2013

An Alu element-based model of human genome instability

George Wyndham Cook, Jr.
Louisiana State University and Agricultural and Mechanical College

AN *ALU* ELEMENT-BASED MODEL OF HUMAN GENOME INSTABILITY

A Dissertation

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

in

The Department of Biological Sciences

by

George Wyndham Cook, Jr.
B.S., University of Arkansas, 1975
May 2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AAI	*Alu-Alu* Insertion
APE	Alu Pair Exclusion
APSN	*Alu* Pair Sequence Number
ARMD	*Alu* Recombination Mediated Deletion
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
CAC	Catenated *Alu* Cluster
CLIQUE	Catenated LINE1 Endonuclease Induced Queues of Uninterrupted *Alu*, LINE1 and SVA Elements
DDJ	Doomsday Junction
DNA	Deoxyribonucleic Acid
DSB	Double-Strand Break
FAP	Full Length *Alu* pair
hg18	Human Genome Assembly 18
hg19	Human Genome Assembly 19
I:D	Ratio of Inverted to Direct Oriented *Alu* Pairs
L1	LINE1 Element
L1EN	LINE1 Endonuclease
L1RT	LINE1 Reverse Transcriptase
NAHR	Non-Allelic Homologous Recombination
ORF2p	Protein from the Second Open Reading Frame in LINE1 Elements
PanTro2	Second Genome Assembly for the Common Chimpanzee, *Pan troglodytes*

PCR	Polymerase Chain Reaction
Poly(A)	Poly-Adenine
RNA	Ribonucleic Acid
SINE	Short Interspersed Element
SSA	Single Strand Annealing Repair of a Double-Strand DNA Break
SVA	SINE-r; VNTR; HERV-like Region
TPRT	Target Primed Reverse Transcription
TSD	Target Site Duplication
UCSC	University of California, Santa Cruz

ABSTRACT

The human genome is strewn with repetitive sequence. An early estimate derived from the draft human genome sequence placed this repetitive content at ~45%. More detailed recent analyses have advanced the idea that the human repetitive and repeat derived contribution to the genome may be closer to 66-69%. The most commonly repeated sequence in the human genome is the *Alu* element. *Alus* make up 10.6 percent of all human DNA and have expanded to over one million copies in the human genome reproducing through a copy and paste mechanism.

New *Alu* germline insertions are estimated to occur at a rate of 1 in 20 human births. In addition to their insertional impact, *Alus* have also been associated with various forms of genomic sequence disruptions including inversions, rearrangements, translocations and deletions. Chimeric *Alus* are frequently located at the breakpoints of these various forms of structural variations. This observation has led to the putative conclusion that chimeric *Alus* primarily result from the non-allelic homologous recombination between *Alu* elements. However, little proof is available regarding the actual mechanism(s) that catalyze this activity.

This dissertation reveals a newly recognized pattern among human *Alu* pairs that may provide additional insight into the mechanism(s) driving chimeric *Alu* formation. After adjusting for directional biases associated with clustering, *Alu* pairs in the same orientation (direct) outnumber *Alu* pairs in the opposite orientation (inverted pairs) by over two percent ($p<0.05$). If this imbalance was generated by deletions resulting from interactions between inverted *Alu* elements, many chimeric *Alus* may have formed from the homologous repair of these deletions.

This dissertation characterizes the human *Alu* pair imbalance and constructs an *Alu*-based model of human genome instability. This model was used to compare the relative instabilities of 50 human deletion-prone cancer genes and 50 randomly chosen genes. Taken as separate groups, the 50 deletion-prone cancer genes were estimated to be 58% more unstable than the 50 randomly chosen genes.

This approach to estimating human gene instability may lay the foundation for comparing genetic risks unique to specific individuals, families and people groups.

CHAPTER ONE:
BACKGROUND

The repetitive nature of the human genome was first reported in 1975 (Schmid and Deininger 1975). This discovery was soon followed by the identification of the *Alu* element as a ubiquitous contributor to this repetition (Houck et al. 1979). The completion of the draft human genome sequence permitted the quantification of the fraction of repetitive sequence within the human genome at approximately 45% (Lander et al. 2001). Recent advanced analyses reveal that the repeat related portion of the genome may be as high as 69% (de Koning et al. 2011). *Alu* elements make up 10.6 percent of all human DNA with a copy number of over one million.

Several descriptors have been ascribed to human *Alu* elements. The past three decades have witnessed *Alu* elements being alternately referred to as junk DNA, genomic parasites, drivers of evolution, facilitators of transcription and progenitors of new genes (Doolittle and Sapienza 1980; Orgel and Crick 1980; Jurka and Milosavljevic 1991; Deininger et al. 2003; Schmid 2003; Kreahling and Graveley 2004; Hasler et al. 2007).

The most fitting descriptor of *Alu* elements may prove to be “antagonist to human health”. *Alu* element interactions appear to be involved in much of human structural variation-related genomic disease (Xing et al. 2009; Gonzaga-Jauregui et al. 2012; Pang et al. 2012). The presence of chimeric *Alu* elements at structural variation breakpoints reinforces this view (Sen et al. 2006; de Smith et al. 2008; Hastings et al. 2009; Kitada et al. 2013). Prescient researchers have recognized the risk that *Alu* insertions pose to human health, and potentially more significantly, their post-insertion

interactions (Deininger and Batzer 1999; Hedges and Deininger 2007; Lupski 2010). During the past 15 years, over 100 studies have linked *Alu* elements to various deletion-associated diseases, including cancer (Table S3.1)**.** Evidence of *Alu-Alu* interactions is provided by the presence of human specific chimeric *Alu* elements (Sen et al. 2006). Further support for the view that human *Alu-Alu* interactions occur comes from *Alu* gene conversion events (Kass et al. 1995; Roy et al. 2000). Increased sequence homology among neighboring *Alu* elements reinforces the position that human *Alu-Alu* interactions are not uncommon events (Zhi 2007; Aleshin and Zhi 2010).

The mechanism(s) behind the formation of chimeric *Alus* remain(s) elusive. The presence of a chimeric *Alu* element at the boundary of structural variation provides little evidence for the etiology of its formation. The putative view is that chimeric *Alu*s arise as a result of non-allelic homologous recombination, NAHR, between two *Alu* elements. However, chimeric *Alu*s can also be generated by the single-strand annealing, SSA, repair of a double-strand break, DSB. Unless a rational case can be made for a DSB occurring within the spacer of a pre-chimeric *Alu* pair, NAHR appears to be a more reasonable mechanism than SSA for catalyzing the formation of chimeric *Alu* elements.

In support of the SSA route to chimeric *Alu* formation is the observation that human *Alu* pairs in opposite orientation (inverted pairs) are found statistically less frequently than *Alu* pairs having the same orientation (direct pairs) (Stenger et al. 2001; Cook et al. 2011). After removing *Alu* pairs that are subject to directional clustering biases, a total of 115,185,079 human *Alu* pairs exist with spacer sizes <350,000 bp. Within this spacer size window, direct oriented *Alu* pairs outnumber inverted oriented *Alu* pairs by 1,269,263 ($p<0.05$).

Two mechanisms, ectopic invasion and annealing of 1) complementary DNA breathing bubbles and/or 2) replication forks, have been proposed which may explain this loss of over one million inverted *Alu* pairs (Cook et al. 2011). Both of these mechanisms are also thought to be potential sources of segmental duplications and inversions (Cook et al. 2011). The second mechanism has recently been demonstrated in a yeast experimental model to produce both duplications and deletions (Mizuno et al. 2012).

Chapter two in this dissertation provides an initial characterization of the human imbalance between inverted and direct *Alu* pairs. Chapter three characterizes this *Alu* pair imbalance phenomenon in detail. This detailed characterization is then used to construct a model that predicts relative gene stabilities based upon the unique *Alu* element landscape architecture for each gene.

References

Aleshin A, Zhi D. 2010. Recombination-associated sequence homogenization of neighboring *Alu* elements: signature of nonallelic gene conversion. *Molecular biology and evolution* **27**(10): 2300-2311.

Cook GW, Konkel MK, Major JD, 3rd, Walker JA, Han K, Batzer MA. 2011. *Alu* pair exclusions in the human genome. *Mobile DNA* **2**: 10.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**(12): e1002384.

de Smith AJ, Walters RG, Coin LJ, Steinfeld I, Yakhini Z, Sladek R, Froguel P, Blakemore AI. 2008. Small deletion variants have stable breakpoints commonly associated with *Alu* elements. *PloS one* **3**(8): e3104.

Deininger PL, Batzer MA. 1999. *Alu* repeats and human disease. *Mol Genet Metab* **67**(3): 183-193.

Deininger PL, Moran JV, Batzer MA, Kazazian HH, Jr. 2003. Mobile elements and mammalian genome evolution. *Curr Opin Genet Dev* **13**(6): 651-658.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* **284**(5757): 601-603.

Gonzaga-Jauregui C, Lupski JR, Gibbs RA. 2012. Human genome sequencing in health and disease. *Annual review of medicine* **63**: 35-61.

Hasler J, Samuelsson T, Strub K. 2007. Useful 'junk': Alu RNAs in the human transcriptome. *Cellular and molecular life sciences : CMLS* **64**(14): 1793-1800.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. *Nat Rev Genet* **10**(8): 551-564.

Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation research* **616**(1-2): 46-59.

Houck CM, Rinehart FP, Schmid CW. 1979. A ubiquitous family of repeated DNA sequences in the human genome. *J Mol Biol* **132**(3): 289-306.

Jurka J, Milosavljevic A. 1991. Reconstruction and analysis of human *Alu* genes. *Journal of molecular evolution* **32**(2): 105-121.

Kass DH, Batzer MA, Deininger PL. 1995. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Molecular and cellular biology* **15**(1): 19-25.

Kitada K, Aikawa S, Aida S. 2013. *Alu-Alu* fusion sequences identified at junction sites of copy number amplified regions in cancer cell lines. *Cytogenet Genome Res* **139**(1): 1-8.

Kreahling J, Graveley BR. 2004. The origins and implications of Aluternative splicing. *Trends Genet* **20**(1): 1-4.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lupski JR. 2010. Retrotransposition and structural variation in the human genome. *Cell* **141**(7): 1110-1112.

Mizuno K, Miyabe I, Schalbetter SA, Carr AM, Murray JM. 2012. Recombination-restarted replication makes inverted chromosome fusions at inverted repeats. *Nature*.

Orgel LE, Crick FH. 1980. Selfish DNA: the ultimate parasite. *Nature* **284**(5757): 604-607.

Pang AW, Migita O, Macdonald JR, Feuk L, Scherer SW. 2012. Mechanisms of Formation of Structural Variation in a Fully Sequenced Human Genome. *Hum Mutat*.

Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL. 2000. Potential gene conversion and source genes for recently integrated Alu elements. *Genome research* **10**(10): 1485-1495.

Schmid CW. 2003. Alu: a parasite's parasite? *Nat Genet* **35**(1): 15-16.

Schmid CW, Deininger PL. 1975. Sequence organization of the human genome. *Cell* **6**(3): 345-358.

Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, Dyer M, Cordaux R, Liang P, Batzer MA. 2006. Human genomic deletions mediated by recombination between *Alu* elements. *American journal of human genetics* **79**(1): 41-53.

Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. 2001. Biased distribution of inverted and direct *Alus* in the human genome: implications for insertion, exclusion, and genome stability. *Genome research* **11**(1): 12-27.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA et al. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res* **19**(9): 1516-1526.

Zhi D. 2007. Sequence correlation between neighboring *Alu* instances suggests post-retrotransposition sequence exchange due to Alu gene conversion. *Gene* **390**(1-2): 117-121.

CHAPTER TWO:
ALU PAIR EXCLUSIONS IN THE HUMAN GENOME*

Introduction

Retrotransposons are mobile DNA elements that populate genomes via their respective RNA transcripts. The retrotransposon with the highest copy number in the human genome is the *Alu* element (Lander et al. 2001). *Alu* elements lack the necessary repertoire of enzymes to effect their independent insertion and are thus classified as non-autonomous mobile elements. For recent reviews, see (Belancio et al. 2008; Cordaux and Batzer 2009).

Following transcription, *Alu* RNA is thought to require the assistance of the LINE1 open reading frame 2 protein (ORF2p) both for nicking the genome at the insertion site and for reverse transcription of the *Alu* RNA transcript (Mathias et al. 1991; Luan et al. 1993). The endonuclease and reverse transcriptase functions of ORF2p are referred to as L1EN and L1RT, respectively. While L1EN has been shown to have some tolerance for target site variation, it most frequently cleaves at the T/A transition within the sequence, 5′-TTTTAA-3′ (Fogedby and Metzler 2007; Repanas et al. 2007; Konkel and Batzer 2010). Following cleavage, the poly-T sequence of the target site becomes accessible to the complementary poly(A) tail of *Alu* RNA. Hybridization of these two sequences results in a short RNA-DNA hybrid that both orients the RNA transcript and primes reverse transcription of the *Alu* RNA by L1RT. Identical sequences flanking the insertion are characteristic of most *Alu*

* Portions of this chapter previously appeared as Cook GW, Konkel MK, Major JD, 3rd, Walker JA, Han K, Batzer MA. 2011. *Alu* pair exclusions in the human genome. *Mobile DNA* **2**:10. The permission from the publisher to republish this article is available in Appendix B, page 149.

elements (Batzer et al. 1990). These flanking sequences are referred to as target site duplications (TSDs) (Grimaldi and Singer 1982; Cordaux and Batzer 2009).

The presence of TSDs suggests that a nick occurs on the complementary strand of DNA 3′ to the L1EN cleavage site on the first strand. However, little is known of the mechanisms associated with this second nick or the eventual insertion of the 5′ end of the *Alu* element (Batzer and Deininger 2002; Goodier and Kazazian 2008). This process of *Alu* element mobilization and insertion is commonly referred to as target primed reverse transcription (TPRT) (Luan and Eickbush 1995; Kazazian 1998). TPRT also occurs with two additional non-long terminal repeat (LTR) retrotransposons, LINE1 and SVA (SINE-R, variable number of tandem repeats and *Alu*) elements, within the human lineage (Konkel and Batzer 2010). While recognizing rare exceptions (Morrish et al. 2002; Srikanta et al. 2009), the majority of non-LTR retrotransposon insertions are dependent upon the activity of L1EN. As with Alu elements, LINE1 and SVA element insertions are typically characterized by TSDs that flank each element.

Alu elements also possess several features that provide directionality. Including the poly(A) tail, full-length *Alu* elements are approximately 300 bp in length (Figure 2.1) and are dimeric structures with two adenine-rich regions flanking the 3′ monomer (Weiner et al. 1986; Cordaux and Batzer 2009). The middle adenine-rich region separates the two monomers and the 3′ adenine-rich region forms the variable length poly(A) tail. Additionally, the 5′ monomer possesses the A and B boxes required for the transcription by RNA Polymerase III and the 3′ monomer contains a 31-bp insert not present in the 5′ monomer (Watson and Sutcliffe 1987; Quentin 1992).

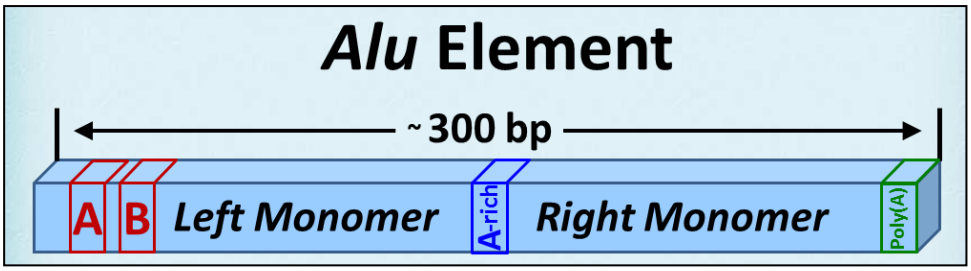


Figure 2.1 – Full-length *Alu* element A full length *Alu* element is approximately 300 bp in length and is contains two monomers of similar length. These two monomers are separated by an adenine-rich region. The 5'monomer is characterized by A and B boxes which function as promoters for RNA Polymerase III transcription. A poly(A) tail is located at the 3' end of the 3' monomer.

Inverted pairs of full-length *Alu* elements form near-palindromic sequences that are separated by spacers of other DNA sequences of varying size and composition. Palindromic sequences have been shown to be unstable in *Escherichia coli* (Collins 1981), yeast (Lengsfeld et al. 2007) and mice (Lewis et al. 1999). The genomic instability of inverted *Alu* pairs has also been demonstrated in a yeast experimental system (Lobachev et al. 2000). Other previous research has reported that inverted *Alu* pairs are potential sources of chromosomal instability when separated by ≤650 bp in humans (Stenger et al. 2001). The ability of *Alu* sequences to interact is directly correlated with the degree of sequence identity between the copies (Lobachev et al. 2000). It is estimated that the majority of full-length human *Alu* elements share sequence identity ranging between 65 and 85 percent (Stenger et al. 2001).

Alu element insertions have been linked to several genetic diseases including hemophilia, hypercholesterolemia and various cancers (Deininger and Batzer 1999; Belancio et al. 2008). While multiple diseases have been attributed to *Alu* element insertions, their most important role may be in shaping human genome architecture through various post-insertion interactions. Such interactions could result in

deletions, duplications, inversions and a host of other complex genomic structural changes (Hedges and Deininger 2007; Durbin et al. 2010). *Alu* element interactions with each other have been found to generate recombination mediated deletions and inversions (Sen et al. 2006; Han et al. 2007). In addition, *Alu* elements have been associated with multiple deletions related to various cancers (Franke et al. 2009; Konkel and Batzer 2010) and copy number variation breakpoints (Xing et al. 2009; Conrad et al. 2010; Durbin et al. 2010; Mills et al. 2011).

It has also been shown in humans that closely spaced adjacent *Alu* pairs in opposing orientation (inverted pairs) are found less frequently than *Alu* pairs having the same orientation (direct pairs) (Stenger et al. 2001). However, this imbalance has previously only been investigated for *Alu* pairs separated by ≤650 bp in a study conducted prior to the completion of the draft human genome sequence. Here, we have performed a comprehensive analysis of all (>800,000) full-length *Alu* elements (275 to 325 bp) in the public human genome assembly (hg18). Using the large data set of full-length *Alu* elements enabled us to detect small imbalances in the ratio between inverted and direct *Alu* pairs (I:D). We report a potential new insight into human genomic instability, a non-random depression in the I:D ratio for full-length *Alu* pairs whose elements are separated by up to 350,000 bp ($P<0.05$). Over 50 million (59,357,435) full-length *Alu* pairs reside within this I:D imbalance window. This phenomenon of full-length *Alu* pair I:D imbalance is hypothesized to reflect the activity of four separate mechanisms which result in *Alu* pair exclusions (APEs).

Results

The size distribution of the human genomic *Alu* element population is shown in Figure 2.2. Full-length *Alu* elements, having lengths between 275 and 325 bp,

account for approximately 69 percent of all human *Alu* elements. Slightly over two percent of human *Alu* elements have lengths greater than 325 bp with 29 percent being truncated (<275 bp). Sequences of less than 30 bp cannot be reliably determined to be actual *Alu* elements and are therefore excluded from this study (P<0.05). *Alu* element length constraints provide a full-length *Alu* element sample size of 806,880 (Methods, page 45).

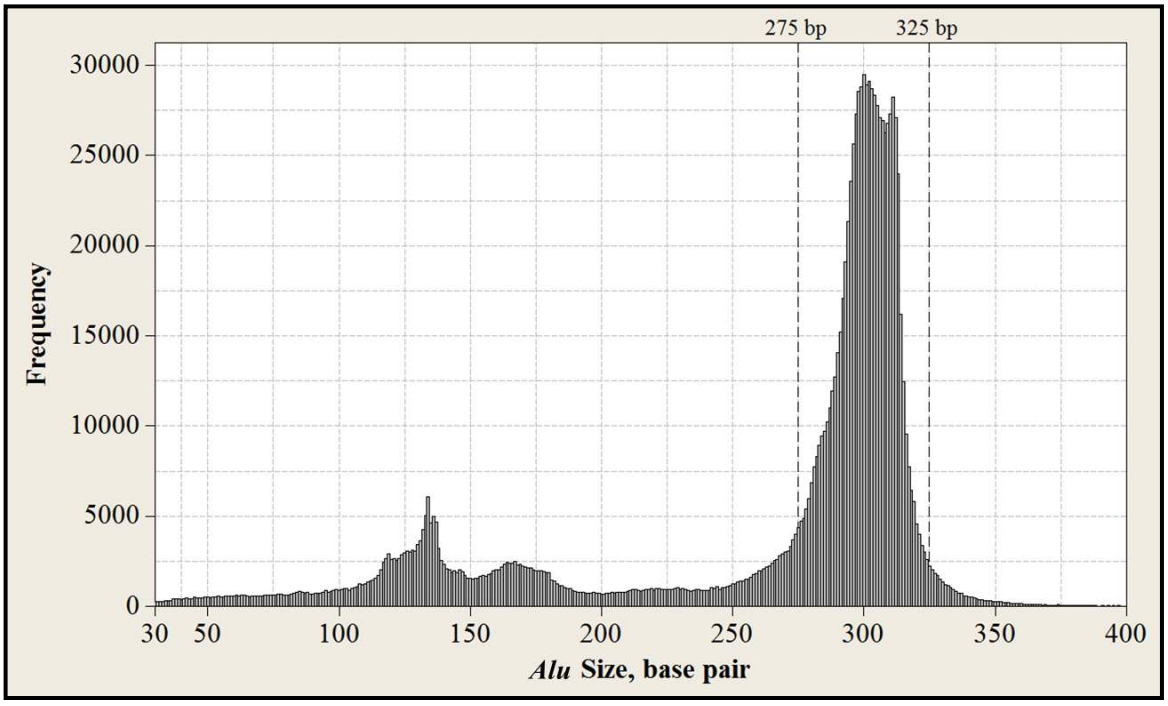


Figure 2.2 - Size Distribution of *Alu* elements in the human genome A total of 1,172,576 Alu elements (non-random) are present in the RepeatMasker scan of the hg18 genome assembly. Approximately 29.0% of these *Alu* elements have lengths less than 275 bp, 68.8% have lengths between 275 bp and 325 bp, and 2.2% have lengths greater than 325 bp. The lower limit of 30 bp is set by certainty that a given sequence is an actual *Alu* element (p<0.05).

The directionality of *Alu* elements creates four possible types of *Alu* pairs (Figure 2.3). Two of these four configurations share both elements in the same (or direct) orientation and two share elements in the opposite (or inverted) orientation. A pair of *Alu* elements in which both members of the pair are positioned on the positive strand are in the 'forward' orientation. Conversely, when both members in the pair

are positioned on the negative strand, the pair is defined as being in the 'reverse' orientation. Throughout this manuscript, the sequence separating each pair is referred to as the spacer. When an inverted *Alu* pair is oriented with the poly(A) tails pointing toward each other, the pair is termed as being in the 'tail-to-tail' orientation, and when an inverted pair is oriented with the poly(A) tails pointing away from each other, it is termed as being in the 'head-to-head' orientation.

Imbalance between the sense and antisense full-length *Alu* elements

The departure from unity in the I:D ratio for adjacent FAPs is, in part, the result of a non-random imbalance between sense and antisense orientations for full-length human *Alu* elements. The 806,880 full-length human *Alu* elements do not appear to be randomly distributed with respect to orientation. The orientational breakdown of this population is 49.80 % in the sense and 50.20% in the anti-sense orientations, respectively ($p = 0.0044$). This distribution would be expected to fall within 49.89% to 50.11% for a random distribution ($p = 0.05$). It should be noted that the human adjacent FAP population is less than the full-length *Alu* element population (560,485 and 860,880, respectively). The adjacent FAP population is smaller than the full- length *Alu* element population because of the interspersion of fragmented *Alu* elements (<275 bp) within the full-length *Alu* population.

The insertional bias associated with full-length *Alu* elements appears to affect only clustered *Alu* elements. Removal of clustered elements from the full-length *Alu* element data set returns the sense/anti-sense ratio to a range that would be expected with random insertions. There are 442,187 non-clustered adjacent human FAPs. The fraction of sense and anti-sense *Alu* elements within this group is 49.90% and 50.10%, respectively ($p=0.22$).

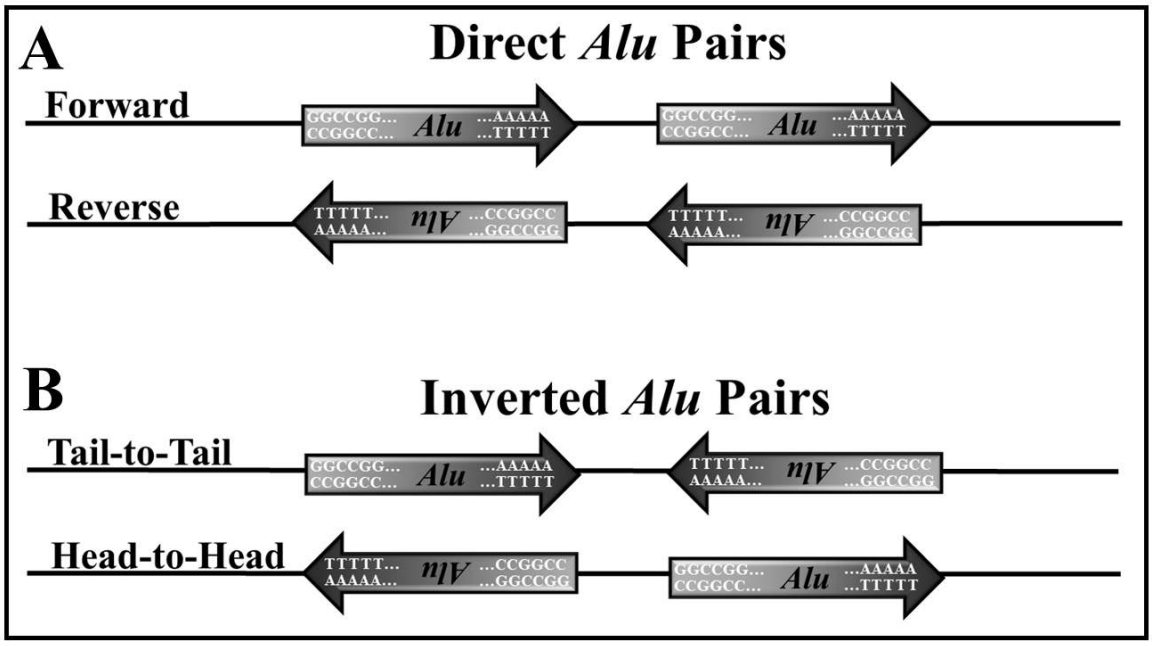


Figure 2.3 - Four types of *Alu* pairs Because of the directionality of *Alu* elements, four orientational combinations are possible for *Alu* pairs. **(A)** Direct *Alu* pairs exist when both elements are in the same orientation. When each *Alu* element is in the sense direction, the pair is defined as being in the "Forward" orientation. When both *Alu* elements in the pair are in antisense orientation, the pair is defined as being in the "Reverse" orientation. **(B)** Inverted *Alu* pairs are defined as those pairs which have the two elements in opposite orientations. When an inverted *Alu* pair is oriented with the poly(A) tails pointing toward each other, the pair is defined as being in the "Tail-to-Tail" orientation and when an inverted pair is oriented with the poly(A) tails pointing away from each other, it is defined as being in the "Head-to-Head" orientation.

I:D ratio for adjacent full-length *Alu* pairs departs from unity

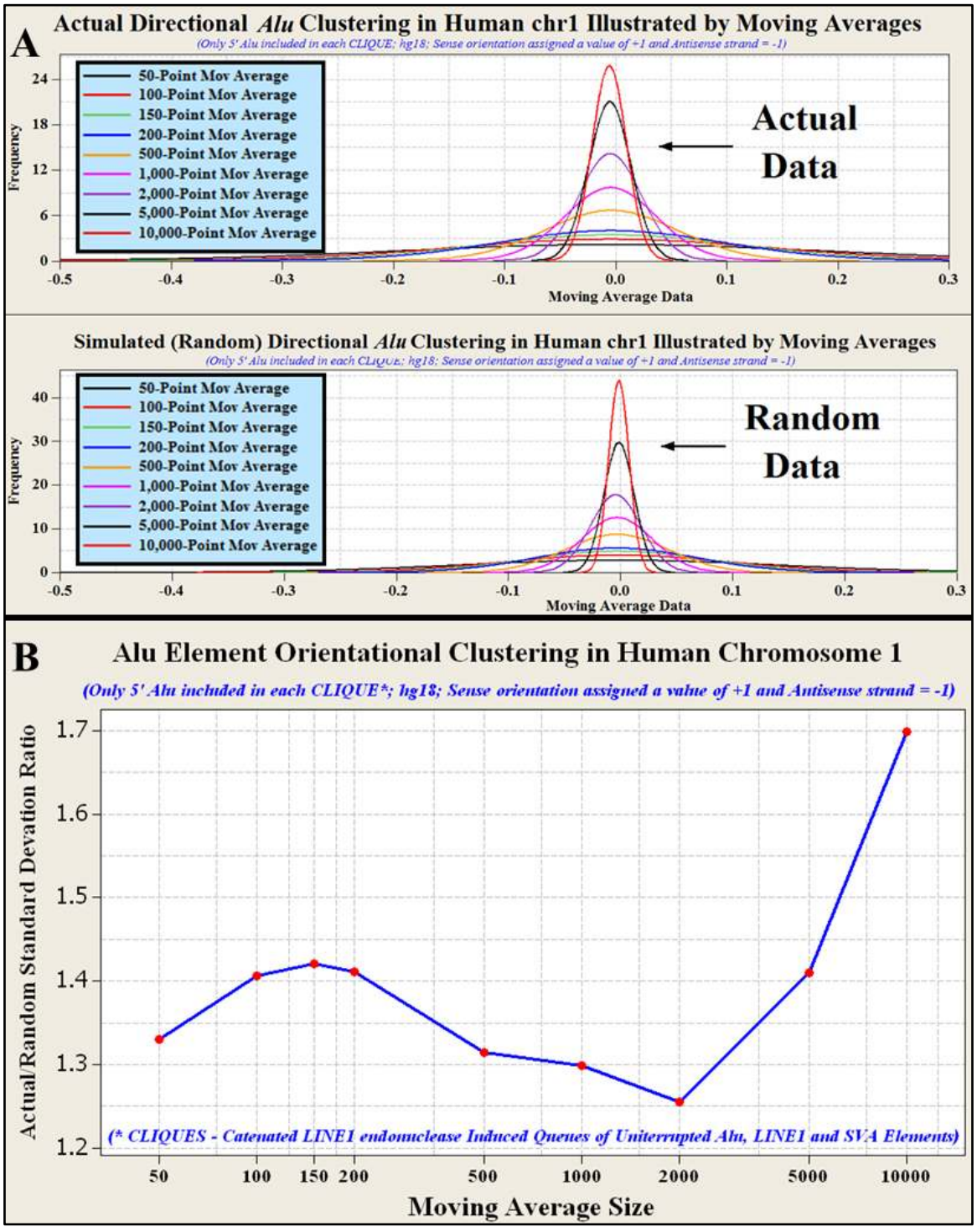
Departures from unity in the full-length *Alu* pair (FAP) I:D ratio may be suggestive of non-random insertion or deletion of *Alu* elements within the human genome. Testing for randomness was performed using binomial distributions assuming an equal probability for *Alu* insertions to occur on both the positive and negative strands (Methods, page 45). Adjacent FAPs contain no *Alu* elements within the spacer. The human adjacent FAP population of 560,485 contains 252,748 inverted pairs and 307,737 direct pairs. The I:D ratio for this population is 0.8213. Any I:D ratio outside of 0.9947 to 1.0053 reflects a non-random distribution ($P<0.05$).

The I:D ratio for adjacent FAPs of 0.8213 represents a *P*-value of <0.000001 and therefore falls well outside of the 95 percent confidence interval for randomness.

Furthermore, the adjacent FAP I:D ratio departure from unity appears to be a function of the FAP spacer size. The median spacer size for adjacent FAPs is 930 bp (mean spacer length = 921 bp). Adjacent FAPs with less than and greater than this median spacer length possess I:D ratios of 0.7105 and 0.9477, respectively. The expected I:D range for a random distribution of these half-size FAP populations is 0.9925 to 1.0075 ($P<0.05$). A more thorough analysis of the variation of FAP I:D ratio versus spacer size requires adjustment of the data set and is provided later in this section (see CLIQUES, catenated L1EN induced queues of uninterrupted *Alu*, LINE1 and SVA elements).

The adjacent FAP I:D imbalance calculation reported above provides a macroscopic view of the entire human genome. Human chromosome one was chosen to determine if a similar I:D bias (non-random distributions of *Alu* elements with respect to orientation) was evident across a smaller region of the genome. A comparison of the actual distribution versus a simulated random distribution of *Alu*

Figure 2.4 – Orientational clustering of *Alu* elements in human chromosome 1
Using the RepeatMasker scan of the hg18 human genome assembly, human chromosome 1 is home to 102,592 *Alu* elements and 34,916 CLIQUEs. CLIQUEs form the typical motif for *Alu* clustering (see related heading in this section). *Alu* elements are present in 26,277 of these CLIQUEs. Removing all but the 5' *Alu* element in CLIQUES reduces the data set to 76,539 *Alu*s. *Alu* element orientation was converted to +1 for sense *Alu*s and -1 for antisense elements, and moving averages across chr1 were calculated. **A)** Distribution of moving average values for actual and random *Alu* clustering data. Note that moving average distributions are less variable for random than for actual data. **B)** Actual/random standard deviation ratios from the distributions shown in Figure 2.4A. Note that except for the extreme cases of moving averages above 2,000, the greatest orientational clustering occurs between APSNs of 100-200. (Figure 2.4 continues on the following page.)



elements on chromosome one indicated that orientational clustering of *Alu* elements occurs over 40 percent more frequently than would be expected if *Alu* insertions were orientationally random (Figure 2.4).

Three patterns of I:D ratio

Figure 2.5 illustrates the I:D ratio for adjacent human FAPs which are separated by ≤500 bp. This range includes over one-third of the human adjacent FAP population and is the first breakdown of this I:D parameter by individual spacer length. Three distinct patterns of FAP density and I:D ratio are evident from Figure 2.5.

The first pattern is the combined high FAP density and low I:D ratio (0.073) for spacer lengths of ≤24 bp. An unexpected inflection point in the frequency of direct FAPs occurs after as spacer size of 6 bp (Figure 2.5). This pattern may be indicative of a potential orientational insertion preference for *Alu* elements within the TSD of an existing *Alu* element. The second FAP I:D ratio pattern evident in Figure 2.5, pane A (magnified in Figure 2.5, pane B) is the 13 bp span of elevated FAPs in the head-to-head orientation within the spacer size range of 24 to 36 bp. This span contains 1.6 percent of adjacent human FAPs and is the only spacer size range within the human genome where the FAP I:D ratio exceeds unity (I:D = 1.053). Previous research identified an elevated presence of *Alu* pairs (>275 bp) in this orientation for the spacer size range of 21 to 40 bp (Stenger et al. 2001). As can be seen in Figure 2.5, pane B, the most accentuated head-to-head frequencies occur between spacer lengths of 24 to 36 bp. For this span of spacer sizes, head-to-head (inverted) FAPs outnumber either forward or reverse (direct) FAPs. Although the most elevated head-to-head frequencies reside within the spacer size range of 24 to 36 bp, Figure 2.5, pane B also reveals that an attenuated elevation of head-to-head FAPs over tail-to-tail inverted FAPs is present within the spacer size range of 37 to 50 bp.

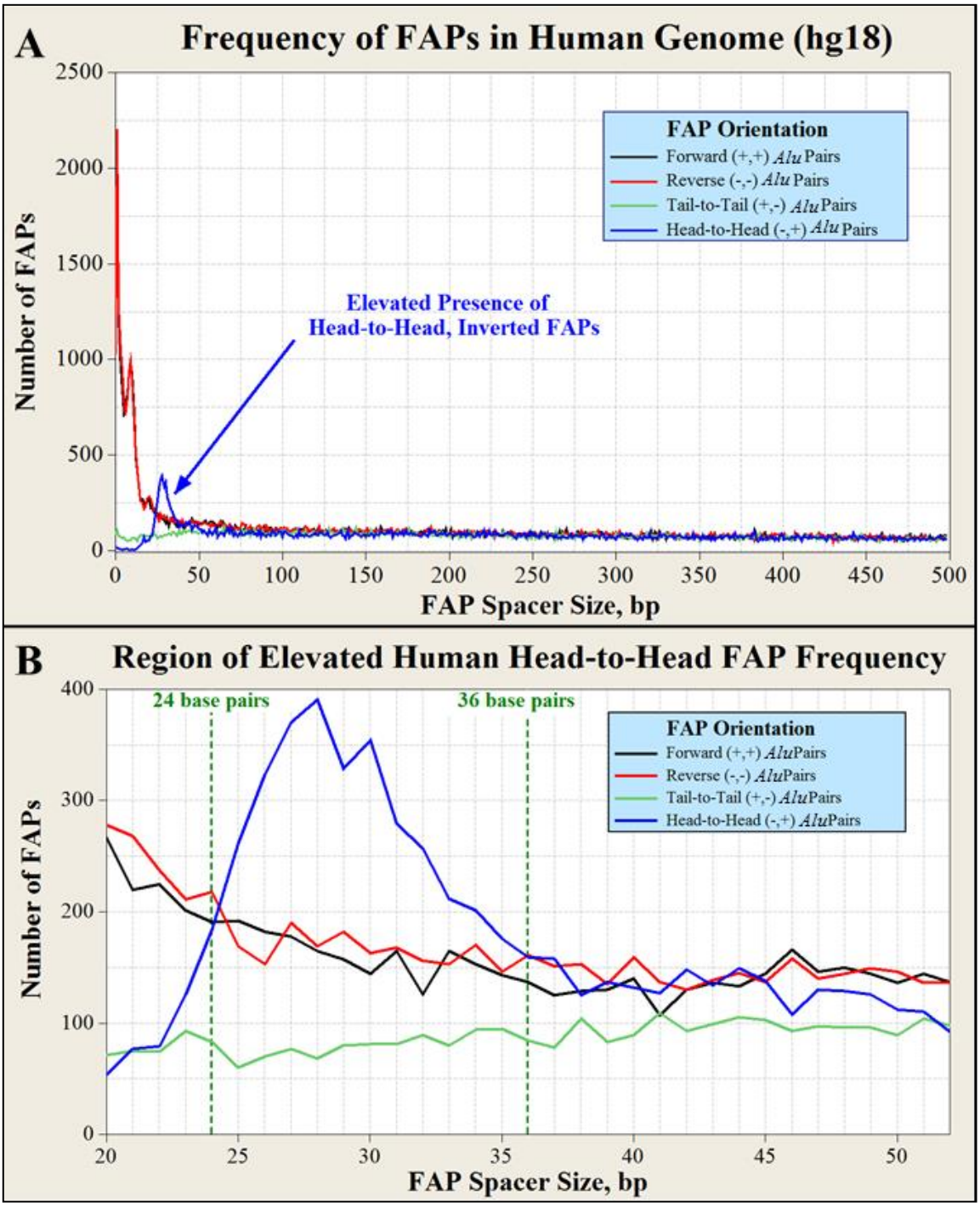


Figure 2.5 - Frequency of closely-spaced, full-length *Alu* pairs, FAPs
(A) Human adjacent FAP frequency versus the spacer size (bp) separating the two members of the FAP. The number of inverted pairs (blue and green lines) is much lower than the number of direct pairs (red and black lines) when the spacer has a size ≤24 bp (I:D = 0.076). **(B)** Spacer lengths within 24 to 36 bp define the only region within the human genome where head-to-head (inverted) FAPs outnumber either type of direct oriented FAPs.

The third FAP density and I:D ratio pattern is evident in Figure 2.5, pane A. It is characterized by similar FAP frequencies among the four *Alu* pair types between spacer sizes of 51 to 500 bp. This third pattern persists for adjacent FAPs with spacer sizes of >500 bp (data not shown).

CLIQUEs, catenated L1EN induced queues of uninterrupted *Alu*, LINE1 and SVA elements

The common dependence of *Alu*, L1NE1, and SVA insertions upon L1 enzymes raises the possibility that the clustering of closely spaced *Alu* elements (≤50 bp) observed in Figure 2.5, pane A is also associated with various combinations of all three element types. A total of 412,380 various combinations of these *Alu*-LINE1-SVA clusters are present within the human genome. These clusters comprise 16.6 percent of all human DNA and contain 52.6 percent of the *Alu*, LINE1 and SVA sequence within the human genome. Retrotransposons residing within these L1EN-induced clusters can exist in both orientations but exhibit a clear bias for one orientation. These clusters are characterized by this orientational bias as the I:D ratio for adjacent FAPs within these clusters is 0.3847. These clusters are enriched with potential L1EN target sites because of their shared TPRT insertion mechanism creating L1EN-induced TSDs flanking these three types of retrotransposons, as well as by the adenine-rich region within *Alu* elements (see Discussion, APE mechanisms). This enrichment of potential L1EN target sites inherently increases the likelihood of future *Alu*, LINE1 and SVA elements within these clusters. The common participation of *Alu*, L1NE1, and SVA elements within catenated clusters is consistent with L1EN activity. These catenated L1EN induced queues of uninterrupted *Alu*, L1NE1, and SVA elements are hereafter referred to as CLIQUEs.

The potential for TPRT-related insertion bias within TSDs makes CLIQUE identification an important consideration in evaluating deviations from unity in the FAP I:D ratio. The potential for L1EN orientational bias to propagate within CLIQUEs could conceivably result in FAPs separated by more than 10 kb to be orientationally related. As an example, CLIQUE number 397,134 (chrX:74,530,726-74,548,236) is 17,511 bp in length and contains two full-length *Alu* elements which form a FAP in the forward orientation with a spacer size of 11,870 bp. This potential for orientational bias between *Alu* elements residing within the same CLIQUE has resulted in their exclusion for determination of genome-wide FAP I:D ratios. The adjacent FAP I:D ratio, excluding FAPs generated within the same CLIQUE, reduces the FAP sample size from 560,485 to 460,588. This correction increases the adjacent FAP I:D ratio from 0.821 to 0.955. The smaller sample size for CLIQUE corrected adjacent FAPs slightly decreases the precision for detection of non-random I:D ratios from 0.9947 to 1.0053 to 0.9942 to 1.0058 ($P<0.05$). However, the CLIQUE-adjusted adjacent I:D ratio (0.955) remains statistically different from random ($P<0.00001$) even though it varies with spacer size. The most closely spaced 10 percent of human adjacent FAPs (spacer size = 51-205 bp) have an I:D ratio of 0.898 while the most distantly spaced 10 percent (spacer size = approximately 7,400-50,000 bp) have an I:D ratio of 0.989. This relationship is illustrated in Figure 2.6.

A calculated 52.6 percent of human LINE1, *Alu* and SVA sequences reside in CLIQUEs. The average CLIQUE is 1,169 bp in length and is occupied by 3.3 elements. The median CLIQUE length is 638 bp and 95 percent of all CLIQUEs have lengths less than 4,100 bp. The most CLIQUE-rich chromosome is the chromosome 19 (0.252 CLIQUES per kb) and the least rich is chromosome Y (0.061

CLIQUEs per kb). Over half of the longest 100 CLIQUEs are found on chromosome X, with the longest being over 55,000 bp at locus chrX:75,592,945-75,648,671 (Figure 2.7).

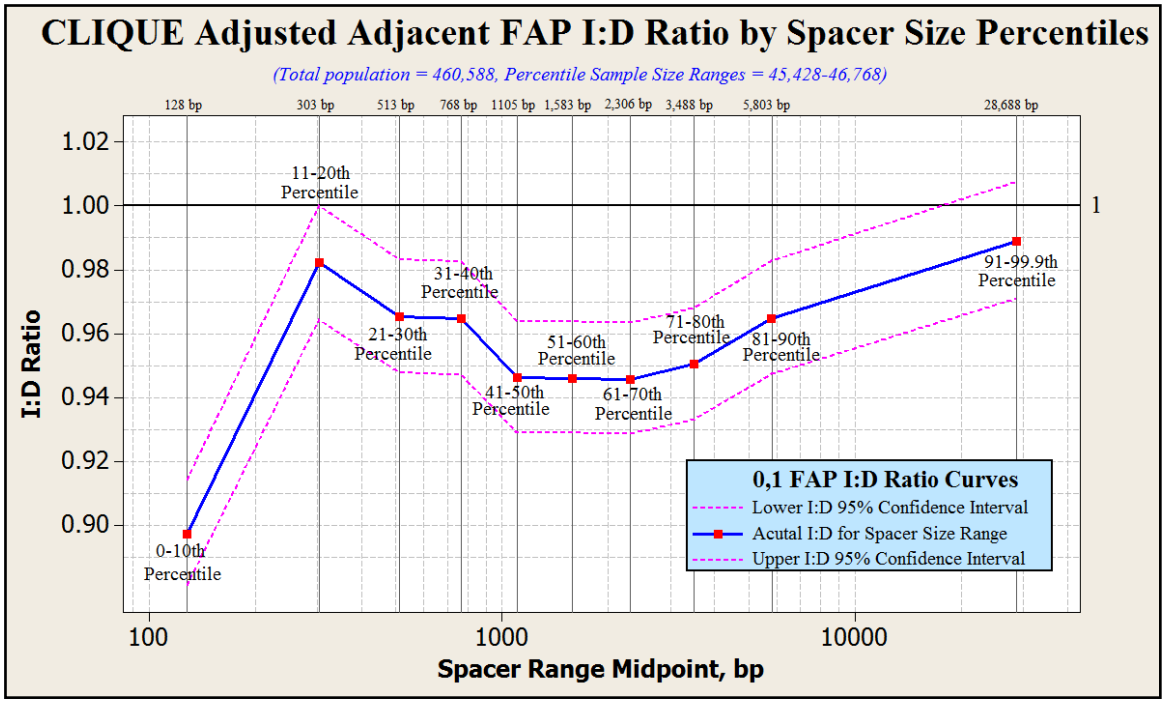


Figure 2.6 - CLIQUE adjusted adjacent FAP I:D ratios versus spacer size The CLIQUE adjusted adjacent FAP population is 460,588. This population was broken down into 10 approximately equally-sized groups (size range = 45,428-46,768) based on spacer size. The midpoints of each range are shown along the top border of the graph. The actual I:D ratio for each percentile range is shown (blue) along with the upper and lower boundaries of the 95% confidence interval (red).

