok, P.Y., Mardis, E.R., Yeh, R.T., Schultz, B., Cook, L., Davenport, R., Dante, M. Fulton, L., Hillier, L., Waterston, R.H., McPherson, J.D., Gilman, B., Schaffner, S., Van Etten, W.J., Reich, D., Higgins, J., Daly, M.J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M.C., Linton, L., Lander, E.S., Altshuler, D.; International SNP Map Working Group. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Sarrowa, J., Chang, D.Y., Maraia, R.J. (1997) The decline in human Alu retroposition was accompanied by an asymmetric decrease in SRP9/14 binding to dimeric Alu RNA and increased expression of small cytoplasmic Alu RNA. *Mol Cell Biol* **17**: 1144–1151.

Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., Kazazian, H.H. (1997) Many human L1 elements are capable of retrotransposition. *Nat Genet* **16**: 37-43.

Sayah, D.M., Sokolskaja, E., Berthoux, L., Luban, J. (2004) Cyclophilin A retrotransposition into TRIM5 explains owl monkey resistance to HIV-1. *Nature* **430**: 569-573.

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

Seleme, M.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., Kazazian, H.H. Jr. (2006) Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. *Proc Natl Acad Sci U S A* **103**: 6611-6616.

Shaikh, T.H.and Deininger, P.L. (1996) The role and amplification of the HS Alu subfamily founder gene. *J Mol Evol* **42**: 15-21.

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

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

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

Sinnett, D., Richer, C., Deragon, J.M., Labuda, D. (1991) Alu RNA secondary structure consists of two independent 7 SL RNA-like folding units. *J Biol Chem* **266**: 8675–8678.

Slagel, V., Flemington, E., Traina-Dorge, V., Bradshaw, H., Deininger, P.L. (1987) Clustering and relationships of the Alu family in the human genome. *Mol Biol Evol* **4**: 19-29.

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

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

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

Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T.,Stanley, S.E. Jiang, R., Messer, C.J., Chew, A., Han, J.H., Duan, J., Carr, J.L., Lee, M.S., Koshy, B., Kumar, A.M., Zhang, G., Newell, W.R., Windemuth, A., Xu, C., Kalbfleisch, T.S., Shaner, S.L., Arnold, K., Schulz, V., Drysdale, C.M., Nandabalan, K., Judson, R.S., Ruano, G., Vovis, G.F. (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* **293**: 489-493.

Strichman-Almashanu, L.Z., Lee, R.S., Onyango, P.O., Perlman, E., Flam, F., Frieman, M.B., Feinberg, A.P. (2001) A genome-wide screen for normally methylated human CpG islands that can identify novel imprinted genes. *Genome Res* **12**: 543-554.

Strub, K. and Walter, P. (1990) Assembly of the Alu domain of the signal recognition particle (SRP): dimerization of the two protein components is required for efficient binding to SRP RNA. *Mol Cell Biol* **10**: 777-784.

Sun, F.J., Fleurdépine, S., Bousquet-Antonelli, C., Caetano-Anollés, G., Deragon, J.M. (2007) Common evolutionary trends for SINE RNA structures. *Trends Genet* **23**: 26-33.

Taillon-Miller, P. and Kwok, P.Y. (2000) A high-density single-nucleotide polymorphism map of Xq25-q28. *Genomics* **65**: 195-202.

Taillon-Miller, P., Piernot, E.E., Kwok, P.Y. (1999) Efficient approach to unique single-nucleotide polymorphism discovery. *Genome Res* **9**: 499-505.

Terzi, L., Pool, M.R., Dobberstein, B., Strub, K. (2004) Signal recognition particle Alu domain occupies a defined site at the ribosomal subunit interface upon signal sequence recognition. *Biochemistry* **43**:107-117.

Teugels, E, De Brakeleer, S., Goelen, G., Lissens, W., Sermijn, E., De Grève, J. (2005) De novo Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes. *Hum Mutat* **26**: 284.

The Chimpanzee Sequencing and Analysis Consortium. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **426**: 789–796.

The International HapMap Consortium. (2003) The International HapMap Project. *Nature* **426**: 789-796.

The International HapMap Consortium. (2005) A Haplotype Map of the Human Genome. *Nature* **437**: 1299-1320.

The International HapMap Consortium. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**: 851-861.

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933.

Tsui, C., Coleman, L.E., Griffith, J.L., Bennett, E.A., Goodson, S.G., Scott, J.D., Pittard, W.S., Devine, S.E. (2003) Single nucleotide polymorphisms (SNPs) that map to gaps in the human SNP map. *Nucleic Acids Res* **31**: 4910-4916.

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

Ullu, E. and Weiner, A.M. (1985) Upstream sequences modulate internal promoter of the human 7SL RNA gene. *Nature* **318**: 371–374.

Ustyugova, S.V., Lebedev, Y.B., Sverdlov, E.D. (2006) Long L1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts. *Genetica* **128**: 261-272.

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

Volff, J.N. and Brosius, J. (2007) Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. *Genome Dyn* **3**: 175-190.

Walker, S.C., Avis, J.M., Conn, G.L. (2003) General plasmids for producing RNA in vitro transcripts with homogeneous ends. *Nucleic Acids Res* **31**: e82. doi: 10.1093/nar/gng082.

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

Walter, P. and Blobel, G. (1983) Disassembly and reconstitution of signal recognition particle. *Cell* **34**: 525–533.