Non-adjacent *Alu* pairs

One of the findings in this study is that the FAP I:D imbalance is not limited to adjacent FAPs. Intervening *Alu* elements within the spacer of a FAP also generate non-random FAP I:D ratios. This non-random I:D imbalance ($P<0.05$) was detected in FAPs whose spacer contains up to 106 intervening *Alu* elements and > 350,000 bp. Taken at the whole human genome level, the human FAP I:D imbalance window encompasses ± 107 of an *Alu*'s neighboring *Alu* elements (Methods, page 45). No size constraint was placed upon intervening *Alu* elements. Therefore, while the

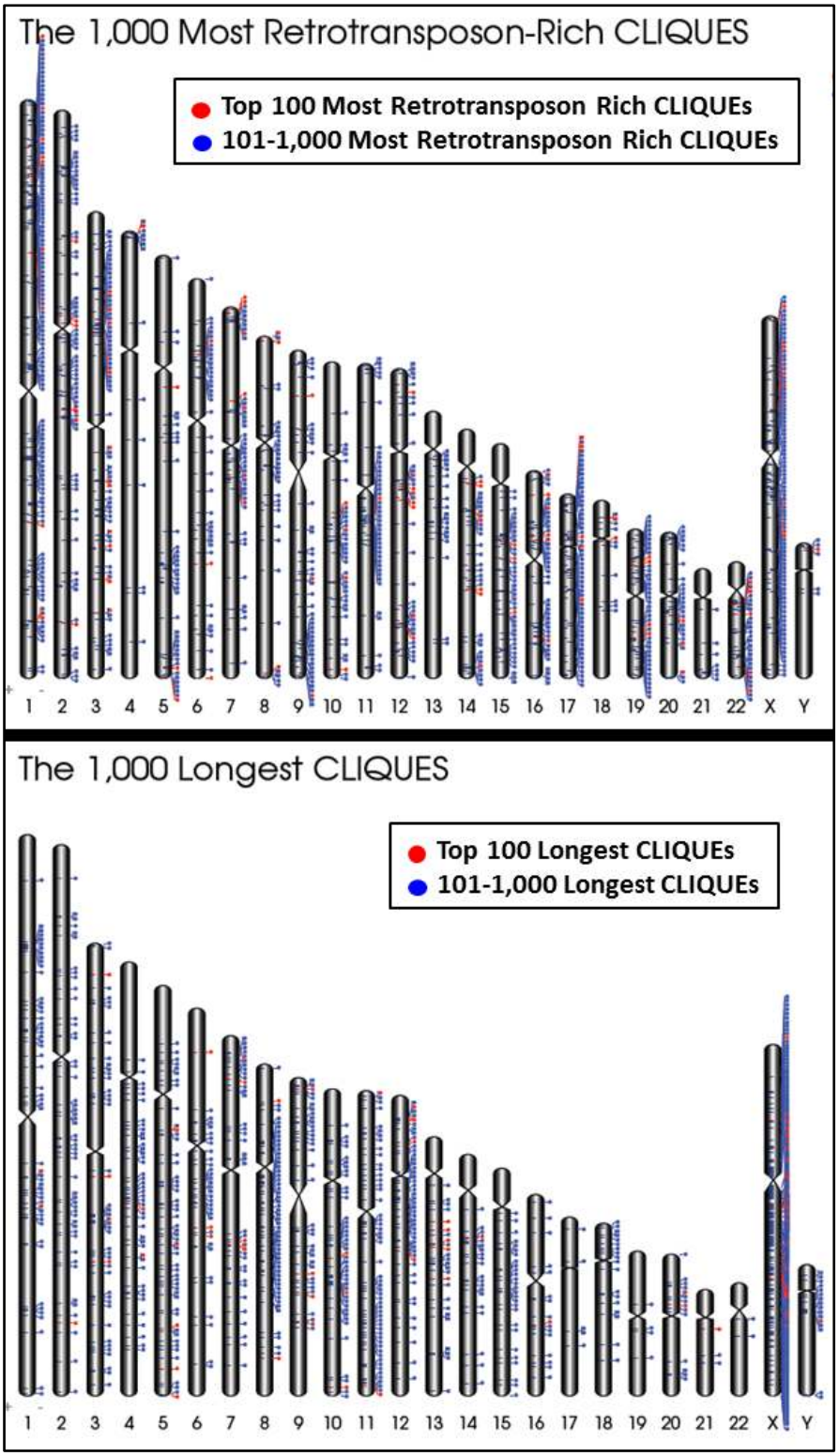


Figure 2.7 - CLIQUE density across the human genome (**Top pane**) The 1,000 most retrotransposon-rich CLIQUEs and (**Bottom pane)** the 1,000 CLIQUES with the longest sequence. Note that the top 100 most retrotransposon-rich and longest CLIQUEs are denoted in red in each ideogram.

entire inventory of human *Alu* elements is used in this study, only I:D ratios for FAPs are reported. The smallest CLIQUE adjusted FAP sample size (460,588) occurs for adjacent FAPs. Sample size ranges of 551,764 to 557,454 exist for all FAP families with more than three intervening *Alu* elements within the spacer (Table 2.1, page 24). The inclusion of FAPs with intervening *Alu* elements requires terminology for defining different FAP types (Figure 2.8 and Methods, page 45).

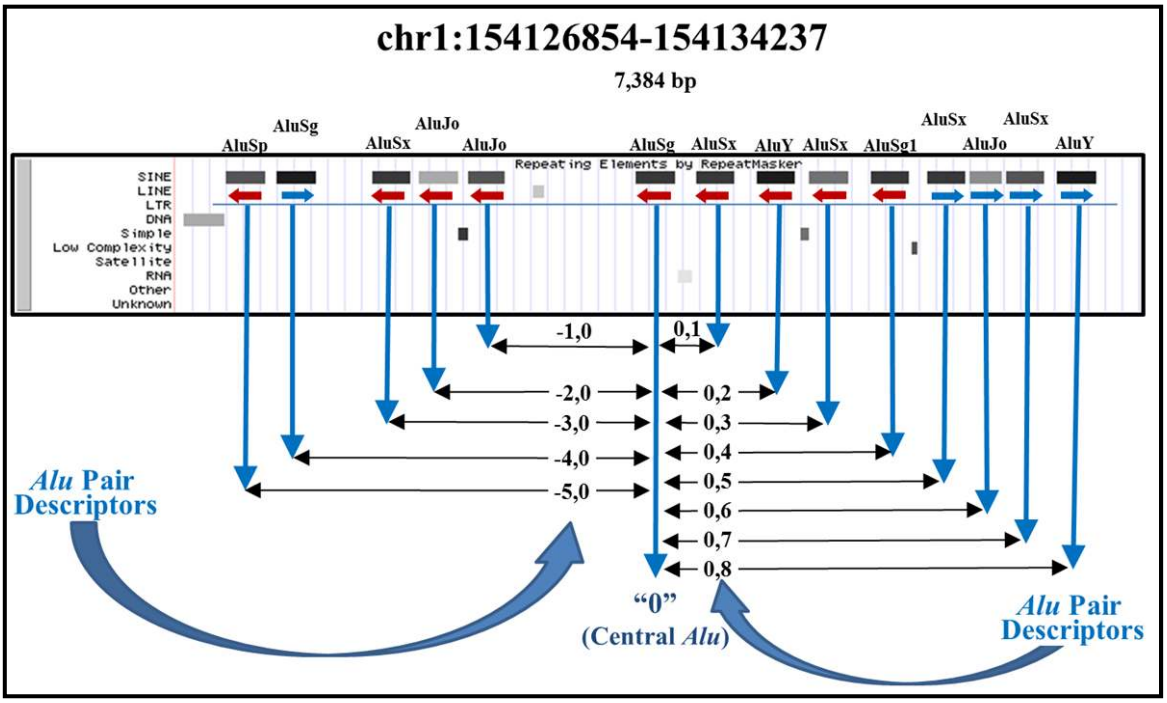


Figure 2.8 - Naming convention for FAPs This example from chr1:154,126,854-154,134,237 (7,384 bp) illustrates the FAP naming convention. The central *Alu* is always the element being evaluated and the second member of the pair is designated by its sequential separation from the central *Alu*. The central *Alu* is designated with the number '0'. The absolute value of the sequential separation of a given *Alu* element from the central *Alu* is defined as its APSN. Additionally, *Alu* elements located 5′ of the central *Alu* are assigned a negative value and with a positive value if located 3′ of the central *Alu*. APSN: *Alu* pair sequence number; FAP: full-length *Alu* pair.

I:D ratio versus *Alu* pair sequence number

Adjusting the adjacent (0,1) FAP population for CLIQUEs increases its median spacer size from 930 to 1,296 bp. The CLIQUE-adjusted I:D ratios for the

smaller and larger spacer sizes about this new median are 0.951 and 0.959, respectively. Both of these I:D ratios are outside of the 0.9918 to 1.0082 range which would be expected for a random distribution ($P<0.05$). The small difference between these I:D ratios raises the possibility that FAPs with much larger spacers may also be subject to an FAP I:D imbalance. Unfortunately, this hypothesis is difficult to measure using only adjacent FAPs as 95 percent of this population has spacer sizes of less than 11,005 bp. The inclusion of intervening *Alu* elements within FAP spacers permits identification of the boundaries of the FAP I:D imbalance phenomenon . The FAP I:D ratio as a function of *Alu* pair sequence number (APSN) is shown in Figure 2.9. Both unadjusted and CLIQUE-corrected I:D curves are provided in this figure. Figure 2.9, pane A shows FAP I:D ratios across APSN values of ± 1,000 and reveals that the FAP I:D ratio depression appears to be limited to APSNs of ≤100. Further refinement of this I:D depression boundary was accomplished by grouping 10 consecutive APSNs together. This increased the FAP sample size from approximately 555,000 to over 5.5 million. The larger sample size improved the precision of detection of the I:D depression boundary to an APSN value of ± 107 (Methods, page 45).

Over 50 million FAPs reside within the CLIQUE-adjusted FAP I:D imbalance window. Based on the CLIQUE-adjusted I:D values illustrated in Figure 2.9, human direct FAPs outnumber inverted FAPs by 629,027 (Table 2.1, page 24). Random variation reduces this difference to 613,924 ($P<0.05$). Figure 2.9, pane C magnifies pane A in the same figure to APSN values of ± 15 and illustrates that the greatest departure between CLIQUE-adjusted and unadjusted FAP I:D ratios occurs for APSNs of less than five. The largest APSN for a FAP residing within a single human

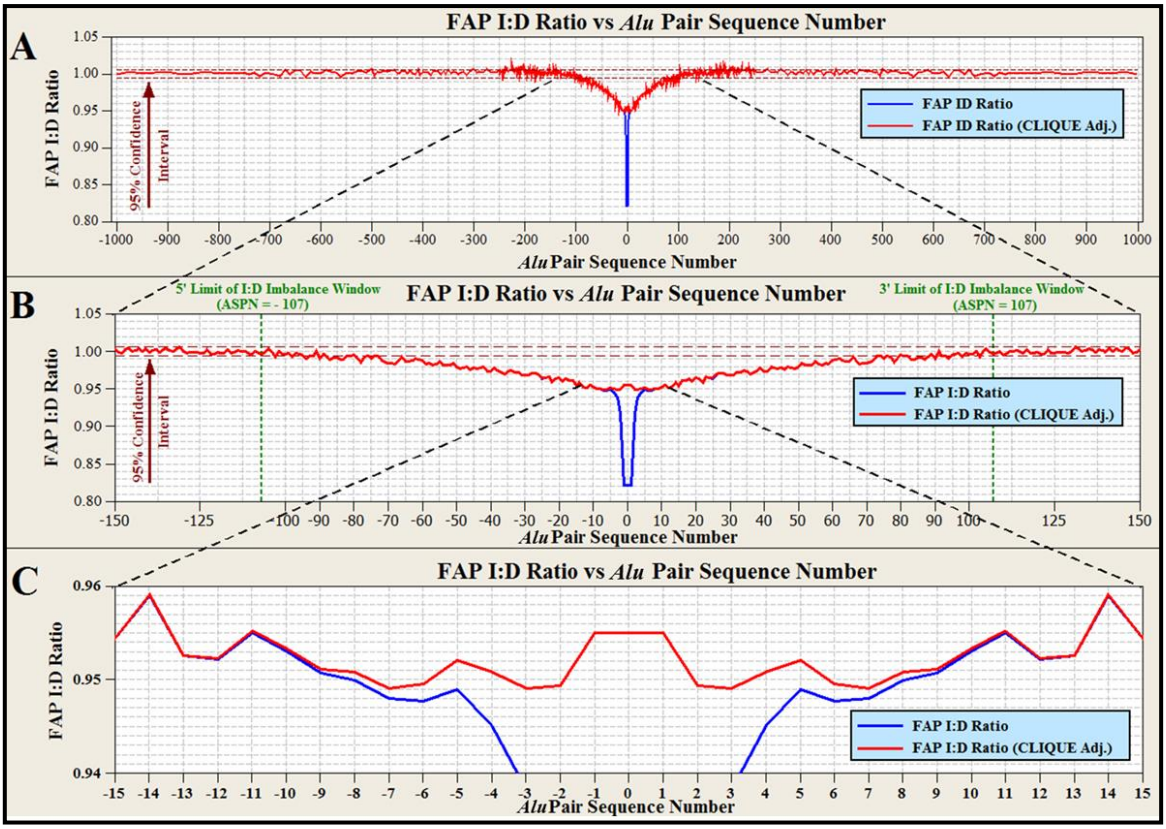


Figure 2.9 - FAP I:D ratio versus *Alu* pair sequence number with and without adjusting for CLIQUEs **(A)** The I:D ratio of full length *Alu* pairs for APSNs of ± 1,000 *Alu* elements. Note that a bubble of depressed I:D ratio exists for those elements within about ± 100 *Alu* elements of the central *Alu* element. **(B)** A closer view of the I:D imbalance bubble. The 95% confidence for each value is estimated ± 0.6%. Therefore, the bubble of I:D imbalance extends for an approximately APSN = ± 85 around the central *Alu*. A more rigorous treatment of the data (see text) extends this I:D imbalance boundary to an APSN = ± 107. **(C)** Over 99% of the impact of CLIQUEs on the FAP I:D ratio dissipates after the APSN = 5. The largest CLIQUEs, while rare, contain up to 32 *Alu* elements. No CLIQUE impact exists on the FAP I:D ratio for an APSN > 31. APSN: *Alu* pair sequence number; CLIQUE: catenated LINE1 endonuclease induced queue of uninterrupted *Alu*, LINE1 and SVA elements; FAP: full-length *Alu* pair; I:D Ratio: ratio between inverted and direct *Alu* pairs.

CLIQUE is 0,31. Consequently, no CLIQUE adjustments to the FAP I:D ratio are required for APSN values greater than 31.

PCR evidence of *Alu* pair exclusions in the chimpanzee genome

We have presented computational evidence for a significant FAP I:D ratio imbalance in the human genome. To investigate our hypothesis that this imbalance

Table 2.1 - CLIQUE adjusted FAP sample sizes and I:D ratios, hg18

APSN Type	Total Number	I:D	APSN Type	Total Number	I:D	APSN Type	Total Number	I:D	APSN Type	Total Number	I:D
0,1	460,588	0.9550	0,30	556,475	0.9690	0,59	556,158	0.9830	0,88	555,764	0.9928
0,2	526,986	0.9494	0,31	556,217	0.9684	0,60	556,035	0.9887	0,89	555,471	0.9972
0,3	540,117	0.9491	0,32	556,631	0.9718	0,61	556,044	0.9897	0,90	556,080	0.9900
0,4	547,346	0.9508	0,33	556,424	0.9723	0,62	556,041	0.9899	0,91	555,560	1.0000
0,5	551,764	0.9521	0,34	556,949	0.9744	0,63	556,373	0.9884	0,92	555,753	0.9945
0,6	554,173	0.9496	0,35	557,086	0.9733	0,64	556,142	0.9869	0,93	555,742	0.9942
0,7	554,928	0.9491	0,36	556,551	0.9702	0,65	556,181	0.9865	0,94	555,439	0.9907
0,8	555,811	0.9508	0,37	556,800	0.9727	0,66	555,964	0.9929	0,95	555,643	0.9952
0,9	556,349	0.9511	0,38	556,785	0.9743	0,67	556,033	0.9876	0,96	555,501	0.9965
0,10	556,963	0.9533	0,39	556,512	0.9782	0,68	555,737	0.9837	0,97	555,354	0.9984
0,11	556,857	0.9552	0,40	556,742	0.9737	0,69	555,962	0.9848	0,98	555,539	0.9933
0,12	557,454	0.9523	0,41	556,808	0.9729	0,70	555,822	0.9843	0,99	555,980	0.9978
0,13	557,033	0.9526	0,42	556,642	0.9795	0,71	555,873	0.9859	0,100	555,392	0.9966
0,14	557,023	0.9591	0,43	556,820	0.9787	0,72	556,065	0.9877	0,101	555,340	0.9961
0,15	556,948	0.9545	0,44	556,216	0.9776	0,73	555,935	0.9942	0,102	555,491	1.0001
0,16	557,239	0.9615	0,45	556,359	0.9782	0,74	555,555	0.9945	0,103	555,697	0.9930
0,17	556,970	0.9620	0,46	556,046	0.9762	0,75	555,763	0.9900	0,104	555,014	0.9987
0,18	557,002	0.9640	0,47	556,704	0.9798	0,76	556,130	0.9938	0,105	555,082	1.0034
0,19	556,886	0.9597	0,48	556,660	0.9782	0,77	556,214	0.9926	0,106	555,165	0.9986
0,20	557,127	0.9649	0,49	556,488	0.9774	0,78	555,611	0.9857	0,107	555,588	0.9971
0,21	556,925	0.9642	0,50	555,988	0.9799	0,79	555,694	0.9912	0,108	555,104	0.9977
0,22	557,364	0.9587	0,51	556,457	0.9839	0,80	555,716	0.9957	0,109	555,298	1.0009
0,23	556,997	0.9660	0,52	556,370	0.9816	0,81	555,617	0.9946	0,110	555,168	0.9959
0,24	556,822	0.9651	0,53	556,147	0.9826	0,82	555,764	0.9945	0,111	555,536	0.9973
0,25	556,542	0.9645	0,54	556,423	0.9820	0,83	555,703	0.9891	0,112	555,117	1.0007
0,26	557,104	0.9700	0,55	556,245	0.9873	0,84	555,973	0.9895	0,113	555,699	0.9997
0,27	556,690	0.9706	0,56	556,205	0.9837	0,85	555,822	0.9918	0,114	554,985	1.0013
0,28	556,952	0.9707	0,57	556,331	0.9819	0,86	555,846	0.9915	0,115	555,514	0.9994
0,29	556,469	0.9689	0,58	556,164	0.9845	0,87	555,393	0.9898			

may be due to the increased instability of inverted *Alu* pairs, resulting in APEs, we compared the human genome (hg18) to the chimpanzee genome (panTro2) to identify potential APE deletions. A total of 58 APE deletion candidate loci were identified for evaluation by PCR (Methods, page 45) in the chimpanzee genome through comparison of the human, chimpanzee, orangutan and rhesus macaque genome draft sequences.

Fourteen of these loci were selected for PCR examination. These validations confirmed that 10 of these 14 loci had undergone chimpanzee-specific deletions consistent with inverted FAP instability. PCR primer design was problematic for the remaining four loci. No instances of false positive identification of chimpanzee-specific deletions were observed. The characteristics of the 10 loci confirmed as chimpanzee-specific deletions are summarized in Table 2.2. Images of gel

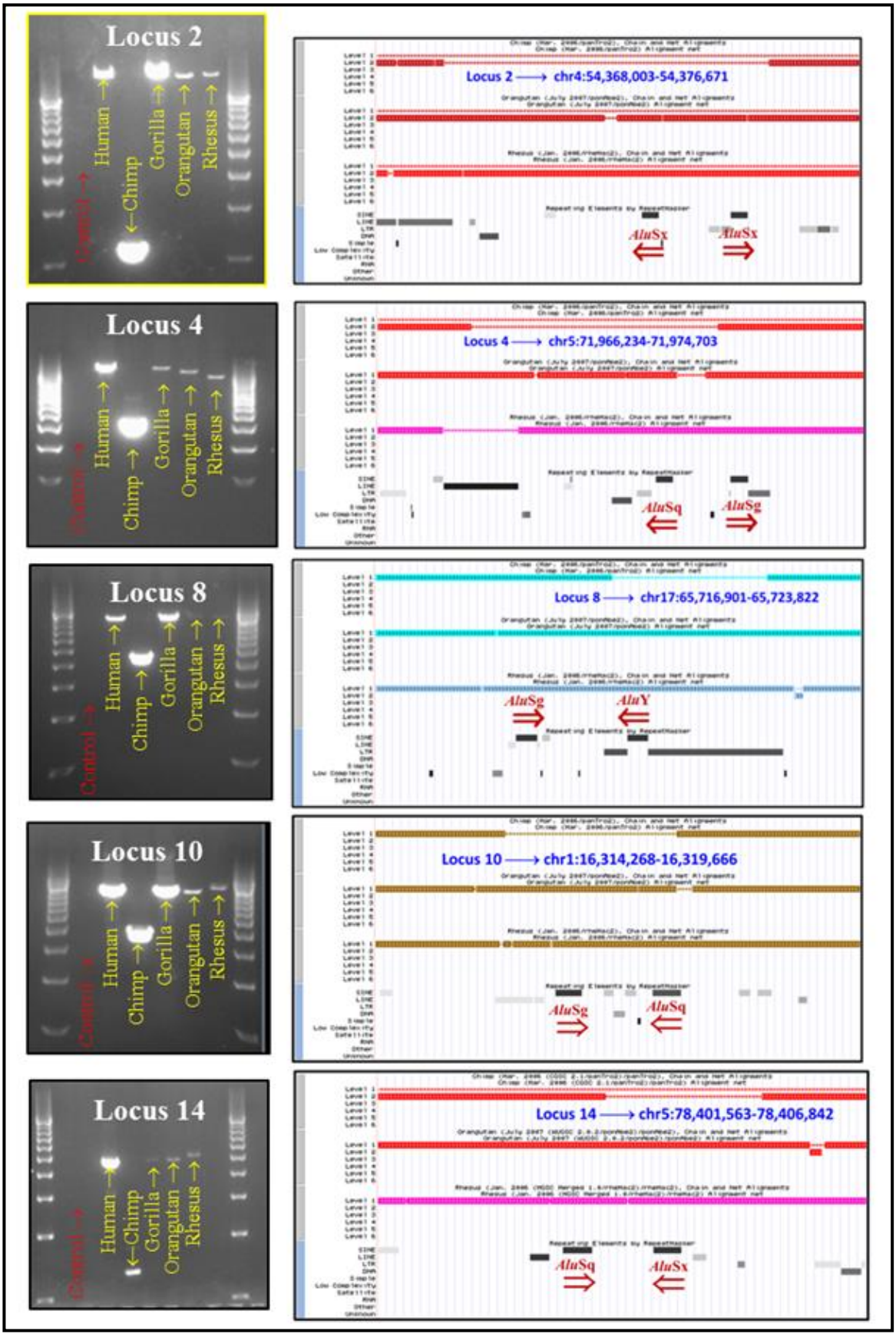
chromatographs of the experimental interrogation of five of the loci are shown in Figure 2.10.

Table 2.2 - Chimpanzee specific APEs characterized by PCR

Locus ID	Position (hg18)	5' *Alu* Element			Spacer (bp)	3' *Alu* Element			Chimpanzee Deletion Size (bp)
		Subfamily	Length (bp)	Orientation		Subfamily	Length (bp)	Orientation	
1	chr1:105842254-105848252	*AluY*	300	Positive	1,407	*AluJb*	291	Negative	4,896
2	chr4:54368003-54376671	*AluSx*	297	Negative	1,292	*AluSx*	310	Positive	5,829
3	chr2:68246922-68253405	*AluY*	312	Negative	1,237	*AluY*	304	Positive	3,413
4	chr5:71966234-71974703	*AluSq*	293	Negative	1,012	*AluSg*	310	Positive	4,307
5	chr13:64130795-64137788	*AluJo*	297	Positive	1,312	*AluSx*	292	Negative	2,776
8	chr17:65716901-65723822	*AluSg*	303	Positive	1,285	*AluY*	300	Negative	5,585
9	chr8:53032075-53037664	*AluSx*	309	Positive	973	*AluSx*	307	Negative	2,340
10	chr1:16314268-16319666	*AluSg*	296	Positive	793	*AluSq*	309	Negative	1,907
14	chr5:78401563-78406842	*AluSq*	313	Positive	665	*AluSx*	301	Negative	1,656
15	chr4:68494452-68500177	*AluY*	318	Negative	1,121	*AluSx*	286	Positive	1,654

A secondary purpose of these PCR examinations was to assess the accuracy of the hg18 and panTro2 genome assemblies at loci involved in APE deletions. If we broadly assume that the combined hg18/panTro2 genome assemblies provide at least 50% accuracy in identification of inverted APE deletion loci, the probability of successfully validating five of these events in five consecutive PCR evaluations would be P=0.03125 (0.5^5). The fact that we were able to validate 10 such APEs from the ectopic invasion and annealing of high-homology bubbles events in 10 consecutive PCR reactions with no evidence of false positives provides over 95%

Figure 2.10 – Chimpanzee specific APE deletions PCR analysis confirmed chimpanzee-specific APE deletions in orthologous human, chimpanzee, gorilla, orangutan and rhesus macaque loci. Human adjacent inverted FAP loci were chosen with spacer sizes between 651 and 1500 bp and a minimum of 1,000 bp of *Alu*-free flanking sequence. PCR loci were selected for which the chimpanzee loci were >350 bp shorter than the human ortholog. Using identical primers, PCRs were then prepared for human, chimpanzee, gorilla, orangutan and rhesus macaque. (Figure 2.10 continues on the following page.)


Locus 2
Human →
Chimp →
Gorilla →
Orangutan →
Rhesus →
Control →
Locus 2 ⟶ chr4:54,368,003-54,376,671
AluSx
AluSx
Locus 4
Human →
Chimp →
Gorilla →
Orangutan →
Rhesus →
Control →
Locus 4 ⟶ chr5:71,966,234-71,974,703
AluSq
AluSg
Locus 8
Human →
Chimp →
Gorilla →
Orangutan →
Rhesus →
Control →
Locus 8 ⟶ chr17:65,716,901-65,723,822
AluSg
AluY
Locus 10
Human →
Chimp →
Gorilla →
Orangutan →
Rhesus →
Control →
Locus 10 ⟶ chr1:16,314,268-16,319,666
AluSg
AluSq
Locus 14
Human →
Chimp →
Gorilla →
Orangutan →
Rhesus →
Control →
Locus 14 ⟶ chr5:78,401,563-78,406,842
AluSq
AluSx

confidence that these two assemblies are at least 74 percent accurate (0.74^{10} = 0.04924). When we compared the PCR-based estimate of chimpanzee-specific inverted APE deletions to the computationally derived estimate of human inverted APE deletions for this same data set, we found these results to be within 15 percent of each other (108 versus 94). The computation was based upon the human FAP I:D ratio (0.931) for loci satisfying the original PCR criteria (Methods, page 45). Thus, these data provide strong evidence for the existence of APE-induced genomic deletions. The characteristics of the 10 loci confirmed as chimpanzee-specific deletions are summarized in Table 2.3. Images of gel chromatographs of the experimental interrogation of five of the loci are shown in Figure 2.10. Chimpanzee-specific APE deletions within these (human) orthologous loci were estimated to have occurred during the six million years following the divergence between human and chimpanzee lineages (Xing et al. 2009).

Comparison of orthologous human-chimpanzee direct and inverted FAP loci

An effort was made to better compare the characteristics of deletions within direct and inverted FAP loci. Loci selection criteria for this evaluation were identical to those used for PCR validation with two exceptions: direct FAP loci were included and chimpanzee loci were limited to those that were 1,000 to 2,000 bp shorter than their human orthologs. The second constraint was applied to avoid lengthy deletions that could be more difficult to analyze and also to provide a reasonable sample size for manual analysis. Surprisingly, these criteria generated an almost equal number of shorter direct (193) and inverted (187) chimpanzee orthologs. A subsequent examination of the shorter direct chimpanzee FAP loci revealed that inverted APE-related deletions can plausibly be attributed to 93 (48%) of these shorter orthologous

loci. These deletions are consistent with an interaction between a member of the direct FAP and a flanking *Alu* element in the opposite orientation. Furthermore,

Table 2.3 - Primers for selected APE loci listed in Table 2.2
(Orthologous in Human, Chimpanzee, Gorilla, Orangutan and Rhesus macaque)

Loci ID	hg18 Position	Primers	Temperatures Anneal	Extend	Inverted *Alu* Pair	FAP Orientation(1)	Spacer Size (bp)
1	chr1:105842254-105848252	Forward: GGAAAGTGGATATCCTTTGGG Reverse: TTGTTCATTGTTCCTTTTAATT	50°C	68°C	*AluY- AluJb*	Tail-to-Tail	1,407
2	chr4:54368003-54376671	Forward: CCTCATGTCCTCCCCTTTAC Reverse: CACCATGAGCTCATCCTATGC	50°C	68°C	*AluSx- AluSx*	Head-to-Head	1,292
3	chr2:68246922-68253405	Forward: CATCGAGTTCTCTTCCATAGC Reverse: CCTGAAAAGGGTGAAATGGAG	50°C	68°C	*AluY- AluY*	Head-to-Head	1,237
4	chr5:71966234-71974703	Forward: GGCAAATCCTGTTTCACCACC Reverse: GGAAACGAGGCTAAATAATGGC	62°C	68°C	*AluSq-AluSg*	Head-to-Head	1,012
5	chr13:64130795-64137788	Forward: CTACATAAGCTTGCACTTCTTTG Reverse: AGTAAGAAAGCTGGTTCTGAAGA	50°C	68°C	*AluJo- AluSx*	Tail-to-Tail	1,312
8	chr17:65716901-65723822	Forward: GGGAAAATTGTTTCTGTACAGGG Reverse: CACATGCTGAGAAGCCACTAC	50°C	68°C	*AluSg- AluY*	Tail-to-Tail	1,285
9	chr8:53032075-53037664	Forward: GTCAGTCCACCAAGGTGGTTA Reverse: CCCTTAAAACATATCTGGAATCATC	50°C	68°C	*AluSx- AluSx*	Tail-to-Tail	973
10	chr1:16314268-16319666	Forward: GATCTGGCCCTAGATTTGACAG Reverse: GCCTGTTCCTAGAGGAGTTGC	62°C	68°C	*AluSg- AluSq*	Tail-to-Tail	793
14	chr5:78401563-78406842	Forward: GGTAGTTAGAATAGCAGTGAAGG Reverse: GCAGAAAGGAGTTTAATATTGAG	55°C	68°C	*AluSq- AluSx*	Tail-to-Tail	665
15	chr4:68494452-68500177	Forward: GGAATGGTTTCTCTTAGCAGC Reverse: GTGAGATCCTGAGCAGAAAGC	60°C	68°C	*AluY- AluSx*	Head-to-Head	1,121

(1) When an inverted *Alu* pair is oriented with the poly(A) tails pointing toward each other, the pair is defined as being in the "Tail-to-Tail" orientation, and when an inverted pair is oriented with the poly(A) tails pointing away from each other it is defined as being in the "Head-to-Head" orientation.

excluding chimpanzee orthologs that are shorter because of a human-specific retrotransposon insertion, fully 75 percent of the balance of the shorter chimpanzee loci can be plausibly attributed to have resulted from a flanking inverted APE-related deletion (Table 2.4, page 30). The attribution of shorter chimpanzee orthologs to possible inverted APE-related deletions is based upon the hypothesized APE deletion mechanism involving the resolution of *Alu*-induced double-strand breaks. These double-strand breaks are theorized to arise from the ends of *Alu* elements involved in an inverted Alu pair interaction. This mechanism is discussed in detail in the Discussion section of this report in Figures 2.13 through 2.15 (pages 37-42). This hypothesized APE deletion pattern applies to interactions between inverted FAPs with spacer sizes over 50 bp.

Comparison of Direct and Inverted FAPs in Orthologous Chimpanzee/Human Loci

Further examination into the APE phenomenon was made by examination of orthologous direct and inverted FAP loci in the chimpanzee genome (panTro2) and the human genome (hg18). The results of this examination are shown in Table 2.4. As with PCR comparisons (Figure 2.10, pages 25-26), the selection criteria for these FAP loci were a spacer size of 651-1,500 bp with 1,000 base pair of *Alu*-free flanking sequence. Once identified, these initial loci were filtered using the LiftOver feature in BLAT. All chimpanzee loci which were 1,000 -2,000 bp shorter than their human ortholog were chosen for manual examination.

The total direct and inverted FAP loci selected for individual examination were 193 and 186 loci, respectively. Evidence for shorter chimpanzee sequences fell into three categories; A) human specific retrotransposon insertion or repetitive DNA insertions (116 loci), B) possible APE-related deletions (254 loci) and C) possible non-*Alu* related sequence deletions (8 loci). The focus of this examination was category B. Category B is further broken down into three sub-categories. The first sub-category (201 direct and inverted FAP loci) contained an orthologous human inverted FAP which could be reasonably associated with an APE-related deletion in chimpanzee. The second sub-category (53 direct plus inverted FAP loci) contained patterns that did not conform to what would be expected from an inverted *Alu* related deletion. Specifically, these 53 deletions did not include sequences that included the ends of an *Alu* element. However, each of these 53 loci were found to contain (within the human indel) at least one consensus L1EN target sequence in the orientation required to form an inverted *Alu* pair. In the case of this second subcategory, the insertion of a chimpanzee specific *Alu* element within the indel

could potentially generate an inverted APE deletion event. Such an *Alu* insertion would have the potential to eliminate the new *Alu* insertion from detection.

Table 2.4 Comparison of orthologous direct and inverted FAP loci[1]

Loci Characteristics	Direct FAP Loci (Number,%)	Inverted FAP Loci (Number,%)
Orthologous panTro2/hg18 FAP Loci [2]		
Total orthologous FAP loci (100% of FAP population)	14,680	13,664
PanTro2 loci 1,000-2,000 bp shorter than hg18 ortholog	193, 1.2%	186,1.4%
Examination of Shorter Chimp Loci		
1 – Human-Specific Retrotransposon or Repetitive DNA Insertions	72, 37.3%	45, 24.2%
2 - Possible APE-Related Deletions		
A-Possible interaction of inverted *Alu* pair associated with indel [3]	95, 49.2%	106, 57.0%
B-Inverted L1EN Target Site(s) within human/chimp indel[4]	22, 11.4%	31, 16.7%
3 – Possible non-*Alu* Inverted Sequence Deletions		
C-Palindrome (with spacer) within human/chimp indel [5]	4, 2.1%	4, 2.2%
Potential APEs Resulting in *Alu-Alu* SSA[6] Repair, % of APEs	15[7], 16.1%	5[7], 4.7%

(1) panTro2 loci which are 1,000-2,000 bp shorter than the orthologous loci in hg18.
(2) Orthologous loci have hg18 spacer sizes between 651-1,500 bp and 1,000 bp of 5' and 3' "*Alu* element free," flanking sequence.
(3) Approximately half of the shorter chimpanzee direct FAP loci had deletion patterns characterized by deletions proceeding from the end of one of the two elements making up the pair (i.e., a deletion pattern consistent with the predicted inverted APE deletions as illustrated in Figure 2.15, diagram C, page 42). These potential APE deletions could result from the instability of a second inverted *Alu* pair formed by a flanking *Alu* element and one of the *Alu* elements within the FAP being evaluated.
(4) One or more L1EN target site sequences (5'-TTTTAA-3') is/are present in the orthologous human sequence of the chimpanzee deletion. These orthologous target sites are in the inverted orientation relative to an existing *Alu* present within the loci window. The presence of L1EN inverted target site(s) within this human/chimpanzee orthologous indel opens the possibility that the indel may be the result of a chimpanzee-specific APE deletion catalyzed by a chimpanzee-specific *Alu* insertion.
(5) A palindrome of minimum length of 7 bp was present in the orthologous human sequence of the chimpanzee deletion. This palindrome could create a potential region of instability within the deletion. This instability could possibly occur by a mechanism similar to those outlined in Figures 2.13 and 2.14 (pages 37-40**.)**
(6) SSA – Single Strand Annealing repair (Hedges et al., 2007).
(7) The incorporation of a direct-oriented *Alu* pair into the SSA repair of a deletion event can produce a chimeric *Alu* element (Sen et al., 2006). The examination of these direct and inverted FAP loci revealed that several chimeric *Alu* elements apparently resulted from these potential chimpanzee APE-related deletions. The number of chimeric *Alu* elements produced from these events is shown here along with the percentage as a total of potential APE-related deletions see heading below entitled, "Potential for ARMD Masking of APE Deletions" and Figure 2.11.

An unexpected finding in the third sub-category of Table 2.4 was the presence of a perfect inverted sequence (from 7 to 22 bp) separated by a spacer within the

human indel. This self-contained inverted sequence could potentially create inherent genomic instability within the indel sequence (Lewis et al. 1999). This inverted sequence could also be subject to genomic interactions similar to those reviewed in detail in the Discussion section (Figures 2.13 and 2.14, pages 37-40). Summing subcategories one and two, the potential fraction of APE deletions in these direct versus inverted FAP loci was 60.6 percent and 73.7 percent, respectively.

Potential for *Alu* mediated recombination deletions (ARMDs) masking APE deletions

One of the patterns which may be associated with an inverted APE-related deletion can be generated by a DNA double-strand break repair process known as single-strand annealing, SSA. SSA, which utilizes high-homology direct repeats as a repair template, can create a repair pattern that mimics an intra-chromosomal slippage and recombination event. Direct-oriented *Alu* elements in the vicinity of an inverted APE-related deletion could possibly be used as templates in the SSA repair process (Hedges et al. 2007). APE deletions which are repaired by SSA could produce a chimeric *Alu* element which would appear as *Alu* recombination mediated deletions, ARMDs (Sen et al. 2006; Han et al. 2007). It is interesting to note from Table 2.4 that 16.1 percent of the direct FAP loci that were identified as possible APE deletions were also associated with an ARMD pattern of repair. Similar inverted loci had five percent of deletions associated with the ARMD pattern of repair. It is not possible to determine whether these ARMDs were formed by inter-chromosomal slippage/recombination or SSA associated with an unknown deletion. The 3X disparity in the percentage of ARMDs between direct and inverted APEs appears to be attributable to the opportunity to form ARMDs between the members of direct FAPs that is absent in the inverted FAP loci. An examination of the 15

ARMDs associated with direct loci showed that ten were associated with the originally identified direct FAP and five were associated with *Alu* elements flanking the direct pair. Thus, the number of flanking ARMD repairs (ARMDs between one element in the target FAP and a second flanking *Alu* of identical orientation) was identical in both inverted and direct loci.

ARMDs occasionally skip over one or more *Alus* before recombining with another *Alu* (Han et al. 2007). This same *Alu* skipping feature could potentially be associated with an APE-deletion model followed by SSA repair. Unfortunately, SSA destroys the evidence of the original source of a deletion. Therefore, the possibility of SSA repair following an inverted APE-deletion cannot be eliminated as a possible cause of ARMDs.

An attempt was made to evaluate chimpanzee ARMDs as potential APE loci. This was accomplished by evaluating ARMD loci from previous work (Han et al. 2007). The first 100 chimpanzee ARMD loci were evaluated for their closest proximity to an inverted *Alu* element. A histogram of these distances is shown in Figure 2.11, pane A. This figure shows that 95 percent of these ARMD loci contain an inverted full length *Alu* element within 8,500 bp of one of the chimeric elements composing the ARMD. All of the 100 loci fell within 25,000 bp of an inverted element. The 25,000 base pair span of these ARMDs closely matched the range of the APSN5 FAP family. Figure 2.11, pane B is a linear regression of the I:D ratio across the ten spacer percentiles of this family. The total, CLIQUE adjusted, population for the APSN5 family is 551,764 FAPs. Each percentile contains slightly over 50,000 data points and provides a 95 percent confidence interval of $\pm$ 1.7 percent from unity (green dashed line in Figure 2.11, pane B). All of these ARMDs

fall outside the range of this confidence interval, indicating that APE deletions followed by SSA between direct *Alu* pairs may therefore be considered as one possible mechanism for the formation of ARMDs.

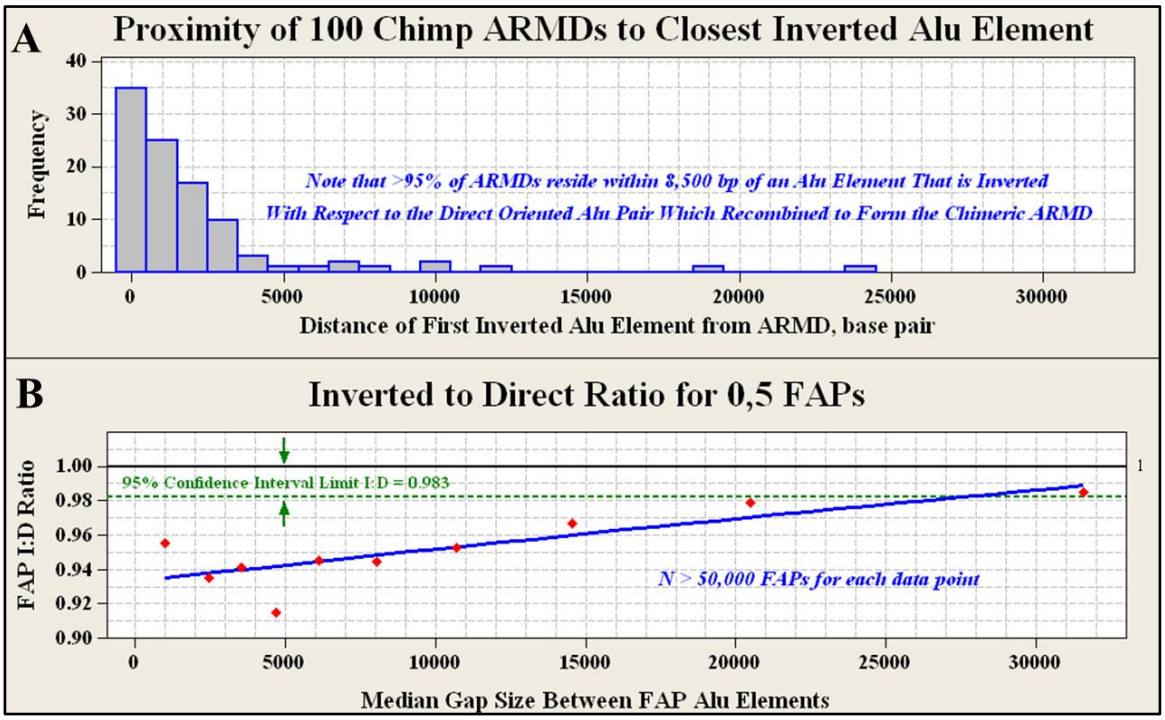


Figure 2.11 - ARMDs in proximity to inverted *Alu* pairs The cause of indels between chimpanzees and humans can be difficult to diagnose. This is especially true of *Alu* recombination mediated deletions, ARMDs. The existence of the chimeric *Alu* element product of an ARMD provides little information regarding the reasons behind its formation. This chimeric element could be generated by non-allelic homologous recombination, NAHR, or because of homologous repair associated with an unknown deletion. **(A)** The closest inverted *Alu* element for 100 random ARMDs is shown in histogram form. Note that 95% of these ARMDs are within 8,500 bp of an inverted *Alu* element. **(B)** 0,5 FAP I:D ratios were distributed most closely to the scatter seen in these ARMDs. Each data point in this chart represents over 50,000 *Alu* pairs. As can be seen in B) the 95% confidence interval for the I:D ratio about unity is ± 1.7 percent for this sample size. The I:D ratio of 0.95 at a spacer size of 8,500 bp reveals that these ARMDs could be the homologous repair product of a deletion caused by a doomsday junction.

Discussion

Non-random differences between direct and inverted FAPs exist for spacer sizes of zero to ≤350,000 bp. These differences may reflect orientation biases for

either *Alu* element insertions or deletions. The instability of *Alu* pairs with spacer sizes below 650 bp has been previously described (Stenger et al. 2001). Our research suggests that additional mechanisms may be operational.

APE mechanisms

Four separate mechanisms are theorized for generating APEs within the human genome (Figure 2.11). Although some overlap likely exists for the spacer size ranges wherein these four mechanisms operate, the first three mechanisms appear to be the first of these small-spacer APEs is identified by the observation that inverted *Alu* pairs form near-palindromic sequences that are vulnerable to hairpin formation and can induce double-strand breaks. This mechanism is termed 'hairpin APE' (Figure 2.12) and is thought to be operational between spacer sizes of 0 and approximately100 bp (Lobachev et al. 2000).

The second mechanism is termed 'TSD APE' and appears to be active for spacer lengths of less than 23 bp (Figure 2.5, page 16). This spacer length only slightly exceeds the 7 to 20 bp size range for TSDs (Cordaux and Batzer 2009). The nexus of high FAP density coupled with low I:D ratio is unique to human FAPs with these spacer lengths. The instability of inverted *Alu* pairs with spacer lengths of ≤100 bp has been demonstrated in a yeast model (Lobachev et al. 2000). This instability would be expected to reduce the FAP I:D ratio. However, the coincident phenomena of high FAP density and low FAP I:D ratio may also be associated with the TPRT insertion mechanism. *Alu* elements inherently provide an increased density of L1EN target sites. These additional target sites are generated by *Alu* TSDs and by the adenine-rich region within *Alu* elements (Levy et al. 2010). The additional L1EN target sites coupled with *Alu* insertion bias associated with the

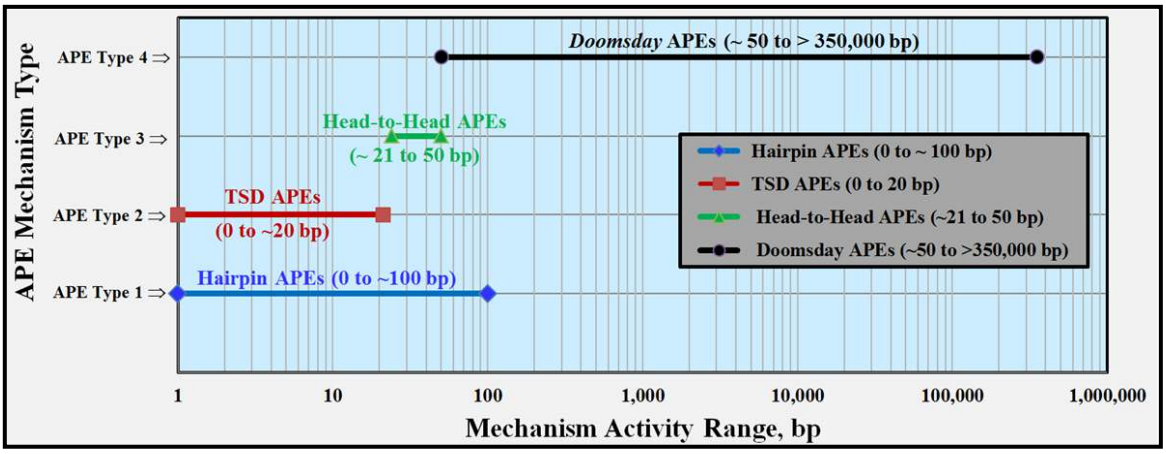


Figure 2.12 - Estimated ranges for four potential mechanisms for generating APEs This semi-log chart illustrates the activity of the one previously identified (Lobachev et al. 2000) and three new APE mechanisms. The APE Type 1 mechanism can also be termed 'hairpin APEs' and has been previously identified as related to *Alu-Alu* hairpin formation with subsequent deletion. The range of this mechanism has been demonstrated to extend up to 100 bp in a yeast model (Lobachev et al. 2000). The APE Type 2 mechanism can be described as 'TSDs APEs' and refers to a potential orientational insertion preference for *Alu* element insertions within the TSD of existing *Alu* elements. This mechanism would preferentially form direct-oriented FAPs. As with TSD APEs (Type 2), the Type 3 APE mechanism appears to reflect an insertional preference for the formation of head-to-head (inverted) FAPs. Type 3 APEs occur approximately within the range of 21 to 50 bp (Figure 2.5, page 16). The proposed mechanism for formation of Type 4 APEs is described in Figures 2.13 and 2.14 (pages 37-40) and is hypothesized to arise through a DNA conformation termed a 'doomsday junction'. APE: *Alu* pair exclusion; bp: base pair; FAP: full-length *Alu* pair.

RNA/DNA hybrid during the TPRT mechanism are consistent with the two super-imposed patterns observed in Figure 2.5, pane A (page 16). The instability of inverted *Alu* pairs almost certainly contributes to the low I:D ratios associated with closely spaced human FAPs. However, total attribution of this instability to the low I:D ratio observed for FAPs with spacer sizes of ≤20 bp may be an overestimate.

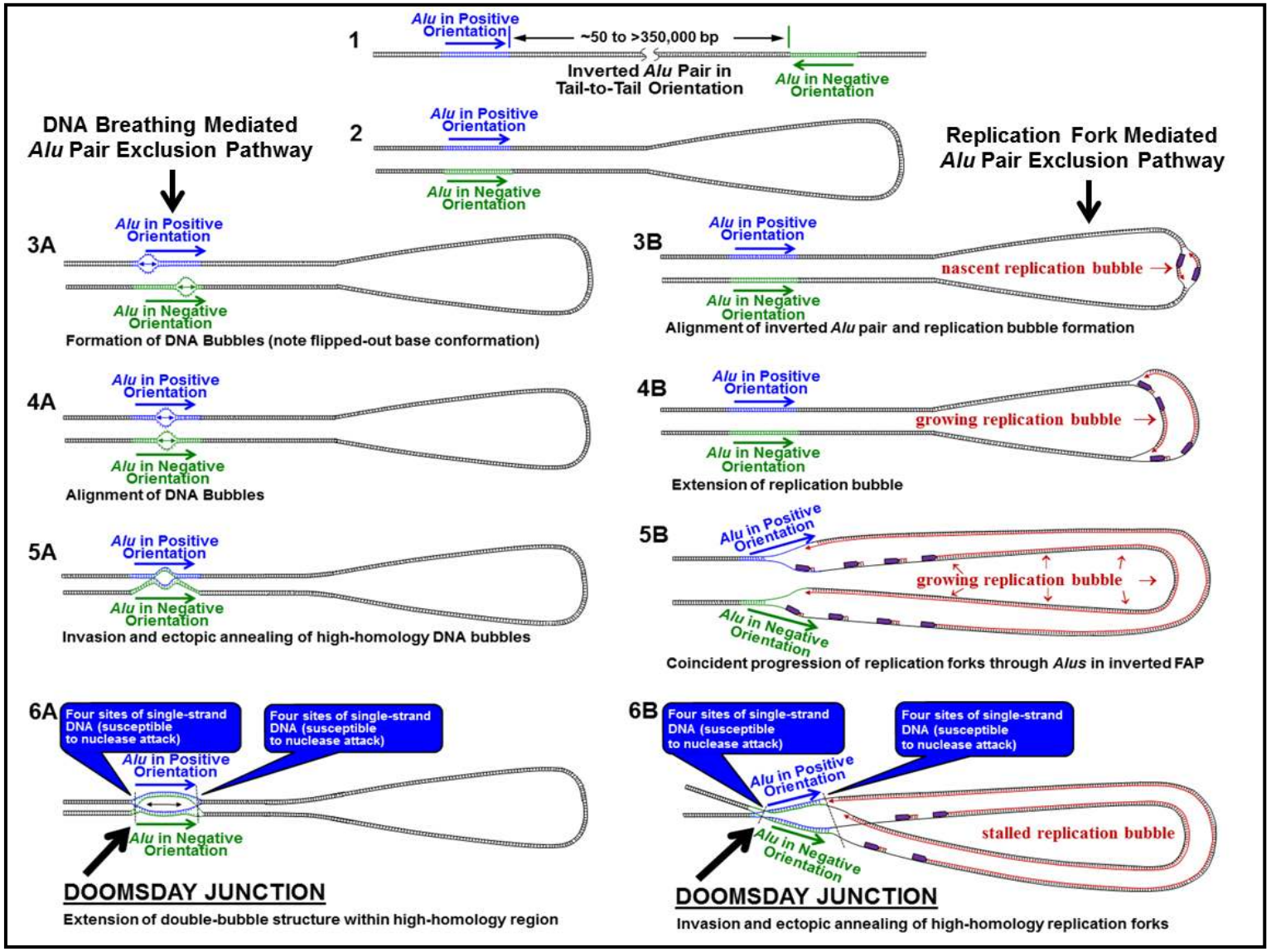
The third small-spacer APE mechanism is termed 'head-to-head APE' and involves the elevated frequency of head-to-head FAPs present between spacer sizes of 23 and 50 bp. This elevated frequency is more pronounced for spacer sizes between 24 and 36 bp and very pronounced for spacer sizes of 27 to 30 bp. Within

the spacer range of 24 to 36 bp, head-to-head (inverted) FAPs outnumber either type of direct-oriented FAPs (Figure 2.5, page 16). For spacer sizes of 27 to 30 bp, head-to-head FAPs actually outnumber the sum of both direct-oriented FAP pair types. If direct-oriented FAPs are relatively stable entities, this region of elevated head-to-head frequency may evidence an insertion-related phenomenon. A more detailed discussion of this possibility is provided below under the heading, "Possible epigenetics associated with head-to-head FAPs with spacer sizes of 24-36 bp." The fourth APE mechanism is very dissimilar from the first three small-spacer APE mechanisms in that it involves the loss of inverted FAPs separated by approximately 50 to ≤350,000 bp. The third APE mechanism overlaps this range up to a spacer size of 100 bp. Over 99 percent of all CLIQUE-corrected FAPs (not residing within the same CLIQUE) have spacer sizes greater than 100 bp. The higher energy state required for formation of single-stranded DNA makes hairpin loop formation a rare event between inverted *Alu* pairs separated by more than 100 bp (Lobachev et al. 2000; SantaLucia and Hicks 2004). Three possible pathways for interactions of distantly separated inverted FAPs are illustrated in Figures 2.13 and 2.14. Each of these pathways results in the ectopic annealing of single-stranded DNA associated with inverted FAPs. This annealing, which is hypothesized to result in a 'double-bubble' type structure, could potentially overcome the thermodynamic hurdle associated with single-stranded large-spacer hairpins. This structure is termed a 'doomsday junction' or DDJ (illustrated in Figure 2.13, Steps 6A and 6B and 2.14, Step 5).

Nuclease attack of DNA hairpins has been found to occur at the base, rather than the loop of DNA hairpins in yeast (Lengsfeld et al. 2007). If DDJs exist, and if single-strand nucleases are active in primates, the eight single-stranded sections of

DNA on the periphery of DDJs (Figure 2.13, steps 6A and 6B and Figure 2.14, step 5) could form attractive nuclease targets. Such nicking could help resolve the DDJ. However, this nicking could potentially result in various combinations of flanking deletions on either side of the two *Alu* elements forming the DDJ. The resultant tell-tale deletion patterns that we would predict from this mechanism are outlined in Figure 2.15 (page 42). The varied repair products from nuclease attack on these single-stranded structures could result in partial or total removal of one or both *Alu* elements. These proposed patterns are consistent with those observed by PCR of possible chimpanzee-specific APE deletions shown in Figure 2.10 (pages 25-26) and diagramed in Figure 2.15 (page 42). The pattern is also consistent with deletion patterns in 199 of 380 orthologous human-chimpanzee FAP loci (51%) where a

Figure 2.13 - Possible pathways for formation of G and S phase DDJs (Steps 1 and 2) This diagram illustrates the structure of an inverted FAP. When the DNA in Step 1 is bent 180°, the two *Alu* elements within the inverted FAP are aligned. Steps 3A-6A and 3B-6B illustrate two possible mechanisms for interactions between inverted *Alu* elements without the formation of a hairpin loop. Steps 3A-6A, Illustrate a DNA breathing (G phase) mediated APE deletion. **(Step 3A)** DNA breathing bubbles are typically < 20 bp are characterized by flipping of the unpaired nucleotide bases away from the center line of the double-helix (Fogedby and Metzler 2007) . A bubble in this conformation could be susceptible to interaction with a bubble of similar sequence. **(Step 4A)** Simultaneous bubbles may arise in identical sections of aligned *Alu* elements. **(Step 5A)** Simultaneous homologous bubble alignment could initiate bubble-bubble interaction with the potential for forming a 'double-bubble' conformation. **(Step 6A)** The ectopic formation of the double-bubble conformation within two aligned breathing bubbles could potentially extend to the entire length of the two aligned *Alu* elements. The high GC content of *Alu* elements would likely increase the stability of the hypothesized doomsday junction. Doomsday junctions, DDJs, likely possess four single-stranded sections of single-stranded DNA at each end which could be susceptible to single-strand nuclease attack. Steps 3B-6B describe a replication fork (S phase) mediated APE deletion. (**Steps 3B-5B)** The initiation and growth of a replication bubble and coincident progression of the DNA replication bubble through an inverted FAP. **(Step 6B)** This diagram describes the invasion and ectopic annealing of high-homology replication forks. (Figure 2.13 continues on the following page.)

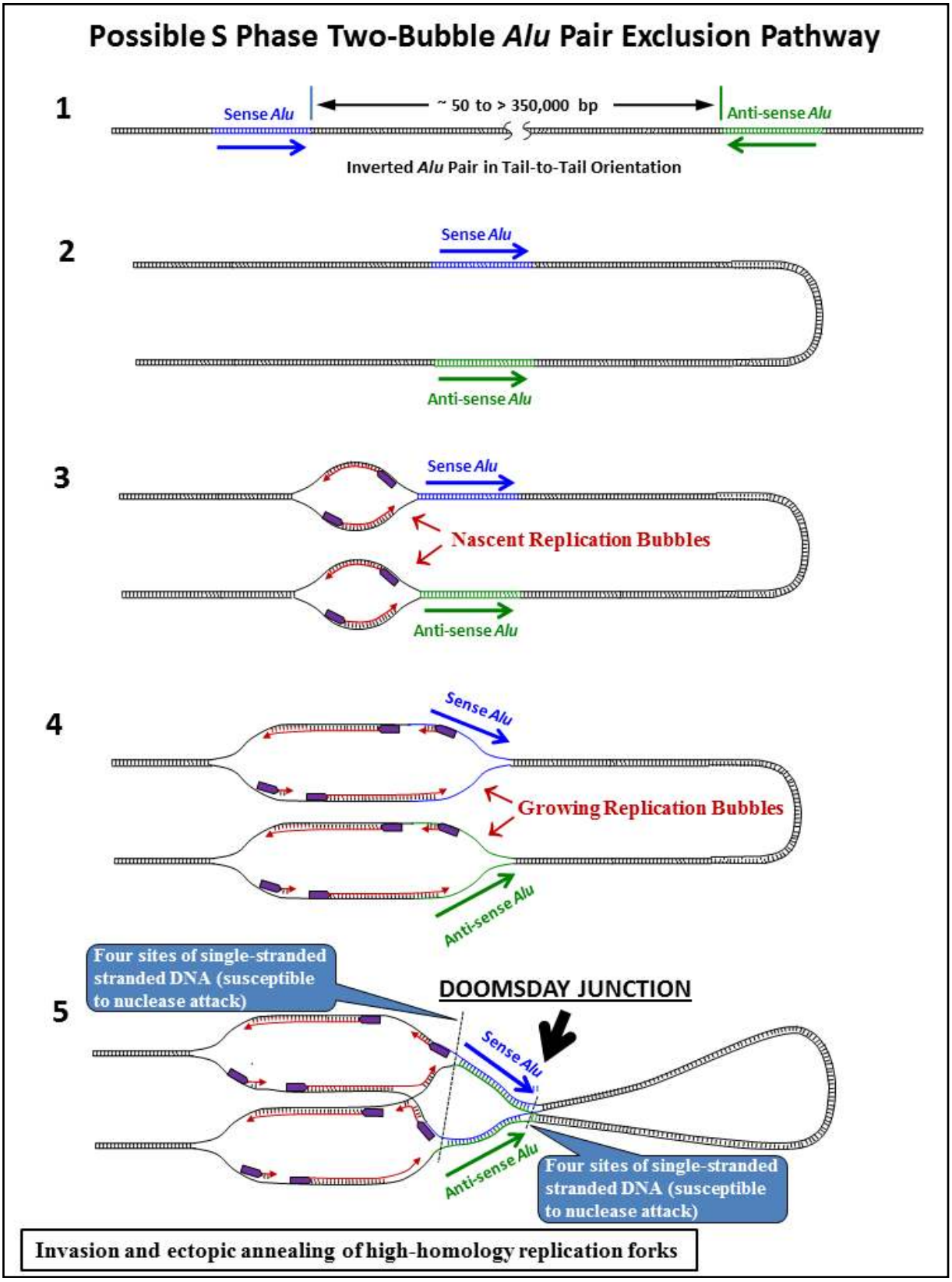

1
Alu in Positive Orientation
~50 to >350,000 bp
Inverted Alu Pair in Tail-to-Tail Orientation
Alu in Negative Orientation
DNA Breathing Mediated Alu Pair Exclusion Pathway
2
Alu in Positive Orientation
Alu in Negative Orientation
Replication Fork Mediated Alu Pair Exclusion Pathway
3A
Alu in Positive Orientation
Alu in Negative Orientation
Formation of DNA Bubbles (note flipped-out base conformation)
3B
Alu in Positive Orientation
nascent replication bubble
Alu in Negative Orientation
Alignment of inverted Alu pair and replication bubble formation
4A
Alu in Positive Orientation
Alu in Negative Orientation
Alignment of DNA Bubbles
4B
Alu in Positive Orientation
growing replication bubble
Alu in Negative Orientation
Extension of replication bubble
5A
Alu in Positive Orientation
Alu in Negative Orientation
Invasion and ectopic annealing of high-homology DNA bubbles
5B
Alu in Positive Orientation
growing replication bubble
Alu in Negative Orientation
Coincident progression of replication forks through Alus in inverted FAP
6A
Four sites of single-strand DNA (susceptible to nuclease attack)
Four sites of single-strand DNA (susceptible to nuclease attack)
Alu in Positive Orientation
Alu in Negative Orientation
DOOMSDAY JUNCTION
Extension of double-bubble structure within high-homology region
6B
Four sites of single-strand DNA (susceptible to nuclease attack)
Four sites of single-strand DNA (susceptible to nuclease attack)
Alu in Positive Orientation
stalled replication bubble
Alu in Negative Orientation
DOOMSDAY JUNCTION
Invasion and ectopic annealing of high-homology replication forks

potential chimpanzee deletion had occurred (Table 2.4, page 30). This deletion pattern increases to 75 percent when the 114 human-specific retrotransposon insertions are removed from the data set.

G-phase doomsday APEs

Figures 2.13 and 2.14 outline separate mechanisms by which DDJs could form during the G and S phases of the cell cycle. We propose that G-phase DDJs result from the ectopic invasion and annealing of high-homology bubbles associated with DNA breathing (Figure 2.13, steps 1-6A). Nucleosomes and other chromatin structures mitigate DNA breathing and thus may reduce the potential for G-phase DDJ formation. Therefore, in addition to their multifarious roles in signaling and protein binding, nucleosomes may also serve to minimize the interaction between high-homology DNA strands. The instability of closely spaced inverted *Alu* elements shown here and noted by previous researchers may be evidence that nucleosomes are either absent from hairpin prone DNA sequences or provide insufficient interference for hairpin formation (Lobachev et al. 2000; Stenger et al. 2001; Lee et al. 2008). The postulated G phase DDJ phenomenon may enjoy this same dominance over nucleosome interference.

Figure 2.14 - Possible S phase dual replication bubble DDJ formation pathway Single-stranded DNA is present at the DNA replication fork during S-phase of the cell cycle. Single-stranded DNA is inherently vulnerable to forming non-canonical binding structures such as hairpins and cruciform structures and thus must be stabilized by single strand binding proteins (Broderick et al. 2010). Figure 2.13, Steps 1-6B describe the creation of a hypothetical DNA configuration termed a "doomsday junction" or DDJ. The coincident passage and proximity of two separate replication forks through an inverted repeat may set the stage for ectopic invasion and annealing of the single-strand DNA associated with these replication forks. The DDJ pathway described above is similar in all aspects to that outlined in Figure 2.13 except that the DDJ formation, above, is generated from replication forks associated with different DNA replication bubbles. (Figure 2.14 continues on the following page.)


Possible S Phase Two-Bubble Alu Pair Exclusion Pathway
1
Sense Alu
~ 50 to > 350,000 bp
Anti-sense Alu
Inverted Alu Pair in Tail-to-Tail Orientation
2
Sense Alu
Anti-sense Alu
3
Sense Alu
Nascent Replication Bubbles
Anti-sense Alu
4
Sense Alu
Growing Replication Bubbles
Anti-sense Alu
5
Four sites of single-stranded stranded DNA (susceptible to nuclease attack)
DOOMSDAY JUNCTION
Sense Alu
Anti-sense Alu
Four sites of single-stranded stranded DNA (susceptible to nuclease attack)
Invasion and ectopic annealing of high-homology replication forks

If simultaneous DNA breathing bubbles were to arise between aligned homologous sequences, the flipped-out conformation of complimentary bases on both strands could provide additional potential for intra-strand interaction (Figure 2.13, step 4A) (Fogedby and Metzler 2007). This altered genomic structure formed by the hypothetical interaction between two homologous DNA bubbles would effectively create the double-bubble conformation associated with DDJs. The initial, smaller double-bubble structure (Figure 2.13, step 5A) could easily expand to form a larger double-bubble which could extend to almost the entire length of the two aligned *Alu* elements (Figure 2.13, step 6A). The high GC content (>60%) of *Alu* elements composing the large bubble conformation would likely enhance the stability of the hypothesized DDJ.

S-phase doomsday APEs

S phase DDJs are proposed to result from invasion and subsequent annealing of high-homology DNA replication forks. This is illustrated in Figure 2.13 (pages 37-38) and in Figure 2.14. Coincident passage of replication forks through inverted FAPs could provide an environment susceptible to formation of an S-phase DDJ. Unlike the chromatin interference present in G phase, replicating S-phase DNA is forced to lift its chromatin kimono and becomes much more vulnerable to ectopic DNA interaction. While single-strand binding proteins stabilize single-stranded portions of the replication fork, they are eventually displaced with a newly replicated strand of single-stranded DNA. This second strand could conceivably be supplied from an invading second replication fork. Notably, upon formation of an S-phase DDJ, the DNA replication apparatus would be completely assembled and could potentially proceed, albeit in an ectopic fashion, and conceivably generate

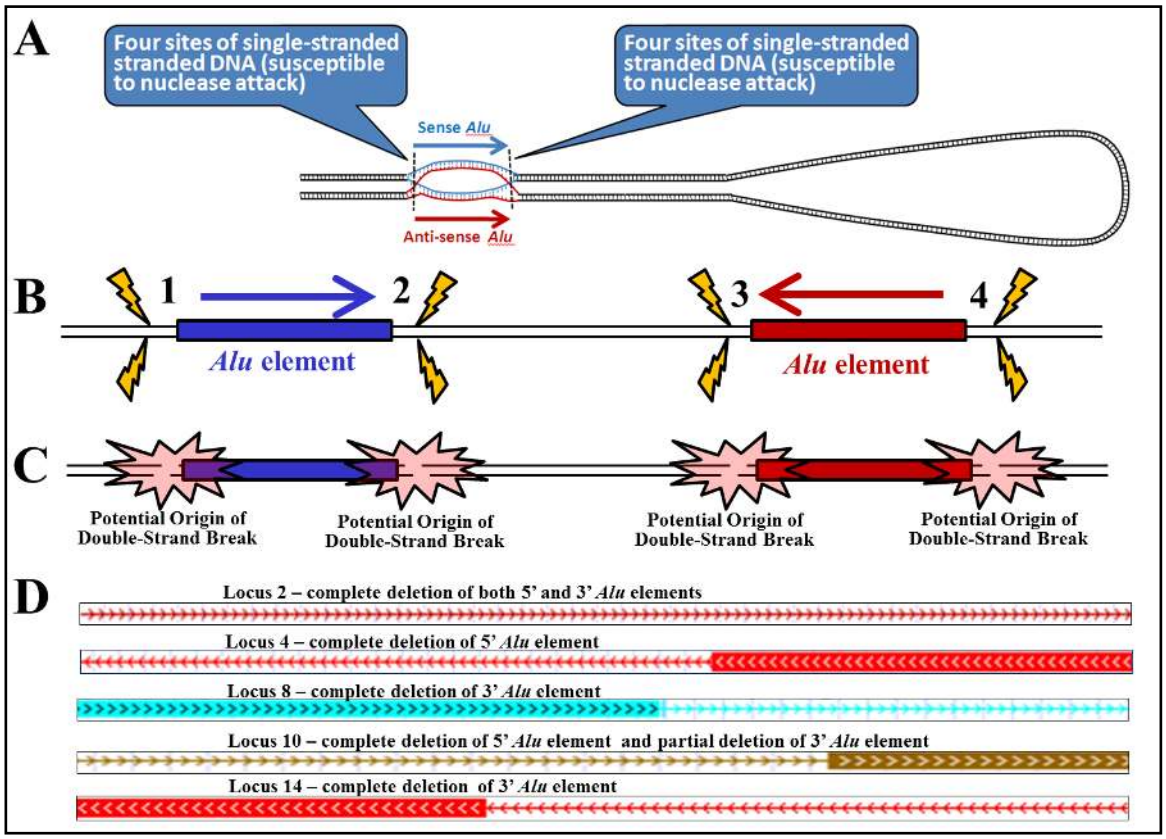


Figure 2.15 - Possible deletion patterns resulting from resolution of DDJs
(A) This doomsday junction, DDJ, is taken from Figure 2.13, step 6A (pages 37-38). Note the eight regions of single-stranded DNA associated with the ends of the DDJ. These regions may be susceptible to single-strand DNA nuclease attack. **(B)** A linear model of an unraveled DDJ illustrating the eight regions of potential single-strand nuclease attack. **(C)** The regions of the DDJ which are most susceptible to a double-strand break are adjacent to both 5' and 3' ends of each *Alu* element (shown as light red starbursts). Using this model, deletion of portions of either *Alu* element or the spacer region would only occur as a result of nuclease attack proceeding from the origin of the double strand break. **(D)** Deletion patterns from PCR chimpanzee loci shown in Figure 2.10 (pages 25-26).

segmental duplications. In addition, the double-bubble binding of near-homologous *Alu* elements within a DDJ could invite the activity of cellular mismatch repair mechanisms. Such mismatch activity could help explain elevated mutation rates which have previously been observed close to deletions (Tian et al. 2008).

Finally, the DDJ mechanisms outlined in Figure 2.13 and Figure 2.14 (pages 37-40) do not preclude interactions between direct-oriented FAPs. However, the

distinctive 'V' shape of replication forks may provide steric hindrance to interactions with direct pairs and thus preferably favor interactions between inverted pairs. Regardless of the mechanism(s) associated with the human FAP I:D ratio imbalance, this metric is not an absolute measure of change in the number of either direct or inverted FAPs, but of the relative change between the two types.

Possible epigenetics associated with head-to-head FAPs with spacer size of 24-36 bp

Head-to-head FAP frequencies are elevated within the spacer size range of 24-50 bp (Figure 2.5, pane B, page 16). More notable is that this FAP frequency exceeds each type of direct oriented FAPs between spacer sizes of 25-35 bp. It is intriguing that *Alu* insertions within *Alu* TSDs predominantly form direct FAPs and yet appear to form inverted FAPs when spacer sizes are between of 24 and 36 bp (Cordaux and Batzer 2009). Assuming that direct FAPs are reasonably stable entities, the latter may be evidence of a previously-uncharacterized inverted *Alu* insertion mechanism.

One explanation for this pattern is that nucleosomes may be attracted to head-to-head FAPs with spacer sizes of 24-36 bp. However, this theory does not explain why head-to-head FAP frequencies within this spacer range exceed the number of either type of direct-oriented FAPs. The fact that head-to-head FAPs within this spacer size range actually exceed either type of direct-oriented FAP may indicate that an insertional mechanism is driving this phenomenon. A second explanation for this pattern of elevated head-to-head FAPs is that L1EN may somehow associate with the 5' end of *Alu* elements. In addition to this association, the mechanism would also require L1EN to cleave its target sequence on the sense strand, approximately 24-36 bp from the 5' end of an existing *Alu* element. This

orientational nicking, coupled with subsequent formation of the TPRT PolyA/PolyT, RNA/DNA hybrid would drive orientation of the new FAP toward the head-to-head orientation.

The GC content of the human genome has been estimated to be 41 percent (Lander et al. 2001). With this GC frequency, the probability of the 5'-TTTTAA-3' L1EN target sequence randomly centering at any locus is one chance in 1,517. With the 806,880 full-length *Alu* elements in the human genome, this target site should randomly occur 6,914 times within the 24-36 bp spacer span for high head-to-head FAPs. The actual number of human head-to-head FAPs possessing spacer sizes within this range is 3,464. This actual number is 50.1 percent of the theoretical 6,914 L1EN target sites that are predicted to be centered randomly within this same 24-36 bp range. The highest incidence of head-to-head FAPs is 74 percent of the theoretical estimate which occurs at a spacer size of 28 bp. Some flexing of DNA between the L1EN anchoring site and cut site could possibly explain the high incidence of head-to-head FAPs spanning across the 13 nucleotides within the 24-36 bp spacer range.

The genetic distance of a 28 bp spacer size is equivalent to approximately three turns of DNA or about 100 Å (in non-bent conformation). The physical size of L1EN is approximately 25 bp, or 80 Å (Weichenrieder et al. 2004). Possibilities for an L1EN association with the 5' end of an *Alu* sequence include 1) direct L1EN binding with DNA flexing, 2) indirect L1EN association through a scaffolding protein, or possibly 3) direct L1EN binding plus dimerization because of the proximity of the two *Alu* elements in the head-to-head FAP orientation. The sustained presence of

L1EN and any associated proteins could also inhibit inverted *Alu* pair instability previously noted by other researchers (Stenger et al. 2001).

Conclusions

Direct and inverted FAPs are distributed non-randomly in the human genome. This non-random pattern exists for APSNs ≤107 bp and for spacer sizes up to 350,000 bp. A total of 59,357,435 FAPs (CLIQUE corrected) reside within this window and direct FAPs outnumber inverted FAPs by 629,027 (over two percent). Random variation only reduces this imbalance to 613,924 ($P<0.05$). Outside of CLIQUEs, no known orientation insertion preferences exist for *Alu* elements. We believe that APE-related deletions may be responsible for a substantial proportion of the imbalance of over 600,000 between inverted and direct human FAPs. Future investigations of the APE phenomenon should better illuminate the mechanisms involved and characterize its extent in primate genomes.

Methods

Data acquisition and management

Data used in the research was obtained from the RepeatMasker (Karolchik et al. 2004) output for the hg18, 2006 Human Genome assembly. This data was downloaded from the UCSC genome BLAT Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables) (Smit A. 1996-2012) and imported into Excel 2010 (Microsoft Corporation; Redmond, Washington). Orthologous chimpanzee, orangutan and rhesus macaque loci were obtained using the panTro2, ponAbe2 and rheMac2 genomes assemblies, respectively. Statistics were calculated using Minitab 15 (Minitab Inc.; State College, Pennsylvania).

Histogram of human *Alu* size distribution

The RepeatMasker scan of the hg18 human genome assembly identifies potential *Alu* fragments as small as 12 bp. Using a haploid genome size of 3.1×10^9 bp, a total of 185 instances of a given 12 bp should randomly occur in human DNA. However, most *Alu* elements have sequence identities between 65 and 85 percent (Stenger et al. 2001). Using the lower sequence identity (65%) increases the number of random instances of a 12 bp target sequence occurring in the human genome from 185 to 32,485 (Figure 2.13, pages 37-38). The target sequence must increase in length to 26bp before statistical significance ($P<0.05$) occurs. This sequence size increases to 29 bp for 60% identity. For this study, only *Alu* sequences of ≥30 bp are used. For perspective, a 30 bp *Alu* fragment length is roughly 10 percent of the length of a full-length *Alu* element. Finally, it should be noted that the 12 bp sequences become significant ($P<0.05$) when a segment of DNA shorter than 4,770 bp is being evaluated.

Sequences of less than 30 bp in length cannot be reliably determined to be actual *Alu* elements and are therefore excluded from this truncated percentage. A lower size limit of 275 bp is set to avoid I:D ratio directional bias caused by fragmented elements that can be generated by *Alu* insertions into a preexisting *Alu* element(Levy et al. 2010). The upper *Alu* element size limit of 325 bp is set to avoid the potential for confounding results by inclusion of the smaller population of larger elements.

Terminology for non-adjacent *Alu* pairs

The central *Alu* in this naming convention is always designated with the number ‘0’. The second member of the pair is designated by its sequential

separation from the central *Alu*. If this second member of a pair is located 5′ of the central *Alu* element, it is designated by a negative number and by a positive number if it is located 3′ of the central *Alu* element. The value of the sequential separation of a given *Alu* element from the central *Alu* is defined as its APSN. For adjacent elements, these FAP pairs are described as -1,0 and 0,1. Similarly, FAPs separated by 25 intervening *Alu* elements are described as -26,0 and 0,26 pairs, respectively.

Determination of 95% confidence interval for FAP I:D ratios

FAP sample sizes used in this study range from 555,354 to 567,242 (APSNs 0,1 to 0,107). These sample sizes are retrieved by counting functions within the *Alu* element Excel spreadsheet. Following removal of FAPs residing within the same CLIQUE (CLIQUE-adjusted), these data set sizes are reduced to between 460,588 and 557,364. CLIQUE-adjusted samples sizes below 550,000 only exist for APSNs ≤4. For a FAP sample size of 550,000, the number of direct and inverted FAPs should range between 274,272 and 275,728 ($P<0.05$). Any imbalance in direct or inverted FAPs is offset by an equal and opposite imbalance in the other FAP type. Therefore, the I:D ratio for a sample size of 550,000 is expected to range from 0.9947 to 1.0053 ($P\leq0.05$). This range increases to between 0.9942 and 1.0058 for the lowest sized (0,1) FAP family of 460,588.

Determination of maximum APSN within the FAP I:D ratio imbalance window

Determination of the limits of the FAP I:D ratio imbalance boundary beyond an APSN of approximately 85 (Figure 2.9, pane A, page 23) was accomplished by increasing the precision of the method. This added precision was achieved by increasing the FAP sample size. This larger sample size was acquired by calculating a 10-point moving average of the FAP I:D ratio across consecutive

APSNs beyond the ± 85 range. This approach increased the FAP sample size from a value of approximately 550,000 to 5.55 million and reduced the 95 percent confidence interval for randomness from 1 ± 0.0053 to 1 ± 0.0017. The highest ten consecutive APSNs which had an I:D average outside of these new confidence limits was the APSN range 103 to 112. The midpoint of this range is the APSN value of 107.

Determination of maximum spacer size within the FAP I:D ratio imbalance window

Approximately 90 percent of the adjacent FAPs have spacer sizes below 6,400 bp. In addition, the I:D ratio for the upper 10 percent of this family is 0.9838 which is lower than the statistically significant I:D ratio of 1 ± 0.995. Consequently, determination of the boundary of the FAP I:D imbalance bubble (Figure 2.9, pane B, page 23) requires examination of larger APSN families. The number of FAPs within a given size range can be summed across various APSNs. This summation was used to determine the spacer size boundaries for the FAP I:D imbalance window.

APSN families smaller than 0,25 contain very few members with spacer sizes between 300,000 and 400,000 bp. However, 3,541,238 FAPs reside within this spacer range for APSN's of 0,25 to 0,107. This spacer size range was divided into two separate ranges of 300,000 to 350,000 and 351,000 to 400,000. The number of FAPs within these spacer ranges was determined as 1,974,605 (I:D = 0.9951) and 1,566,633 (I:D = 0.9956), respectively. The expected ranges for FAP I:D ratios for these two spacer size ranges are 0.9972 to 1.0028 and 0.9969 to 1.0031, respectively ($P<0.05$). These two I:D ratios are outside of these ranges and thus show that the FAP I:D imbalance window extends beyond ± 350,000 bp.

Selection of loci for validation of APE deletions in the chimpanzee genome

The methodology employed for selection of potential APE deletion loci utilized five criteria. These criteria were pair orientation, APSN, *Alu* element size, spacer size and *Alu*-free flanking sequence 5′ and 3′ of the pair being evaluated. Only inverted *Alu* pairs were chosen as potential experimental loci as they have been previously demonstrated to be unstable (Lobachev et al. 2000). The second criterion, APSN, was limited to 0,1 (adjacent) FAPs as any intervening *Alu* element necessarily forms a second, more closely spaced inverted pair with one of the two elements of that FAP. Therefore, any deletion associated with this locus could reasonably be attributed to interactions associated with the intervening element. For this reason, only the pool of adjacent human FAPs (APSN = 0,1) was used to identify candidate APE deletion loci.

The third criterion, *Alu* element size, was limited to the 275 to 325 bp constraints set for FAPs. The fourth criterion, spacer length separating the two FAP elements, was limited to those elements separated by 651 to 1,500 bp. The lower spacer size limit was set by the upper limit of previous work (Stenger et al. 2001) and upper limit was set to provide an acceptable number of candidate loci. The fifth criterion, 5′ and 3′ *Alu*-free flanking sequence around a 0,1 FAP, was set to a minimum of 1,000 bp. This constraint was necessary to avoid attribution of an APE deletion to nearby elements. These criteria created locus sizes between 3,201 and 4,150 bp.

A total of 13,664 human loci were identified which satisfied these five criteria. This sample size was approximately 0.03 percent of the approximately 50 million CLIQUE-adjusted FAPs within the I:D imbalance window shown in Figure 2.9, pane

B (page 23). These loci were then compared to the chimpanzee panTro2 genome assembly using the LiftOver feature of the USCS Genome Browser (Kent 2002; Gibbs et al. 2007). This screening identified 715 (or slightly over five percent) of the chimpanzee loci that were over 350 bp smaller than their human ortholog. The less than 350 bp lower limit was set to reduce the number of false-positive loci (in other words, human specific *Alu* insertions can be flagged as potential sites for chimpanzee APE-related deletions). The 715 loci were individually inspected using the UCSC genome browser for the human, chimpanzee, orangutan and rhesus macaque genomes (Kent 2002; 2005; Gibbs et al. 2007; Locke et al. 2011). These inspections reduced the number of PCR candidate loci to 58. Four criteria accounted for approximately 90 percent of this reduction. These four criteria, in order of magnitude, were as follows.

1. The presence of N's in the chimpanzee genome assembly (382 loci)
2. The insertion of a human specific transposable element as the cause of the smaller chimpanzee loci (141 loci)
3. A deletion present, but so large that it encompassed an adjacent *Alu* element making the deletion non-diagnostic (56 loci)
4. Complementary deletions were also present in orangutan or rhesus (38 loci).

The remaining 58 loci were selected as potential candidates for further examination with PCR.

Estimation of APE deletions in chimpanzee genome by observation

Although only 58 of the 715 loci were accepted for further examination by PCR, an additional 94 of these loci showed considerable evidence of being potential APE deletions (criterions 3 and 4, above). Adding these 94 loci to the 58 PCR

candidate loci increases the number of APE-related deletion loci to 152. It was also assumed that the 382 loci which contained N's in the chimpanzee (rejection criterion 1) were indeterminate and could neither be accepted nor rejected regarding detection of APE-related deletions. Separating these 382 loci (which contained N's in the chimpanzee deletion) from the original set of 715 loci reduces the total number of individually inspected loci to 333. It is estimated that 152 likely APE-related deletion loci exist out of these 333 loci (45.6%). Of the 14 loci evaluated by PCR, 10 were informative (71.4%). The PCR results from the remaining four loci were uninformative and no false positive instances of chimpanzee-specific deletions were observed. Combining these two probabilities provides an estimate that 32.6 percent (108) of the 333 loci were likely APE-related deletions. Therefore, within these 13,664 inverted FAP loci, a total of 108 APE-type deletions are estimated to have occurred in chimpanzee (by observation) since the human-chimpanzee divergence.

Primer design for PCR

Candidate PCR amplicon sequences were obtained with the BLAT feature of the UCSC genome browser. These sequences were aligned using the BioLign software (developed by Tom Hall and available from the Buckler Lab website: http://www2.maizegenetics.net/bioinformatics). These alignments were manually inspected for common identity between the four primate species. Forward and reverse oligonucleotide primers were selected from regions of common alignment. Primer sequences are shown in supplementary information in, Table 2.3 (page 28).

PCR amplification

All PCR amplifications were conducted in 27.5 μL reactions using 25 ng DNA template, 0.2 μM oligonucleotide primer, 1.25 units TaKaRa LA Taq™, 0.4mM

dNTPs, and 1X TaKaRa LA Taq™ buffer containing 2.3 uM $MgCl_2$. A list of primers is provided in Table 2.4 (page 30). The primate panel contained templates from *Homo sapiens* (HeLa; cell line ATTCC CCL-2); *Pan troglodytes* (common chimpanzee "Clint", cell line Coriell Cell repositories NS06006), *Gorilla gorilla* (Western lowlands gorilla; cell line Coriell Cell Repositories NG05251); *Pongo abelii* (Sumatran orangutan; cell line Coriell Cell Repositories NG06209); and *Macaca mulatta* (rhesus macaque; cell line Coriell cell Repositories NG07110). PCRs were run for 80 sec for initial denaturation at 94°C. Denaturing, annealing and extension times and temperatures were 20 sec at 94°C, 20 sec at optimum temperatures (Table 2.4, page 30) and 8 min 30 sec at 68°C, respectively, for 32 cycles. The 32 cycles were followed by a final extension time of 10 min at 68°C. Following amplification, all PCR products were electrophoresed on 1.5% agarose gels stained with ethidium bromide at a concentration of 1 μl per 50 mL of gel solution. Gels were run for 45 to 55 min at 175 volts. Finally, fragments were visualized using UV fluorescence.

Comparison of APE deletions in chimpanzee genome by computation and observation

Using the original criteria for isolating potential experimental loci, 13,664 inverted FAP and 14,680 direct FAPs were identified. The I:D ratio for these FAPs is 0.931 and the difference between these inverted and direct FAPs is 1,016, which we believe correspond to APE-associated deletion events. All *Alu* element insertions have occurred over the 65 million years of primate evolution. It is estimated that the most recent common ancestor of humans and chimpanzees lived approximately six million years ago (Xing et al. 2009). Consequently, approximately 12 million years of genome evolution are estimated to have occurred between extant humans and

chimpanzees. For this 12-million year period of evolution to be incorporated into calculated APE rate estimates, both orthologous chimpanzee-specific and human-specific APE-related deletions must be estimated. Only chimpanzee-specific APE-related deletions are measured in this study. Therefore, only half of the 12-million years of evolution are used (six million years) in this estimate. Therefore, a conservative estimate of 94 chimpanzee-specific APE deletions would be expected over the 6 million years since the human-chimpanzee divergence (1016 × 6 ÷ 65 = 94). This number is concordant with the 108 APE deletions previously estimated to have occurred by observational methods (discussed under heading, 'Estimation of APE Deletions in Chimpanzee Genome by Observation', page 50).

Moving average distributions of actual and random *Alu* clustering

The RepeatMasker scan of the hg18 human chromosome assembly recovers 102,592 *Alu* elements in chromosome 1. Since orientational clustering bias has been shown to occur within CLIQUEs, only the 5′ *Alu* element in each CLIQUE was included in this evaluation. Chromosome 1 contains 50,262 *Alu* elements that do not reside within a CLIQUE. Human chromosome 1 contains 34,916 CLIQUEs, of which 26,277 contain at least one *Alu* element. Consequently, only 76,539 (50,262 + 26,277) *Alu* elements were used in this clustering evaluation. A value of +1 was assigned to each *Alu* on the positive strand and a value of -1 was assigned to each *Alu* on the negative strand. Moving average data was calculated for the 50, 100, 200, 500 and 1,000 sequential directional data points in Excel.

Five sets of 76,539 random +1 and -1 data (equivalent to the revised data set of *Alu* elements in human chromosome 1, above) were generated using Minitab15. This data was transferred to Excel and moving averages were calculated for each

set of random data for 50, 100, 200, 500, 1,000, 2,000, 5,000 and 10,000 sequential directional data points. These 48 sets of moving average data (one set of actual data and five sets of random data for eight separate moving averages) were then transferred back to Minitab. Individual mean and standard deviations for each set of random distributions were determined using the Mintab15 histogram 'with fit and groups' algorithm. The five individual means and standard deviations were then averaged for each set of random moving averages. The random data curves were generated using these average mean and standard deviations (Figure 2.4, pages 13-14).

References

2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**(7055): 69-87.

Batzer MA, Deininger PL. 2002. *Alu* repeats and human genomic diversity. *Nature reviews Genetics* **3**(5): 370-379.

Batzer MA, Kilroy GE, Richard PE, Shaikh TH, Desselle TD, Hoppens CL, Deininger PL. 1990. Structure and variability of recently inserted *Alu* family members. *Nucleic Acids Res* **18**(23): 6793-6798.

Belancio VP, Hedges DJ, Deininger P. 2008. Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. *Genome Res* **18**(3): 343-358.

Broderick S, Rehmet K, Concannon C, Nasheuer HP. 2010. Eukaryotic single-stranded DNA binding proteins: central factors in genome stability. *Subcell Biochem* **50**: 143-163.

Collins J. 1981. Instability of palindromic DNA in Escherichia coli. *Cold Spring Harb Symp Quant Biol* **45 Pt 1**: 409-416.

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P et al. 2010. Origins and functional impact of copy number variation in the human genome. *Nature* **464**(7289): 704-712.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nature reviews Genetics* **10**(10): 691-703.

Deininger PL, Batzer MA. 1999. *Alu* repeats and human disease. *Mol Genet Metab* **67**(3): 183-193.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467**(7319): 1061-1073.

Fogedby HC, Metzler R. 2007. Dynamics of DNA breathing: weak noise analysis, finite time singularity, and mapping onto the quantum Coulomb problem. *Physical review E, Statistical, nonlinear, and soft matter physics* **76**(6 Pt 1): 061915.

Franke G, Bausch B, Hoffmann MM, Cybulla M, Wilhelm C, Kohlhase J, Scherer G, Neumann HP. 2009. *Alu-Alu* recombination underlies the vast majority of large VHL germline deletions: Molecular characterization and genotype-phenotype correlations in VHL patients. *Human mutation* **30**(5): 776-786.

Gibbs RA Rogers J Katze MG Bumgarner R Weinstock GM Mardis ER Remington KA Strausberg RL Venter JC Wilson RK et al. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316**(5822): 222-234.

Goodier JL, Kazazian HH, Jr. 2008. Retrotransposons revisited: the restraint and rehabilitation of parasites. *Cell* **135**(1): 23-35.

Grimaldi G, Singer MF. 1982. A monkey *Alu* sequence is flanked by 13-base pair direct repeats by an interrupted alpha-satellite DNA sequence. *Proc Natl Acad Sci U S A* **79**(5): 1497-1500.

Han K, Lee J, Meyer TJ, Wang J, Sen SK, Srikanta D, Liang P, Batzer MA. 2007. *Alu* recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet* **3**(10): 1939-1949.

Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation research* **616**(1-2): 46-59.

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ. 2004. The UCSC Table Browser data retrieval tool. *Nucleic acids research* **32**(Database issue): D493-496.

Kazazian HH, Jr. 1998. Mobile elements and disease. *Curr Opin Genet Dev* **8**(3): 343-350.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**(4): 656-664.

Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Seminars in cancer biology* **20**(4): 211-221.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lengsfeld BM, Rattray AJ, Bhaskara V, Ghirlando R, Paull TT. 2007. Sae2 is an endonuclease that processes hairpin DNA cooperatively with the Mre11/Rad50/Xrs2 complex. *Mol Cell* **28**(4): 638-651.

Levy A, Schwartz S, Ast G. 2010. Large-scale discovery of insertion hotspots and preferential integration sites of human transposed elements. *Nucleic Acids Res* **38**(5): 1515-1530.

Lewis S, Akgun E, Jasin M. 1999. Palindromic DNA and genome stability. Further studies. *Ann N Y Acad Sci* **870**: 45-57.

Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA. 2000. Inverted Alu repeats unstable in yeast are excluded from the human genome. *The EMBO journal* **19**(14): 3822-3830.

Locke DP Hillier LW Warren WC Worley KC Nazareth LV Muzny DM Yang SP Wang Z Chinwalla AT Minx P et al. 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* **469**(7331): 529-533.

Luan DD, Eickbush TH. 1995. RNA template requirements for target DNA-primed reverse transcription by the R2 retrotransposable element. *Mol Cell Biol* **15**(7): 3882-3891.

Luan DD, Korman MH, Jakubczak JL, Eickbush TH. 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell* **72**(4): 595-605.

Mathias SL, Scott AF, Kazazian HH, Jr., Boeke JD, Gabriel A. 1991. Reverse transcriptase encoded by a human transposable element. *Science* **254**(5039): 1808-1810.

Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK et al. 2011. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**(7332): 59-65.

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* **31**(2): 159-165.

Quentin Y. 1992. Fusion of a free left *Alu* monomer and a free right *Alu* monomer at the origin of the *Alu* family in the primate genomes. *Nucleic Acids Res* **20**(3): 487-493.

Repanas K, Zingler N, Layer LE, Schumann GG, Perrakis A, Weichenrieder O. 2007. Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res* **35**(14): 4914-4926.

SantaLucia J, Jr., Hicks D. 2004. The thermodynamics of DNA structural motifs. *Annual review of biophysics and biomolecular structure* **33**: 415-440.

Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, Dyer M, Cordaux R, Liang P, Batzer MA. 2006. Human genomic deletions mediated by recombination between Alu elements. *American journal of human genetics* **79**(1): 41-53.

Smit AHR, Green P. 1996-2012. RepeatMasker Open-3.0. http://wwwrepeatmaskerorg.

Srikanta D, Sen SK, Conlin EM, Batzer MA. 2009. Internal priming: an opportunistic pathway for L1 and *Alu* retrotransposition in hominins. *Gene* **448**(2): 233-241.

Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. 2001. Biased distribution of inverted and direct *Alus* in the human genome: implications for insertion, exclusion, and genome stability. *Genome research* **11**(1): 12-27.

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ. 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**(7209): 105-108.

Watson JB, Sutcliffe JG. 1987. Primate brain-specific cytoplasmic transcript of the Alu repeat family. *Mol Cell Biol* **7**(9): 3324-3327.

Weichenrieder O, Repanas K, Perrakis A. 2004. Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure* **12**(6): 975-986.

Weiner AM, Deininger PL, Efstratiadis A. 1986. Nonviral retroposons: genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu Rev Biochem* **55**: 631-661.

Xing J, Zhang Y, Han K, Salem AH, Sen SK, Huff CD, Zhou Q, Kirkness EF, Levy S, Batzer MA et al. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res* **19**(9): 1516-1526.

CHAPTER THREE:
A COMPARISON OF 100 HUMAN GENES USING AN *ALU* ELEMENT-BASED INSTABILITY MODEL

Introduction

The draft human genome is interspersed with approximately 45% of mobile element related repetitive sequence (Lander et al. 2001). Advanced sequence analyses indicate that the repeat related portion of the genome may be as high as 69% (de Koning et al. 2011). Retrotransposons, which reproduce through a copy and paste mechanism, have generated the majority of this repetition. The human retrotransposon with the highest copy number is the *Alu* element. *Alu* elements have populated the human genome with over one million copies and account for over 10 percent of all human DNA (Batzer and Deininger 2002).

Both by insertion and by recombination, *Alu* elements spawn genetic disease (Deininger and Batzer 1999; Sen et al. 2006; Witherspoon et al. 2009; Konkel and Batzer 2010). Over 100 studies link *Alu* elements to deletion-related diseases (Table A3.1, page 102). It has been suggested that the most damaging impact of mobile elements may not be their insertion into genes, but their potential interactions with each other. Such interactions could result in deletions, duplications, inversions and a host of more complex genomic structural changes (Hedges and Deininger 2007; Lupski 2010). *Alus* have also been associated with copy number variation breakpoints (de Smith et al. 2008; Kitada et al. 2012). The incidence of *Alu-Alu* interactions is further supported by studies highlighting *Alu-Alu* gene conversion events (Kass et al. 1995; Roy et al. 2000). The homogenization of neighboring *Alu* sequences in ostensibly healthy subjects is

consistent with the theory that *Alu-Alu* interactions routinely occur in healthy cells (Zhi 2007; Aleshin and Zhi 2010).

Recombinant inverted *Alu* pairs have been shown to be unstable in genetically engineered yeast experiments when separated by up to 100 base pair (bp) and are potential sources of chromosome instability when separated by up to 350,000 bp in humans (Lobachev et al. 2000; Stenger et al. 2001; Cook et al. 2011). Furthermore, fusions of inverted *Alu* pairs previously separated by 1-5 kb have been recently identified at the breakpoints of high copy number loci in cancer cells (Kitada et al. 2012).

Previously we reported that full-length inverted (I) *Alu* pairs were statistically underrepresented in the human genome when compared to full-length direct (D) oriented *Alu* pairs (Cook et al. 2011). The term, *Alu* pair exclusions (APEs), was used to describe this human I:D *Alu* pair imbalance. In this study we provide evidence that the inverted APE phenomenon applies to all combinations of human *Alu* sizes. Additionally, we characterize human APEs and construct a model for estimating relative human genome instability based upon the premise that inverted APEs are generated as a consequence of inverted *Alu* pair instability.

This newly developed *Alu* induced instability model was used to compare the relative instabilities of 50 human cancer genes with 50 randomly selected genes from the human genome to experimentally validate the model. The cancer genes considered in this study were selected for their potential susceptibility to deletions (Forbes et al. 2011; Solimini et al. 2012; Stephens et al. 2012). This selection criterion was adopted in order to maximize the model's opportunity to distinguish between these two groups of

genes**.** Taken together, the model estimates that the deletion-prone cancer genes are 58% more unstable than the randomly chosen genes.

Results

Each human gene resides within a unique landscape of *Alu* elements. The structures of these landscapes vary in attributes which include *Alu* density, clustering and orientation. Adding further to *Alu* landscape complexity is the exon number and spacing of each gene. Within these backdrops inverted *Alu* pairs are statistically less numerous than direct oriented *Alu* pairs. It has been hypothesized that this imbalance is the consequence of deletions generated by interactions between inverted *Alu* pairs (Cook et al. 2011).

This hypothesis was tested by construction of an algorithm to estimate the risk that a gene's *Alu* landscape could potentially impose upon its coding sequence. The coding sequence risk was estimated by multiplying two independent probabilities. The first probability, the *Alu*-induced deletion risk, is the probability of the occurrence of an *Alu*-induced deletion. This probability was estimated using an algorithm that characterizes the human I:D imbalance. The second probability, the *Alu*-induced deletion size risk, is the risk that once a deletion is formed, it will be of sufficient size to extend into the coding region of the gene being evaluated. Deletion size risk is estimated using an algorithm constructed from recent studies describing the human indel size frequency distribution. Each of these two probabilities is discussed in greater detail later in this section.

This *Alu* element-based instability model was used to compare the relative stabilities of 50 human cancer genes with 50 randomly selected genes from the human genome. The cancer genes considered in this study were selected for their potential susceptibility to deletions (Forbes et al. 2011; Solimini et al. 2012; Stephens et al. 2012). This methodology was utilized to increase the likelihood for the model to discriminate between these two groups of genes**.**

Two-hit potential of *Alu* elements

The instability model assumes that each end of an *Alu* element is vulnerable to a double-strand break, DSB. These DSB sites are identified from the proposed DNA conformations associated with two mechanisms that have been suggested to explain human inverted *Alu* pair instability. These two mechanisms are characterized by the ectopic invasion and annealing of single-stranded DNA between high-homology DNA bubbles and/or replication forks (Cook et al. 2011). Coincident DNA bubbles passing through aligned *Alu* elements may expose their complementary “flipped out” bases to one another (Jeon et al. 2006; Fogedby and Metzler 2007). Complementary replication forks may also be susceptible to this type of interaction. Each pathway may result in the formation and subsequent resolution of a DNA conformation referred to as a doomsday junction. These two mechanisms are illustrated in Figures 3.1 and 3.2, respectively.