Wang, H., Xing, J., Grover, D., Hedges, D.J., Han, K., Walker, J.A., Batzer, M.A. (2005) SVA elements: a hominid-specific retroposon family. *J Mol Biol* **354**: 994-1007.

Wang, J., Song, L., Gonder, M.K., Azrak, S., Ray, D.A., Batzer, M.A., Tishkoff, S.A., Liang, P. (2006) Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms. *Gene* **365**: 11-20.

Wang, J., Song, L., Grover, D., Azrak, S., Batzer, M.A., Liang, P. (2006) dbRIP: A highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum Mutat* **27**: 323–329.

Wang, J., Wang, W., Li, R., Li, Y., Tian, G., Goodman, L., Fan, W., Zhang, J., Li, J., Zhang, J., Guo, Y., Feng, B., Li, H., Lu, Y., Fang, X., Liang, H., Du, Z., Li, D., Zhao, Y., Hu, Y., Yang, Z., Zheng, H., Hellmann, I., Inouye, M., Pool, J., Yi, X., Zhao, J., Duan, J., Zhou, Y., Qin, J., Ma, L., Li, G., Yang, Z., Zhang, G., Yang, B., Yu, C., Liang, F., Li, W., Li, S., Li, D., Ni, P., Ruan, J., Li, Q., Zhu, H., Liu, D., Lu, Z., Li, N., Guo, G., Zhang, J., Ye, J., Fang, L., Hao, Q., Chen, Q., Liang, Y., Su, Y., San, A., Ping, C., Yang, S., Chen, F., Li, L., Zhou, K., Zheng, H., Ren, Y., Yang, L., Gao, Y., Yang, G., Li, Z., Feng, X., Kristiansen, K., Wong, G.K., Nielsen, R., Durbin, R., Bolund, L., Zhang, X., Li, S., Yang, H., Wang, J. (2008) The diploid genome sequence of an Asian individual. *Nature* **456**: 60-65.

Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T.D., Toyoda, A., Kuroki, Y., Noguchi, H., BenKahla, A., Lehrach, H., Sudbrak, R., Kube, M., Taenzer, S., Galgoczy, P., Platzer, M., Scharfe, M., Nordsiek, G., Blöcker, H., Hellmann, I., Khaitovich, P., Pääbo, S., Reinhardt, R., Zheng, H.J., Zhang, X.L., Zhu, G.F., Wang, B.F., Fu, G., Ren, S.X., Zhao, G.P., Chen, Z., Lee, Y.S., Cheong, J.E., Choi, S.H., Wu, K.M., Liu, T.T., Hsiao, K.J., Tsai, S.F., Kim, C.G., OOta, S., Kitano, T., Kohara, Y., Saitou, N., Park, H.S., Wang, S.Y., Yaspo, M.L., Sakaki, Y. (2004) DNA sequence and comparative analysis of chimpanzee chromosome 22. *Nature* **429**: 382-388.

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

Weichenrieder, O., Kapp, U., Cusack, S., Strub, K. (1997) Identification of a minimal Alu RNA folding domain that specifically binds SRP9/14. *RNA* **3**: 1262–1272.

Weichenrieder, O., Stehlin, C., Kapp, U., Birse, D.E., Timmins, P.A., Strub, K., Cusack, S. (2001) Hierarchical assembly of the Alu domain of the mammalian signal recognition particle. *RNA* **7**: 731–740.

Weichenrieder, O., Wild, K., Strub, K., Cusack, S. (2000) Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature* **408**: 167–173.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A., Rothberg, J.M. (2008) The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872-826.

Wicks, S.R., Yeh, R.T., Gish, W.R., Waterston, R.H., Plasterk, R.H.A. (2001) Rapid gene mapping in Caenorhabditis elegans using a high density polymorphism map. *Nat Genet* **28**: 160-164.

Wilund, K.R., Yi, M., Campagna, F., Arca, M., Zuliani, G., Fellin, R., Ho, Y.K., Garcia, J.V., Hobbs, H.H., Cohen, J.C. (2002) Molecular mechanisms of autosomal recessive hypercholesterolemia. *Hum Mol Genet* **11**: 3019-3030.

Xing, J., Wang, H., Belancio, V.P., Cordaux, R., Deininger, P.L., Batzer, M.A. (2006) Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci USA* **103**: 17608–17613.

Xing, J., Witherspoon, D.J., Ray, D.A., Batzer, M.A., Jorde, L.B. (2007) Mobile DNA elements in primate and human evolution. *Am J Phys Anthropol* Suppl **45**: 2-19.

Yohn, C.T., Jiang, Z., McGrath, S.D., Hayden, K.E., Khaitovich, P., Johnson, M.E., Eichler, M.Y., McPherson, J.D., Zhao, S., Pääbo, S., Eichler, E.E.(2005) Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans. *PLoS Biol* **3**: e110.

Zhang, Y., Dipple, K.M., Vilain, E., Huang, B.L., Finlayson, G., Therrell, B.L., Worley, K., Deininger, P., McCabe, E.R. (2000) AluY insertion (IVS4-52ins316alu) in the

glycerol kinase gene from an individual with benign glycerol kinase deficiency. *Hum Mutat* **15**: 316-323.

Zhu, Z.B., Jian, B., Volanakis, J.E. (1994) Ancestry of SINE-R.C2 a human-specific retroposon. *Hum Genet* **93**: 545-551.

Zillig, W., Prangishvilli, D., Schleper, C., Elferink, M., Holz, I., Albers, S., Janekovic, D., Götz, D. (1996) Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. *FEMS Microbiol* **18**: 225-236.