Figures 3.1, diagrams E-F and 3.2, diagram D identify the eight potential sites where a single-strand break could occur during the resolution of a doomsday junction. These sites (illustrated by yellow lightning bolts) are created at the periphery of the doomsday junction where each single strand of DNA transitions from the original DNA

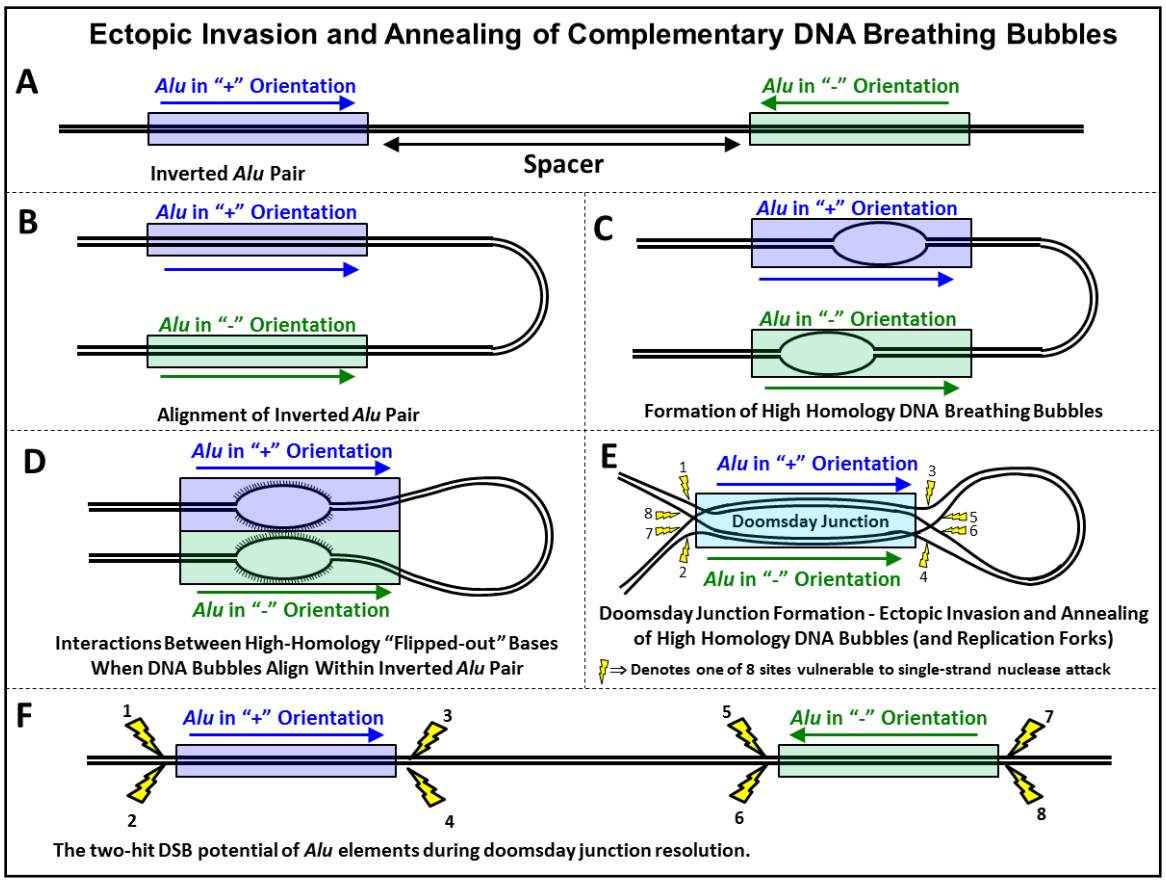


Figure 3.1 - Proposed mechanism for formation and resolution of doomsday junction formed by the ectopic invasion and annealing of complementary DNA breathing bubbles **(A)** Two *Alu* elements in opposite orientations form an inverted *Alu* pair. **(B)** These inverted *Alu* pairs can align as high-homology regions. **(C)** DNA bubbles create short-lived sections of single-stranded DNA (Jeon et al. 2006). **(D)** The unbound bases within these bubbles are characterized by their flipping out from the centerline of the DNA strand (Fogedby and Metzler 2007). Coincident passage of these bubbles within aligned *Alu* elements can create the opportunity for interactions between the flipped-out bases of the complementary DNA strands. **(E)** The ectopic invasion and annealing of single-stranded DNA associated with high-homology DNA bubbles could potentially extend to the entire length of the *Alu* elements. The hypothetical conformation created by this interaction is termed a doomsday junction. A similar interaction may also occur between high-homology replication forks and is described in Figure 3.2 and (Cook et al. 2011). Eight segments of single-stranded DNA formed at the boundary of doomsday junctions create the opportunity for single-strand nuclease attack. These sites are illustrated as yellow lightning bolts. **(F)** As again illustrated by the yellow lightning bolts, each end of each *Alu* element involved in the doomsday junction is vulnerable to a double-strand break. This two-hit hypothesis for each *Alu* element was incorporated into the model's algorithm.

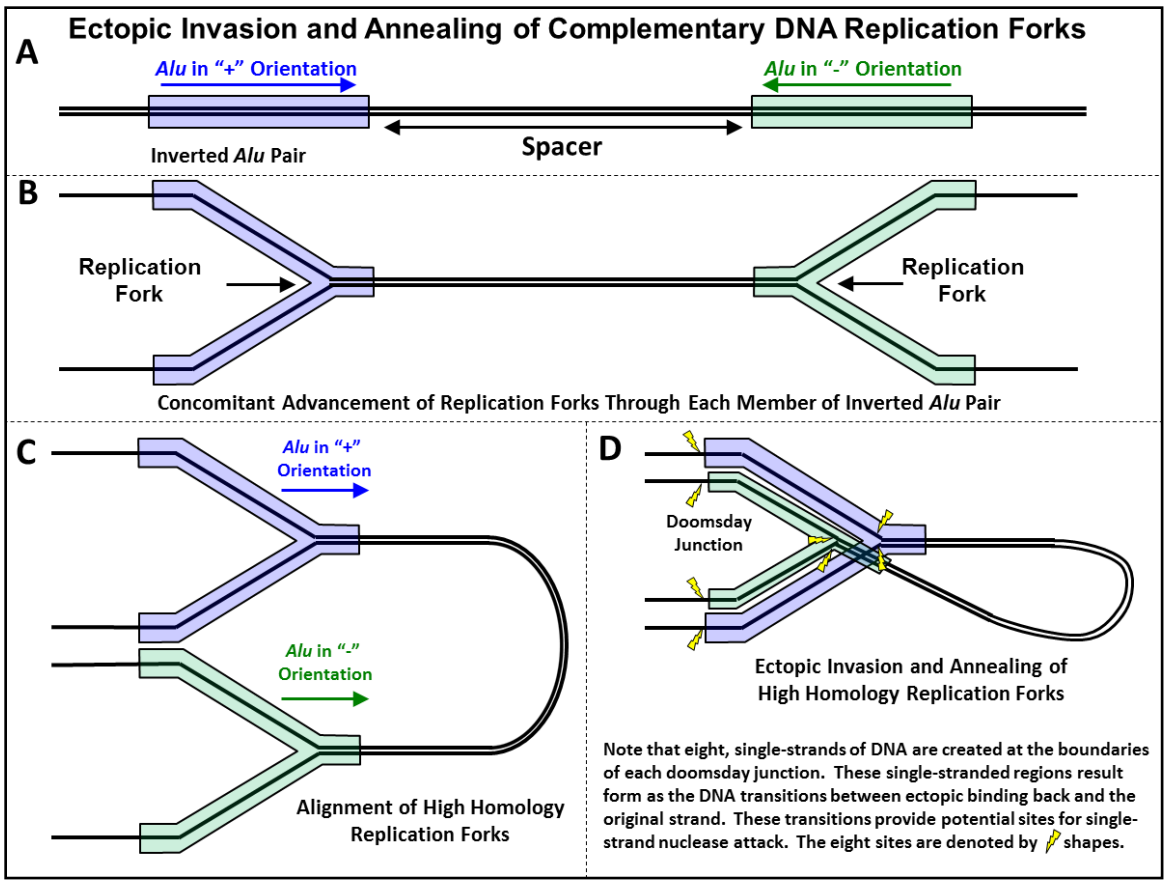


Figure 3.2 - Proposed mechanism for the formation of a doomsday junction formed by the ectopic invasion and annealing of complementary replication forks **(A)** Two *Alu* elements in opposite orientations form an inverted *Alu* pair. **(B)** Concomitant advancement of replication forks through each member of an inverted *Alu* pair. **(C)** Bending of the DNA to permit alignment of the complementary replication forks. **(D)** Ectopic invasion and annealing of single-stranded DNA associated between high-homology replication forks could potentially extend to the entire length of the *Alu* elements. The hypothetical conformation created by this interaction is termed a doomsday junction. As also illustrated in Figure 3.1, eight segments of single-stranded DNA are formed at the boundary of the doomsday junction and create the opportunity for single-strand nuclease attack. These sites are illustrated as yellow lightning bolts.

double-helix to the ectopic conformation of the doomsday junction. These regions of single-stranded DNA may be susceptible to attack by single strand nucleases. If only one strand at the end of each *Alu* element is cut, the doomsday junction can likely resolve itself without damage to the original sequence. However, if both strands at the

same end of either of the two inverted *Alu* elements are cut, a DSB can occur (Figure 3.1, diagram F). This potential for a DSB at each end of an *Alu* element forms the basis for the "two-hit hypothesis" for each *Alu* element considered by this instability model.

Probability One – *Alu*-induced deletion risk

The *Alu*-induced deletion risk is the likelihood of a deletion arising from the resolution of a doomsday junction. The two-hit deletion potential of each *Alu* element results in the number of potential *Alu*-induced deletion sites within a given *Alu* landscape being twice the number of *Alu* elements. Three variables were found to significantly correlate with the *Alu* pair I:D ratio; 1) spacer size, 2) the number of *Alu* elements within the spacer and 3) the clustering state of the each *Alu* pair (discussed in more detail, below). Figures 3.3 and 3.4 express the human inverted to direct *Alu* pair ratio, I:D ratio, as a function of these three variables. The *Alu* pair I:D ratio was not found to significantly correlate with *Alu* length (Methods, page 84).

The shape of the curves in these three Figures 3.3 and 3.4 illustrate that the *Alu* pair I:D ratio is not a smooth function across the full range of spacer sizes. These curves are plotted along the medians of ten spacer size percentile groupings for each of the respective *Alu* pair sequence numbers (APSNs). The APSN is the parameter that describes the number of *Alus* within the spacer of an *Alu* pair. The APSN for an *Alu* pair is the n+1 number of *Alu* elements residing with the spacer (Methods, page 84).

Three possible mechanisms may explain the unusual shape of the human *Alu* pair I:D ratio versus spacer size curves. Using the APSN1 curve in Figure 3.3 as a reference, these three mechanisms may be as follows; 1) between the 0th and 5th

spacer size percentiles (centered at ~100 bp), hairpin formation may be the predominant form of *Alu-Alu* interaction, 2) for the 10th (5th-15th) and 20th (15th-25th) spacer size percentiles (centered between ~100 and ~5,000 bp) DNA persistence (stiffness), may hinder inverted *Alu-Alu* interactions and 3) for spacer sizes between the 25th and 95th percentiles, DNA persistence appears to wane and the curve begins to progress toward unity.

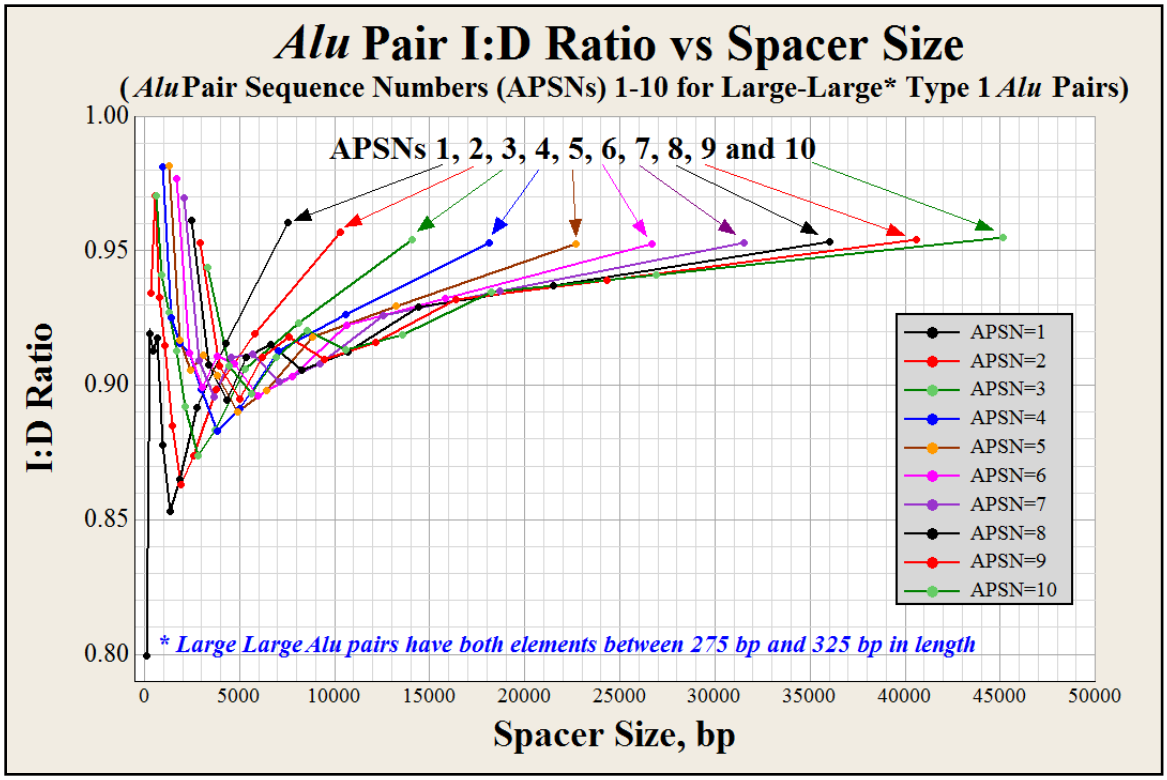
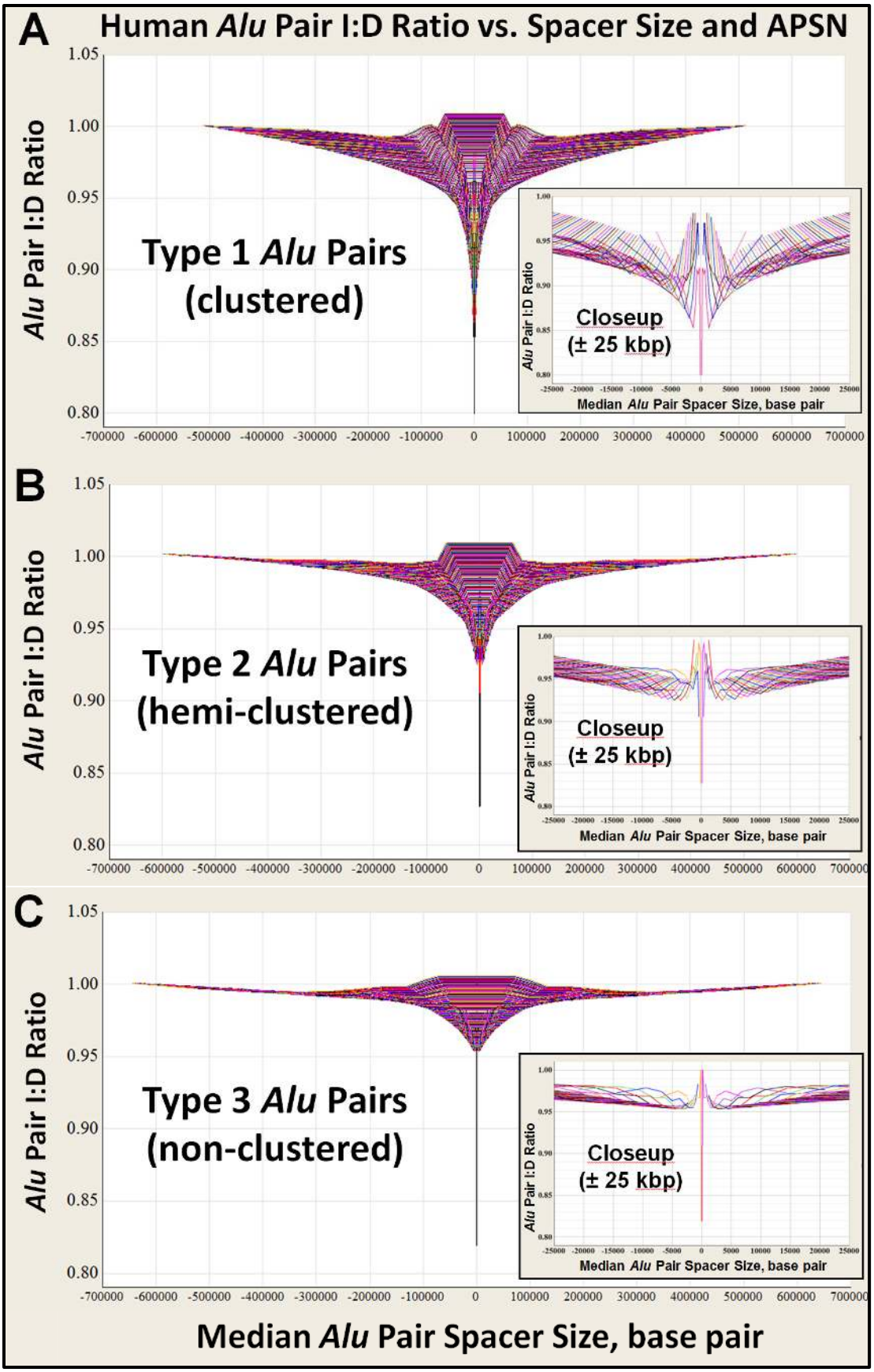


Figure 3.3 - *Alu* pair I:D ratio versus spacer size for Type 1 *Alu* pairs for APSNs 1-10 Each of the ten points which make up each curve are the composite I:D ratios for the ten spacer size percentile ranges (Methods, page 84) plotted against their median spacer size for the ten respective APSN families. This plot illustrates that the I:D ratio is not a smooth function across the full range of adjacent spacer sizes, but instead varies with *Alu-Alu* interaction mechanisms (see text). These ten curves, along with their 5' mirror images, make up ten of the 220 (APSNs ± 110) curves which are shown together in Figure 3.4.

Human *Alu*, LINE1 and SVA elements, frequently cluster together in groups where adjacent elements are separated by ≤50 bp (Cook et al. 2011). Using this definition of clustering, four types of clustered *Alu* pairs can be described. These are identified as Types 0, Type 1, Type 2 and Type 3 *Alu* pairs. Type 0 *Alu* pairs (clustered together) have both *Alu* elements residing within the same cluster, Type 1 *Alu* pairs (clustered separately) have both *Alu* elements residing within different clusters, Type 2 *Alu* pairs (hemi-clustered) have only one of the two elements residing within a cluster and Type 3 *Alu* pairs (non-clustered) have neither element residing within a (Methods, page 84). Type 1, 2 and 3 *Alu* pairs exhibit distinctly different I:D ratios and their stabilities must therefore be estimated separately (Figure 3.4). Type 0 *Alu* pairs are subject to strong orientational insertion bias and their instability is estimated via experimental studies of *Alu* elements in yeast (Methods, page 84, and (Lobachev et al. 2000)).

Figure 3.4, pane A, illustrates the I:D ratio for Type 1 large-large (275-325 bp) *Alu* pairs for APSNs 1-10. Figure 3.4 is similar to Figure 3.3 and includes all APSNs (±110) containing at least one spacer size percentile with an I:D ratio <0.995. I:D ratios ≥0.995 do not provide statistical confidence that the I:D ratio is below unity (Methods, page 84). Figures 3.4, pane B and 3.4, pane C are similar to Figure 3.3 and

Figure 3.4 - *Alu* pair I:D ratio versus *Alu* pair type, spacer size, and APSN. Panes A, B and C of this figure illustrate the human *Alu* pair I:D ratio versus spacer size for the ±110 APSN curves for full-length (275-325 bp), Type 1, 2 and 3 *Alu* pairs. These 220 curves shown in each of these three panes are so closely spaced that they collectively appear as surfaces. Expanded views showing individual curves (spacer sizes ±25,000 bp) are shown in the inset in each pane and for APSNs 1-10 for Type 1 *Alu* pairs in Figure 3.3. (Figure 3.4 continues on the following page.)


A
Human Alu Pair I:D Ratio vs. Spacer Size and APSN
Alu Pair I:D Ratio
1.05
1.00
0.95
0.90
0.85
0.80
Type 1 Alu Pairs
(clustered)
Closeup
(± 25 kbp)
Alu Pair I:D Ratio
Median Alu Pair Spacer Size, base pair
-700000 -600000 -500000 -400000 -300000 -200000 -100000 0 100000 200000 300000 400000 500000 600000 700000
B
Alu Pair I:D Ratio
Type 2 Alu Pairs
(hemi-clustered)
Closeup
(± 25 kbp)
Alu Pair I:D Ratio
Median Alu Pair Spacer Size, base pair
C
Alu Pair I:D Ratio
Type 3 Alu Pairs
(non-clustered)
Closeup
(± 25 kbp)
Alu Pair I:D Ratio
Median Alu Pair Spacer Size, base pair
Median Alu Pair Spacer Size, base pair

show the I:D ratio versus spacer size relationships for Type 2 and Type 3 *Alu* pairs, respectively.

Using the I:D ratio relationships illustrated in Figures 3.3 and 3.4, the model generates a predicted stability for each *Alu* element within a gene's *Alu* landscape. The predicted I:D ratio is the predicted stability for the *Alu* pair. The contribution that an inverted *Alu* pair makes to the overall stability of each *Alu* element of the pair is obtained by taking the square root of that pair's predicted I:D ratio. Likewise, the contribution that an inverted *Alu* pair makes to the overall stability for one end of an *Alu* element of the pair is the fourth root of that *Alu* pair's predicted I:D ratio. The overall stability of one end of an *Alu* element is the product of the fourth roots of all the predicted I:D ratios for each of the potential 220 inverted *Alu* pairs (i.e., grand product) that an *Alu* element might form with its ± 110 *Alu* neighbors (Methods, page 84).

Figure 3.4 reveals an unexpected excursion of the I:D ratio above unity for the highest *Alu* density genomic regions. This excursion only exists for APSNs ≥65 and only for the most *Alu* dense regions of the genome (0-5th spacer size percentile, Table A3.2, page 114). This high I:D ratio may indicate that direct *Alu* pair recombination in these high *Alu* density regions of the genome may outpace the activity of inverted APE events.

Alu Landscapes

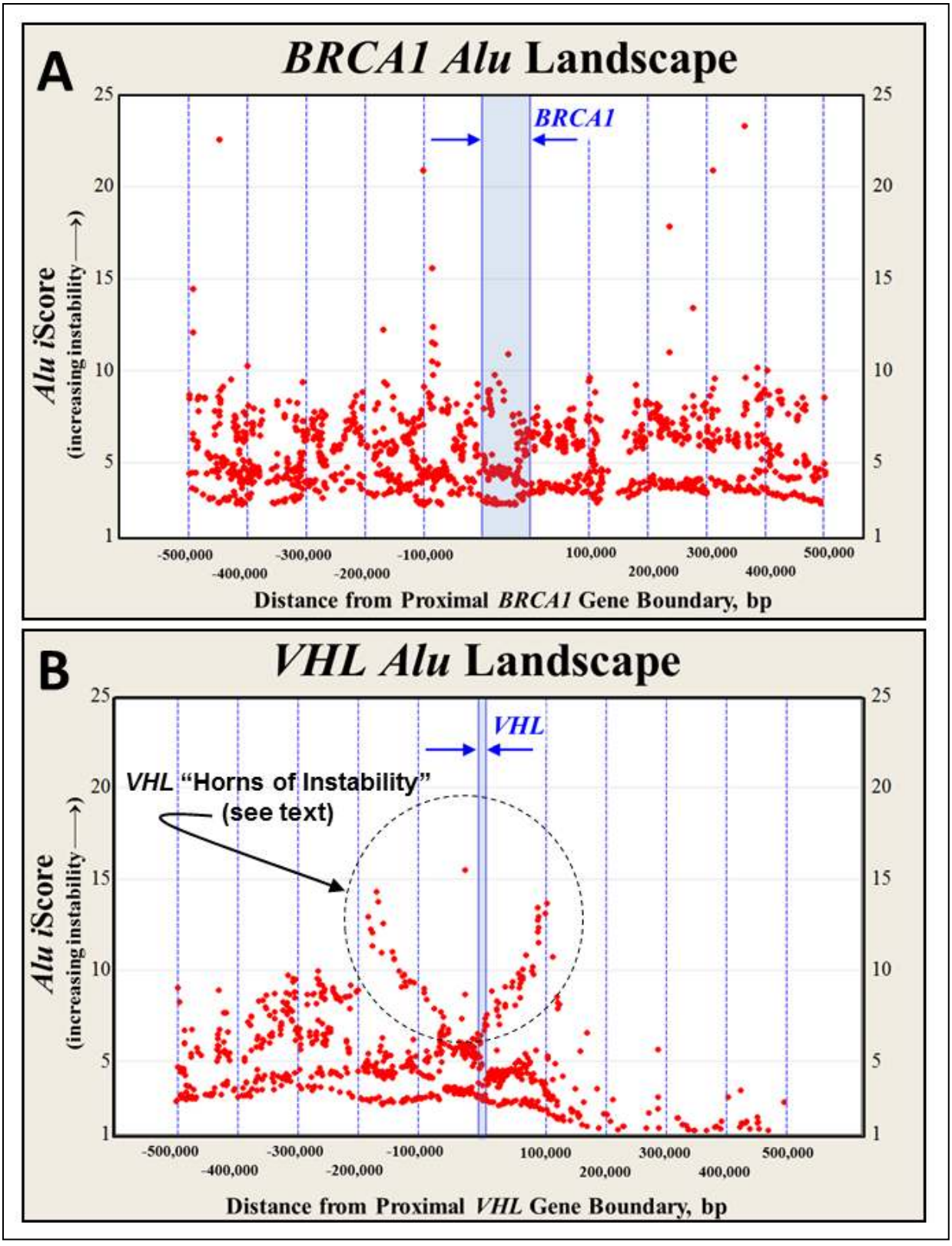
Each of the genes considered in this study were evaluated using the backdrop of *Alu* elements in which they reside. These *Alu* backdrops are referred to as *Alu* landscapes. Figure 3.5 illustrates the *Alu* landscapes around two of the deletion-prone

cancer genes evaluated in this study, *BRCA1* and *VHL*. The vertical blue lines in each figure demarcate 100,000 bp distances from the respective end of each gene and the light blue region in the center of each diagram encompasses the respective gene's coding locus.

The respective instability score (*i*Score) of each *Alu* element is plotted on the vertical axis. These *i*Score values are the inverse of the *Alu* stabilities calculated using the algorithms developed from Figure 3.4. Higher *i*Score values represent higher *Alu* instabilities. The red dots signify the locus versus the iScore value for each element within the *Alu* landscape.

The *Alu* landscapes illustrated in Figure 3.5 span ± 500,000 bp from the end of each gene. Similar landscapes are shown for eight additional genes in Figure A3.1 (page 137). The instability model only includes those *Alus* residing within ± 250,000 bp from the end of each gene (discussed in more detail, below). The larger landscapes provided in Figures 3.5 and A3.1 (page137) are shown to illustrate the ebb and flow of *Alu* instabilities across the genome. Approximately 0.3% of the human genome is represented in the ten panes shown in these two figures. The panes in Figure A3.1

Figure 3.5 - *Alu* landscapes for *BRCA1* and *VHL* This figure characterizes the *Alu* landscape within and 500,000 bp 5' and 3' of A) *BRCA1* and B) *VHL*. The locus for each *Alu* element is plotted against its respective instability score, *i*Score. Larger *i*Score values represent higher predicted *Alu* element instabilities. Similar *Alu* landscapes for eight additional genes examined in this study are shown in Figures A3.1A-A3.1G. The span about each respective gene for these landscapes is ± 500 kbp. These spans are twice the size of the ± 250 kbp flanking landscapes which are considered to pose a risk for an exon damaging deletion (see text). These larger spans better illustrate the ebb and flow of *Alu*-related instability around each respective gene.
(Figure 3.5 continues on the following page.)


A
BRCA1 Alu Landscape
BRCA1
Alu iScore
(increasing instability ⟶)
25
20
15
10
5
1
-500,000
-400,000
-300,000
-200,000
-100,000
100,000
200,000
300,000
400,000
500,000
Distance from Proximal BRCA1 Gene Boundary, bp
B
VHL Alu Landscape
VHL
VHL "Horns of Instability"
(see text)
Alu iScore
(increasing instability ⟶)
Distance from Proximal VHL Gene Boundary, bp

illustrate the *Alu* landscapes for the five deletion-prone cancer genes, *APC*, *ATM, MLH1, MSH2,* and *TP53*. Panes F-H in Figure A3.1 (pages 137) describe the *Alu* landscapes for randomly chosen genes, *GDPD2, KEAP1 and SF3B3.* Among the 100 genes examined in this study, only two of the top 10 highest *Alu* density landscapes are associated with deletion-prone cancer genes, *ARID1A* and *BRCA1*. These two genes rank 8th and 10th this list with *Alu* landscape densities of 1,322 and 1,309 *Alus* per mega base, respectively (see Table A3.3, page 116). The *Alu* element density across the human genome averages 381 *Alus* per mega base. The top five most *Alu* dense landscapes (all randomly selected genes) belong to *KEAP1, NCF1, NANOS3, OPRD1,* and *SET1* with *Alu* densities of 1,916, 1,783, 1,644, 1,534 and 1,525 *Alus* per mega base, respectively (see Table A3.4, page 117).

Probability Two – *Alu*-induced deletion size risk

Human genome indel size frequency distributions from two previous studies provide a glimpse into the shape of the overall human deletion size frequency distribution (Wheeler et al. 2008; Mills et al. 2011). A hybrid deletion size frequency model was developed from these studies and is shown in Figure 3.6. The sum of the 500,000 individual deletion size probabilities in this figure equals 1.0. This hybrid model is used to estimate the relative deletion size risks which arise from inverted *Alu*-induced DSBs (Methods, page 84). The shape of the curve in Figure 3.6 reflects a deletion size frequency distribution where 95 percent of deletions are ≤50 bp. The maximum deletion size of 500,000 bp in Figure 3.6 was chosen because this size deletion has a risk of occurrence that is less than one billionth of the risk predicted for a 1 bp deletion. This model assumes that deletions extend equidistant from an initiating DSB. Consequently,

the maximum distance from which an individual *Alu* element is considered to pose a deletion risk to a coding exon is 250,000 bp (250,000 bp x 2 = 500,000 bp). In addition to considerations for maximum deletion size, additional flanking sequence must be examined within an *Alu* landscape to accommodate for the possibility that inverted *Alu* pairs can interact when separated by up to 421,000 bp. This is the spacer size (in Figure 3.4, pane A, pages 66-67) that intersects with an I:D ratio of 0.995. This I:D ratio is statistically lower than unity ($p \leq 0.05$; Methods, page 84). Therefore, an *Alu* element that is separated by as much as 671,000 bp from a coding exon could potentially

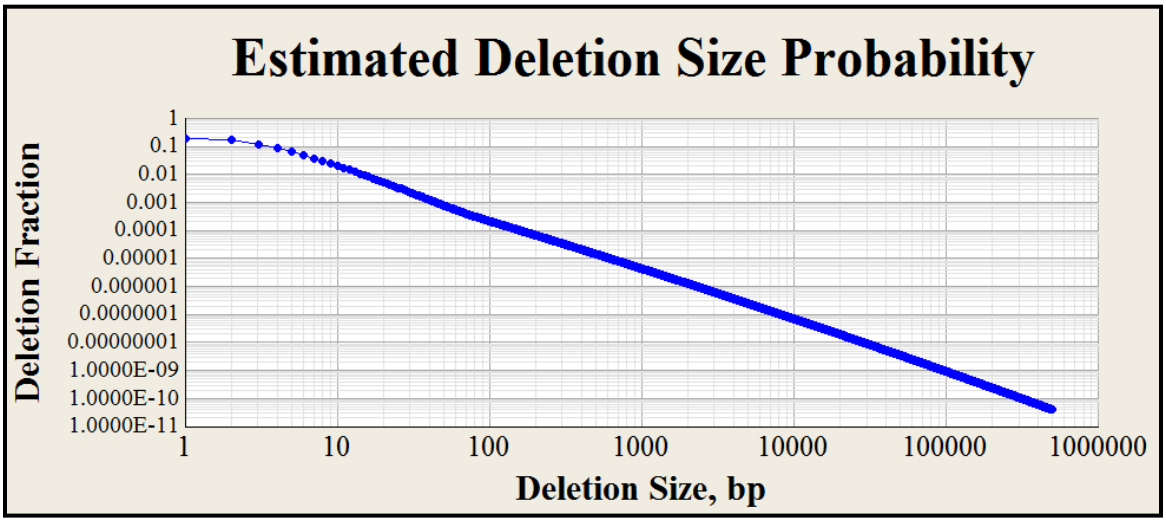


Figure 3.6 - Estimated human deletion size frequency distribution This log-log (base 10) plot estimates the relative distribution of deletion sizes within the human genome. The sum of the 500,000 individual deletion size probabilities in this figure equals 1.0. The curve was constructed from two different studies and predicts that 95% of deletions are ≤50 bp in size and 99% of deletions are ≤445 bp (Wheeler et al. 2008; Mills et al. 2011). When combined with the two-hit hypothesis for *Alu* elements, this curve suggests that the two ends of an *Alu* element pose specific and different risks to an exon's coding region.

threaten the coding integrity of that exon. At this distance from a coding exon, an *Alu* element could conceivably interact with a second *Alu* separated by only 250,000 bp

from the same exon (spacer size between the two *Alus* = 671,000 bp - 250,000 bp = 421,000 bp). This interaction could potentially generate a DSB at the second *Alu* that could possibly extend into the coding exon.

Relative gene stabilities

The relative stability of a gene for the purpose of this study is defined as the relative likelihood that a coding exon will not be breached by a deletion. The determination of this stability must consider the collective deletion risks along with the respective deletion size risks posed by all potential DSB sites generated within a gene's *Alu* landscape. More specifically, the overall stability of a gene is the multiplied product (grand product) of the individual *Alu* element contributions to that gene's stability within its *Alu* landscape (Methods, page 84). The required calculations to determine this stability are extensive. Estimation of the stability of *BRCA1*, because of its large *Alu* landscape, requires 171,225 consecutive calculations. As seen in Table A3.4 (page 117) *BRCA1* has 761 *Alu* elements residing within its intronic regions and the 250,000 bp flanking regions, 5' and 3' of the gene. The majority of these calculations are associated with the 220 potential *Alu* pair interactions for each of these 761 *Alu* elements. The sheer number of required consecutive calculations raised concerns that significant adjustments would be required for proper interpretation of the raw output from the model. This concern did not materialize. The individual gene stabilities plotted in Figure 3.7 are the unadjusted output stability values from the model.

The uppermost histogram in Figure 3.7 is a distribution of the raw stabilities of the 50 deletion-prone genes taken directly from the model. The bottom histogram is a

distribution of the raw stabilities of the 50 randomly selected genes. Lower values represent greater instability. Table A3.3 and Table A3.4 (pages 116-117) list the individual gene stabilities. For reference, this instability model would generate a stability

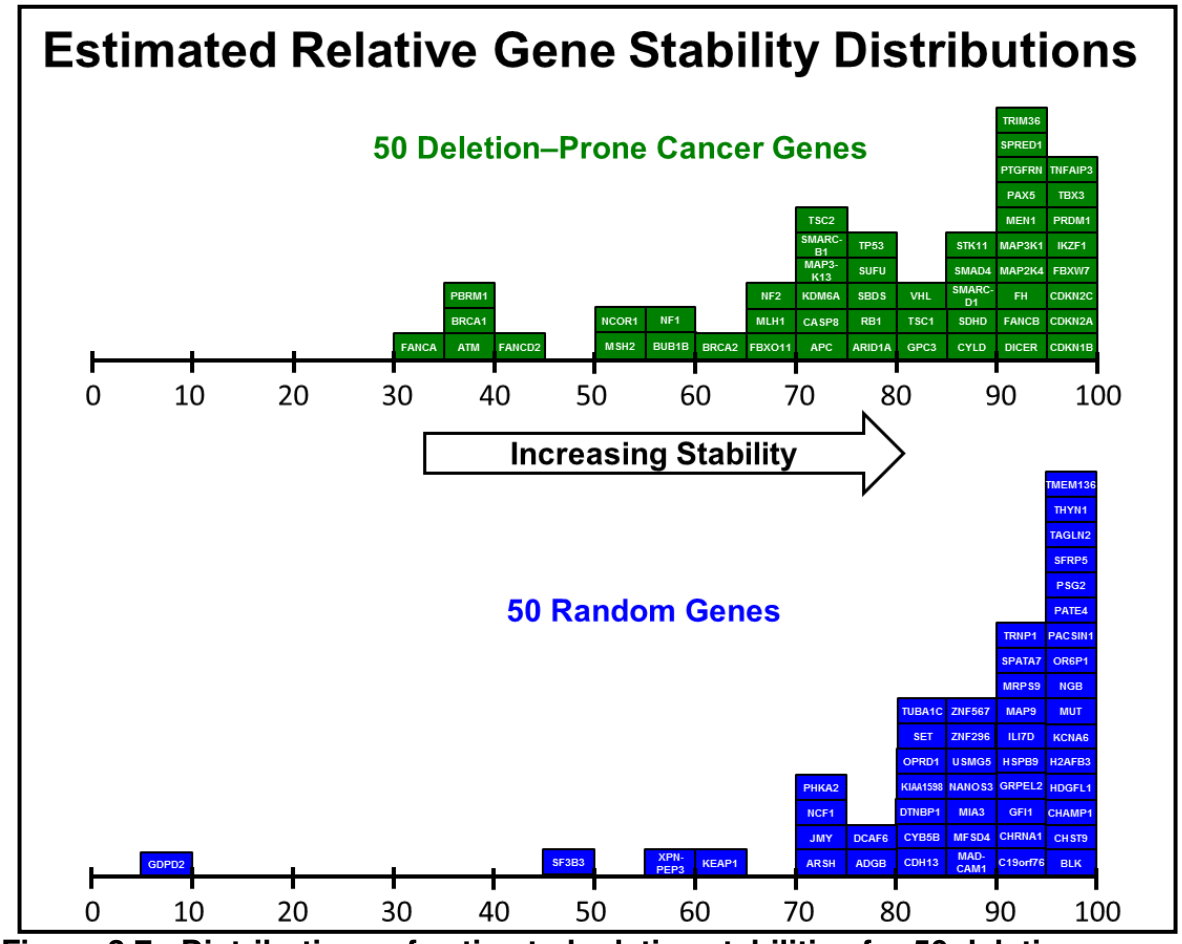


Figure 3.7 - Distributions of estimated relative stabilities for 50 deletion-prone cancer genes and 50 randomly chosen genes The two histograms in this figure describe the relative estimated stabilities of the 50 deletion-prone genes and the 50 randomly selected genes, respectively. The stability values in these histograms are the unadjusted outputs from the *Alu* instability model algorithm. These stabilities are also provided in Table A3.3 and Table A3.4, respectively (pages 116-117). Note that the least stable of all 100 genes is the randomly selected gene, *GDPD2*. This low stability springs from the putative exonized *Alu* that occurs in variant 1 of *GDPD2's* 12th exon.

of 100 for any gene residing within an *Alu*-free landscape. The average unadjusted stabilities of the deletion-prone cancer genes and randomly chosen genes from Tables

A3.3 and A3.4 (pages 116-117) are 77.7% and 85.9%, respectively**.** The deletion-prone cancer genes, therefore, have a 58% greater likelihood of a deletion insult than that of the randomly chosen genes. The equation for this difference in stability between these two sets of genes is as follows.

$$\frac{[(1-0.777)-(1-0.859)]}{(1-0.859)} \times 100 = 58\%$$

This likelihood increases to 78% when *GDPD2,* the randomly chosen gene with an exonized *Alu* element, is excluded from the list of random genes (discussed in more detail, below).

Only one cancer gene, *IKZF1*, was among the most stable 10% of the 100 genes analyzed, while seven deletion-prone cancer genes, *FANCA, NCOR1, BRCA1, PBRM1, ATM, FANCD2* and *MSH2* were among the most unstable 10% (10) of the 100 genes analyzed (Tables A3.3 and A3.4, pages 116-117). The top 10% most stable genes contain an average of 4 coding exons, versus an average coding exon count of 31 for the 10% most unstable genes.

The least stable of all 100 genes is the randomly selected gene, *GDPD2*. The low relative stability of *GDPD2* (7.1%, see Table A3.4) results from a putative exonized *Alu* that occurs in variant 1 of *GDPD2's* 12th exon. Four different variants of this gene are represented in the UCSC genome browser. The absence of this exon in the other three variants is consistent with this predicted instability. This *Alu* element-based instability model considers an exonized *Alu* element as the most unstable form of structural variation within a gene's coding region Therefore, in addition to the disruption

of coding sequence associated with an *Alu* insertion into an exon, subsequent disruption may also ensue because of the high potential for small deletions to occur at the ends of the *Alu* element. Both of these mechanisms may help explain the scarcity of exonized *Alus*. The potential risk of an exon-damaging deletion originating from the end of a nearby *Alu* element is consistent with the observed scarcity of *Alu* elements within 50 bp of human exons (Lev-Maor et al. 2008; Zhang et al. 2011).

An examination of the variation in relative gene instabilities with respect to variation in the deletion size frequency distribution was also conducted. This evaluation was performed by varying the ≤50 bp deletion size frequency between 90 and 99 percent in increments of one percent (Figure A3.3, page 148). While this analysis resulted in significant changes in absolute gene instabilities, the relative instabilities between most genes was unaltered. Exceptions to this observation occurred for *ATM* and *CASP8.* These have the two closest *Alu* elements located within 5 and 7 bp of exons 14 and 8, respectively. The next closest *Alu* to a deletion-prone cancer gene exon occurs at exon 19 of *FANCD2* with a separation of 20 bp. *ATM* and *CASP8* disproportionately increase in relative instability (compared to the other 48 genes in the deletion-prone cancer gene group) as the fraction of deletions ≤50 bp was increased (Methods, page 84).

Relative exon stabilities

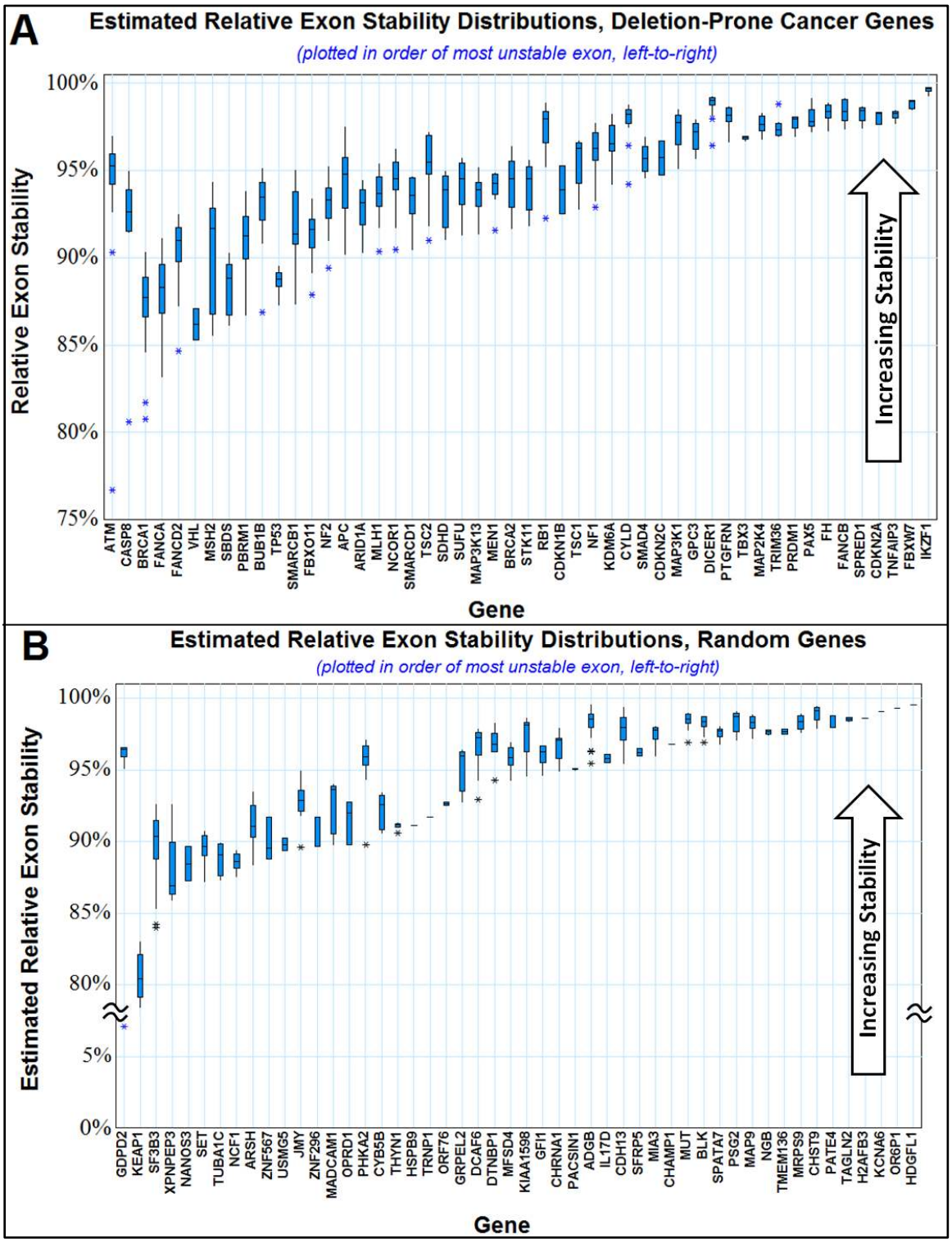
The relative stabilities of the 1,287 coding exons which make up the 100 genes evaluated in this study were also compared. Figure 3.8, pane A is a boxplot of the individual exon stabilities for the 50 deletion-prone cancer genes. Figure 3.8, pane B is

a similar boxplot for the 50 randomly selected genes. The two figures are constructed left-to-right based upon each gene's most unstable exon. These two figures illustrate that relative exon stability values tend to cluster in a gene specific manner. Within the deletion-prone cancer gene group, the two, left-most genes, *ATM* and *CASP8* have moderate mean exon stability values. However, the presence of exons with outlying high instabilities within *ATM* and *CASP8* puts these two genes first and second place of the most unstable among the deletion-prone cancer genes. These two genes have *Alu* elements that are within 5 and 7 bp of their 14th and 8th exons, respectively. When average exon instability is used as the sorting criterion (illustrated by the bold black line through each respective boxplot), *VHL*, *BRCA1*, FANCA, *TP53* and *SBDS* make up the top 10% most unstable genes among the 50 deletion-prone cancer genes. Finally, Figure 3.8, pane B illustrates the very low stability value (7.2) determined for the exon containing the putative exonized *Alu* in *GDPD2.*

Deletion sizes in VHL cancer deletion families do not recapitulate Figure 3.6

Figure 3.6 (page 72) is constructed upon the premise that >95% of deletions in the human genome are less than 50 bp in length (Wheeler et al. 2008; Mills et al. 2011).

Figure 3.8 - Estimated relative exon stability distributions for the 50 deletion-prone cancer genes and 50 randomly chosen genes (A) Boxplot of the individual exon stabilities for the 50 deletion-prone cancer genes. The genes in this figure are ordered left-to-right on the basis of each gene's least stable exon. Note that the while exon stabilities vary within and between genes, these stabilities tend to cluster in a gene specific manner. The presence of a single, outlying low stability exon within *ATM* and *CASP8* puts these two genes first and second place of lowest stability among these 50 genes. **(B)** Boxplot of the individual exon stabilities for the 50 randomly selected genes. Note that a broken Y-axis scale is required to capture the low stability of the putative exonized *Alu* in the 12th exon of *GDPD2*. (Figure 3.8 continues on the following page.)


A
Estimated Relative Exon Stability Distributions, Deletion-Prone Cancer Genes
(plotted in order of most unstable exon, left-to-right)
Relative Exon Stability
100%
95%
90%
85%
80%
75%
Increasing Stability
ATM CASP8 BRCA1 FANCA FANCD2 VHL MSH2 SBDS PBRM1 BUB1B TP53 SMARCB1 FBXO11 NF2 APC ARID1A MLH1 NCOR1 SMARCD1 TSC2 SDHD SUFU MAP3K13 MEN1 BRCA2 STK11 RB1 CDKN1B TSC1 NF1 KDM6A CYLD SMAD4 CDKN2C MAP3K1 GPC3 DICER1 PTGFRN TBX3 MAP2K4 TRIM36 PRDM1 PAX5 FH FANCB SPRED1 CDKN2A TNFAIP3 FBXW7 IKZF1
Gene
B
Estimated Relative Exon Stability Distributions, Random Genes
(plotted in order of most unstable exon, left-to-right)
Estimated Relative Exon Stability
100%
95%
90%
85%
80%
5%
0%
Increasing Stability
GDPD2 KEAP1 SF3B3 XPNPEP3 NANOS3 SET TUBA1C NCF1 ARSH ZNF567 USMG5 JMY ZNF296 MADCAM1 OPRD1 PHKA2 CYB5B THYN1 HSPB9 TRNP1 ORF76 GRPEL2 DCAF6 DTNBP1 MFSD4 KIAA1598 GFI1 CHRNA1 PACSIN1 ADGB IL17D CDH13 SFRP5 MIA3 CHAMP1 MUT BLK SPATA7 PSG2 MAP9 NGB TMEM136 MRPS9 CHST9 PATE4 TAGLN2 H2AFB3 KCNA6 OR6P1 HDGFL1
Gene

In contrast, 25% of the deletions resulting in *VHL* cancer are greater than 10,000 bp (Franke et al. 2009). This apparent conflict in deletion size frequency may arise from ascertainment bias as only those deletions that result in VHL cancer are detected. The *Alu* landscape flanking the *VHL* gene in Figure 3.5, pane B (page 70) reveals two regions of high *Alu* instability (*i*Scores shaped as horns) that extend in both 5' and 3' directions from the base of the VHL gene. As can be seen from the diagram, the 5' and 3' regions extend approximately 150,000 bp and 100,000 bp, respectively from the gene. Based on genome-wide derived deletion size frequencies in Figure 3.6 (page 72) most of the deletions arising within these "horns of *Alu* instability" would be much shorter than the distances required to damage the VHL coding integrity and would likely go undetected.

Discussion

Evolution is a slow process. Clues to its activity reside in the subtle patterns that it leaves behind. Two of these patterns, chimeric *Alus* and the instability of cancer genomes are consistent with this study's model of inverted *Alu* pair instability. The potential implications of these two evolutionary patterns are discussed below.

Chimeric *Alus* may camouflage the instability of inverted *Alu* pairs

It is generally accepted that most chimeric *Alu* elements are formed by non-allelic homologous recombination (NAHR) between two direct oriented *Alu* elements (Sen et al. 2006). However, chimeric *Alu* elements can also be generated by single-strand annealing repair of DSBs that occur within the spacer sequence separating a direct oriented *Alu* pair. However, single-strand annealing repair is only possible when high-

homology sequences flank the DSB. Satisfying this homology requirement entails sufficient resection of the intervening spacer sequence separating the *Alu* pair (Hedges and Deininger 2007).

The presence of a chimeric *Alu* element at the boundary, or breakpoint, of structural variation provides little evidence regarding the etiology of its formation. As a result, the mechanistic details behind this type of structural variation are difficult to ascertain. Without supporting evidence for an intervening deletion mechanism in the pre-chimeric spacer, the putative NAHR route is the most reasonable explanation for the formation of chimeric *Alu* elements.

This study's *Alu* element-based stability algorithm was constructed upon the premise that DSBs can be generated from the interaction between inverted *Alu* pairs. It is possible that a fraction of these inverted *Alu* pair generated DSBs could be repaired through single-strand annealing repair of direct-oriented *Alu* pairs. This type of repair would generate a chimeric *Alu* element. The chimeric *Alu* element would effectively mask the inverted *Alu* pair as the source of the DSB. Further adding to this camouflage is the possibility that the chimeric *Alu* breakpoint (repair point) can be thousands of base pair removed from the initiating DSB (Sen et al 2006, Han et al. 2007).

Both non-allelic homologous recombination and single strand annealing repair likely contribute to the human chimeric *Alu* population. However, to our knowledge, the strongest evidence in support of either theory is the imbalance in the human *Alu* pair I:D ratio (Stenger et al. 2001; Cook et al. 2011). Chimeric *Alu* elements appear to result

from repair of approximately 10 percent of inverted APE deletions (see Table 2.4, page 30).

Oncogenesis could also be a passenger mutation to genome-wide instability

As mentioned previously in the Results section, the *Alu* element-based instability model predicts that deletion-prone cancer genes are ~58% more unstable than randomly selected genes. This 58% difference between cancer and random gene deletion rates is not sufficiently large to preclude the possibility that both rates may be common products of an insidious process that damages the genomes of somatic cells. Prior to senescence, the trillions of cells in our bodies likely provide multiple occasions for an unfortunate combination of cancer-prone genetic damage to occur (Serrano 2010).

Most of the mutations in a cancer cell are passenger mutations that do not appear to contribute to the cancer cell's fitness it is generally assumed that the vast majority of these passenger mutations are byproducts of oncogenesis. While passenger mutations may be more likely to occur subsequent to the oncogenic driver mutation, the assumption that somatic cell genomes are stable prior to oncogenesis has not been proven.

In final support of a model suggesting general somatic cell instability is the observation that deletion size frequencies observed in VHL cancer (see Results) do not conform to the deletion size frequency distribution which has been observed in healthy cells (Figure 3.6, page 72). The disproportionate number of large deletions (relative to

Figure 3.6) observed among various VHL cancer families suggests that many smaller, non-cancerous deletions occur but go undetected within healthy cell populations.

The human *Alu* pair I:D ratio may underrepresent inverted *Alu* pair interactions

As previously stated, a premise of this study is that the imbalance in the human *Alu* pair I:D ratio is a consequence of genomic instability. The human *Alu* pair I:D imbalances illustrated in Figure 3.4 (pages 66-67) may under estimate inverted *Alu* pair instability for two reasons. 1) The depression of the I:D ratio does not include inverted *Alu* pair deletions that have been lost through negative selection pressure and genetic drift. 2) The instability estimates derived from the I:D ratio assumes no instability between direct oriented pairs. Several studies have shown that both inter-chromosomal and intra-chromosomal recombination occurs between *Alu* elements (Elliot et al. 2005, Sen et al. 2006, Han et al. 2007).

The development of this genomic instability model is just one approach to finding tangible risk factors associated with mobile element-related threats to the genome. Unfortunately, we are far from a complete understanding of the entire puzzle. However, the fundamentals provided by the algorithm used in this study may lay the foundation for other computational approaches to comparing genetic risks posed by structural variations which are unique to specific individuals, families and people groups. With the advancement of genome sequencing technologies and the emergence of whole genome analyses, sophisticated modeling systems such as this *Alu*-element based instability model will likely be essential to the future of genomics research.

Conclusions

Interactions between highly homologous *Alu* elements and their potential to result in deletions, duplications, inversions and gene conversion events has been well documented (Kass et al. 1995; Roy et al. 2000; Bailey et al. 2003; Sen et al. 2006; Lee et al. 2008). Various forms of structural variation have been shown to account for a large proportion of human genetic diversity (Lupski 2010; Girirajan et al. 2011; Mills et al. 2011). Recent studies have suggested that common types of *Alu* induced structural variation may be just the tip of the iceberg, with far more complex mechanisms for *Alu* induced genome instability being possible (Lobachev et al. 2000; Stenger et al. 2001; Lupski 2010; Cook et al. 2011). The model developed in this study estimates relative human genome instability based upon the premise that inverted *Alu* pair exclusions are generated as a consequence of genomic instability.

Assuming that the basic concepts for this *Alu* element-based gene stability model are correct, the following five conclusions are evident from this study. 1) *Alu* landscapes create regions of genomic instability that are unique for each human gene. The majority of this instability resides within the ± 250,000 bp regions flanking each gene. 2) Genes with higher exon counts are potentially more vulnerable to coding deletions. Additional exons provide more opportunities for *Alu* elements to reside in close proximity to coding regions. 3) Exonized *Alu* elements are a particularly unstable class of structural variation. This instability is inherent in exonized *Alus* because any deletion resulting from an *Alu-Alu* interaction is more likely to result in loss of coding sequence. 4) The human deletion size frequency curve predicts that large deletions detected through a cancer phenotype may be evidence that many smaller deletions also

occur at the same locus, but go undetected. 5) This *Alu*-based human genome instability model may be used to evaluate the genetic risk posed by *Alu* structure-based variation which is unique to specific individuals, families, and people groups.

Methods

Data acquisition and flow

The hg19, 2009 Human Genome Assembly was used for this study. Retrotransposon data was obtained from RepeatMasker (Smit et al. 1996-2010) and downloaded from the UCSC genome BLAT Table Browser (http://genome.ucsc.edu/cgi-bin/hgTables?). This data was imported into Excel 2010 (Microsoft Corporation; Redmond, Washington). Statistics were calculated using Excel 2010 output using Minitab 15 and Minitab 16 (Minitab Inc.; State College, Pennsylvania).

Identification of the key variables that correlate with the human *Alu* pair I:D ratio

Three variables were found to significantly correlate with the *Alu* pair I:D ratio. These three variables are 1) the spacer size separating the two members of the *Alu* pair, 2) the number of *Alu* elements within the spacer separating the two members of the *Alu* pair and 3) the clustering state (clustered or not clustered) of the each member of the *Alu* pair.

The *Alu* pair I:D ratio was not found to correlate strongly with *Alu* size. The only exception to this observation occurs between the first 10 immediate *Alu* neighbors of small-small and small-medium *Alu* pairs. Small *Alus* are between 30 and 135 bp in length and medium *Alus* are between 136 and 274 bp in length. This anomaly involves less than 0.2 percent of the *Alu* pair population. Manual inspection of several of these

loci suggests that this phenomena results from these smaller *Alu* fragments being incorporated into tandem repeats (data not shown). Incorporation of *Alu* fragments into tandem repeats lowers the I:D ratio for pairs of this size.

Description of key variables – spacer size

The spacer is the intervening sequence between the two *Alu* elements which make up an *Alu* pair. Spacer size is the number of base pairs within this intervening sequence. Additional *Alu* elements may be present within the spacer sequence.

Description of key variables – *Alu* Pair Sequence Number (APSN)

The parameter describing the number of *Alu* elements incorporated within an *Alu* pair is termed the *Alu* pair sequence number (APSN). The APSN would ideally be defined as the number of *Alu* elements within the spacer sequence separating an *Alu* pair. However, the APSN uses either a positive or negative value to discriminate between pairs formed by *Alus* located either 5' (negative) or 3'(positive) of each *Alu* being evaluated. As a result, mathematical confounding of 5' and 3' adjacent pairs precludes the use of zero to describe this parameter. The APSN is consequently defined as the "n+1" number of *Alu* elements within the spacer.

Description of key variables – clustering

Human retrotransposons*, Alu* LINE and SVA elements, frequently cluster together in groups we previously defined as CLIQUEs, catenated *L*INE1 endonuclease induced *q*ueues of uninterrupted *Alu*, LINE1 and SVA elements (Cook et al. 2011). Building on our original work, this study found that the *Alu* pair I:D ratio is a strong function of the clustering state of *Alu* pairs (Figure 3.4, pages 66-67). Four types of

clustered *Alu* pairs exist and are identified as types 0, 1, 2 and 3. Type 0 and Type 1 *Alu* pairs are located within CLIQUEs. Type 0 *Alu* pairs are formed when both members of the pair reside within the same CLIQUE and Type 1 *Alu* pairs are formed when both members of the pair reside within different CLIQUEs. Type 0 pairs are rare (<0.5 percent of human *Alu* pair population) and because of inherent orientational *Alu* biases within a CLIQUE, require a different methodology than I:D ratio to determine instability (Cook et al. 2011). This methodology is discussed separately under the heading entitled, "Determination of *Alu* pair instability within CLIQUEs"**,** below**.** Type 2 *Alu* pairs are hemi-clustered. This category of *Alu* pairs occurs where only one of the two *Alus* making up the pair resides within a CLIQUE. Type 3 *Alu* pairs are non-clustered. Figure 3.4, pages 66-67, illustrates the relationship of I:D ratio among different clustering conformations within the human *Alu* pair population.

Algorithm development for estimating *Alu* pair I:D ratio from key variables

Segregation of the separate contributions of spacer size, APSN and clustering to the *Alu* pair I:D ratio was accomplished with a five-step methodology.

Step one in *Alu* pair I:D ratio algorithm development was determination of the full-size *Alu* pair population (275-325 bp) with its associated I:D ratio for each APSN (from APSN=±1 through APSN=±110). This information is available from previously published work (for APSNs 1-107) that utilized the human genome assembly hg18 as its resource (Cook et al. 2011). This study updated the earlier work using improved techniques and the most recent human genome assembly, hg19. The improved techniques permitted extending the number of statistically significant APSNs from 107 to 110.

Step two in I:D ratio algorithm development was accomplished by stepping through each of the populations of APSNs 1-110 in small (0.03-0.05%) spacer size increments. The population of *Alu* pair types 1, 2 and 3 (clustered, hemi-clustered and non-clustered) are determined within each increment. The resultant data set for each APSN and *Alu* pair type was then sorted into ten percentile groups. The first percentile accounts for the smallest five percent of the spacer sizes and the remaining nine percentiles capture sequential groupings of approximately ten percent of the APSN's *Alu* pair population. Each of these final nine percentiles is identified by its respective median point; 10^{th}, 20^{th}, 30^{th} etc., through 90^{th} percentiles. The spacer size boundaries for these final nine percentile groupings include $\pm$ 5% of the *Alu* pair population for the APSN being evaluated. As examples, the 10^{th} percentile describes the grouping that includes spacer sizes ranging between the 5^{th} and 15^{th} percentiles, the 20^{th} percentile describes the spacer sizes falling between the 15^{th} and 25^{th} percentiles, etc. The *Alu* pair sample size for most APSN populations falls between 550,000 and 560,000. The only exceptions are the APSNs 1-4. These APSN families increase in population size from 461,054 to 548,606 because of CLIQUE (clustering) effects. An *Alu* pair population size above 507,000 is required to provide statistical confidence that an I:D value $\leq$0.995 is below unity ($p<0.05$)

The percentile groupings are further reduced in size by subdividing them into their respective *Alu* pair types. The median spacer sizes along with actual and fitted I:D ratios for Type 1 *Alu* pairs are shown in Table A3.2 (page 114). As shown in Table A3.5, page118) sample sizes across these spacer size percentile groupings reduce the sample size to as low as 2,611 for the 0-5^{th} percentile grouping for Type 1 *Alu* pairs for

APSN=1. The average sample size for the larger percentiles (APSN>1) is 18,574. This sample size problem for measuring the I:D ratio for individual APSNs within percentiles and *Alu* pair types is addressed in step three of this five-step methodology.

Step three in *Alu* I:D ratio algorithm development plots each of the ten percentile groupings for APSNs 1 through 115 against its median spacer size. This approach increases the population size for each percentile grouping by approximately 115X and permits more accurate estimation of the actual I:D ratio at each APSN (see Figure A3.2, page 145). The smallest of these 115X sample sizes is 693,930 for the 2.5th percentile of Type 1 *Alu* pairs. This sample size is larger than the 507,000 minimum sample size (see step two, above) required for I:D values of <0.995 to be statistically less than unity ($p<0.05$). Examination of these 115 groupings revealed that for APSNs >110, no percentile grouping dropped below the minimum statistically significant I:D value of 0.995 ($p\leq0.05$). Consequently, only APSNs of 1 through 110 were used in the construction of the instability model algorithm.

A total of 30 regression curves are generated; 10 for Type 1 *Alu* pairs (clustered; 13,364,142 total full-length pairs), 10 for Type 2 *Alu* pairs (hemi-clustered; 28,537,478 total full-length pairs) and 10 for Type 3 *Alu* pairs (non-clustered; 18,836,832 total full-length pairs). Each set of percentile data is then regressed versus median spacer size. The resultant algorithm(s) which describe(s) the data for each respective percentile is then identified. In several instances the best fit for the data is accomplished by using a composite of two or more regressions for one set of percentile data. Examples of these curve fits are shown in Figure A3.2 (page 145) for the 2.5th percentile curves for Type 1, 2 and 3 *Alu* pairs.

Step four in development of the *Alu* I:D ratio prediction algorithm was the extraction of the respective I:D ratios for each of the ten percentiles for each APSN for *Alu* pair types 1, 2 and 3. Each regressed I:D ratio value was plotted for each APSN against its median spacer size. This step produces 345 different I:D curves, 115 curves for each *Alu* pair type. As mentioned previously, only APSN curves 1-110 had at least one point along the spacer size percentiles with an I:D ratio that was statistically below unity (0.995 = $p<0.05$). This technique excludes *Alu* pair type zero, which was treated separately (see heading, **"**Determination of *Alu* pair instability within CLIQUEs", below). An example of regressed data extracted from this step for Type 1 *Alu* pairs for APSNs 1-10 is shown in Figure 3.3 (page 65). Figure 3.4 (pages 66-67) shows the complete set of regressed I:D data (APSNs = ± 1-110) for Type 1, 2 and 3 *Alu* pairs.

Step five in development of the *Alu* pair instability algorithm development was the regression of the ten percentile data points derived from step four (above) for each of the 345 graphs. The shape of these curves often requires more than one regression equation to accurately portray these regressed values. In addition, median spacer size values below the 2.5th percentile and above the 90th percentile fall outside of the regressed region for these curves. Spacer sizes that are smaller than the median spacer size for the 2.5th percentile are assigned the I:D value of the 2.5th percentile. Straight lines connect the 2.5th percentile midpoints for the 5' and 3' curves for each APSN for each of the three *Alu* pair types shown in Figure 3.4 (page 65). Spacer sizes that are larger than the median spacer size for the 90th percentile are fit along a straight line from the I:D value at the 90th percentile to unity at the 99th percentile. The equation

types and associated coefficients for the ± 110 APSN curves associated with Type 1 *Alu* pairs are provided in Table A3.6, page 123.

Determination of *Alu* pair instability within CLIQUEs

Type 0 *Alu* pairs possess inherent *Alu* orientational insertion biases. This is reflected by the low CLIQUE I:D ratio = 0.460. These biases preclude the direct estimation of *Alu* pair instability from I:D measurements. However, less than 0.5% of human *Alu* pairs reside within the same CLIQUE. Most of these Type 0 *Alu* pairs have spacer sizes of ≤50 bp (Cook et al. 2011). Although these pairs represent a relatively small fraction of the total *Alu* pair population, their small spacer size may make a disproportionately large contribution to the total inverted *Alu* pair instability within the genome.

Type 0 *Alu* pairs possess inherent *Alu* orientational insertion biases. These biases preclude the direct estimation of *Alu* pair instability from I:D measurements (Cook et al. 2011). These directional biases are illustrated by comparing the CLIQUE I:D ratio versus the I:D ratio of the 2.5th spacer size percentile I:D ratio for Type 1 *Alu* pairs (0.460 versus 0.799, respectively). A solution to this stability prediction dilemma for Type 0 *Alu* pairs was resolved using data from previous work performed with a yeast experimental system. This system measured the instability of inverted *Alu* pairs when separated by 12, 20, 30 and 100 bp for homologies of 94% and 100% (Lobachev et al. 2000). Typical human *Alu* pair homologies are 85% (Stenger et al. 2001).

Fortunately, the median spacer size for adjacent Type 1 (clustered) *Alu* pairs in 0th-5th percentile range was 100 bp (Table A3.2, page 114). This data point,

representing 2,611 *Alu* pairs (Table A3.5, page 118), is one of the four spacer sizes evaluated for its inverted *Alu* pair instability in the experimental yeast system. This data point was used to anchor the 85% *Alu* homology curve to the 94% and 100% homology curves used in the yeast experiments. (Lobachev et al. 2000). The resultant Type 0 *Alu* pair algorithm for estimating inverted *Alu* pairs with 85% homologies is as follows.

$$0.7804 - (3.0271 \times e^{(-0.164251 \times \text{SpacerSize,bp})})$$

This algorithm is used to predict the I:D ratio for Type 0 *Alu* spacer sizes ≤50 bp. The algorithms developed for Type 1 *Alu* pairs were used to estimate Type 0 *Alu* pairs with spacer sizes >50 bp.

Instability estimate for individual *Alu* elements within an *Alu* pair

The I:D ratio is the stability of an *Alu* pair, not the stability of an individual *Alu* element. The instability of an individual *Alu* element within an *Alu* pair is estimated as the square root of the I:D ratio estimated for that pair. Depending upon the single-strand cleavage pattern at its eight potential cleavage sites, the resolution of the hypothetical doomsday junction can result in some level of gene conversion and/or from zero to four DSBs ((Cook et al. 2011) and Figures 3.1 and 3.2, pages 62-63).

Each of the *Alu* pair types represented in Figure 3.4 (pages 66-67) is composed of ±110 APSN versus spacer size curves. Each of these curves contain at least one percentile along their spacer size interval where the I:D ratio is <0.995. The I:D<0.995 cutoff represents the statistical confidence interval for full-length *Alu* pair families (P<0.05). These curves permit the maximum inverted *Alu* pair interaction distance to be

increased from the previously reported value of APSN=±107 to APSN=±110 (Cook et al. 2011). Any predicted I:D ratio that is >0.995 is assigned a value of 1.0.

Alu element stability and *i*Score determination

The stability of an *Alu* element is the grand product of the square root of the I:D ratios calculated for each of the *Alu* pairs formed by its ±110 immediately flanking (5' and 3') *Alu* elements. This stability is expressed by the following equation.

$$\text{Stability of an } Alu \text{ element} = \prod_{APSN=-110}^{APSN=110} \sqrt{I:D(APSN)}$$

The stability of each of these 220 flanking *Alu* pairs is determined from the previously developed I:D versus spacer size versus APSN algorithms. Direct oriented *Alu* pairs are considered stable and assigned a value of 1. The *i*Score is the inverse of the estimated stability of an *Alu* element and is used only in Figure 3.5 (pages 69-70) and in Figure A3.1 (page 137) to illustrate the relative stabilities of the various *Alu* elements located within a gene's *Alu* landscape.

Since each end of an *Alu* element is subject to a potential deletion, the stability of only one end of each *Alu* element is the grand product of the fourth root of the I:D ratio for all 220 potential *Alu-Alu* interactions. This stability is expressed as follows.

$$\text{Stability of either end of an } Alu \text{ element} = Alu_{End} = \prod_{APSN=-110}^{APSN=110} \sqrt[4]{I:D(APSN)}$$

Estimation of deletion size probability

Two studies provided insight into the human deletion frequency distribution (Wheeler et al. 2008; Mills et al. 2011). Recent cancer studies also provide similar information. However, the unique nature of cancer cells precludes the use of this data in the characterization of DNA stability in healthy cells. In this study, human indel size frequency curves are treated as having the same shape as the corresponding human deletion size frequency curve.

The deletion size frequency curve in Figure 3.6 (page 72) was prepared from a composite of data provided in the two studies mentioned, above. The first study, Wheeler et al., 2008, provides a deletion size frequency curve that was used to estimate the deletion size frequency for deletion sizes ≤75 bp. The second study (Mills et al. 2011), is used to estimate the deletion size frequency for deletion sizes >75 bp. Modeling of the deletion/indel size frequency data from both studies excluded the *Alu* insertion perturbation present between 250 and 350 bp. This permitted smoothing of the respective regression fits.

In the first study, deletion frequency data was regressed between 1 and 400 bp and for the second study, the indel frequency data was regressed between 50 and 10,000 bp. In both studies over 95% of deletions/indels were ≤50 bp. The second study (Mills et al. 2011) used a higher number of individuals (79) and thus supplied additional data for the more rare larger deletion sizes.

The sum of the 500,000 individual deletion size probabilities illustrated in Figure 3.6 (page 72) equal 1.0. The probability of a specific deletion size occurring is lower

than the probability of that same or larger deletion size occurring. This latter probability of a "minimum required deletion size or larger" required for loss of coding sequence is used in the model's algorithm.

The model's algorithm considers each end of each *Alu* element separately in its determination of exon and gene stability. Estimation of the risk that an *Alu* end poses to an exon coding sequence first requires that the distance between the end of the *Alu* element and the proximal end of the exon be determined. This distance is defined as D_{Min}. The formula that describes the probability of a minimum deletion size is as follows.

$$D_{Min} = \text{Probability of a specific deletion size (or larger)} = P_{Deletion}$$

$$P_{deletion} = \sum_{d=D_{(min)}}^{d=500,000} \text{deletion fraction}(d)^{*}$$

* deletion fractions are taken from Figure 3.6

Determination of relative exon instability

Individual exon instabilities are calculated through a five-step process. Step one is calculating the DSB risk posed by each end of each *Alu* element ($Risk_{End}$) within a gene's $\pm$ 250,000 bp *Alu* landscape. Step two is determining the potential deletion size risk, $P_{Deletion}$, posed by each end of each *Alu* element within this landscape, to the coding exon of interest. Step three is multiplying each individual $Risk_{End}$ value by its respective $P_{Deletion}$ value. Step four calculates the grand product of these "$Risk_{End}$ x

$P_{Deletion}$" products. This estimated relative exon stability, $Exon_{RS}$, is expressed by the following formula.

$$Exon_{RS} = \prod_{N=3'\text{end of the 3' most } Alu \text{ in } +/-250{,}000\text{bp flanking landscape}}^{N=5'\text{end of the 5' most } Alu \text{ in } +/-250{,}000\text{bp flanking landscape}} Alu_{End}(N)\, P_{Deletion}(N)$$

Step five was determining exon instability. Since exon stability plus exon instability equals one, the exon instability is one minus the estimated exon stability derived from the formula above.

Determination of relative gene instability

Relative gene instability is defined as the relative likelihood of a deletion occurring at some location within a gene's coding exons. This is determined through a four-step process. The first three steps are identical to the first three steps described under the "Determination of relative exon instability" heading above. Step three in this procedure is only performed for the closest exon to each *Alu* element end. This step determines the highest risk, $Risk_{Max}$, that one end of an *Alu* element can pose to a gene. Step four multiplies each of these, $Risk_{Max}$, values determined for each *Alu* end. This grand product produces the estimate of that gene's relative stability, $Gene_{RS}$.

$$Gene_{RS} = \prod_{N=3'\text{end of the 3' most } Alu \text{ in } +/-250{,}000\text{bp flanking landscape}}^{N=5'\text{end of the 5' most } Alu \text{ in } +/-250{,}000\text{bp flanking landscape}} Risk_{Max}$$

Step five determines the gene instability. Since the stability of a gene plus its instability equals 1.0, gene instability is one minus the estimated gene stability derived from the formula above.

Gene selection

The 50 random human genes used in this study were selected from the list of 19,026 human protein-coding genes provided by the HUGO Gene Nomenclature Committee, HGNC. The source file containing these genes was downloaded from the HGNC website (Seal et al. 2011). The 50 random genes were selected from this list using Minitab 16.

The 50 deletion-prone cancer genes were selected from (Solimini et al. 2012; Stephens et al. 2012) and the Catalogue of Somatic Mutations in Cancer (Forbes et al. 2011) web page entitled, "Cancer genes that have deletion mutations", http://www.sanger.ac.uk/genetics/CGP/Census/large_deletion.shtml. Only coding exons were selected for each gene. Exon loci were obtained from the RefSeq CDS Fasta Alignment page on the UCSC genome browser, http://genome.ucsc.edu/cgi-bin/hgPal. Variant 1 isoforms of all genes were selected when more than one gene was listed under RefSeq genes.

References

Aleshin A, Zhi D. 2010. Recombination-associated sequence homogenization of neighboring Alu elements: signature of nonallelic gene conversion. *Molecular biology and evolution* **27**(10): 2300-2311.

Bailey JA, Liu G, Eichler EE. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. *American journal of human genetics* **73**(4): 823-834.

Batzer MA, Deininger PL. 2002. *Alu* repeats and human genomic diversity. *Nature reviews Genetics* **3**(5): 370-379.

Cook GW, Konkel MK, Major JD, 3rd, Walker JA, Han K, Batzer MA. 2011. *Alu* pair exclusions in the human genome. *Mobile DNA* **2**: 10.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet* **7**(12): e1002384.

de Smith AJ, Walters RG, Coin LJ, Steinfeld I, Yakhini Z, Sladek R, Froguel P, Blakemore AI. 2008. Small deletion variants have stable breakpoints commonly associated with *Alu* elements. *PloS one* **3**(8): e3104.

Deininger PL, Batzer MA. 1999. *Alu* repeats and human disease. *Mol Genet Metab* **67**(3): 183-193.

Fogedby HC, Metzler R. 2007. Dynamics of DNA breathing: weak noise analysis, finite time singularity, and mapping onto the quantum Coulomb problem. *Physical review E, Statistical, nonlinear, and soft matter physics* **76**(6 Pt 1): 061915.

Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, Jia M, Shepherd R, Leung K, Menzies A et al. 2011. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids research* **39**(Database issue): D945-950.

Franke G, Bausch B, Hoffmann MM, Cybulla M, Wilhelm C, Kohlhase J, Scherer G, Neumann HP. 2009. *Alu-Alu* recombination underlies the vast majority of large VHL germline deletions: Molecular characterization and genotype-phenotype correlations in VHL patients. *Human mutation* **30**(5): 776-786.

Girirajan S, Campbell CD, Eichler EE. 2011. Human copy number variation and complex genetic disease. *Annual review of genetics* **45**: 203-226.

Hedges DJ, Deininger PL. 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. *Mutation research* **616**(1-2): 46-59.

Jeon JH, Sung W, Ree FH. 2006. A semiflexible chain model of local denaturation in double-stranded DNA. *The Journal of chemical physics* **124**(16): 164905.

Kass DH, Batzer MA, Deininger PL. 1995. Gene conversion as a secondary mechanism of short interspersed element (SINE) evolution. *Molecular and cellular biology* **15**(1): 19-25.

Kitada K, Aikawa S, Aida S. 2012. *Alu-Alu* Fusion Sequences Identified at Junction Sites of Copy Number Amplified Regions in Cancer Cell Lines. *Cytogenetic and genome research*.

Konkel MK, Batzer MA. 2010. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Seminars in cancer biology* **20**(4): 211-221.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lee J, Han K, Meyer TJ, Kim HS, Batzer MA. 2008. Chromosomal inversions between human and chimpanzee lineages caused by retrotransposons. *PloS one* **3**(12): e4047.

Lev-Maor G, Ram O, Kim E, Sela N, Goren A, Levanon EY, Ast G. 2008. Intronic *Alus* influence alternative splicing. *PLoS genetics* **4**(9): e1000204.

Lobachev KS, Stenger JE, Kozyreva OG, Jurka J, Gordenin DA, Resnick MA. 2000. Inverted *Alu* repeats unstable in yeast are excluded from the human genome. *The EMBO journal* **19**(14): 3822-3830.

Lupski JR. 2010. Retrotransposition and structural variation in the human genome. *Cell* **141**(7): 1110-1112.

Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, Kemeza DM, Strassler DS, Ponting CP, Webber C et al. 2011. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome research* **21**(6): 830-839.

Roy AM, Carroll ML, Nguyen SV, Salem AH, Oldridge M, Wilkie AO, Batzer MA, Deininger PL. 2000. Potential gene conversion and source genes for recently integrated *Alu* elements. *Genome research* **10**(10): 1485-1495.

Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. 2011. genenames.org: the HGNC resources in 2011. *Nucleic acids research* **39**(Database issue): D514-519.

Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, Dyer M, Cordaux R, Liang P, Batzer MA. 2006. Human genomic deletions mediated by recombination between *Alu* elements. *American journal of human genetics* **79**(1): 41-53.

Serrano M. 2010. Cancer: a lower bar for senescence. *Nature* **464**(7287): 363-364.

Smit AFA, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0 In *[http://www repeatmaskerorg]*.

Solimini NL, Xu Q, Mermel CH, Liang AC, Schlabach MR, Luo J, Burrows AE, Anselmo AN, Bredemeyer AL, Li MZ et al. 2012. Recurrent hemizygous deletions in cancers may optimize proliferative potential. *Science* **337**(6090): 104-109.

Stenger JE, Lobachev KS, Gordenin D, Darden TA, Jurka J, Resnick MA. 2001. Biased distribution of inverted and direct *Alus* in the human genome: implications for insertion, exclusion, and genome stability. *Genome research* **11**(1): 12-27.

Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, Nik-Zainal S, Martin S, Varela I, Bignell GR et al. 2012. The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**(7403): 400-404.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT et al. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189): 872-876.

Witherspoon DJ, Watkins WS, Zhang Y, Xing J, Tolpinrud WL, Hedges DJ, Batzer MA, Jorde LB. 2009. *Alu* repeats increase local recombination rates. *BMC genomics* **10**: 530.

Zhang Y, Romanish MT, Mager DL. 2011. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS computational biology* **7**(5): e1002046.

Zhi D. 2007. Sequence correlation between neighboring *Alu* instances suggests post-retrotransposition sequence exchange due to *Alu* gene conversion. *Gene* **390**(1-2): 117-121.

CHAPTER FOUR:
CONCLUSIONS

The development of this DNA structure-based model for assessing relative human gene instabilities is a testimony to the value of comparative genomics. The past decade has witnessed the completion of the Human Genome Project, the sequencing of the genomes of all great apes as well as the sequencing of several other members of the primate family. These achievements have made it possible to identify chimpanzee specific deletions in orthologous inverted *Alu* pair loci in human, gorilla, orangutan and rhesus macaque genomes. This ability facilitiated the validation of the *Alu* pair I:D imbalance observed through bioinformatics analysis of the human genome. This potential for advanced analysis of the genome has made the research for this dissertation possible.

The continued development of genome sequencing technologies should permit further improvements in genomic modeling systems that will be critical to the future of genomics research. It is hoped that the techniques and principles used in the development of this *Alu* element-based model of human genome instability will provide the groundwork for more advanced computational approaches to recognizing human genome instability. The value of personal genomics will be substantially enhanced by the increased capability of researchers to evaluate the genetic risks posed by structural variations which are unique to specific individuals, families and people groups.

APPENDIX A:
SUPPLEMENTAL INFORMATION

Table A3.1
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
1	F. Duraturo	2013	2p21	LS	Contribution of Large Genomic Rearrangements in Italian Lynch Syndrome Patients:Characterization of a Novel *Alu*-Mediated Deletion	BRI	'13	1
2	K. Kitada	2013	7q22.1	C	*Alu-Alu* Fusion Sequences Identified at Junction Sites of Copy Number Amplified Regions in Cancer Cell Lines	CGR	139	1
3	C. Vaughn	2013	7p22.1	LS	The Frequency of Previously Undetectable Deletions Involving 3' Exons of the *PMS2* Gene	GCC	52	1
4	M. Barbaro	2012	3q11.2	HCP	Identification of an *AluY* -mediated deletion of exon 5 in C*POX* gene by MLPA analysis in patients with hereditary coproporphyria	CG	81	3
5	N. Bondurand	2012	22q13.1	WS IV	*Alu*-mediated deletion of *SOX10* regulatory elements in Waardenburg syndrome type 4	EJHG	20	9
6	V. Chanavat	2012	11p11.2	HCM	Molecular characterization of a large *MYBPC3* rearrangement in a cohort of 100 unrelated patients with hypertrophic cardiomyopathy	EJMG	55	3
7	M. Coutinho	2012	12q23.2	ML II	*Alu-Alu* Recombination Underlying the First Large Genomic Deletion in GlcNAc-Phosphotransferase Alpha/Beta (*GNPTAB*) Gene in a MLII Alpha/Beta Patient	JIMD	4	1
8	A. Eiden-Plach	2012	8p11.23	LCAH	*Alu Sx* repeat-induced homozygous deletion of the StAR gene causes lipoid congenital adrenal hyperplasia	JSBMB	130	1-2
9	A. Gonçalves	2012	17p13.1	LFS	Li-Fraumeni-like syndrome associated with a large BRCA1 intragenic deletion	BMCC	12	1
10	A. Jelassi	2012	19p13.2	ADH	Genomic characterization of two deletions in the *LDLR* gene in Tunisian patients with familial hypercholesterolemia	CCA	414	-

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
11	H. Mahmoudi	2012	13q14.2	HY	Identification of an *Alu*-mediated 12.2-kb deletion of the complete *LPAR6 (P2RY5)* gene in a Turkish family, hypotrichosis and wooly hair	ED	21	6
12	M. Pereira	2012	15q21.1	SPG 11	*Alu* elements mediate large *SPG11* gene rearrangements: further spatacsin mutations	GM	14	1
13	L. Pezzoli	2012	11p11.2	HCM	A new mutational mechanism for hypertrophic cardiomyopathy	GE	507	2
14	M. Vlckova	2012	6q	M	Mechanism and Genotype-Phenotype Correlation of Two Proximal 6q Deletions Characterized Using mBAND, FISH, Array CGH, and DNA Sequencing	CGR	136	1
15	T. Arai	2011	Xq22.1	XLA	Genetic analysis of contiguous X-chromosome deletion syndrome encompassing the BTK and TIMM8A genes	JHG	56	8
16	P. Boone	2011	2p22.3	SPG IV	*Alu*-specific microhomology-mediated deletion of the final exon of SPAST in three unrelated subjects with hereditary spastic paraplegia	GM	13	6
17	G. Borck	2011	6p24.3-2	CC	An *Alu* repeat-mediated genomic GCNT2 deletion underlies congenital cataracts and adult i blood group	HG	131	2
18	M. Cozar	2011	1q22	GD	Molecular characterization of a new deletion of the GBA1 gene due to an inter *Alu* recombination event	MGM	102	2
19	I. Guella	2011	1q24.2	FVD	Identification of the first *Alu*-mediated large deletion involving the F5 gene in a compound heterozygous patient with severe factor V deficiency	JTH	106	2
20	X. Guo	2011	22q11.2	DGS	Characterization of the past and current duplication activities in the human 22q11.2region	BMCG	12	71
21	I. Jennes	2011	8q24.11	MO	Breakpoint characterization of large deletions in EXT1 or EXT2 in 10 Multiple Osteochondromas families	BMCMG	12	85
22	R. Kuiper	2011	2p21	LS	Recurrence and Variability of Germline EPCAM Deletions in Lynch Syndrome	HGVS	32	4

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
23	M. Kurnikova	2011	1q21.3	SCN	*Alu*-Mediated Recombination in the HAX1 Gene as the Molecular Basis of Severe Congenital Neutropenia	AJMG	155A	3
24	M. Legarda	2011	16q22.2	T II	Large TAT deletion in tyrosinaemia type II patient	MGM	104	3
25	J. Oshima	2011	Xq28	MPS II	LCR-initiated rearrangements at the IDS locus, completed with *Alu*-mediated recombination or non-homologous end joining	JHG	56	7
26	L. Perez-Cabornero	2011	2p21	LS	Characterization of New Founder Alu-Mediated Rearrangements in MSH2 Gene Associated with a Lynch Syndrome Phenotype	CPR	4	10
27	H. Raef	2011	11q13.1	MEN I	A novel deletion of the MEN1 gene in a large family of multiple endocrine neoplasia type 1 (MEN1) with aggressive phenotype	CE	75	6
28	A. Rose	2011	19q13.42	RP	A 112kb deletion in chromosome 19q13.42 leads to retinitis pigmentosa	IOVS	52	9
29	M. Sluiter	2011	17q21.31	BC	Large genomic rearrangements of the BRCA1 and BRCA2 genes: review of the Literature and report of a novel BRCA1 mutation	BCRT	125	2
30	M. Soejima/ Y. Koda	2011	19q13.33	BP	TaqMan-based real-time polymerase chain reaction for detection of FUT2 copy number Variations: identification of novel *Alu*-mediated deletion	T	51	4
31	J. Wan	2011	19p13.2	EA II	Large genomic deletions in CACNA1A cause episodic ataxia type 2	FN	2	-
32	K. Champion	2010	17q21.2	SS B	Identification and characterization of a novel homozygous deletion in the α-N-acetyl -glucosaminidase gene in a patient with Sanfilippo type B syndrome	MGM	100	1
33	M. DeRosa	2010	19p13.3	PJS	*Alu*-Mediated Genomic Deletion of the Serine/Threonine Protein Kinase 11 (STK11) Gene in Peutz-Jeghers Syndrome	G	138	7
34	M. Gentsch	2010	1q25.3	CGD	*Alu*-Repeat--Induced Deletions Within the NCF2 Gene Causing p67-phoxi-Deficient Chronic Granulomatous Disease (CGD)	HGVS	31	2
35	A. Janecke	2010	11q23.1	PGL	Identification of a 4.9-kilo base-pair *Alu*-mediated founder SDHD deletion in two extended paraganglioma families From Austria	JHG	55	3

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
36	M. Kleppe	2010	18p11.21	T-ALL	Deletion of protein tyrosine phosphatase gene PTPN2 in T-cell acute NG lymphoblastic leukemia	NG	42	6
37	A. Lindstrand	2010	10p14	HDR	Molecular and Clinical Characterization of Patients with AJMG Overlapping 10p Deletions	152A	5	-
38	M. McCabe	2010	19p13.3	PJS	Homozygous Deletion of the STK11/LKB1 Locus and the Generation of Novel Fusion Transcripts in Cervical Cancer Cells	CGC	197	2
39	M. Phylipsen	2010	16p13.3	αT	A new α0-thalassemia deletion found in a Dutch family (- AW)	BCMD	45	2
40	V. Picard	2010	1q25.1	AT I	Detection and characterization of large SERPINC1 deletions in type I inherited antithrombin deficiency	HG	127	1
41	N. Resta	2010	19p13.3	PJS	Breakpoint determination of 15 large deletions in Peutz-Jeghers subjects	HG	128	4
42	Z. Yang	2010	Xq24	DD	LAMP2 Microdeletions in Patients with Danon Disease	CCG	3	2
43	F. Zhang	2010	17p12	N	Mechanisms for Nonrecurrent Genomic Rearrangements Associated with CMT1A or HNPP: Rare CNVs as a Cause for Missing Heritability	AJHG	86	6
44	L. Desviat	2009	13q32.3	PA	High frequency of large genomic deletions in the PCCA gene causing propionic acidemia	MGM	96	4
45	A. Erez	2009	Xp22.13	RTT	*Alu*-specific microhomology-mediated deletions in CDKL5 in females with early-onset seizure disorder	N	10	4
46	G. Franke	2009	3p25.3	VHL	*Alu-Alu* Recombination Underlies the Vast Majority of Large VHL Germline Deletions:Molecular Characterization and Genotype--Phenotype Correlations in VHL Patients	HM	30	5
47	C. Oliveria	2009	16q22.1	HDGC	Germline CDH1 deletions in hereditary diffuse gastric cancer families	HMG	18	9
48	A. Pangrazio	2009	11q13.2	ARO	Characterization of a Novel *Alu-Alu* Recombination-Mediated Genomic Deletion in the TCIRG1 Gene in Five Osteopetrotic Patients	JBMR	24	1

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
49	R. Quental	2009	Xp11.4	OTCD	Molecular mechanisms underlying large genomic deletions in ornithine transcarbamylase (OTC) gene	CG	75	5
50	H. Singh	2009	Xq28	BS	A Novel *Alu*-Mediated Xq28 Microdeletion Ablates TAZ and Partially Deletes DNL1L in a Patient with Barth Syndrome	AJMG	149A	5
51	A. Mohl	2008	12p13.31	VWD	An *Alu*-mediated novel large deletion is the most frequent cause of type 3 von Willebrand disease in Hungary	JTH	6	10
52	S. Quental	2008	19q13.2	MSUD	Maple syrup urine disease due to a new large deletion at BCKDHA caused by non-homologous recombination	JIMD	31	2
53	M. Zikan	2008	17q21.31	BC	Novel complex genomic rearrangement of the *BRCA1* gene	MR	637	1-2
54	S. Armaou	2007	17q21.31	BC	Novel genomic rearrangements in the *BRCA1* gene detected in Greek breast/ovarian cancer patients	EJC	43	2
55	E. Costa	2007	7q11.21	SDS	Identification of a novel *AluSx*-mediated deletion of exon 3 in the SBDS gene in a patient with Shwachman-Diamond syndrome	BCMD	39	1
56	T. Fukao	2007	Xp22.13	XLG	Identification of *Alu*-mediated, large deletion-spanning introns 19-26 in PHKA2 in a patient with X-linked liver glycogenosis (hepatic phosphorylase kinase deficiency)	MGM	92	1-2
57	B. Hayward	2007	2p22.3	L	Extensive Gene Conversion at the PMS2 DNA Mismatch Repair Locus	HM	28	5
58	M. Okubo	2007	8p21.3	LPL	A novel complex deletion--insertion mutation mediated by *Alu* repetitive elements leads to lipoprotein lipase deficiency	MGM	92	3
59	M. Smyk	2007	Xp21.2	AHC	Male-to-female sex reversal associated with ~250 kb deletion upstream of *NR0B1 (DAX1)*	HG	122	1
60	E. Di Pierro	2006	11q23.3	AIP	A large deletion on chromosome 11 in acute intermittent porphyria	BCMD	37	1
61	A. Fukuuchi	2006	11q13.1	MEN I	A Whole MEN1 Gene Deletion Flanked by *Alu* Repeats in a Family with Multiple JJCO Endocrine Neoplasia Type 1	36	11	-

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
62	C. Has	2006	20p12.3	KS	Molecular Basis of Kindler Syndrome in Italy: Novel and Recurrent *Alu/Alu* Recombination, Splice Site, Nonsense, and Frameshift Mutations in the KIND1 Gene	JID	126	8
63	G. Humbert	2006	15q26.1	RPA	Homozygous Deletion Related to *Alu* Repeats in RLBP1 Causes Retinitis Punctata Albescens	IOVS	47	11
64	V. Matejas	2006	17p12	HNPP	Identification of *Alu* elements mediating a partial PMP22 deletion	N	7	2
65	S. Preisler-Adams	2006	17q21.31	BC	Gross rearrangements in BRCA1 but not BRCA2 play a notable role in predisposition to breast and ovarian cancer in high-risk families of German origin	CGC	168	1
66	F. Xie	2006	12p13.31	VWD	A novel *Alu*-mediated 61-kb deletion of the von Willebrand factor (VWF) gene whose breakpoints co-locate with putative matrix attachment regions	BCMD	36	3
67	G. Zhang	2006	6q27	T2D	Identification of *Alu*-mediated, large deletion-spanning exons 2-4 in a patient with mitochondrial acetoacetyl-CoA thiolase deficiency	MGM	89	3
68	S. Agata	2005	13q13.1	BC	Large genomic deletions inactivate the BRCA2 gene in breast cancer families	JMG	42	10
69	C. Bergmann	2005	6p12.3-2	ARPKD	Multi-exon deletions of the PKHD1 gene cause autosomal recessive polycystic kidney disease (ARPKD)	JMG	42	10
70	F. Charbonnier	2005	2p21	HNPCC	The 5' Region of the MSH2 Gene Involved in Hereditary Non-Polyposis Colorectal Cancer Contains a High Density of Recombinogenic Sequences	HM	26	3
71	F. del Castillo	2005	13q12.11	ARNSHI	A novel deletion involving the connexin-30 gene, del(GJB6-d13s1854), found in trans with mutations in the *GJB2* gene (connexin-26) in subjects with DFNB1 non-syndromic hearing impairment	JMG	42	7

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
72	C. Dobson-Stone	2005	9q21.2	ChAc	Identification of a VPS13A founder mutation in French Canadian families with chorea-acanthocytosis	N	6	3
73	J. Douglas	2005	5q35.2-3	SS	Partial NSD1 deletions cause 5% of Sotos syndrome and are readily identifiable by multiplex ligation dependent probe amplification	JMG	42	9
74	C. Eng	2005	Xq22.1	FD	Molecular Basis of Fabry Disease: Mutations and Polymorphisms in the Human α-Galactosidase A Gene	HM	3	2
75	C. Giunta	2005	1p36.22	EDS	Mutation analysis of the PLOD1 gene: An efficient multistep approach to the molecular diagnosis of the kyphoscoliotic type of Ehlers-Danlos syndrome (EDS VIA)	MGM	86	1-2
76	S. Hsieh	2005	1p36	HCC	High-freq. *Alu*-mediated recomb./del. within the hCAD in hepatoma	O	24	43
77	H. van der Klift	2005	2p21	HNPCC	Molecular Characterization of the Spectrum of Genomic Deletions in the Mismatch Repair Genes MSH2, MLH1, MSH6, and PMS2 Responsible for HNPCC(1)	GCC	44	2
78	B. Baysal	2004	11q23.1	PGL	An *Alu*-mediated partial SDHC deletion causes familial and sporadic paraganglioma	JMG	41	9
79	U. Guenther	2004	11q13.3	SMARD1	Genomic rearrangements at the IGHMBP2 gene locus in two patients with SMARD1	HG	115	4
80	C. Hartmann	2004	17q21.31	BC	Large *BRCA1* Gene Deletions Are Found in 3% of German High-risk Breast Cancer Families	HM	24	6
81	F. Laccone	2004	Xq28	RS	Large Deletions of the *MECP2* Gene Detected by Gene Dosage Analysis in Patients With Rett Syndrome	HM	23	3
82	M. Mitchell	2004	4q35.2	FXID	An *Alu*-mediated 31.5-kb deletion as the cause of factor XI deficiency in 2 unrelated patients	B	104	8
83	S. Nakaya	2004	Xq28	HA	Severe HA(1) due to a 1.3 kb factor VIII gene deletion including exon 24: homologous recombination between 41 bp within an *Alu* repeat sequence in introns 23 and 24	JTH	2	11

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
84	L. Rossetti	2004	Xq28	HA	Homologous Recombination Between AluSx-Sequences as a Cause of Hemophilia	HM	24	
85	C. Silao	2004	1p21.2	MSUD	A novel deletion creating a new terminal exon of the dihydrolipoyl transacylase gene is a founder mutation of Filipino maple syrup urine disease	MGM	81	2
86	I. Tournier	2004	13q13.1	BC	Significant Contribution of Germline BRCA2 Rearrangements in Male Breast Cancer Familes	CR	64	22
87	M. Venturin	2004	17q11.2	NF1	Evidence for non-homologous end joining and non-allelic homologous recombination in atypical NF1 microdeletions	HG	115	1
88	C. Bergmann	2003	Xq12	XMR	Oligophrenin 1 (OPHN1) gene mutation causes syndromic BJN X-linked mental retardation with epilepsy, rostral ventricular enlargement and cerebellar hypoplasia	126	7	-
89	E. Jo	2003	Xq22.1	XLA	Identification of mutations in the Bruton's tyrosine kinase gene, including a novel genomic rearrangement resulting in large deletion, in Korean XLA(1) patients	JHG	48	6
90	V. Ricci	2003	Xq28	HD	An *Alu*-mediated rearrangement as cause of exon skipping in Hunter disease	HG	112	4
91	R. Shaji	2003	16p13.3	HbH	Determination of the breakpoint and molecular diagnosis of a common α-thalassaemia-1 deletion in the Indian population	BJH	123	5
92	Y. Wang	2003	2p21	HNPCC	Hereditary Nonpolyposis Colorectal Cancer: Frequent Occurrence of Large Genomic Deletions in *MSH2* and *MLH1* Genes	IJC	103	5
93	W. Balemans	2002	17q21.31	VBD	Identification of a 52 kb deletion downstream of the SOST gene in patients with van Buchem disease	JMG	39	2
94	Z. Guo	2002	9q31.1	TD	Double deletions and missense mutations in the first nucleotide-binding fold of the ATP-binding cassette transporter A1 (ABCA1) gene in Japanese patients with TD(1)	JHG	47	6
95	M. Huber	2002	10q24-25	EB	Deletion of the Cytoplasmatic Domain of BP180/Collagen XVII Causes a Phenotype with Predominant Features of Epidermolysis Bullosa Simplex	JID	118	1

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
96	M. Lutskiy	2002	Xp11.23	WAS	An *Alu*-mediated deletion at Xp11.23 leading to Wiskott-Aldrich syndrome	HG	110	5
97	K. Staehling-Hampton	2002	17q12-q21	VBD	A 52-kb Deletion in the SOST-MEOX1 Intergenic Region on 17q12-q21 is Associated With van Buchem Disease in the Dutch Population	AJMG	110	2
98	F. Vidal	2002	Xq28	HA	First Molecular Characterization of an Unequal Homologous *Alu*-mediated Recombination Event Responsible for Hemophilia	JTH	88	1
99	T. Yabe	2002	6p21.32	BLS	A subject with a novel type I bare lymphocyte syndrome has tapasin deficiency due to deletion of 4 exons by *Alu*-mediated recombination	B	100	4
100	X. Cao	2001	5q22.2	FAP	Topoisomerase-I- and *Alu*-mediated genomic deletions of the APC gene in familial adenomatous polyposis	HG	108	5
101	F. Ringpfeil	2001	16p13.11	PE	Compound Heterozygosity for a Recurrent 16.5-kb *Alu*-Mediated Deletion Mutation and Single-Base-Pair Substitutions in the ABCC6 Gene Results in PE(1)	AJHG	68	3
102	T. Wang	2001	13q13.1	BC	A Deletion/Insertion Mutation in the *BRCA2* Gene in a Breast Cancer Family: A GCC Possible Role of the *Alu*-polyA Tail in the Evolution of the Deletion	31	1	
103	S. Dabora	2000	16p13.3	TSC	Characterisation of six large deletions in *TSC2* identified using long range PCR suggests diverse mechanisms including *Alu* mediated recombination	JMG	37	11
104	M. Hiltunen	2000	14q24.2	EOAD	Identification of novel 4.6-kb genomic deletion in presenilin-1 gene which results in exclusion of exon 9 in a Finnish early onset core Alzheimer's disease family: an *Alu* sequence-stimulated recombination?	EJHG	8	4
105	Y. Koda	2000	19q13.33	BP	An *Alu*-mediated large deletion of the *FUT2* gene in individuals with the ABO-Bombay phenotype	HG	106	1

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

No.	First Author	Year	Locus	Pheno-type[1]	Title	Journal[2]	Vol	Issue
106	E. Rohlfs	2000	17q21.31	BC	An *Alu*-Mediated 7.1 kb Deletion of *BRCA1* Exons 8 and 9 in Breast and Ovarian GCC Cancer Families That Results in Alternative Splicing of Exon 10	28	3	-
107	Y. Saikawa	2000	22q12.3	HO-1	Structural Evidence of Genomic Exon-Deletion Mediated by *Alu-Alu* Recombination in a Human Case with Heme Oxygenase-1 Deficiency	HM	16	2
108	R. Suminaga	2000	Xp21.1-2	DMD	Non-homologous recombination between *Alu* and LINE-1 repeats caused a 430-kb deletion in the dystrophin gene: a novel source of genomic instability	JHG	45	6

(1) **Phenotype Abbreviations**

ADH - Autosomal Dominant Hypercholesterolemia
AHC - Congenital Adrenal Hypoplasia
AIP - Acute Intermittent Porphyria
ARNSHI - Autosomal Recessive Non-Syndromic Hearing Impairment
ARO - Autosomal Recessive Osteopetrosis
ARPKD - Autosomal Recessive Polycystic Kidney Disease
AT I - Antithrombin Deficiency Type I
BC - Breast Cancer
BLS - Type I Bare Lymphocyte Syndrome
BP - Bombay Phenotype
BS - Barth Syndrome
C - Cancer
CC - Congenital Cataracts
CGD - Chronic Granulomatous Disease
ChAc - Chorea-acanthocytosis
DD - Danon Disease
DGS - DiGeorge Syndrome (in paper #20, LCR22's are related to 3 other phenotypes)

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

(1)	**Phenotype Abbreviations**
DMD -	Duchenne Muscular Dystrophy
EA II -	Episodic Ataxia Type 2
EB -	Epidermolysis Bullosa Simplex
EDS -	Ehlers-Danlos Syndrome
EOAD -	Early Onset Alzheimer's Disease
FAP -	Familial Adenomatous Polyposis
FD -	Fabry Disease
FVD -	Factor V Deficiency
FXID -	Factor XI Deficiency
GD -	Gaucher Disease
HA -	Hemophilia A
HbH -	Haemoglobin H Disease
HCC -	Hepatocellular Carcinoma
HCM -	Hypertrophic Cardiomyopathy
HCP -	Hereditary Coproporphyria (other phenotypes mentioned in paper 34)
HD -	Hunter Disease
HDGC -	Hereditary Diffuse Gastric Cancer
HDR -	HDR Sydrome (Hypoparathyroidism, Sensorineural Deafness, Renal Dysplasia)
HNPCC -	Hereditary Non-Polyposis Colorectal Cancer
HNPP -	Hereditary Neuropathy with Liability to Pressure Palsies
HY -	Hypotrichosis
KS -	Kindler Syndrome
L -	Leukemia
LCAH -	Lipoid Congenital Adrenal Hyperplasia
LFS -	Li-Fraumeni Syndrome
LPL -	Lipoprotein Lipase Deficiency
LS -	Lynch Syndrome
M -	Microcephaly (in paper #14, other phenotypes related to 6q deletions are mentioned)
MEN I -	Multiple Endocrine Neoplasia Type I
ML II -	Mucolipidosis Type II α/β
MO -	Multiple Osteochondromas

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

(1) **Phenotype Abbreviations, continued**

MPS II -	Mucopolysaccharidosis Type II
MSUD -	Maple Syrup Urine Disease
N -	Neuropathy
NF1 -	Neurofibromatosis Type I
OTCD -	Ornithine Transcarbamylase Deficiency
PA -	Propionic Acidemia
PE -	Pseudoxanthoma Elasticum
PGL -	Paraganglioma
PJS -	Peutz-Jeghers Syndrome
RP -	Retinitis Pigmentosa
RPA -	Retinitis Punctata Albescens
RS -	Rett Syndrome
RTT -	Rett Syndrome
SCN -	Severe Congenital Neutropenia
SDS -	Shwachman-Diamond Syndrome
SMARD1 -	Spinal Muscular atrophy with Respiratory Distress Type I
SPG 11 -	Spastic Paraplegia Type 11
SPG IV -	Spastic Paraplegia Type IV
SS -	Sotos Syndrome
SS B -	Sanfilippo Syndrome Type B
T II -	Tyrosinaemia Type II
T2D -	T2-Defiency
T-ALL -	T-cell Acute Lymphoblastic Leukemia
TD -	Tangier Disease
TSC -	Tuberous Sclerosis Complex
VBD -	van Buchem Disease
VHL -	Von Hippel-Lindau Disease
VWD -	von Willebrand Disease
WAS -	Wiskott-Aldrich syndrome
WS IV -	Waardenburg Syndrome type IV
XLA -	X-linked Agammaglobulinemia

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

(1) **Phenotype Abbreviations, continued**

XLG -	X-linked Liver Glycogenosis
XMR -	X-Linked Mental Retardation
αT -	Alpha-thalassemia

(2) **Journal Titles**

AJHG -	The American Journal of Human Genetics
AJMG -	*American Journal of Medical Genetics*
B -	*Blood*
BCMD-	*Blood Cells, Molecules, and Diseases*
BCRT -	*Breast Cancer Research and Treatment*
BJH -	*British Journal of Haematology*
BJN -	*Brain. A Journal of Neurology*
BMCC -	*BioMed Central Cancer*
BMCG -	*BioMed Central Genomics*
BMCMG -	*BioMed Central Medical Genetics*
BRI -	*BioMed Research International*
CCA -	*Clinica Chimica Acta*
CCG -	*Circulation. Cardiovascular Genetics.*
CE -	*Clinical Endocrinology*
CG -	*Clinical Genetics*
CGC -	*Cancer Genetics and Cytogenetics*
CGR -	*Cytogenetic and Genome Research*
CPR -	*Cancer Prevention Research*
CR -	*Cancer Research*
ED -	*Experimental Dermatology*
EJC -	*European Journal of Cancer*
EJHG -	*European Journal of Human Genetics*
EJMG -	*European Journal of Medical Genetics*
FN -	*Frontiers in Neurology*
G -	*Gastroenterology*
GCC -	*Genes, Chromosomes & Cancer*

Table A3.1, continued
Studies Linking *Alu*-related Deletions to Human Disease Phenotypes

(2) **Journal Titles, continued**

GE -	Gene
GM -	*Genetics in Medicine*
HG -	*Human Genetics*
HGVS -	*Human Genome Variation Society*
HM -	*Human Mutation*
HMG -	*Human Molecular Genetics*
IJC -	*International Journal of Cancer*
IOVS -	*Investigative Ophthalmology and Visual Science*
JBMR -	*Journal of Bone and Mineral Research*
JHG -	*Journal of Human Genetics*
JID -	*The Journal of Investigative Dermatology*
JIMD -	*Journal of Inherited Metabolic Disease*
JJCO -	*Japanese Journal of Clinical Oncology*
JMG -	*Journal of Medical Genetics*
JSBMB -	*Journal of Steroid Biochemistry and Molecular Biology*
JTH-	*Journal of Thrombosis and Haemostasis*
MGM -	*Molecular Genetics and Metabolism*
MR -	*Mutation Research*
N -	*Neurogenetics*
NG -	*Nature Genetics*
O -	*Oncogene*
T -	*Transfusion*

Table A3.2

Actual and Fitted *Alu* Pair I:D Ratios Across Ten Spacer Percentiles, APSNs 1-115

(Type 1, Large-Large (275-325 bp) *Alu* Pairs; hg19 Human Genome Assembly)

APSN	0-5th Percentile			6-15th Percentile			16-25th Percentile			26-35th Percentile			36-45th Percentile			46-55th Percentile			56-65th Percentile			66-75th Percentile			76-85th Percentile			86-95th Percentile		
	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio
1	100	0.8007	0.7994	265	0.9194	0.9194	470	0.9117	0.9129	691	0.9236	0.9176	965	0.8848	0.8780	1,337	0.8339	0.8530	1,890	0.8452	0.8652	2,755	0.9074	0.8916	4,280	0.8999	0.9159	7,538	1.0042	0.9604
2	344	0.9194	0.9344	550	0.9768	0.9706	809	0.9462	0.9328	1,092	0.9098	0.9150	1,447	0.9006	0.8851	1,926	0.8759	0.8634	2,614	0.8698	0.8738	3,758	0.9140	0.8986	5,821	0.9224	0.9191	10,299	0.9039	0.9569
3	632	0.9751	0.9704	935	0.9471	0.9413	1,291	0.8922	0.9271	1,676	0.9063	0.9131	2,148	0.8994	0.8923	2,796	0.8808	0.8740	3,732	0.8952	0.8834	5,279	0.9232	0.9062	8,126	0.9163	0.9231	14,072	0.9453	0.9544
4	972	1.0094	0.9814	1,391	0.9282	0.9254	1,853	0.9307	0.9155	2,354	0.9040	0.9119	2,978	0.8930	0.8986	3,812	0.8909	0.8830	5,041	0.8837	0.8916	7,030	0.9192	0.9127	10,569	0.9123	0.9265	18,160	0.9478	0.9533
5	1,324	0.9920	0.9816	1,866	0.9177	0.9170	2,447	0.9276	0.9059	3,081	0.9063	0.9113	3,851	0.8841	0.9037	4,909	0.8876	0.8904	6,417	0.9132	0.8982	8,867	0.9283	0.9180	13,265	0.9346	0.9297	22,676	0.9473	0.9528
6	1,694	0.9800	0.9770	2,355	0.9019	0.9121	3,060	0.9086	0.8995	3,819	0.9466	0.9109	4,742	0.8843	0.9080	5,988	0.8822	0.8962	7,786	0.8836	0.9035	10,670	0.9062	0.9222	15,844	0.9783	0.9324	26,666	0.9630	0.9529
7	2,091	0.9520	0.9699	2,869	0.8808	0.9093	3,692	0.8774	0.8958	4,585	0.9074	0.9106	5,677	0.8782	0.9118	7,104	0.8843	0.9013	9,218	0.8825	0.9082	12,563	0.9117	0.9260	18,722	0.9099	0.9350	31,551	0.9687	0.9532
8	2,500	0.9419	0.9616	3,398	0.8949	0.9078	4,343	0.8935	0.8944	5,381	0.9144	0.9105	6,650	0.9143	0.9151	8,299	0.9086	0.9059	10,711	0.9287	0.9123	14,431	0.9257	0.9292	21,497	0.9601	0.9372	36,034	0.9431	0.9537
9	2,912	0.9458	0.9529	3,931	0.9183	0.9072	5,009	0.8873	0.8950	6,185	0.9025	0.9105	7,609	0.9236	0.9181	9,472	0.9221	0.9099	12,189	0.9159	0.9159	16,401	0.8822	0.9322	24,354	0.9214	0.9393	40,625	0.9563	0.9542
10	3,337	0.9453	0.9439	4,475	0.8944	0.9072	5,664	0.8943	0.8971	6,945	0.8927	0.9105	8,538	0.9392	0.9206	10,597	0.9166	0.9132	13,547	0.9368	0.9189	18,262	0.9191	0.9347	26,899	0.9569	0.9410	45,150	0.9901	0.9549
11	3,766	0.9259	0.9350	5,039	0.9030	0.9077	6,368	0.9037	0.9006	7,801	0.9058	0.9105	9,517	0.9310	0.9231	11,769	0.9277	0.9164	15,042	0.9244	0.9218	20,186	0.9313	0.9370	29,726	0.9652	0.9428	49,917	0.9537	0.9556
12	4,219	0.9162	0.9258	5,602	0.9175	0.9084	7,065	0.8998	0.9051	8,640	0.9234	0.9106	10,521	0.9039	0.9253	12,984	0.8858	0.9193	16,561	0.9217	0.9246	22,228	0.9263	0.9393	32,717	0.9783	0.9445	54,434	0.9788	0.9563
13	4,672	0.9483	0.9300	6,163	0.8912	0.9094	7,738	0.9248	0.9103	9,437	0.9359	0.9106	11,486	0.9006	0.9273	14,175	0.8995	0.9220	18,024	0.8899	0.9270	24,094	0.9330	0.9412	35,384	0.9567	0.9460	58,961	0.9295	0.9570
14	5,134	0.9526	0.9329	6,757	0.9286	0.9105	8,432	0.9042	0.9163	10,259	0.8857	0.9107	12,439	0.9281	0.9292	15,280	0.9368	0.9243	19,363	0.9153	0.9290	25,948	0.9463	0.9429	38,135	0.9429	0.9474	63,302	0.9540	0.9577
15	5,574	0.9543	0.9354	7,329	0.9159	0.9118	9,144	0.9335	0.9230	11,136	0.9016	0.9108	13,507	0.9338	0.9311	16,557	0.9112	0.9267	20,875	0.9241	0.9311	27,999	0.9263	0.9447	41,024	0.9391	0.9488	68,492	0.9530	0.9585
16	6,041	0.9793	0.9379	7,894	0.9139	0.9131	9,871	0.9196	0.9269	11,963	0.9080	0.9130	14,480	0.9502	0.9332	17,745	0.9157	0.9289	22,424	0.9082	0.9332	30,051	0.9477	0.9464	43,919	0.9371	0.9501	73,690	0.9884	0.9594
17	6,506	0.9395	0.9402	8,489	0.9234	0.9146	10,574	0.9588	0.9291	12,797	0.9374	0.9157	15,489	0.9676	0.9338	18,951	0.9335	0.9309	23,880	0.9720	0.9350	31,811	0.9392	0.9477	46,623	0.9508	0.9513	77,558	0.9515	0.9600
18	6,980	0.9216	0.9424	9,107	0.9518	0.9161	11,299	0.9336	0.9313	13,652	0.8990	0.9183	16,523	0.9586	0.9345	20,144	0.9517	0.9327	25,443	0.9350	0.9368	33,841	0.9495	0.9492	49,711	0.9412	0.9525	82,510	0.9600	0.9607
19	7,463	0.9439	0.9445	9,679	0.9483	0.9176	11,988	0.9168	0.9332	14,459	0.9178	0.9206	17,476	0.9471	0.9351	21,315	0.9467	0.9345	26,885	0.9503	0.9383	35,649	0.9377	0.9504	52,505	0.9256	0.9536	86,868	0.9513	0.9614
20	7,960	0.9581	0.9465	10,300	0.9201	0.9192	12,734	0.8894	0.9352	15,345	0.9479	0.9230	18,481	0.9499	0.9358	22,571	0.9653	0.9362	28,444	0.9401	0.9400	37,705	0.9704	0.9517	55,237	0.9489	0.9547	91,524	0.9744	0.9621
21	8,412	0.9655	0.9482	10,877	0.9287	0.9207	13,419	0.9379	0.9369	16,165	0.9619	0.9251	19,415	0.9592	0.9363	23,715	0.9382	0.9378	29,851	0.9528	0.9414	39,509	0.9453	0.9528	58,081	0.9552	0.9557	96,502	0.9542	0.9629
22	8,890	0.9337	0.9499	11,484	0.9294	0.9223	14,123	0.9369	0.9386	17,008	0.9679	0.9272	20,430	0.9537	0.9370	24,913	0.9383	0.9393	31,346	0.9227	0.9428	41,534	0.9694	0.9540	61,037	0.9402	0.9567	101,169	0.9289	0.9636
23	9,390	0.9700	0.9516	12,102	0.9079	0.9240	14,900	0.9399	0.9403	17,870	0.9579	0.9292	21,474	0.9645	0.9377	26,141	0.9591	0.9408	32,902	0.8980	0.9442	43,486	0.9750	0.9551	63,784	0.9693	0.9577	105,759	0.9578	0.9642
24	9,853	1.0072	0.9531	12,711	0.9088	0.9256	15,630	0.9521	0.9419	18,746	0.9510	0.9311	22,481	0.9045	0.9383	27,421	0.9422	0.9423	34,365	0.9605	0.9454	45,434	0.9578	0.9562	66,616	0.9562	0.9586	110,468	0.9720	0.9649
25	10,343	0.9742	0.9546	13,314	0.9271	0.9272	16,345	0.9597	0.9434	19,600	0.9423	0.9329	23,494	0.9409	0.9390	28,596	0.9517	0.9436	35,885	0.9472	0.9467	47,402	0.9691	0.9572	69,655	0.9426	0.9595	115,375	0.9615	0.9656
26	10,836	0.9437	0.9561	13,905	0.9491	0.9287	17,004	0.9609	0.9447	20,385	0.9427	0.9345	24,434	0.9238	0.9396	29,704	0.9343	0.9447	37,271	0.9360	0.9478	49,231	0.9643	0.9581	72,041	0.9836	0.9603	119,496	0.9716	0.9662
27	11,319	0.9724	0.9575	14,535	0.9323	0.9304	17,812	0.9484	0.9462	21,342	0.9469	0.9364	25,564	0.9555	0.9403	31,058	0.9468	0.9461	38,970	0.9391	0.9491	51,524	0.9603	0.9592	75,575	0.9705	0.9613	124,295	0.9759	0.9668
28	11,853	0.9424	0.9589	15,174	0.9074	0.9320	18,530	0.9327	0.9475	22,210	0.9429	0.9381	26,592	0.9505	0.9409	32,227	0.9551	0.9473	40,453	0.9683	0.9501	53,395	1.0016	0.9600	77,892	0.9552	0.9620	128,840	0.9598	0.9674
29	12,320	0.9546	0.9601	15,760	0.9395	0.9335	19,295	0.9320	0.9489	23,064	0.9486	0.9396	27,608	0.9533	0.9416	33,544	0.9723	0.9485	41,859	0.9515	0.9511	55,386	0.9549	0.9609	80,957	0.9637	0.9628	133,782	1.0005	0.9681
30	12,812	0.9736	0.9614	16,379	0.8980	0.9351	20,006	0.9578	0.9501	23,870	0.9236	0.9410	28,534	0.9417	0.9422	34,717	0.9526	0.9496	43,327	0.9834	0.9521	57,351	0.9385	0.9617	83,520	0.9786	0.9635	138,266	0.9860	0.9687
31	13,376	0.9364	0.9627	17,001	0.9277	0.9367	20,764	0.9715	0.9513	24,788	0.9055	0.9426	29,600	0.9644	0.9429	36,026	0.9545	0.9508	44,999	0.9774	0.9532	59,459	0.9486	0.9626	86,892	0.9764	0.9644	142,612	0.9877	0.9692
32	13,839	0.9130	0.9638	17,576	0.9289	0.9381	21,508	0.9521	0.9525	25,696	0.9350	0.9441	30,670	0.9598	0.9435	37,149	0.9452	0.9517	46,298	0.9427	0.9541	61,336	0.9630	0.9634	89,401	0.9647	0.9651	147,465	0.9772	0.9699
33	14,386	0.9118	0.9650	18,214	0.9506	0.9397	22,215	0.9606	0.9536	26,568	0.9420	0.9454	31,712	0.9324	0.9442	38,465	0.9801	0.9528	47,904	0.9449	0.9551	63,159	0.9320	0.9641	92,576	0.9514	0.9659	152,367	0.9439	0.9705
34	14,915	0.9817	0.9662	18,854	0.9527	0.9413	23,018	0.9633	0.9548	27,396	0.9269	0.9467	32,736	0.9318	0.9449	39,692	0.9928	0.9538	49,588	0.9665	0.9561	65,606	0.9418	0.9650	95,918	0.9838	0.9667	157,912	0.9601	0.9712
35	15,381	0.9854	0.9671	19,427	0.9451	0.9427	23,697	0.9526	0.9557	28,188	0.9336	0.9479	33,652	0.9489	0.9454	40,711	0.9719	0.9546	50,941	0.9564	0.9569	67,198	0.9561	0.9655	97,871	0.9805	0.9671	160,843	0.9526	0.9715
35	15,381	0.9854	0.9671	19,427	0.9451	0.9427	23,697	0.9526	0.9557	28,188	0.9336	0.9479	33,652	0.9489	0.9454	40,711	0.9719	0.9546	50,941	0.9564	0.9569	67,198	0.9561	0.9655	97,871	0.9805	0.9671	160,843	0.9526	0.9715
36	15,962	0.9464	0.9683	20,073	0.9821	0.9442	24,472	0.9799	0.9568	29,152	0.9329	0.9493	34,750	0.9453	0.9461	42,063	0.9327	0.9556	52,591	0.9768	0.9578	69,346	0.9808	0.9663	101,068	1.0127	0.9679	166,334	0.9711	0.9722
37	16,385	0.9368	0.9691	20,702	0.9416	0.9458	25,245	0.9856	0.9579	30,034	0.9427	0.9505	35,740	0.9249	0.9468	43,230	0.9218	0.9565	54,069	0.9777	0.9586	71,329	0.9696	0.9670	104,361	0.9587	0.9686	171,223	0.9884	0.9727
38	16,900	0.9381	0.9701	21,303	0.9174	0.9472	25,978	0.9565	0.9588	30,926	0.9334	0.9517	36,808	0.9439	0.9475	44,572	0.9208	0.9575	55,738	0.9575	0.9595	73,504	0.9822	0.9677	107,615	0.9963	0.9693	176,317	0.9880	0.9733
39	17,439	0.9640	0.9711	21,928	0.9334	0.9487	26,719	0.9511	0.9598	31,745	0.9339	0.9528	37,726	0.9267	0.9480	45,680	0.9480	0.9582	57,150	0.9931	0.9602	75,361	0.9641	0.9683	110,293	0.9599	0.9699	181,108	0.9752	0.9739
40	17,947	0.9706	0.9720	22,604	0.9619	0.9502	27,608	0.9386	0.9609	32,761	0.9444	0.9541	38,915	0.9593	0.9488	46,889	0.9380	0.9591	58,752	1.0055	0.9610	77,490	0.9762	0.9690	113,133	0.9810	0.9705	185,418	0.9328	0.9744
41	18,400	0.9748	0.9728	23,180	0.9375	0.9516	28,215	0.9530	0.9616	33,483	0.9652	0.9550	39,798	0.9703	0.9494	48,019	0.9624	0.9598	59,927	0.9881	0.9616	79,083	0.9597	0.9695	115,843	0.9671	0.9711	189,553	0.9917	0.9748
42	18,989	1.0036	0.9739	23,842	0.9659	0.9531	29,005	0.9641	0.9625	34,449	0.9468	0.9562	40,929	0.9828	0.9501	49,328	0.9818	0.9607	61,700	0.9742	0.9625	81,220	0.9926	0.9701	119,142	1.0021	0.9718	195,023	1.0023	0.9754
43	19,505	0.9617	0.9747	24,439	0.9594	0.9544	29,831	0.9570	0.9635	35,435	0.9287	0.9574	42,025	1.0115	0.9508	50,578	0.9743	0.9614	63,297	0.9586	0.9632	83,306	0.9809	0.9707	122,210	0.9941	0.9724	200,344	0.9532	0.9760
44	20,052	0.9538	0.9756	25,095	0.9461	0.9559	30,522	0.9496	0.9642	36,284	0.9374	0.9584	42,953	0.9623	0.9514	51,786	0.9630	0.9622	64,852	0.9458	0.9639	85,096	0.9917	0.9712	125,176	0.9728	0.9729	205,283	0.9707	0.9765
45	20,584	0.9689	0.9764	25,751	0.9514	0.9573	31,331	0.9337	0.9652	37,127	0.9578	0.9594	44,020	0.9569	0.9521	53,015	0.9669	0.9629	66,314	1.0026	0.9646	86,886	0.9791	0.9718	127,473	0.9448	0.9734	209,826	1.0201	0.9770
46	21,075	0.9892	0.9772	26,383	0.9608	0.9587	32,169	0.9582	0.9662	38,157	1.0058	0.9605	45,229	0.9696	0.9528	54,474	0.9819	0.9638	68,095	0.9621	0.9654	89,353	0.9630	0.9724	130,922	0.9444	0.9740	213,098	0.9751	0.9774
47	21,636	0.9668	0.9780	26,961	0.9832	0.9600	32,833	0.9617	0.9669	39,008	0.9593	0.9614	46,205	0.9498	0.9535	55,658	0.9775	0.9645	69,523	0.9691	0.9660	91,147	0.9558	0.9729	133,329	0.9506	0.9745	218,419	0.9629	0.9779
48	22,156	0.9593	0.9788	27,685	0.9687	0.9615	33,575	0.9826	0.9678	39,894	0.9655	0.9624	47,313	0.9612	0.9542	56,849	0.9699	0.9652	71,081	0.9748	0.9666	93,208	0.9609	0.9735	135,860	0.9647	0.9749	223,001	0.9634	0.9784
49	22,656	0.9893	0.9795	28,241	0.9612	0.9627	34,299	0.9724	0.9685	40,794	0.9865	0.9633	48,357	0.9804	0.9548	58,117	0.9770	0.9659	72,562	0.9585	0.9673	95,145	0.9790	0.9740	138,712	0.9422	0.9755	227,791	0.9722	0.9789
50	23,145	0.9552	0.9802	28,951	0.9597	0.9642	35,196	0.9521	0.9695	41,787	0.9785	0.9643	49,373	0.9546	0.9555	59,363	0.9835	0.9665	74,011	0.9625	0.9678	97,043	0.9765	0.9744	141,714	0.9662	0.9760	231,295	0.9616	0.9792
51	23,715	0.9828	0.9810	29,513	0.9671	0.9654	35,824	0.9563	0.9701	42,624	0.9747	0.9652	50,318	0.9730	0.9561	60,496	1.0023	0.9671	75,766	0.9665	0.9685	99,487	0.9784	0.9750	145,126	0.9740	0.9766	236,850	0.9958	0.9798
52	24,244	0.9873	0.9817	30,213	0.9773	0.9667	36,626	0.9881	0.9710	43,498	0.9655	0.9660	51,349	0.9737	0.9567	61,799	0.9828	0.9678	77,413	0.9344	0.9692	101,423	0.9997	0.9755	147,374	0.9652	0.9770	240,314	0.9845	0.9801
53	24,701	0.9834	0.9823	30,799	0.9838	0.9675	37,335	0.9589	0.9717	44,392	0.9486	0.9669	52,451	0.9645	0.9574	63,167	0.9452	0.9685	78,937	0.9879	0.9697	103,601	0.9866	0.9760	151,159	0.9825	0.9776	246,671	0.9911	0.9807
54	25,257	0.9500	0.9830	31,467	0.9649	0.9684	38,129	0.9670	0.9724	45,279	0.9428	0.9677	53,560	0.9793	0.9582	64,453	0.9194	0.9691	80,596	0.9834	0.9704	105,698	0.9612	0.9765	153,856	0.9811	0.9780	249,886	0.9769	0.9810
55	25,769	0.9464	0.9837	32,061	0.9368	0.9691	38,877	0.9730	0.9732	46,099	0.9650	0.9685	54,519	0.9540	0.9588	65,578	0.9736	0.9697	81,904	1.0057	0.9708	107,442	1.0034	0.9769	156,369	0.9888	0.9784	254,587	1.0018	0.9815
56	26,281	0.9882	0.9843	32,702	0.9895	0.9700	39,619	0.9859	0.9739	46,945	0.9353	0.9692	55,423	0.9419	0.9593	66,581	0.9722	0.9702	83,084	0.9463	0.9713	109,173	0.9696	0.9773	159,235	0.9790	0.9789	260,229	0.9548	0.9820
57	26,735	0.9805	0.9848	33,377	0.9876	0.9708	40,449	0.9712	0.9746	47,943	0.9876	0.9701	56,580	0.9460	0.9601	68,025	0.9671	0.9709	84,905	0.9592	0.9719	111,271	0.9850	0.9778	161,968	0.9702	0.9793	263,552	0.9841	0.9823
58	27,372	0.9876	0.9856	34,008	0.9660	0.9716	41,236	1.0003	0.9753	48,826	0.9499	0.9709	57,570	0.9665	0.9607	69,271	0.9875	0.9715	86,607	0.9497	0.9725	113,370	1.0311	0.9782	164,906	0.9746	0.9798	267,196	0.9790	0.9826
59	27,911	0.9519	0.9862	34,635	0.9947	0.9723	42,065	0.9679	0.9761	49,792	0.9801	0.9717	58,698	0.9892	0.9614	70,493	0.9667	0.9720	88,297	0.9605	0.9731	115,913	1.0005	0.9788	167,848	0.9692	0.9802	272,890	0.9879	0.9832
60	28,405	0.9747	0.9868	35,267	0.9806	0.9731	42,836	0.9745	0.9768	50,667	0.9734	0.9725	59,656	0.9936	0.9620	71,601	0.9414	0.9725	89,430	1.0083	0.9734	117,934	0.9853	0.9792	171,349	0.9503	0.9808	277,612	0.9824	0.9836

Table A3.2, continued

Actual and Fitted *Alu* Pair I:D Ratios Across Ten Spacer Percentiles, APSNs 1-115

(Type 1, Large-Large (275-325 bp) *Alu* Pairs; hg19 Human Genome Assembly)

APSN	0-5th Percentile			6-15th Percentile			16-25th Percentile			26-35th Percentile			36-45th Percentile		
	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio
60	28,405	0.9747	0.9868	35,267	0.9806	0.9731	42,836	0.9745	0.9768	50,667	0.9734	0.9725	59,656	0.9936	0.9620
61	29,039	0.9722	0.9875	36,025	0.9767	0.9740	43,638	0.9737	0.9774	51,687	0.9758	0.9733	60,795	0.9496	0.9628
62	29,467	0.9918	0.9880	36,587	1.0015	0.9746	44,347	0.9733	0.9780	52,437	0.9781	0.9739	61,616	0.9585	0.9633
63	30,086	0.9866	0.9887	37,288	0.9701	0.9754	45,185	0.9695	0.9787	53,370	0.9814	0.9747	62,830	0.9966	0.9641
64	30,658	0.9846	0.9893	37,957	0.9945	0.9762	45,949	0.9568	0.9794	54,276	0.9897	0.9754	63,880	0.9775	0.9647
65	31,107	1.0214	0.9898	38,500	0.9982	0.9767	46,732	0.9890	0.9800	55,188	0.9781	0.9761	64,878	0.9994	0.9654
66	31,659	1.0249	0.9903	39,162	1.0124	0.9775	47,461	1.0032	0.9806	56,248	0.9906	0.9769	66,021	0.9889	0.9661
67	32,183	1.0053	0.9909	39,894	0.9590	0.9782	48,269	0.9578	0.9812	57,058	0.9973	0.9775	67,002	0.9899	0.9667
68	32,699	1.0324	0.9914	40,587	0.9894	0.9783	49,050	0.9401	0.9818	58,012	0.9914	0.9783	68,138	0.9836	0.9675
69	33,240	0.9484	0.9919	41,139	0.9792	0.9785	49,787	0.9730	0.9824	58,806	0.9695	0.9788	69,103	0.9803	0.9681
70	33,653	0.9787	0.9923	41,809	0.9561	0.9789	50,577	0.9400	0.9829	59,707	0.9755	0.9795	70,108	0.9459	0.9687
71	34,268	0.9301	0.9929	42,528	0.9705	0.9793	51,397	0.9807	0.9835	60,706	0.9576	0.9802	71,294	0.9744	0.9695
72	34,814	0.9599	0.9934	43,126	0.9797	0.9796	52,214	0.9748	0.9841	61,568	0.9139	0.9808	72,264	0.9619	0.9701
73	35,371	1.0125	0.9939	43,752	0.9714	0.9800	52,945	0.9918	0.9847	62,301	0.9606	0.9813	73,247	0.9639	0.9707
74	36,019	1.0167	0.9945	44,489	0.9668	0.9804	53,818	1.0108	0.9853	63,434	0.9616	0.9821	74,408	0.9605	0.9715
75	36,523	1.0070	0.9950	45,180	0.9571	0.9808	54,554	1.0103	0.9858	64,251	1.0003	0.9826	75,632	1.0038	0.9722
76	37,040	0.9908	0.9954	45,765	0.9447	0.9812	55,355	0.9987	0.9863	65,236	0.9828	0.9833	76,631	0.9877	0.9729
77	37,547	0.9402	0.9959	46,466	0.9713	0.9816	56,127	0.9881	0.9868	66,176	0.9890	0.9839	77,653	0.9457	0.9735
78	38,207	0.9635	0.9965	47,156	1.0037	0.9821	56,996	0.9937	0.9874	67,316	0.9702	0.9846	78,978	0.9887	0.9744
79	38,725	0.9713	0.9969	47,901	1.0091	0.9825	57,939	1.0003	0.9880	68,284	1.0083	0.9852	80,171	0.9998	0.9751
80	39,258	1.0129	0.9973	48,480	1.0145	0.9829	58,539	0.9885	0.9884	69,025	1.0044	0.9857	81,066	0.9586	0.9757
81	39,829	1.0420	0.9978	49,095	0.9607	0.9833	59,309	1.0159	0.9889	69,854	0.9853	0.9862	82,084	0.9507	0.9764
82	40,446	1.0123	0.9983	49,882	0.9871	0.9838	60,065	1.0481	0.9894	70,550	0.9958	0.9867	82,853	0.9858	0.9769
83	40,909	1.0379	0.9987	50,464	0.9740	0.9842	61,008	1.0177	0.9900	71,722	0.9734	0.9874	84,141	0.9905	0.9777
84	41,422	1.0246	0.9991	51,055	0.9930	0.9846	61,628	1.0193	0.9904	72,574	0.9688	0.9879	85,243	0.9733	0.9784
85	42,022	0.9879	0.9996	51,760	0.9765	0.9851	62,460	0.9966	0.9909	73,442	1.0026	0.9884	86,319	0.9647	0.9791
86	42,609	1.0545	1.0000	52,514	0.9868	0.9857	63,284	0.9778	0.9914	74,415	0.9951	0.9890	87,380	0.9848	0.9797
87	43,253	0.9964	1.0005	53,160	0.9666	0.9862	64,112	0.9745	0.9918	75,378	0.9820	0.9895	88,596	0.9749	0.9805
88	43,714	0.9594	1.0009	53,884	0.9391	0.9867	65,017	0.9708	0.9924	76,415	0.9642	0.9901	89,675	1.0111	0.9812
89	44,282	0.9957	1.0013	54,505	0.9801	0.9872	65,756	0.9670	0.9928	77,203	0.9972	0.9905	90,691	0.9787	0.9819
90	44,988	0.9865	1.0018	55,192	0.9992	0.9877	66,544	0.9779	0.9932	78,082	1.0180	0.9910	91,806	0.9606	0.9826
91	45,471	1.0173	1.0022	55,803	1.0000	0.9882	67,418	0.9858	0.9937	79,007	1.0110	0.9916	92,956	0.9567	0.9833
92	45,861	1.0089	1.0024	56,460	1.0015	0.9887	68,198	0.9756	0.9942	79,880	0.9697	0.9920	94,101	0.9707	0.9840
93	46,544	1.0090	1.0029	57,200	0.9731	0.9893	68,962	0.9880	0.9946	80,854	0.9977	0.9926	95,249	0.9848	0.9848
94	47,049	1.0184	1.0033	57,860	0.9848	0.9898	69,862	1.0103	0.9951	81,902	1.0023	0.9931	96,439	0.9997	0.9855
95	47,591	1.0258	1.0037	58,587	1.0140	0.9904	70,575	0.9888	0.9955	82,732	0.9821	0.9935	97,487	1.0049	0.9862
96	48,190	1.0220	1.0041	59,295	1.0123	0.9910	71,392	0.9946	0.9959	83,568	1.0016	0.9940	98,260	0.9824	0.9867
97	48,722	1.0187	1.0044	59,878	1.0094	0.9915	72,272	1.0161	0.9964	84,682	0.9987	0.9946	99,495	0.9800	0.9875
98	49,261	1.0234	1.0048	60,567	0.9855	0.9920	72,941	0.9438	0.9967	85,673	0.9994	0.9951	100,752	0.9706	0.9883
99	49,873	1.0258	1.0052	61,121	1.0019	0.9925	73,600	0.9590	0.9971	86,353	1.0080	0.9954	101,667	0.9981	0.9889
100	50,467	1.0053	1.0056	61,947	0.9943	0.9933	74,470	0.9780	0.9975	87,305	1.0112	0.9959	102,711	0.9875	0.9895
101	50,941	0.9703	1.0059	62,550	0.9909	0.9938	75,366	1.0184	0.9980	88,183	1.0054	0.9963	103,620	0.9849	0.9901
102	51,480	1.0425	1.0063	63,155	1.0046	0.9943	76,010	1.0103	0.9983	88,982	1.0069	0.9967	104,954	0.9622	0.9910
103	52,127	1.0776	1.0067	63,899	0.9898	0.9950	76,911	1.0115	0.9987	90,061	1.0046	0.9972	106,098	0.9739	0.9917
104	52,692	1.0426	1.0070	64,634	1.0297	0.9957	77,745	1.0166	0.9991	90,960	0.9925	0.9977	107,111	0.9661	0.9923
105	53,308	1.0232	1.0074	65,292	1.0067	0.9963	78,348	1.0325	0.9994	91,969	1.0010	0.9982	108,514	0.9809	0.9932
106	53,986	1.0386	1.0078	66,027	0.9609	0.9970	79,207	1.0474	0.9998	92,777	0.9778	0.9985	109,307	1.0024	0.9938
107	54,350	1.0328	1.0081	66,685	0.9901	0.9976	80,271	1.0250	1.0004	93,900	0.9870	0.9991	110,739	0.9739	0.9947
108	55,056	1.0123	1.0085	67,513	0.9985	0.9984	81,238	0.9911	1.0008	95,050	1.0065	0.9996	112,024	0.9861	0.9955
109	55,664	1.0050	1.0088	68,214	1.0090	0.9991	81,890	0.9771	1.0011	95,922	0.9951	1.0000	113,097	0.9874	0.9962
110	56,093	0.9892	1.0091	68,835	0.9695	0.9997	82,664	0.9638	1.0015	96,905	1.0172	1.0004	114,292	0.9954	0.9969
111	56,738	0.9382	1.0095	69,581	0.9919	1.0005	83,588	0.9989	1.0019	97,890	1.0244	1.0009	115,337	0.9846	0.9976
112	57,274	0.9784	1.0098	70,140	1.0231	1.0010	84,194	1.0090	1.0022	98,630	1.0112	1.0012	116,417	0.9739	0.9983
113	57,809	0.9735	1.0101	70,816	0.9953	1.0017	85,051	1.0147	1.0026	99,479	0.9969	1.0016	117,458	0.9799	0.9990
114	58,352	1.0438	1.0104	71,660	1.0095	1.0026	85,867	1.0155	1.0029	100,450	1.0047	1.0020	118,392	1.0200	0.9996
115	58,993	1.0299	1.0108	72,408	0.9977	1.0034	86,782	1.0031	1.0033	101,605	1.0036	1.0025	119,926	0.9995	1.0005

APSN	46-55th Percentile			56-65th Percentile			66-75th Percentile			76-85th Percentile			86-95th Percentile		
	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio	Median Spacer Size (bp)	Actual I:D Ratio	Fitted I:D Ratio
60	71,601	0.9414	0.9725	89,430	1.0083	0.9734	117,934	0.9853	0.9792	171,349	0.9503	0.9808	277,612	0.9824	0.9836
61	73,053	0.9773	0.9732	91,030	1.0044	0.9740	119,994	0.9813	0.9796	174,285	0.9880	0.9812	283,196	0.9663	0.9841
62	74,075	0.9827	0.9736	92,366	1.0305	0.9744	121,898	0.9724	0.9800	178,426	0.9618	0.9818	288,350	0.9817	0.9845
63	75,626	0.9576	0.9743	94,643	0.9732	0.9751	124,437	0.9676	0.9805	180,410	0.9647	0.9821	294,569	0.9824	0.9851
64	76,820	0.9816	0.9748	96,107	1.0042	0.9756	126,382	0.9709	0.9809	182,987	0.9721	0.9825	298,371	0.9571	0.9854
65	78,121	0.9858	0.9753	97,678	0.9664	0.9761	128,375	0.9702	0.9813	185,872	0.9795	0.9829	302,140	0.9857	0.9857
66	79,433	0.9768	0.9758	99,678	1.0058	0.9767	131,166	0.9890	0.9818	189,691	0.9795	0.9834	302,140	0.9857	0.9857
67	80,725	0.9709	0.9764	101,167	0.9832	0.9771	132,891	1.0120	0.9821	191,507	0.9733	0.9837	312,866	1.0039	0.9866
68	82,012	0.9675	0.9769	102,746	0.9951	0.9776	135,023	0.9693	0.9825	195,962	0.9843	0.9843	316,456	0.9745	0.9869
69	83,169	0.9869	0.9773	104,124	1.0024	0.9780	136,584	0.9436	0.9828	197,717	1.0186	0.9845	320,322	0.9868	0.9872
70	84,324	0.9557	0.9778	105,546	0.9606	0.9784	138,379	0.9631	0.9831	199,732	0.9905	0.9848	323,337	0.9722	0.9875
71	86,008	0.9643	0.9784	107,531	0.9690	0.9789	141,165	0.9932	0.9836	203,925	0.9705	0.9853	329,838	1.0034	0.9880
72	87,263	0.9561	0.9789	109,132	0.9947	0.9794	143,076	0.9972	0.9839	207,236	0.9759	0.9857	335,591	1.0219	0.9885
73	88,314	0.9507	0.9793	110,279	0.9794	0.9797	144,582	1.0015	0.9842	209,233	0.9737	0.9860	340,254	0.9982	0.9888
74	89,678	0.9694	0.9797	111,874	0.9836	0.9801	147,602	0.9709	0.9847	213,721	0.9817	0.9866	345,504	1.0353	0.9892
75	91,197	0.9806	0.9803	113,775	0.9662	0.9806	149,397	0.9633	0.9850	215,569	0.9855	0.9868	349,203	0.9388	0.9895
76	92,345	1.0157	0.9807	115,256	0.9648	0.9810	150,796	0.9710	0.9852	218,176	0.9721	0.9871	354,183	1.0052	0.9899
77	93,588	0.9874	0.9811	117,109	0.9984	0.9815	153,598	0.9712	0.9857	221,347	0.9560	0.9875	357,570	0.9855	0.9902
78	95,045	0.9801	0.9816	118,928	0.9740	0.9820	155,814	0.9937	0.9860	225,080	0.9877	0.9879	363,649	0.9877	0.9906
79	96,519	1.0083	0.9821	120,891	0.9509	0.9825	158,861	0.9945	0.9865	228,225	0.9814	0.9883	367,386	1.0062	0.9909
80	97,659	1.0065	0.9825	122,175	0.9500	0.9828	160,301	0.9841	0.9867	230,850	0.9943	0.9886	373,220	0.9953	0.9914
81	98,767	0.9865	0.9829	123,542	0.9916	0.9831	162,870	0.9885	0.9871	234,246	1.0000	0.9890	377,485	0.9727	0.9917
82	99,684	0.9914	0.9832	124,590	0.9719	0.9834	163,922	0.9871	0.9873	236,341	1.0192	0.9893	382,122	0.9974	0.9920
83	101,227	0.9620	0.9837	126,740	0.9836	0.9839	166,589	1.0226	0.9876	238,424	0.9873	0.9895	385,942	0.9713	0.9923
84	102,557	0.9688	0.9841	127,762	1.0015	0.9841	167,851	0.9769	0.9878	241,254	1.0003	0.9898	389,382	1.0158	0.9925
85	103,787	0.9637	0.9845	129,791	0.9729	0.9846	171,012	1.0147	0.9883	245,783	0.9992	0.9903	397,082	0.9759	0.9931
86	105,150	0.9484	0.9849	131,131	0.9754	0.9849	173,164	0.9956	0.9886	249,264	1.0195	0.9907	400,738	1.0085	0.9933
87	106,434	0.9683	0.9853	133,021	0.9628	0.9853	175,438	0.9864	0.9889	252,220	0.9795	0.9910	405,615	0.9845	0.9937
88	108,073	0.9835	0.9858	135,478	1.0018	0.9859	177,907	0.9908	0.9893	254,939	0.9984	0.9913	409,825	1.0093	0.9940
89	108,969	1.0062	0.9861	136,424	0.9990	0.9861	179,335	0.9671	0.9895	257,231	1.0139	0.9916	412,400	1.0052	0.9942
90	110,363	1.0158	0.9865	138,129	0.9743	0.9865	181,723	0.9922	0.9898	260,670	0.9791	0.9920	416,677	0.9831	0.9945
91	111,667	0.9867	0.9868	139,787	0.9665	0.9868	183,793	1.0064	0.9901	263,134	0.9669	0.9922	421,700	1.0103	0.9948
92	112,986	0.9780	0.9872	141,339	0.9767	0.9872	185,892	1.0054	0.9904	266,091	1.0114	0.9925	427,910	0.9821	0.9952
93	114,255	0.9723	0.9876	143,090	1.0026	0.9875	188,028	0.9997	0.9906	269,367	1.0142	0.9929	431,187	0.9872	0.9954
94	115,674	0.9589	0.9880	144,931	0.9854	0.9879	190,717	0.9885	0.9910	274,291	0.9711	0.9934	435,478	0.9885	0.9957
95	117,213	0.9980	0.9884	146,674	0.9539	0.9883	193,895	1.0010	0.9914	277,831	1.0070	0.9937	441,046	0.9901	0.9961
96	117,868	1.0194	0.9886	147,596	0.9880	0.9885	194,524	1.0348	0.9915	278,704	0.9872	0.9938	445,716	1.0047	0.9964
97	119,295	0.9881	0.9890	149,230	0.9728	0.9888	196,061	1.0116	0.9917	282,427	0.9819	0.9942	451,274	0.9715	0.9968
98	120,916	0.9788	0.9894	151,379	0.9604	0.9892	199,041	0.9742	0.9920	286,673	0.9676	0.9946	457,397	0.9974	0.9972
99	121,983	0.9743	0.9897	152,304	0.9840	0.9894	199,956	0.9632	0.9922	287,219	0.9961	0.9947	458,686	1.0165	0.9973
100	123,315	1.0154	0.9901	153,963	1.0148	0.9897	203,125	0.9645	0.9925	291,252	1.0005	0.9951	464,514	1.0174	0.9976
101	124,736	0.9899	0.9904	155,672	0.9876	0.9901	205,475	0.9643	0.9928	294,781	0.9878	0.9954	466,383	0.9585	0.9978
102	126,052	0.9925	0.9908	157,081	0.9922	0.9903	207,472	1.0079	0.9931	297,343	0.9900	0.9956	470,459	0.9674	0.9980
103	127,459	0.9663	0.9912	159,023	0.9928	0.9907	209,846	0.9935	0.9933	301,686	1.0093	0.9961	475,743	1.0067	0.9984
104	128,680	0.9744	0.9915	160,782	1.0137	0.9911	212,053	1.0277	0.9936	304,235	1.0184	0.9963	481,410	1.0015	0.9987
105	130,125	0.9993	0.9918	162,498	1.0029	0.9914	214,400	0.9935	0.9939	307,044	1.0028	0.9966	485,652	0.9773	0.9990
106	131,198	1.0069	0.9921	163,838	0.9929	0.9916	216,805	0.9580	0.9942	310,171	1.0289	0.9968	492,716	0.9919	0.9994
107	132,945	0.9887	0.9925	166,082	1.0028	0.9920	219,641	0.9833	0.9945	314,599	1.0271	0.9972	495,613	1.0033	0.9996
108	134,541	0.9610	0.9929	168,249	0.9846	0.9924	221,532	0.9583	0.9947	317,489	1.0144	0.9975	501,529	1.0250	1.0000
109	135,797	0.9945	0.9932	169,911	0.9878	0.9927	223,532	1.0083	0.9949	320,531	0.9977	0.9978	507,110	0.9908	1.0003
110	137,335	1.0021	0.9936	171,842	0.9582	0.9931	226,478	0.9744	0.9952	325,507	0.9962	0.9982	512,027	1.0142	1.0006
111	138,247	0.9716	0.9938	173,061	0.9676	0.9933	227,824	1.0094	0.9954	327,396	1.0168	0.9984	514,519	1.0128	1.0007
112	139,678	0.9927	0.9941	174,322	0.9772	0.9935	229,663	0.9841	0.9956	329,214	0.9843	0.9985	521,729	1.0191	1.0012
113	140,562	1.0415	0.9944	176,396	0.9715	0.9939	232,264	1.0091	0.9959	333,027	0.9771	0.9989	525,322	1.0059	1.0014
114	141,963	1.0038	0.9947	177,460	0.9778	0.9940	233,572	0.9896	0.9960	334,927	0.9899	0.9990	525,735	0.9909	1.0014
115	143,872	1.0274	0.9951	180,216	1.0018	0.9945	237,700	1.0016	0.9964	340,248	1.0151	0.9995	536,310	1.0194	1.0020

Table A3.3
Characteristics of 50 deletion-prone human cancer genes

Gene	Locus	Coding Region Length (bp)	Coding Exons	*Alu* Population Across *Alu* Landscape				Raw Stability Scores		
				250 kbp 5' Flanking	Gene	250 kbp 3' Flanking	Total	Coding Exon Scores Lowest	Coding Exon Scores Highest	Gene Scores
APC	chr5:112,073,556-112,181,936	136,409	14	77	68	191	336	0.902	0.975	0.709
ARID1A	chr1:27,022,522-27,108,601	84,353	20	375	76	324	775	0.903	0.944	0.755
ATM	chr11:108,093,559-108,239,826	137,884	62	262	82	193	537	0.767	0.970	0.359
BRCA1	chr17:41,196,312-41,277,500	78,419	22	298	138	325	761	0.808	0.903	0.357
BRCA2	chr13:32,889,617-32,973,809	82,310	26	89	55	182	326	0.917	0.964	0.610
BUB1B	chr15:40,453,210-40,513,337	59,939	23	130	60	127	317	0.869	0.951	0.593
CASP8	chr2:202,122,754-202,152,434	20,108	8	271	19	76	366	0.806	0.949	0.749
CDKN1B	chr12:12,870,302-12,875,305	1,107	2	186	0	269	455	0.925	0.953	0.952
CDKN2A	chr9:21,967,751-21,994,490	6,599	3	51	6	57	114	0.977	0.984	0.971
CDKN2C	chr1:51,435,642-51,440,309	3,902	2	145	0	241	386	0.947	0.967	0.965
CYLD	chr16:50,775,961-50,835,846	46,810	16	77	13	69	159	0.942	0.987	0.886
DICER1	chr14:95,552,565-95,608,085	42,961	26	81	6	82	169	0.964	0.992	0.941
FANCA	chr16:89,803,959-89,883,065	78,015	43	252	125	277	654	0.832	0.911	0.301
FANCB	chrX:14,861,529-14,891,184	21,944	8	63	7	29	99	0.974	0.991	0.949
FANCD2	chr3:10,068,113-10,141,344	70,293	42	264	80	301	645	0.847	0.925	0.410
FBXO11	chr2:48,034,059-48,132,932	31,632	22	116	101	245	462	0.879	0.934	0.671
FBXW7	chr4:153,242,410-153,456,185	88,923	11	132	73	79	284	0.985	0.990	0.968
FH	chr1:241,660,857-241,683,085	21,895	10	116	4	69	189	0.973	0.989	0.949
GPC3	chrX:132,669,776-133,119,673	449,325	10	110	175	190	475	0.957	0.979	0.821
IKZF1	chr7:50,344,378-50,472,798	109,668	7	45	7	54	106	0.993	0.997	0.988
KDM6A	chrX:44,732,423-44,971,845	237,859	29	338	138	62	538	0.942	0.982	0.731
MAP2K4	chr17:11,924,135-12,047,051	120,374	11	134	7	149	290	0.968	0.983	0.904
MAP3K1	chr5:56,110,900-56,191,978	78,107	20	60	28	204	292	0.951	0.985	0.933
MAP3K13	chr3:185,080,836-185,206,882	53,875	13	152	80	263	495	0.913	0.952	0.729
MEN1	chr11:64,570,986-64,578,188	5,776	9	276	2	96	374	0.916	0.948	0.909
MLH1	chr3:37,034,841-37,092,337	57,106	19	131	43	216	390	0.904	0.954	0.695
MSH2	chr2:47,630,263-47,710,360	79,578	16	285	105	169	559	0.856	0.943	0.537
NCOR1	chr17:15,933,408-16,118,874	162,274	45	376	141	208	725	0.905	0.962	0.510
NF1	chr17:29,421,945-29,704,695	278,846	58	327	172	208	707	0.929	0.977	0.592
NF2	chr22:29,999,545-30,094,589	90,804	16	324	86	246	656	0.894	0.952	0.685
PAX5	chr9:36,838,531-37,034,476	193,472	10	171	54	142	367	0.972	0.991	0.910
PBRM1	chr3:52,579,368-52,713,739	131,649	29	150	156	102	408	0.867	0.938	0.376
PRDM1	chr6:106,534,195-106,557,814	20,933	7	119	5	119	234	0.970	0.981	0.954
PTGFRN	chr1:117,452,689-117,532,972	76,764	9	118	18	106	242	0.967	0.986	0.930
RB1	chr13:48,877,883-49,056,026	176,159	27	145	57	72	274	0.922	0.989	0.774
SBDS	chr7:66,452,690-66,460,588	7,047	5	253	9	360	622	0.861	0.902	0.787
SDHD	chr11:111,957,571-111,966,518	8,063	4	195	8	157	360	0.910	0.950	0.888
SMARCB1	chr22:24,129,150-24,176,705	47,011	9	253	61	213	527	0.946	0.969	0.748
SMARCD1	chr12:50,478,983-50,494,494	13,631	13	170	13	373	556	0.874	0.950	0.854
SMAD4	chr18:48,556,583-48,611,411	31,421	11	201	32	154	387	0.905	0.947	0.881
SPRED1	chr15:38,545,052-38,649,450	98,479	7	81	29	67	177	0.974	0.986	0.938
STK11	chr19:1,205,798-1,228,434	19,734	9	261	17	215	493	0.918	0.956	0.874
SUFU	chr10:104,263,719-104,393,214	126,003	12	263	128	286	677	0.913	0.957	0.781
TBX3	chr12:115,108,059-115,121,969	11,360	7	232	0	182	414	0.967	0.969	0.963
TNFAIP3	chr6:138,188,581-138,204,449	10,092	8	82	2	82	166	0.977	0.985	0.969
TP53	chr17:7,571,720-7,590,863	6,986	10	193	33	313	539	0.873	0.895	0.791
TRIM36	chr5:114,460,459-114,516,243	53,535	10	46	22	51	119	0.969	0.988	0.912
TSC1	chr9:135,766,735-135,820,020	32,638	21	174	23	159	356	0.928	0.967	0.838
TSC2	chr16:2,097,990-2,138,713	39,995	41	275	25	218	518	0.910	0.972	0.749
VHL	chr3:10,183,319-10,195,354	8,118	3	292	21	213	526	0.871	0.853	0.812

Table A3.4
Characteristics of 50 randomly chosen human genes

Gene	Locus	Coding Region Length (bp)	Coding Exons	*Alu* Population Across *Alu* Landscape				Raw Stability Scores		
				250 kbp 5' Flanking	Gene	250 kbp 3' Flanking	Total	Coding Exon Scores Lowest	Coding Exon Scores Highest	Gene Score
ADGB	chr6:146,920,136-147,136,597	216,177	36	30	48	48	126	0.954	0.995	0.779
ARSH	chrX:2,924,654-2,951,426	26,773	9	239	28	158	425	0.884	0.935	0.726
BLK	chr8:11,351,521-11,422,108	20,884	12	80	24	101	205	0.969	0.987	0.950
C19orf76	chr19:50,191,942-50,194,247	656	2	361	0	300	661	0.925	0.928	0.925
CDH13	chr16:82,660,399-83,830,215	1,167,938	15	41	317	129	487	0.954	0.994	0.802
CHAMP1	chr13:115,079,965-115,092,803	2,439	1	95	9	6	110	only coding exon = 0.968		
CHRNA1	chr2:175,612,323-175,629,200	16,271	10	99	9	120	228	0.949	0.979	0.911
CHST9	chr18:24,495,595-24,765,289	226,551	5	60	47	82	189	0.979	.994	0.959
CYB5B	chr16:69,458,498-69,500,167	37,835	3	350	38	285	673	0.906	0.934	0.828
DCAF6	chr1:167,905,797-168,045,083	138,524	19	201	67	147	415	0.930	0.978	0.773
DTNBP1	chr6:15,523,032-15,663,289	139,895	10	158	16	129	303	0.943	0.983	0.841
GDPD2	chrX:69,642,881-69,653,241	8,102	16	142	2	185	329	0.072	0.966	0.071
GFI1	chr1:92,940,318-92,951,628	7,459	6	130	3	117	250	0.946	0.967	0.938
GRPEL2	chr5:148,724,977-148,734,146	5,743	4	59	6	236	301	0.928	0.964	0.913
H2AFB3	chrX:154,113,317-154,113,833	348	1	92	0	76	168	only coding exon = 0.986		
HDGFL1	chr6:22,569,678-22,570,750	756	1	60	0	39	99	only coding exon = 0.995		
HSPB9	chr17:40,274,756-40,275,371	480	1	289	0	285	574	only coding exon = 0.911		
IL17D	chr13:21,277,482-21,297,237	17,953	2	194	9	257	460	0.955	0.961	0.940
JMY	chr5:78,531,925-78,623,038	79,657	10	212	92	189	493	0.896	0.949	0.726
KCNA6	chr12:4,918,342-4,960,278	1,590	1	81	9	43	133	only coding exon = 0.991		
KEAP1	chr19:10,596,796-10,614,054	13,382	5	493	43	454	990	0.784	0.830	0.621
KIAA1598	chr10:118,644,306-118,765,088	118,736	15	140	44	96	280	0.946	0.986	0.821
MADCAM1	chr19:496,490-505,343	8,466	5	163	11	268	442	0.897	0.940	0.861
MAP9	chr4:156,263,812-156,298,122	28,068	13	81	6	34	121	0.972	0.989	0.934
MFSD4	chr1:205,538,112-205,572,046	31,392	10	153	14	192	359	0.943	0.969	0.882
MIA3	chr1:222,791,444-222,841,351	47,509	28	132	16	126	274	0.960	0.980	0.885
MRPS9	chr2:105,654,483-105,716,418	61,667	11	64	17	91	172	0.976	0.989	0.942
MUT	chr6:49,398,073-49,431,041	27,739	12	100	8	41	149	0.969	0.990	0.951
NANOS3	chr19:13,988,063-13,991,571	3,255	2	429	6	393	828	0.872	0.896	0.858
NCF1	chr7:74,188,309-74,203,659	15,126	11	432	21	466	919	0.876	0.893	0.723
NGB	chr14:77,731,834-77,737,655	4,402	4	215	1	152	368	0.974	0.978	0.970
OPRD1	chr1:29,138,654-29,190,208	50,900	3	473	72	301	846	0.898	0.928	0.844
OR6P1	chr1:158,532,441-158,533,394	954	1	36	0	42	78	only coding exon = 0.993		
PACSIN1	chr6:34,482,649-34,504,039	6,225	9	323	0	320	643	0.950	0.951	0.973
PATE4	chr11:125,703,211-125,709,967	5,068	3	95	0	110	205	0.980	0.988	0.977
PHKA2	chrX:18,910,416-19,002,480	90,448	33	147	47	187	381	0.898	0.971	0.724
PSG2	chr19:43,568,362-43,586,893	16,013	5	47	3	71	121	0.971	0.990	0.961
SET	chr9:131,451,509-131,458,675	10,769	8	403	2	377	782	0.872	0.907	0.821
SF3B3	chr16:70,557,691-70,611,571	45,157	25	447	64	168	679	0.840	0.926	0.499
SFRP5	chr10:99,526,508-99,531,756	4,320	3	80	1	257	338	0.960	0.965	0.952
SPATA7	chr14:88,851,742-88,904,804	52,604	12	75	20	164	259	0.968	0.980	0.924
TAGLN2	chr1:159,887,903-159,895,284	1,710	4	80	0	58	138	0.984	0.987	0.984
THYN1	chr11:134,118,173-134,123,260	4,445	7	62	0	108	170	0.906	0.913	0.978
TMEM136	chr11:120,195,838-120,204,388	3,140	3	73	2	129	204	0.975	0.979	0.973
TRNP1	chr1:27,320,195-27,327,377	684	1	317	4	334	655	only coding exon = 0.917		
TUBA1C	chr12:49,658,865-49,667,113	8,046	4	333	9	307	649	0.873	0.899	0.817
USMG5	chr10:105,148,809-105,156,270	264	2	168	8	376	552	0.893	0.902	0.892
XPNPEP3	chr22:41,253,085-41,328,823	69,254	10	332	109	416	857	0.859	0.926	0.563
ZNF296	chr19:45,574,758-45,579,688	4,773	3	333	6	407	746	0.896	0.917	0.885
ZNF567	chr19:37,180,303-37,212,225	8,199	3	282	29	270	581	0.888	0.917	0.853

Table A3.5
Spacer sample sizes and groupings used in determination of I:D ratios for Type 1 *Alu* pairs

APSN	Percentile[(1)]									
	2.5th	10th	20th	30th	40th	50th	60th	70th	80th	90th
1	2,611	5,263	5,307	5,263	5,317	5,287	5,255	5,316	5,274	5,299
2	4,643	9,297	9,406	9,358	9,410	9,340	9,360	9,365	9,358	9,375
3	5,475	11,050	11,130	11,047	11,060	11,108	11,051	11,068	11,076	11,061
4	5,765	11,675	11,681	11,704	11,716	11,720	11,690	11,713	11,694	11,685
5	5,952	12,026	12,028	12,029	12,060	12,003	12,082	12,048	12,039	12,042
6	6,027	12,216	12,198	12,223	12,229	12,272	12,140	12,289	12,210	12,186
7	6,102	12,244	12,284	12,240	12,291	12,276	12,306	12,260	12,294	12,271
8	6,084	12,383	12,344	12,342	12,391	12,322	12,363	12,340	12,333	12,372
9	6,139	12,425	12,430	12,444	12,382	12,434	12,388	12,445	12,422	12,436
10	6,112	12,435	12,455	12,422	12,417	12,435	12,407	12,428	12,403	12,440
11	6,163	12,379	12,401	12,485	12,434	12,372	12,455	12,418	12,465	12,371
12	6,195	12,479	12,451	12,460	12,488	12,471	12,520	12,467	12,463	12,496
13	6,141	12,463	12,417	12,409	12,487	12,432	12,439	12,458	12,423	12,455
14	6,141	12,424	12,448	12,453	12,448	12,411	12,390	12,464	12,429	12,453
15	6,164	12,400	12,440	12,412	12,417	12,461	12,420	12,463	12,408	12,427
16	6,108	12,467	12,422	12,488	12,444	12,446	12,491	12,450	12,444	12,479
17	6,121	12,512	12,423	12,463	12,498	12,413	12,475	12,471	12,481	12,462
18	6,107	12,470	12,400	12,446	12,429	12,458	12,392	12,481	12,408	12,434
19	6,131	12,399	12,415	12,414	12,405	12,424	12,357	12,413	12,470	12,387
20	6,176	12,417	12,438	12,478	12,481	12,413	12,477	12,433	12,463	12,470
21	6,150	12,415	12,443	12,413	12,400	12,472	12,441	12,409	12,451	12,423
22	6,095	12,404	12,394	12,431	12,320	12,430	12,357	12,415	12,427	12,351
23	6,113	12,411	12,386	12,425	12,357	12,448	12,360	12,415	12,428	12,391
24	6,150	12,411	12,460	12,422	12,429	12,434	12,398	12,479	12,412	12,479
25	6,120	12,418	12,415	12,363	12,373	12,473	12,359	12,415	12,390	12,438
26	6,115	12,406	12,377	12,383	12,403	12,370	12,406	12,365	12,421	12,360
27	6,142	12,361	12,425	12,388	12,449	12,403	12,383	12,438	12,420	12,361
28	6,103	12,419	12,412	12,417	12,415	12,405	12,400	12,402	12,400	12,421

Table A3.5, continued
Spacer sample sizes and groupings used in determination of I:D ratios for Type 1 *Alu* pairs

APSN	Percentile[1]									
	2.5th	10th	20th	30th	40th	50th	60th	70th	80th	90th
29	6,114	12,347	12,394	12,401	12,386	12,400	12,363	12,345	12,389	12,401
30	6,126	12,356	12,354	12,332	12,359	12,354	12,392	12,302	12,414	12,331
31	6,123	12,368	12,395	12,340	12,425	12,401	12,363	12,387	12,392	12,399
32	6,089	12,364	12,402	12,353	12,376	12,379	12,410	12,353	12,348	12,385
33	6,049	12,470	12,428	12,382	12,354	12,435	12,399	12,394	12,395	12,408
34	6,080	12,427	12,406	12,390	12,352	12,435	12,381	12,445	12,365	12,427
35	6,099	12,336	12,358	12,377	12,366	12,348	12,353	12,337	12,370	12,315
36	6,137	12,315	12,438	12,396	12,349	12,346	12,430	12,374	12,392	12,396
37	6,101	12,393	12,370	12,373	12,385	12,378	12,394	12,395	12,385	12,360
38	6,076	12,398	12,357	12,370	12,396	12,376	12,346	12,371	12,367	12,379
39	6,114	12,374	12,327	12,408	12,333	12,403	12,343	12,413	12,365	12,357
40	6,091	12,362	12,382	12,374	12,369	12,370	12,366	12,359	12,403	12,364
41	6,102	12,398	12,382	12,410	12,391	12,408	12,394	12,393	12,373	12,416
42	6,135	12,407	12,409	12,440	12,450	12,402	12,416	12,426	12,445	12,430
43	6,150	12,411	12,425	12,415	12,393	12,436	12,431	12,416	12,423	12,389
44	6,086	12,422	12,427	12,407	12,384	12,418	12,342	12,456	12,399	12,382
45	6,076	12,419	12,403	12,330	12,393	12,376	12,398	12,383	12,408	12,375
46	6,103	12,355	12,356	12,372	12,312	12,349	12,361	12,367	12,345	12,386
47	6,101	12,393	12,404	12,373	12,354	12,397	12,376	12,396	12,408	12,382
48	6,113	12,381	12,399	12,349	12,399	12,351	12,396	12,422	12,362	12,401
49	6,147	12,379	12,448	12,406	12,417	12,441	12,425	12,438	12,422	12,415
50	6,104	12,358	12,349	12,358	12,349	12,351	12,362	12,381	12,350	12,352
51	6,107	12,389	12,391	12,332	12,353	12,380	12,385	12,373	12,373	12,386
52	6,081	12,380	12,390	12,402	12,365	12,369	12,353	12,416	12,351	12,391
53	6,083	12,341	12,349	12,321	12,333	12,399	12,343	12,335	12,329	12,333
54	6,127	12,312	12,398	12,401	12,329	12,382	12,394	12,371	12,390	12,403
55	6,135	12,341	12,444	12,342	12,404	12,400	12,347	12,393	12,426	12,399
56	6,070	12,293	12,370	12,322	12,298	12,340	12,330	12,312	12,365	12,331

Table A3.5, continued
Spacer sample sizes and groupings used in determination of I:D ratios for Type 1 *Alu* pairs

APSN	Percentile[1]									
	2.5th	10th	20th	30th	40th	50th	60th	70th	80th	90th
57	6,098	12,337	12,399	12,325	12,365	12,383	12,376	12,343	12,371	12,373
58	6,094	12,331	12,362	12,372	12,328	12,386	12,320	12,345	12,383	12,337
59	6,086	12,365	12,337	12,310	12,355	12,329	12,404	12,293	12,388	12,333
60	6,098	12,323	12,317	12,336	12,382	12,295	12,343	12,327	12,334	12,356
61	6,037	12,372	12,369	12,315	12,312	12,289	12,361	12,365	12,308	12,329
62	6,085	12,303	12,272	12,355	12,260	12,281	12,323	12,300	12,316	12,328
63	6,089	12,319	12,341	12,330	12,317	12,325	12,356	12,323	12,358	12,295
64	6,069	12,276	12,330	12,328	12,296	12,308	12,326	12,318	12,288	12,322
65	6,038	12,405	12,342	12,379	12,332	12,322	12,357	12,379	12,378	12,333
66	6,093	12,328	12,404	12,316	12,369	12,351	12,356	12,342	12,378	12,333
67	6,052	12,416	12,350	12,365	12,355	12,332	12,369	12,366	12,400	12,336
68	6,093	12,340	12,312	12,299	12,318	12,395	12,306	12,306	12,360	12,313
69	6,081	12,348	12,349	12,386	12,369	12,315	12,379	12,344	12,382	12,348
70	6,120	12,312	12,356	12,345	12,339	12,323	12,348	12,354	12,323	12,344
71	6,045	12,310	12,314	12,321	12,340	12,285	12,316	12,340	12,296	12,331
72	6,066	12,385	12,305	12,337	12,358	12,288	12,371	12,343	12,314	12,356
73	6,110	12,335	12,383	12,301	12,408	12,350	12,328	12,375	12,383	12,339
74	6,022	12,334	12,342	12,323	12,271	12,301	12,346	12,265	12,362	12,295
75	6,035	12,322	12,293	12,282	12,275	12,276	12,318	12,288	12,284	12,319
76	6,066	12,273	12,290	12,303	12,268	12,302	12,235	12,319	12,298	12,288
77	6,096	12,244	12,356	12,288	12,301	12,302	12,304	12,338	12,325	12,294
78	6,077	12,313	12,383	12,296	12,354	12,340	12,310	12,357	12,332	12,310
79	6,117	12,340	12,394	12,337	12,385	12,387	12,388	12,354	12,352	12,364
80	6,101	12,339	12,327	12,343	12,343	12,324	12,353	12,337	12,347	12,347
81	6,081	12,309	12,291	12,315	12,293	12,354	12,264	12,319	12,330	12,298
82	6,055	12,316	12,346	12,300	12,298	12,319	12,344	12,292	12,307	12,336
83	6,069	12,363	12,308	12,330	12,345	12,296	12,366	12,346	12,343	12,291
84	6,086	12,319	12,342	12,354	12,333	12,366	12,305	12,383	12,326	12,343

Table A3.5, continued
Spacer sample sizes and groupings used in determination of I:D ratios for Type 1 *Alu* pairs

APSN	Percentile[1]									
	2.5th	10th	20th	30th	40th	50th	60th	70th	80th	90th
85	6,067	12,294	12,349	12,312	12,305	12,332	12,309	12,322	12,327	12,298
86	6,069	12,306	12,292	12,262	12,308	12,275	12,281	12,289	12,329	12,300
87	6,071	12,311	12,311	12,298	12,363	12,308	12,281	12,312	12,350	12,276
88	6,031	12,321	12,333	12,274	12,328	12,260	12,291	12,353	12,320	12,283
89	6,091	12,326	12,286	12,307	12,327	12,302	12,358	12,310	12,307	12,338
90	6,013	12,411	12,273	12,308	12,307	12,347	12,290	12,310	12,336	12,325
91	6,052	12,312	12,278	12,279	12,300	12,256	12,312	12,287	12,295	12,295
92	6,081	12,283	12,310	12,295	12,307	12,315	12,313	12,309	12,338	12,289
93	6,027	12,312	12,308	12,292	12,292	12,246	12,308	12,252	12,311	12,311
94	6,037	12,300	12,253	12,244	12,244	12,304	12,236	12,281	12,258	12,303
95	6,051	12,336	12,307	12,287	12,296	12,266	12,302	12,340	12,323	12,295
96	6,072	12,271	12,263	12,258	12,301	12,300	12,256	12,284	12,277	12,327
97	6,040	12,338	12,369	12,270	12,351	12,334	12,328	12,305	12,359	12,316
98	6,044	12,332	12,320	12,320	12,322	12,334	12,313	12,315	12,345	12,298
99	6,041	12,394	12,279	12,355	12,300	12,347	12,287	12,364	12,306	12,341
100	6,080	12,219	12,325	12,258	12,261	12,312	12,266	12,280	12,291	12,274
101	6,047	12,300	12,306	12,249	12,318	12,242	12,319	12,263	12,336	12,274
102	6,060	12,242	12,263	12,232	12,289	12,220	12,260	12,260	12,280	12,269
103	6,048	12,277	12,270	12,286	12,240	12,309	12,236	12,324	12,263	12,249
104	6,042	12,219	12,269	12,212	12,245	12,283	12,221	12,245	12,278	12,239
105	6,023	12,307	12,244	12,256	12,260	12,290	12,252	12,248	12,265	12,267
106	6,024	12,346	12,315	12,310	12,319	12,276	12,304	12,355	12,305	12,312
107	6,066	12,223	12,290	12,258	12,260	12,276	12,209	12,324	12,264	12,276
108	6,049	12,323	12,305	12,294	12,304	12,278	12,281	12,296	12,346	12,292
109	6,055	12,257	12,339	12,262	12,288	12,272	12,269	12,323	12,264	12,319
110	6,065	12,207	12,264	12,222	12,246	12,281	12,213	12,251	12,241	12,226
111	6,020	12,286	12,273	12,258	12,247	12,271	12,252	12,241	12,248	12,308
112	6,042	12,264	12,237	12,254	12,264	12,281	12,239	12,260	12,273	12,254

Table A3.5, continued
Spacer sample sizes and groupings used in determination of I:D ratios for Type 1 *Alu* pairs

APSN	Percentile[(1)]									
	2.5th	10th	20th	30th	40th	50th	60th	70th	80th	90th
113	6,039	12,275	12,296	12,245	12,222	12,304	12,249	12,300	12,262	12,254
114	6,019	12,296	12,228	12,263	12,245	12,263	12,227	12,274	12,272	12,252
115	6,041	12,318	12,325	12,258	12,335	12,270	12,311	12,312	12,302	12,306

(1) The percentile groupings for this table are as follows.

Percentile Name	Lower Limit	Percentile Midpoint	Upper Limit
2.5th	0th	2.5th	5th
10th	5th	10th	15th
20th	15th	20th	15th
30th	25th	30th	35th
40th	35th	40th	45th
50th	45th	50th	55th
60th	55th	60th	65th
70th	65th	70th	75th
80th	75th	80th	85th
90th	85th	90th	95th

Table A3.6
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
1	A	51	99	Constant	0.79941			
	B	100	264	Line (C1= m and C2= b)	0.9194	0.0007272379		
	C	265	469	Line (C1= m and C2= b)	0.91287	-0.0000318732		
	D	470	690	Line (C1= m and C2= b)	0.91759	0.0000213801		
	E	691	7,538	Log10-Log10 Cubic	7.78075	-6.7456467666	1.91217	-0.1783
	F	7,539	21,717	Line (C1= m and C2= b)	0.95928	0.0000028717		
2	A	51	343	Constant	0.93436			
	B	344	549	Line (C1= m and C2= b)	0.9772	0.0002079614		
	C	550	10,299	Log10-Log10 Cubic	2.32578	-1.7777937965	0.42427	-0.0316
	D	10,300	29,292	Line (C1= m and C2= b)	0.96048	0.0000020805		
3	A	51	631	Constant	0.97516			
	B	632	14,072	Log10-Log10 Cubic	0.64634	-0.3094873157	0.00741	0.00689
	C	14,071	37,466	Line (C1= m and C2= b)	0.95926	0.0000017414		
4	A	51	971	Constant	0.97879			
	B	972	18,160	Log10-Log10 Cubic	3.30368	-2.4948824775	0.60811	-0.0483
	C	18,161	46,577	Line (C1= m and C2= b)	0.95435	0.0000016063		
5	A	51	1,323	Constant	0.97511			
	C	1,324	22,676	Log10-Log10 Cubic	5.31725	-4.0530801705	1.01082	-0.083
	D	22,677	55,728	Line (C1= m and C2= b)	0.95187	0.0000014562		
6	A	51	1,693	Constant	0.96868			
	B	1,694	26,666	Log10-Log10 Cubic	6.78571	-5.1303849864	1.27409	-0.1045
	C	26,667	63,564	Line (C1= m and C2= b)	0.95072	0.0000013357		
7	A	51	2,090	Constant	0.96075			
	B	2,091	8,100	Log10-Log10 Cubic	7.67666	-5.7385946167	1.41142	-0.1147
	C	8,101	31,551	Log10-Log10 Quadratic	0.95541	-0.5177208724	0.06707	
	D	31,552	72,428	Line (C1= m and C2= b)	0.96339	0.0000008957		
8	A	51	2,499	Constant	0.95258			
	B	2,500	9,500	Log10-Log10 Cubic	8.03752	-5.9433109647	1.44697	-0.1165
	C	9,501	36,034	Log10-Log10 Quadratic	0.84396	-0.4550433066	0.05843	
	D	36,035	80,576	Line (C1= m and C2= b)	0.96309	0.0000008286		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
9	A	51	2,911	Constant	0.94465			
	B	2,912	10,750	Log10-Log10 Cubic	7.87334	-5.7646659812	1.38977	-0.1108
	C	10,751	40,625	Log10- 118 ratic	0.71591	-0.3868429278	0.04945	
	D	40,626	89,224	Line (C1= m and C2= b)	0.9627	0.0000007675		
10	A	51	3,336	Constant	0.93683			
	B	3,337	12,100	Log10-Log10 Cubic	7.3152	-5.3114112439	1.26928	-0.1003
	C	12,101	45,150	Log10-Log10 Quadratic	0.5754	-0.3149352684	0.04033	
	D	45,151	97,408	Line (C1= m and C2= b)	0.96219	0.0000007235		
11	A	51	3,765	Constant	0.92948			
	B	3,766	13,400	Log10-Log10 Cubic	6.33025	-4.5652093163	1.08231	-0.0848
	C	13,401	49,917	Log10-Log10 Quadratic	0.4264	-0.2407097699	0.03115	
	D	49,918	106,671	Line (C1= m and C2= b)	0.96162	0.0000006763		
12	A	51	4,218	Constant	0.92198			
	B	4,219	14,800	Log10-Log10 Cubic	4.98994	-3.5857176421	0.84491	-0.0657
	C	14,801	54,434	Log10-Log10 Quadratic	0.26715	-0.1634456012	0.02183	
	D	54,435	114,298	Line (C1= m and C2= b)	0.96084	0.0000006541		
13	A	51	4,671	Constant	0.92562			
	B	4,672	12,900	Log10-Log10 Cubic	5.62479	-4.0015748693	0.93521	-0.0722
	C	12,901	58,961	Log10-Log10 Quadratic	0.2831	-0.1684096940	0.02212	
	D	58,962	122,454	Line (C1= m and C2= b)	0.96183	0.0000006011		
14	A	51	5,133	Constant	0.92835			
	B	5,134	13,800	Log10-Log10 Cubic	5.95354	-4.1982490417	0.97354	-0.0746
	C	13,801	63,302	Log10-Log10 Quadratic	0.28973	-0.1691164402	0.02194	
	D	63,303	130,613	Line (C1= m and C2= b)	0.96256	0.0000005562		
15	A	51	5,573	Constant	0.93092			
	B	5,574	15,000	Log10-Log10 Cubic	6.01221	-4.2039766602	0.96714	-0.0736
	C	15,001	68,492	Log10-Log10 Quadratic	0.2881	-0.1658382799	0.02128	
	D	68,493	139,999	Line (C1= m and C2= b)	0.96336	0.0000005562		
16	A	51	6,040	Constant	0.93299			
	B	6,041	20,000	Log10-Log10 Cubic	6.03715	-4.1900057268	0.95715	-0.0723
	C	20,001	73,690	Log10-Log10 Quadratic	0.2771	-0.1584836840	0.02019	
	D	73,691	148,620	Line (C1= m and C2= b)	0.96404	0.0000004799		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
17	A	51	6,505	Constant	0.93515			
	B	6,506	21,000	Log10-Log10 Cubic	6.3595	-4.3846910584	0.9958	-0.0748
	C	21,001	77,558	Log10-Log10 Quadratic	0.28357	-0.1597125440	0.02015	
	D	77,559	155,268	Line (C1= m and C2= b)	0.96466	0.0000004548		
18	A	51	6,979	Constant	0.93723			
	B	6,980	22,700	Log10-Log10 Cubic	6.60112	-4.5212397393	1.02062	-0.0762
	C	22,701	82,510	Log10-Log10 Quadratic	0.28491	-0.1585531188	0.01983	
	D	82,511	163,997	Line (C1= m and C2= b)	0.96547	0.0000004238		
19	A	51	7,462	Constant	0.93911			
	B	7,463	23,900	Log10-Log10 Cubic	6.86596	-4.6745775179	1.04952	-0.078
	C	23,901	86,868	Log10-Log10 Quadratic	0.28819	-0.1584532865	0.01965	
	D	86,869	170,083	Line (C1= m and C2= b)	0.96613	0.0000004070		
20	A	51	7,959	Constant	0.94104			
	B	7,960	25,500	Log10-Log10 Cubic	7.08419	-4.7943952397	1.07051	-0.0791
	C	25,501	91,524	Log10-Log10 Quadratic	0.292	-0.1585728258	0.0195	
	D	91,525	177,631	Line (C1= m and C2= b)	0.96686	0.0000003849		
21	A	51	8,411	Constant	0.94274			
	B	8,412	26,750	Log10-Log10 Cubic	7.22215	-4.8615843620	1.08005	-0.0795
	C	26,751	96,502	Log10-Log10 Quadratic	0.29317	-0.1576419079	0.01925	
	D	96,503	186,522	Line (C1= m and C2= b)	0.9676	0.0000003599		
22	A	51	8,889	Constant	0.9445			
	B	8,890	28,100	Log10-Log10 Cubic	7.4093	-4.9616460858	1.09696	-0.0803
	C	28,101	101,169	Log10-Log10 Quadratic	0.29772	-0.1582725229	0.01918	
	D	101,170	194,649	Line (C1= m and C2= b)	0.96831	0.0000003390		
23	A	51	9,389	Constant	0.94623			
	B	9,390	29,500	Log10-Log10 Cubic	7.57276	-5.0453328033	1.11018	-0.0809
	C	29,501	105,759	Log10-Log10 Quadratic	0.30285	-0.1591522224	0.01914	
	D	105,760	201,592	Line (C1= m and C2= b)	0.96895	0.0000003240		
24	A	51	9,852	Constant	9853			
	B	9,853	31,000	Log10-Log10 Cubic	7.68002	-5.0928725095	1.1157	-0.081
	C	31,001	110,468	Log10-Log10 Quadratic	0.307	-0.1596808311	0.01907	
	D	110,469	207,881	Line (C1= m and C2= b)	0.96962	0.0000003118		
25	A	51	10,342	Constant	0.94941			
	B	10,343	32,300	Log10-Log10 Cubic	7.80893	-5.1544774357	1.12428	-0.0813
	C	32,301	115,375	Log10-Log10 Quadratic	0.31133	-0.1603210874	0.01901	
	D	115,376	217,907	Line (C1= m and C2= b)	0.97031	0.0000002896		
26	A	51	10,835	Constant	0.95089			
	B	10,836	33,500	Log10-Log10 Cubic	8.00558	-5.2620537953	1.14327	-0.0823
	C	33,501	119,496	Log10-Log10 Quadratic	0.31836	-0.1622678128	0.01911	
	D	119,497	225,030	Line (C1= m and C2= b)	0.97087	0.0000002760		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
27	A	51	11,318	Constant	0.95243			
	B	11,319	35,000	Log10-Log10 Cubic	8.06202	-5.2767565134	1.14185	-0.0819
	C	35,001	124,295	Log10-Log10 Quadratic	0.32327	-0.1631763677	0.01909	
	D	124,296	233,993	Line (C1= m and C2= b)	0.97154	0.0000002595		
28	A	51	11,852	Constant	0.95393			
	B	11,853	36,300	Log10-Log10 Cubic	8.22296	-5.3592619123	1.15509	-0.0825
	C	36,301	128,840	Log10-Log10 Quadratic	0.33068	-0.1652284960	0.0192	
	D	128,841	241,529	Line (C1= m and C2= b)	0.97212	0.0000002474		
29	A	51	12,319	Constant	0.95524			
	B	12,320	37,700	Log10-Log10 Cubic	8.24358	-5.3522305553	1.14935	-0.0818
	C	37,701	133,782	Log10-Log10 Quadratic	0.33379	-0.1654797210	0.01912	
	D	133,783	249,163	Line (C1= m and C2= b)	0.97275	0.0000002362		
30	A	51	12,811	Constant	0.95663			
	B	12,812	39,000	Log10-Log10 Cubic	8.35346	-5.4032014313	1.15617	-0.082
	C	39,001	138,266	Log10-Log10 Quadratic	0.3413	-0.1676708908	0.01926	
	D	138,267	239,318	Line (C1= m and C2= b)	0.97333	0.0000002639		
31	A	51	13,375	Constant	0.958			
	B	13,376	40,500	Log10-Log10 Cubic	8.52009	-5.4904964718	1.17077	-0.0828
	C	40,501	142,612	Log10-Log10 Quadratic	0.34866	-0.1697504644	0.01938	
	D	142,613	265,420	Line (C1= m and C2= b)	0.97388	0.0000002127		
32	A	51	13,838	Constant	0.95917			
	B	13,839	41,500	Log10-Log10 Cubic	8.48164	-5.4468271723	1.15756	-0.0816
	C	41,501	147,465	Log10-Log10 Quadratic	0.35274	-0.1704882715	0.01936	
	D	147,466	271,220	Line (C1= m and C2= b)	0.97444	0.0000002066		
33	A	51	14,385	Constant	0.9605			
	B	14,386	43,200	Log10-Log10 Cubic	8.60465	-5.5059347391	1.16614	-0.0819
	C	43,201	152,367	Log10-Log10 Quadratic	0.35969	-0.1724239548	0.01947	
	D	152,368	279,208	Line (C1= m and C2= b)	0.97503	0.0000001969		
34	A	51	14,914	Constant	0.96179			
	B	14,915	44,600	Log10-Log10 Cubic	8.64496	-5.5117456963	1.16331	-0.0815
	C	44,601	157,912	Log10-Log10 Quadratic	0.36526	-0.1737900390	0.01952	
	D	157,913	287,054	Line (C1= m and C2= b)	0.97572	0.0000001880		
35	A	51	15,380	Constant	0.96293			
	B	15,381	45,800	Log10-Log10 Cubic	8.75243	-5.5650925696	1.17158	-0.0818
	C	45,801	160,843	Log10-Log10 Quadratic	0.37522	-0.1771714507	0.0198	
	D	160,844	294,503	Line (C1= m and C2= b)	0.97604	0.0000001792		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
36	A	51	15,961	Constant	0.96414			
	B	15,962	47,300	Log10-Log10 Cubic	8.80507	-5.5789392078	1.17055	-0.0815
	C	47,301	166,334	Log10-Log10 Quadratic	0.38036	-0.1783330903	0.01983	
	D	166,335	303,528	Line (C1= m and C2= b)	0.97666	0.0000001701		
37	A	51	16,384	Constant	0.96533			
	B	16,385	49,000	Log10-Log10 Cubic	8.72895	-5.5142209618	1.1536	-0.0801
	C	49,001	171,223	Log10-Log10 Quadratic	0.38752	-0.1804542509	0.01997	
	D	171,224	310,150	Line (C1= m and C2= b)	0.97723	0.0000001639		
38	A	51	16,899	Constant	0.96641			
	B	16,900	50,000	Log10-Log10 Cubic	8.74565	-5.5081973951	1.149	-0.0795
	C	50,001	176,317	Log10-Log10 Quadratic	0.39136	-0.1811777640	0.01996	
	D	176,318	318,342	Line (C1= m and C2= b)	0.9778	0.0000001563		
39	A	51	17,438	Constant	0.96756			
	B	17,439	51,500	Log10-Log10 Cubic	8.80864	-5.5309815528	1.1504	-0.0794
	C	51,501	181,108	Log10-Log10 Quadratic	0.40054	-0.1841547520	0.02019	
	D	181,109	327,005	Line (C1= m and C2= b)	0.97832	0.0000001486		
40	A	51	17,946	Constant	0.96871			
	B	17,947	52,750	Log10-Log10 Cubic	8.74909	-5.4776029778	1.13609	-0.0782
	C	52,751	185,418	Log10-Log10 Quadratic	0.40868	-0.1866418148	0.02037	
	D	185,419	301,516	Line (C1= m and C2= b)	0.97877	0.0000001828		
41	A	51	18,399	Constant	0.96973			
	B	18,400	54,000	Log10-Log10 Cubic	8.79886	-5.4949973504	1.13697	-0.0781
	C	54,001	189,553	Log10-Log10 Quadratic	0.41743	-0.1895402963	0.0206	
	D	189,554	339,349	Line (C1= m and C2= b)	0.97922	0.0000001387		
42	A	51	18,988	Constant	0.97085			
	B	18,989	55,500	Log10-Log10 Cubic	8.82416	-5.4939691984	1.13342	-0.0776
	C	55,501	195,023	Log10-Log10 Quadratic	0.42413	-0.1914173967	0.02071	
	D	195,024	348,417	Line (C1= m and C2= b)	0.97978	0.0000001318		
43	A	51	19,504	Constant	0.97179			
	B	19,505	57,000	Log10-Log10 Cubic	8.73067	-5.4205615811	1.11521	-0.0762
	C	57,001	200,344	Log10-Log10 Quadratic	0.42797	-0.1921694132	0.02071	
	D	200,345	356,317	Line (C1= m and C2= b)	0.98031	0.0000001263		
44	A	51	20,051	Constant	0.97288			
	B	20,052	58,300	Log10-Log10 Cubic	8.79712	-5.4469249713	1.11771	-0.0761
	C	58,301	205,283	Log10-Log10 Quadratic	0.43743	-0.1952772019	*0.02096*	
	D	205,284	362,245	Line (C1= m and C2= b)	0.98082	0.0000001222		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
45	A	51	20,583	Constant	0.97391			
	B	20,584	59,700	Log10-Log10 Cubic	8.76577	-5.4128755524	1.10782	-0.0753
	C	59,701	209,826	Log10-Log10 Quadratic	0.44597	-0.1979802132	0.02116	
	D	209,827	370,629	Line (C1= m and C2= b)	0.98123	0.0000001167		
46	A	51	21,074	Constant	0.9749			
	B	21,075	61,300	Log10-Log10 Cubic	8.71724	-5.3710822314	1.09697	-0.0744
	C	61,301	213,098	Log10-Log10 Quadratic	0.4517	-0.1995403580	0.02125	
	D	213,099	378,013	Line (C1= m and C2= b)	0.98157	0.0000001118		
47	A	51	21,635	Constant	0.97582			
	B	21,636	62,600	Log10-Log10 Cubic	8.71833	-5.3574844497	1.09137	-0.0738
	C	62,601	218,419	Log10-Log10 Quadratic	0.4586	-0.2016072791	0.0214	
	D	218,420	384,952	Line (C1= m and C2= b)	0.98206	0.0000001077		
48	A	51	22,155	Constant	0.97689			
	B	22,156	64,000	Log10-Log10 Cubic	8.68622	-5.3238643715	1.08179	-0.073
	C	64,001	223,001	Log10-Log10 Quadratic	0.4686	-0.2049163494	0.02166	
	D	223,002	392,076	Line (C1= m and C2= b)	0.98249	0.0000001035		
49	A	51	22,655	Constant	0.97772			
	B	22,656	65,300	Log10-Log10 Cubic	8.61097	-5.2653979547	1.06746	-0.0719
	C	65,301	227,791	Log10-Log10 Quadratic	0.47297	-0.2059762289	0.02171	
	D	227,792	399,869	Line (C1= m and C2= b)	0.98292	0.0000000993		
50	A	51	23,144	Constant	0.97874			
	B	23,145	66,700	Log10-Log10 Cubic	8.51836	-5.1965486331	1.05112	-0.0706
	C	66,701	231,295	Log10-Log10 Quadratic	0.48343	-0.2094683826	0.02199	
	D	231,296	407,231	Line (C1= m and C2= b)	0.98324	0.0000000952		
51	A	51	23,714	Constant	0.97959			
	B	23,715	68,100	Log10-Log10 Cubic	8.58137	-5.2228939903	1.05409	-0.0706
	C	68,101	236,850	Log10-Log10 Quadratic	0.48924	-0.2111573475	0.02211	
	D	236,851	415,131	Line (C1= m and C2= b)	0.9838	0.0000000909		
52	*A*	*51*	*24,243*	*Constant*	*0.98053*			
	B	24,244	69,600	Log10-Log10 Cubic	8.56614	-5.2022787638	1.04775	-0.0701
	C	69,601	240,314	Log10-Log10 Quadratic	0.49771	-0.2138742516	0.02232	
	D	240,315	422,838	Line (C1= m and C2= b)	0.9841	0.0000000871		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
53	A	51	24,700	Constant	0.98118			
	B	24,701	71,100	Log10-Log10 Cubic	8.41163	-5.0969237864	1.02421	-0.0684
	C	71,101	246,671	Log10-Log10 Quadratic	0.4961	-0.2125883238	0.02213	
	D	246,672	428,614	Line (C1= m and C2= b)	0.98467	0.0000000843		
54	A	51	25,256	Constant	0.98196			
	B	25,257	72,500	Log10-Log10 Cubic	8.43964	-5.1044200744	1.02392	-0.0682
	C	72,501	249,886	Log10-Log10 Quadratic	0.50133	-0.2140452025	0.02222	
	D	249,887	437,018	Line (C1= m and C2= b)	0.98494	0.0000000805		
55	A	51	25,768	Constant	0.98259			
	B	25,769	73,700	Log10-Log10 Cubic	8.38483	-5.0609738063	1.01319	-0.0674
	C	73,701	254,587		0.50468	-0.2148067115	0.02225	
	D	254,588	442,936	Line (C1= m and C2= b)	0.98534	0.0000000779		
56	A	51	26,280	Constant	0.98332			
	B	26,281	74,800	Log10-Log10 Cubic	8.30592	-5.0018314423	0.99908	-0.0663
	C	74,801	260,229	Log10-Log10 Quadratic	0.51098	-0.2167404331	0.02239	
	D	260,230	451,854	Line (C1= m and C2= b)	0.98582	0.0000000740		
57	A	51	26,734	Constant	0.98397			
	B	26,735	76,500	Log10-Log10 Cubic	8.16592	-4.9088811922	0.97883	-0.0648
	C	76,501	263,552	Log10-Log10 Quadratic	0.51027	-0.2158201160	0.02224	
	D	263,553	458,089	Line (C1= m and C2= b)	0.98606	0.0000000717		
58	A	51	27,371	Constant	0.9847			
	B	27,372	77,900	Log10-Log10 Cubic	8.26716	-4.9609748182	0.98759	-0.0653
	C	77,901	267,196	Log10-Log10 Quadratic	0.51649	-0.2177068384	0.02238	
	D	267,197	466,019	Line (C1= m and C2= b)	0.98638	0.0000000685		
59	A	51	27,910	Constant	0.98531			
	B	27,911	79,400	Log10-Log10 Cubic	8.14568	-4.8776797063	0.96897	-0.0639
	C	79,401	272,890	Log10-Log10 Quadratic	0.51628	-0.2170096407	0.02225	
	D	272,891	475,040	Line (C1= m and C2= b)	0.98684	0.0000000651		
60	A	51	28,404	Constant	0.98596			
	B	28,405	80,500	Log10-Log10 Cubic	8.07343	-4.8250858639	0.9567	-0.063
	C	80,501	277,612	Log10-Log10 Quadratic	0.52625	-0.2197938338	0.02243	
	D	277,613	481,867	Line (C1= m and C2= b)	0.98725	0.0000000624		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
61	A	51	29,038	Constant	0.98672			
	B	29,039	82,000	Log10-Log10 Cubic	8.02976	-4.7881649218	0.9473	-0.0623
	C	82,001	283,196	Log10-Log10 Quadratic	0.52807	-0.2200778579	0.02242	
	D	283,197	489,064	Line (C1= m and C2= b)	0.98768	0.0000000599		
62	A	51	29,466	Constant	0.98724			
	B	29,467	83,200	Log10-Log10 Cubic	7.96352	-4.7408392322	0.9364	-0.0614
	C	83,201	288,350	Log10-Log10 Quadratic	0.52694	-0.2189759548	0.02225	
	D	288,351	497,233	Line (C1= m and C2= b)	0.98814	0.0000000568		
63	A	51	30,085	Constant	0.98789			
	B	30,086	85,100	Log10-Log10 Cubic	7.85936	-4.6683287516	0.92004	-0.0602
	C	85,101	294,569	Log10-Log10 Quadratic	0.53231	-0.2205501273	0.02236	
	D	294,570	504,300	Line (C1= m and C2= b)	0.98858	0.0000000545		
64	A	51	30,657	Constant	0.98855			
	B	30,658	86,500	Log10-Log10 Cubic	7.83885	-4.6478952633	0.91445	-0.0598
	C	86,501	298,371	Log10-Log10 Quadratic	0.52993	-0.2191099262	0.02218	
	D	298,372	510,396	Line (C1= m and C2= b)	0.98886	0.0000000526		
65	A	51	31,106	Constant	0.98900			
	B	31,107	87,900	Log10-Log10 Cubic	7.70823	-4.5635749962	0.89653	-0.0585
	C	87,901	302,140	Log10-Log10 Quadratic	0.53295	-0.2197758341	0.0222	
	D	302,141	519,956	Line (C1= m and C2= b)	0.98912	0.0000000499		
66	A	51	31,658	Constant	0.98965			
	B	31,659	89,600	Log10-Log10 Cubic	7.81415	-4.6228578772	0.90766	-0.0592
	C	89,601	302,140	Log10-Log10 Quadratic	0.53306	-0.2192671932	0.0221	
	D	302,141	524,828	Line (C1= m and C2= b)	0.98918	0.0000000486		
67	A	51	32,182	Constant	0.99021			
	B	32,183	90,900	Log10-Log10 Cubic	7.5079	-4.4276041070	0.86647	-0.0563
	C	90,901	312,866	Log10-Log10 Quadratic	0.5126	-0.2023614945	0.01973	
	D	312,867	536,209	Line (C1= m and C2= b)	0.98991	0.0000000452		
68	A	51	32,698	Constant	0.9906			
	B	32,699	92,400	Log10-Log10 Cubic	7.47995	-4.4061127434	0.86133	-0.0559
	C	92,401	316,456	Log10-Log10 Quadratic	0.52987	-0.2175731016	0.0219	
	D	316,457	539,552	Line (C1= m and C2= b)	0.99021	0.0000000439		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
69	A	51	33,239	Constant	0.99096			
	B	33,240	93,600	Log10-Log10 Cubic	7.42255	-4.3659427914	0.85227	-0.0553
	C	93,601	320,322	Log10-Log10 Quadratic	0.52894	-0.2167723751	0.02179	
	D	320,323	548,261	Line (C1= m and C2= b)	0.99044	0.0000000419		
70	A	51	33,652	Constant	0.99138			
	B	33,653	94,900	Log10-Log10 Cubic	7.26012	-4.2641052988	0.83118	-0.0538
	C	94,901	323,337	Log10-Log10 Quadratic	0.52708	-0.2155854502	0.02163	
	D	323,338	554,659	Line (C1= m and C2= b)	0.99061	0.0000000406		
71	A	51	34,267	Constant	0.99186			
	B	34,268	96,800	Log10-Log10 Cubic	7.1923	-4.2169301442	0.82058	-0.0531
	C	96,801	329,838	Log10-Log10 Quadratic	0.52351	-0.2136827375	0.0214	
	D	329,839	561,106	Line (C1= m and C2= b)	0.9911	0.0000000385		
72	A	51	34,813	Constant	0.99225			
	B	34,814	98,200	Log10-Log10 Cubic	7.09779	-4.1545002169	0.80707	-0.0521
	C	98,201	335,591	Log10-Log10 Quadratic	0.52197	-0.2126534244	0.02127	
	D	335,592	575,895	Line (C1= m and C2= b)	0.9915	0.0000000354		
73	A	51	35,370	Constant	0.9927			
	B	35,371	99,300	Log10-Log10 Cubic	7.03376	-4.1099144784	0.79705	-0.0514
	C	99,301	340,254	Log10-Log10 Quadratic	0.52539	-0.2135764048	0.02132	
	D	340,255	575,765	Line (C1= m and C2= b)	0.9918	0.0000000348		
74	A	51	36,018	Constant	0.9932			
	B	36,019	100,800	Log10-Log10 Cubic	6.99902	-4.0832060246	0.79067	-0.0509
	C	100,801	345,504	Log10-Log10 Quadratic	0.52529	-0.2130659515	0.02124	
	D	345,505	585,587	Line (C1= m and C2= b)	0.99221	0.0000000325		
75	A	51	36,522	Constant	0.99361			
	B	36,523	102,500	Log10-Log10 Cubic	6.86536	-3.9991967230	0.77326	-0.0497
	C	102,501	349,203	Log10-Log10 Quadratic	0.52159	-0.2111564225	0.02101	
	D	349,204	591,502	Line (C1= m and C2= b)	0.9924	0.0000000314		
76	A	51	37,039	Constant	0.99399			
	B	37,040	103,800	Log10-Log10 Cubic	6.69448	-3.8924377046	0.75121	-0.0482
	C	103,801	354,183	Log10-Log10 Quadratic	0.52094	-0.2104695232	0.02091	
	D	354,184	601,662	Line (C1= m and C2= b)	0.99269	0.0000000295		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
77	A	51	37,546	Constant	0.99443			
	B	37,547	105,300	Log10-Log10 Cubic	6.64494	-3.8591717288	0.74398	-0.0477
	C	105,301	357,570	Log10-Log10 Quadratic	0.51858	-0.2091401787	0.02074	
	D	357,571	604,617	Line (C1= m and C2= b)	0.99294	0.0000000286		
78	A	51	38,206	Constant	0.99489			
	B	38,207	107,000	Log10-Log10 Cubic	6.49795	-3.7661270101	0.72457	-0.0463
	C	107,001	363,649	Log10-Log10 Quadratic	0.51757	-0.2082500928	0.02062	
	D	363,650	617,499	Line (C1= m and C2= b)	0.99332	0.0000000263		
79	A	51	38,724	Constant	0.99532			
	B	38,725	108,700	Log10-Log10 Cubic	6.37182	-3.6878824123	0.70856	-0.0452
	C	108,701	367,386	Log10-Log10 Quadratic	0.51291	-0.2059898343	0.02036	
	D	367,387	621,255	Line (C1= m and C2= b)	0.99356	0.0000000254		
80	A	51	39,257	Constant	0.9957			
	B	39,258	109,900	Log10-Log10 Cubic	6.29415	-3.6365939507	0.69748	-0.0445
	C	109,901	373,220	Log10-Log10 Quadratic	0.51513	-0.2065133822	0.02039	
	D	373,221	631,524	Line (C1= m and C2= b)	0.99394	0.0000000235		
81	A	51	39,828	Constant	0.99612			
	B	39,829	111,200	Log10-Log10 Cubic	6.26297	-3.6137926603	0.69222	-0.0441
	C	111,201	377,485	Log10-Log10 Quadratic	0.51609	-0.2065195574	0.02036	
	D	377,486	639,864	Line (C1= m and C2= b)	0.99424	0.0000000220		
82	A	51	40,445	Constant	0.99665			
	B	40,446	112,100	Log10-Log10 Cubic	6.26441	-3.6082197217	0.68996	-0.0438
	C	112,101	382,122	Log10-Log10 Quadratic	0.52572	-0.2098929314	0.02065	
	D	382,123	645,202	Line (C1= m and C2= b)	0.99454	0.0000000207		
83	A	51	40,908	Constant	0.99695			
	B	40,909	114,000	Log10-Log10 Cubic	5.96683	-3.4303842131	0.65471	-0.0415
	C	114,001	385,942	Log10-Log10 Quadratic	0.51611	-0.2056923349	0.02021	
	D	385,943	652,058	Line (C1= m and C2= b)	0.9947	0.0000000199		
84	A	51	41,421	Constant	0.99731			
	B	41,422	115,200	Log10-Log10 Cubic	5.85731	-3.3622734215	0.64074	-0.0406
	C	115,201	389,382	Log10-Log10 Quadratic	0.51654	-0.2054922329	0.02016	
	D	389,383	655,475	Line (C1= m and C2= b)	0.99489	0.0000000192		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
85	A	51	42,021	Constant	0.99777			
	B	42,022	116,800	Log10-Log10 Cubic	5.78353	-3.3138643315	0.63036	-0.0398
	C	116,801	397,082	Log10-Log10 Quadratic	0.51653	-0.2051295229	0.0201	
	D	397,083	667,679	Line (C1= m and C2= b)	0.99543	0.0000000169		
86	A	51	42,608	Constant	0.99827			
	B	42,609	118,100	Log10-Log10 Cubic	5.71191	-3.2675558679	0.62058	-0.0392
	C	118,101	400,738	Log10-Log10 Quadratic	0.51954	-0.2058833134	0.02014	
	D	400,739	673,287	Line (C1= m and C2= b)	0.99566	0.0000000159		
87	A	51	43,252	Constant	0.99869			
	B	43,253	119,700	Log10-Log10 Cubic	5.59695	-3.1959122189	0.60587	-0.0382
	C	119,701	405,615	Log10-Log10 Quadratic	0.51916	-0.2053154653	0.02005	
	D	405,616	684,223	Line (C1= m and C2= b)	0.99594	0.0000000146		
88	A	51	43,713	Constant	0.99907			
	B	43,714	121,800	Log10-Log10 Cubic	5.38796	-3.0709798288	0.58112	-0.0365
	C	121,801	409,825	Log10-Log10 Quadratic	0.5126	-0.2023614945	0.01973	
	D	409,826	687,322	Line (C1= m and C2= b)	0.99615	0.0000000139		
89	A	51	44,281	Constant	0.99949			
	B	44,282	122,700	Log10-Log10 Cubic	5.31758	-3.0258954677	0.57167	-0.0359
	C	122,701	412,400	Log10-Log10 Quadratic	0.51762	-0.2039370947	0.01985	
	D	412,401	696,136	Line (C1= m and C2= b)	0.9963	0.0000000130		
90	A	51	44,987	Constant	0.99995			
	B	44,988	124,300	Log10-Log10 Cubic	5.30199	-3.0127527785	0.56842	-0.0356
	C	124,301	416,677	Log10-Log10 Quadratic	0.51964	-0.2043220234	0.01986	
	D	416,678	702,364	Line (C1= m and C2= b)	0.99657	0.0000000120		
91	A	51	45,470	Constant	1.0003			
	B	45,471	125,700	Log10-Log10 Cubic	5.06191	-2.8692896607	0.53999	-0.0338
	C	125,701	421,700	Log10-Log10 Quadratic	0.51717	-0.2029876934	0.0197	
	D	421,701	706,736	Line (C1= m and C2= b)	0.99682	0.0000000111		
92	A	51	45,860	Constant	1.00064			
	B	45,861	127,200	Log10-Log10 Cubic	4.80507	-2.7164303587	0.50981	-0.0318
	C	127,201	427,910	Log10-Log10 Quadratic	0.51405	-0.2014490218	0.01952	
	D	427,911	719,371	Line (C1= m and C2= b)	0.99716	0.0000000098		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
93	A	51	46,543	Constant	1.00112			
	B	46,544	128,700	Log10-Log10 Cubic	4.76421	-2.6890651276	0.5039	-0.0314
	C	128,701	431,187	Log10-Log10 Quadratic	0.51694	-0.2021510738	0.01956	
	D	431,188	724,404	Line (C1= m and C2= b)	0.99735	0.0000000090		
94	A	51	47,048	Constant	1.0015			
	B	47,049	130,300	Log10-Log10 Cubic	4.60981	-2.5970888947	0.48576	-0.0302
	C	130,301	435,478	Log10-Log10 Quadratic	0.51295	-0.2002558995	0.01935	
	D	435,479	729,043	Line (C1= m and C2= b)	0.99763	0.0000000081		
95	A	51	47,590	Constant	1.00194			
	B	47,591	131,900	Log10-Log10 Cubic	4.52941	-2.5473245684	0.47562	-0.0295
	C	131,901	441,046	Log10-Log10 Quadratic	0.51258	-0.1997728719	0.01928	
	D	441,047	737,941	Line (C1= m and C2= b)	0.99797	0.0000000068		
96	A	51	48,189	Constant	1.00241			
	B	48,190	132,700	Log10-Log10 Cubic	4.38688	-2.4593100901	0.45769	-0.0283
	C	132,701	445,716	Log10-Log10 Quadratic	0.52324	-0.2035166958	0.01961	
	D	445,717	743,210	Line (C1= m and C2= b)	0.9982	0.0000000061		
97	A	51	48,721	Constant	1.00274			
	B	48,722	134,300	Log10-Log10 Cubic	4.11951	-2.3013317358	0.42672	-0.0263
	C	134,301	451,274	Log10-Log10 Quadratic	0.52083	-0.2022207466	0.01945	
	D	451,275	751,297	Line (C1= m and C2= b)	0.99848	0.0000000051		
98	A	51	49,260	Constant	1.00309			
	B	49,261	136,100	Log10-Log10 Cubic	3.97696	-2.2166865876	0.41008	-0.0252
	C	136,101	457,397	Log10-Log10 Quadratic	0.51606	-0.2000528899	0.01922	
	D	457,398	762,524	Line (C1= m and C2= b)	0.99883	0.0000000038		
99	A	51	49,872	Constant	1.00347			
	B	49,873	137,100	Log10-Log10 Cubic	3.89447	-2.1659010308	0.39979	-0.0245
	C	137,101	458,686	Log10-Log10 Quadratic	0.52107	-0.2016149260	0.01934	
	D	458,687	764,863	Line (C1= m and C2= b)	0.99885	0.0000000038		
100	A	51	50,466	Constant	1.00401			
	B	50,467	138,600	Log10-Log10 Cubic	3.8001	-2.1077345298	0.38799	-0.0237
	C	138,601	464,514	Log10-Log10 Quadratic	0.52555	-0.2029655287	0.01944	
	D	464,515	775,268	Line (C1= m and C2= b)	0.99922	0.0000000025		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
101	A	51	50,940	Constant	1.00436			
	B	50,941	466,383	Log10-Log10 Cubic	3.68304	-2.0388595083	0.37459	-0.0229
	C	466,384	779,185	Line (C1= m and C2= b)	0.99793	0.0000000066		
102	A	51	51,479	Constant	1.00469			
	B	51,480	470,459	Log10-Log10 Cubic	3.50742	-1.9353228578	0.35436	-0.0215
	D	470,460	787,993	Line (C1= m and C2= b)	0.99821	0.0000000056		
103	A	51	52,126	Constant	1.00517			
	B	52,127	475,743	Log10-Log10 Cubic	3.37018	-1.8532160896	0.33812	-0.0205
	C	475,744	791,354	Line (C1= m and C2= b)	0.99857	0.0000000045		
104	A	51	52,691	Constant	1.00561			
	B	52,692	481,410	Log10-Log10 Cubic	3.20418	-1.7543260275	0.31861	-0.0192
	D	481,411	799,404	Line (C1= m and C2= b)	0.99894	0.0000000033		
105	A	51	53,307	Constant	1.00598			
	B	53,308	485,652	Log10-Log10 Cubic	3.03169	-1.6530714656	0.29893	-0.0179
	C	485,653	806,287	Line (C1= m and C2= b)	0.99921	0.0000000025		
106	A	51	53,985	Constant	1.00649			
	B	53,986	492,716	Log10-Log10 Cubic	2.97155	-1.6143284675	0.29078	-0.0174
	D	492,717	813,671	Line (C1= m and C2= b)	0.99967	0.0000000010		
107	A	51	54,349	Constant	1.00679			
	B	54,350	495,613	Log10-Log10 Cubic	2.64157	-1.4246620667	0.2546	-0.0151
	C	495,614	823,601	Line (C1= m and C2= b)	0.9999	0.0000000003		
108	A	51	55,055	Constant	1.00729			
	B	55,056	501,529	Log10-Log10 Cubic	2.4058	-1.2861637866	0.22764	-0.0133
	D	501,530	829,938	Line (C1= m and C2= b)	1.00028	-0.0000000008		
109	A	51	55,663	Constant	1.00729			
	B	55,664	507,110	Log10-Log10 Cubic	2.44434	-1.3070106312	0.23141	-0.0136
	C	507,111	833,015	Line (C1= m and C2= b)	1.00027	-0.0000000008		
110	A	51	56,092	Constant	1.00805			
	B	56,093	512,027	Log10-Log10 Cubic	2.08966	-1.1016563185	0.19195	-0.0111
	D	512,028	842,529	Line (C1= m and C2= b)	1.00097	-0.0000000029		
111	A	51	56,737	Constant	1.00855			
	B	56,738	514,519	Log10-Log10 Cubic	1.91349	-0.9975894506	0.17161	-0.0097
	C	514,520	847,529	Line (C1= m and C2= b)	1.00116	-0.0000000035		

Table A3.6, continued
Coefficients for equations describing the I:D ratio versus spacer size for Type 1 *Alu* pairs

APSN	Spacer Size Range, bp			Equation Type	Equation Coefficients			
	Range ID	Range Start (bp)	Range End (bp)		C1	C2	C3	C4
112	A	51	57,273	Constant	1.00885			
	B	57,274	521,729	Log10-Log10 Cubic	1.71257	-0.8805040568	0.14897	-0.0083
	D	521,730	860,157	Line (C1= m and C2= b)	1.00158	-0.0000000047		
113	A	51	57,808	Constant	1.00928			
	B	57,809	525,322	Log10-Log10 Cubic	1.60704	-0.8187515723	0.13703	-0.0075
	C	525,323	863,992	Line (C1= m and C2= b)	1.00182	-0.0000000054		
114	A	51	58,351	Constant	1.00978			
	B	58,352	525,735	Log10-Log10 Cubic	1.41992	-0.7094195331	0.11587	-0.0062
	D	525,736	870,703	Line (C1= m and C2= b)	1.00189	-0.0000000055		
115	A	51	58,992	Constant	1.0102			
	B	58,993	536,310	Log10-Log10 Cubic	1.18247	-0.5726831283	0.08975	-0.0045
	C	536,311	882,203	Line (C1= m and C2= b)	1.00252	-0.0000000073		

Figure A3.1
***Alu* landscapes for selected genes**

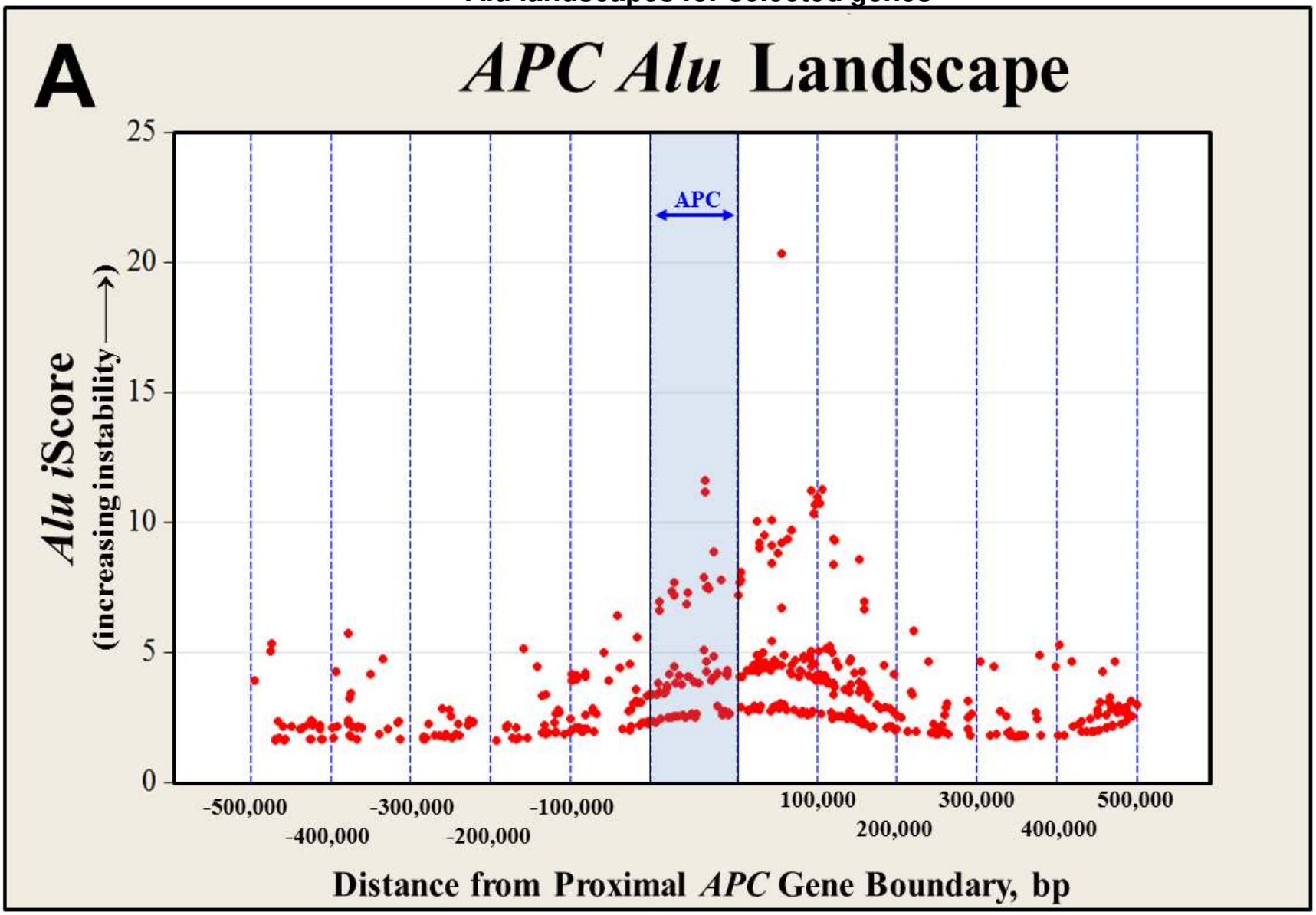


Figure A3.1, continued
***Alu* landscapes for selected genes**

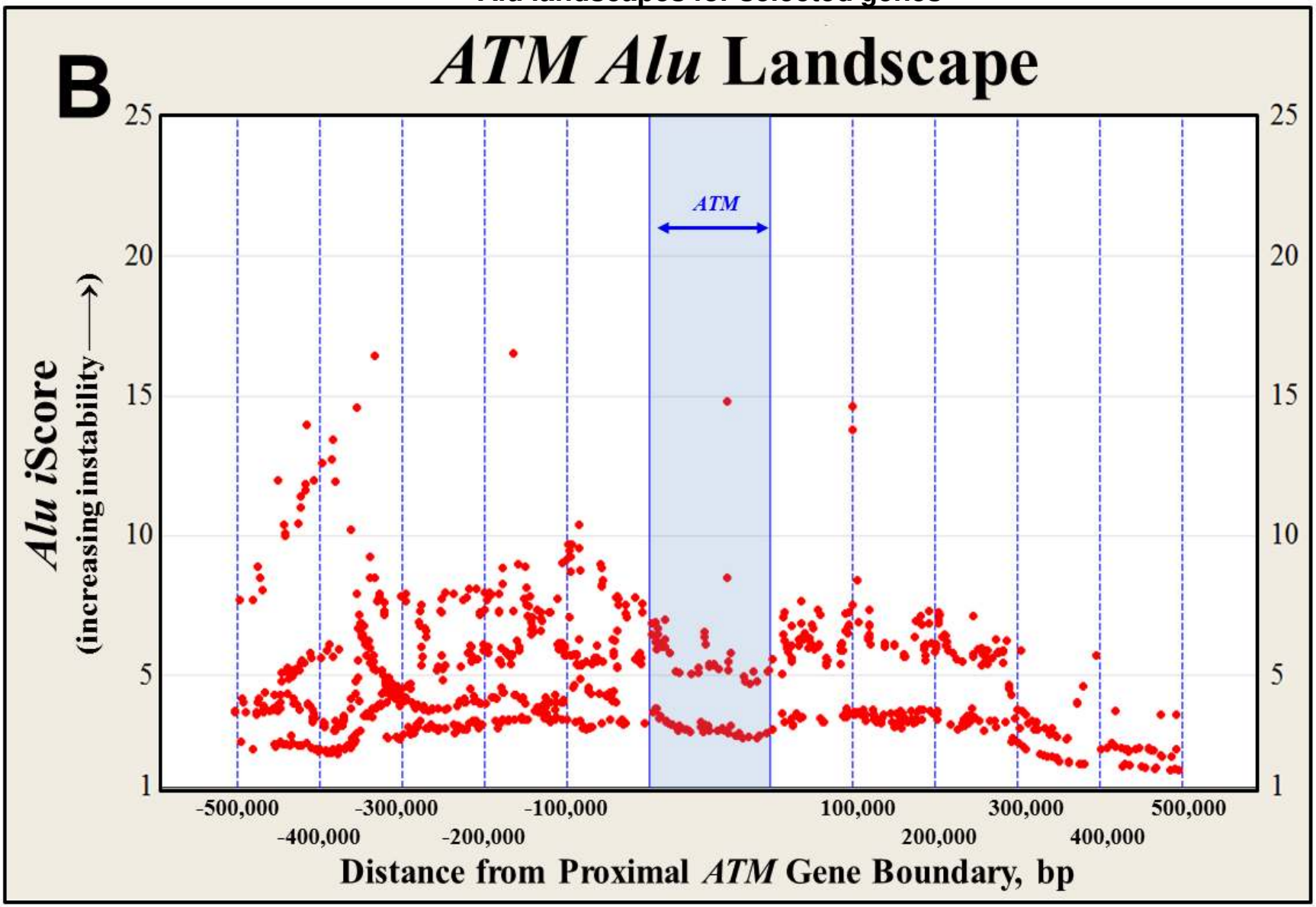


Figure A3.1, continued
***Alu* landscapes for selected genes**

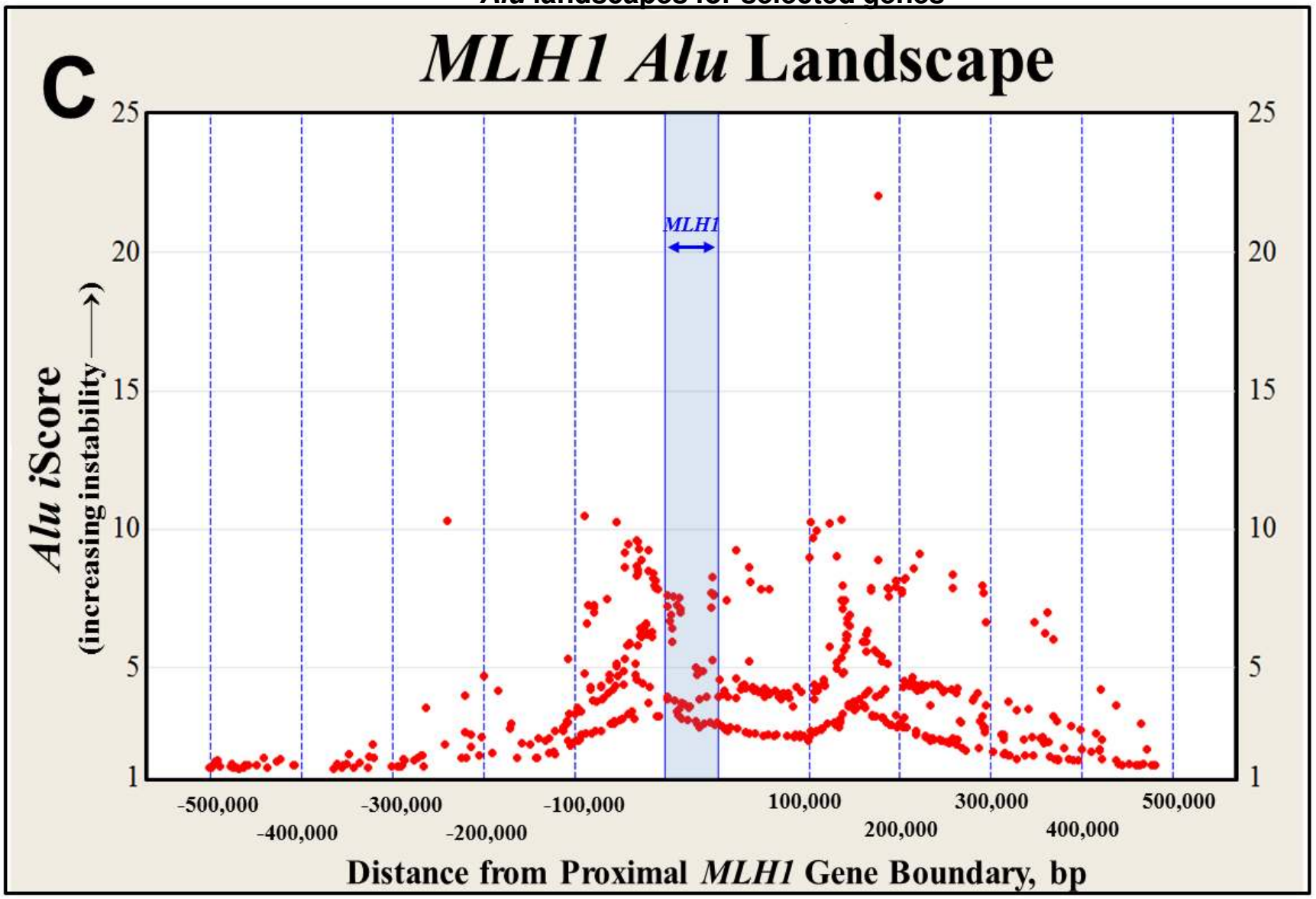


Figure A3.1, continued
***Alu* landscapes for selected genes**

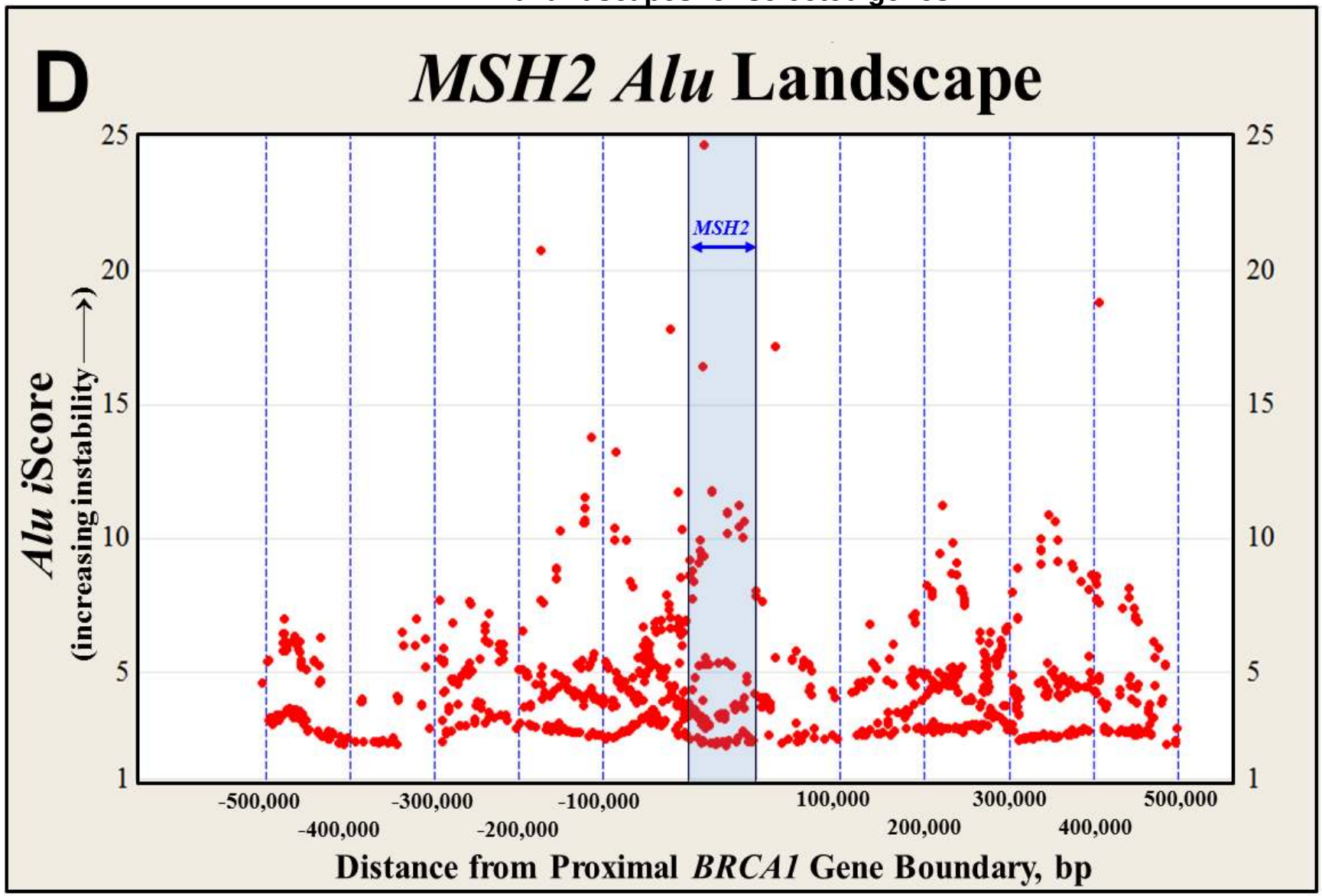


Figure A3.1, continued
***Alu* landscapes for selected genes**

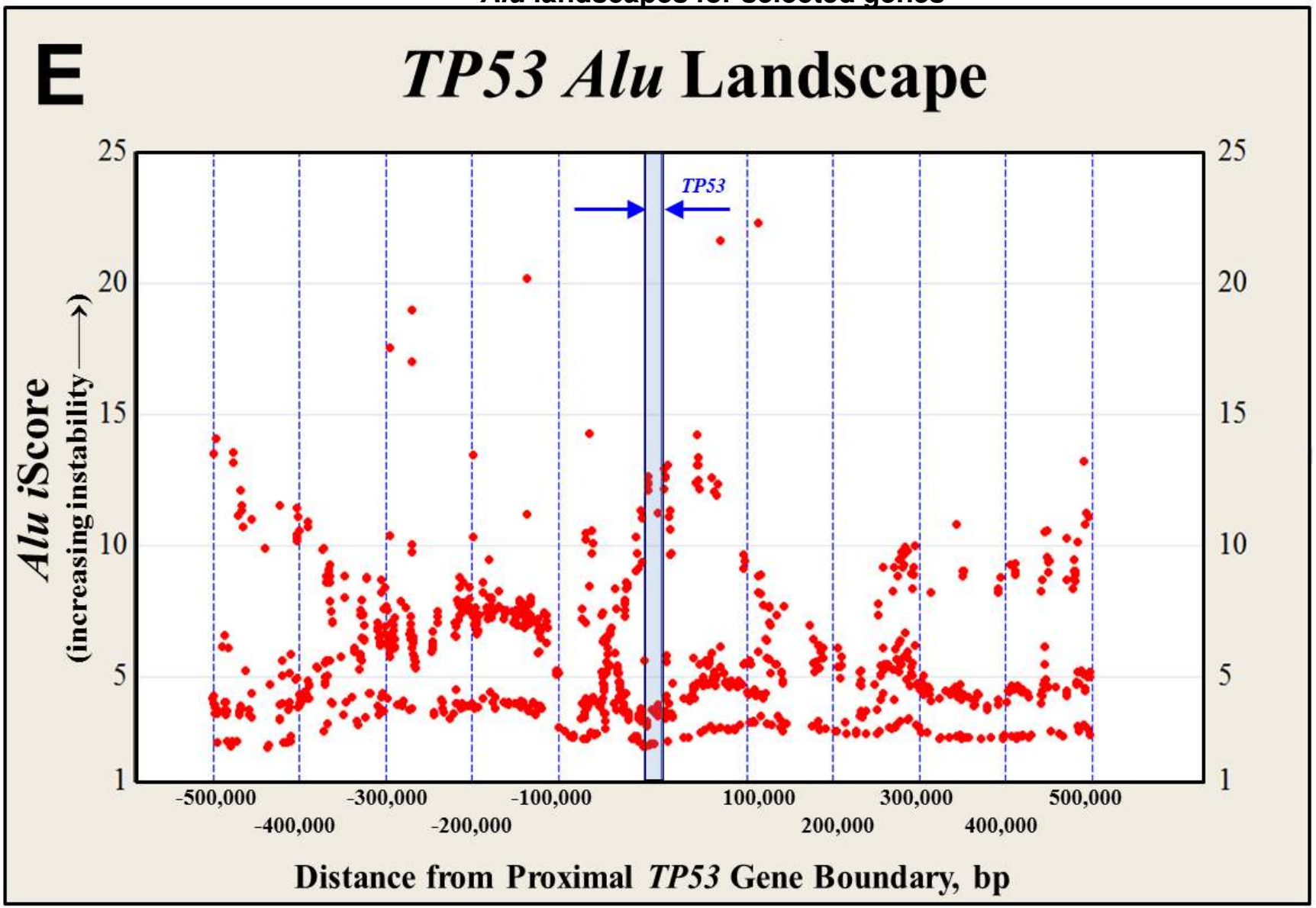


Figure A3.1, continued
***Alu* landscapes for selected genes**

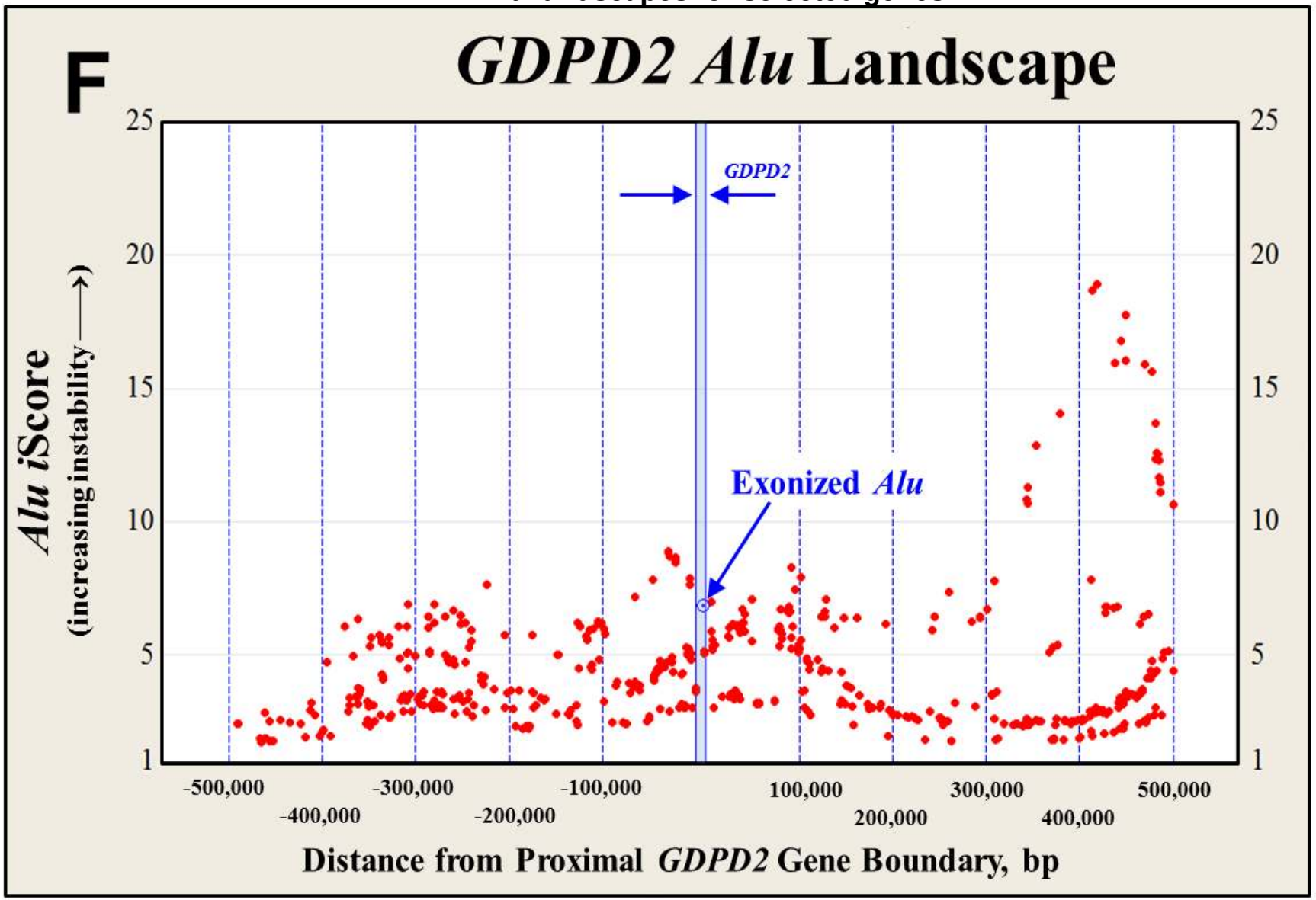


Figure A3.1, continued
***Alu* landscapes for selected genes**

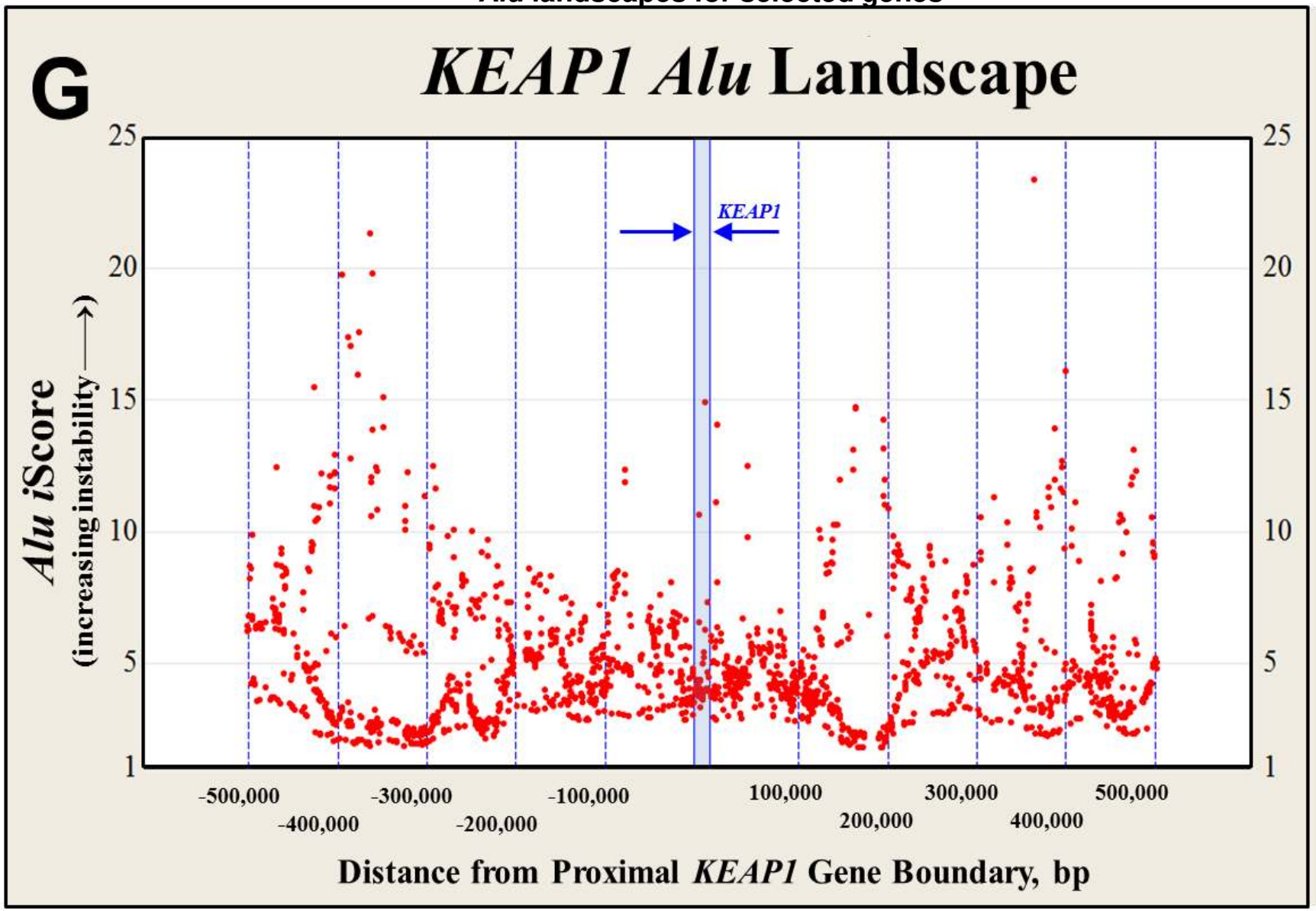


Figure A3.1, continued
***Alu* landscapes for selected genes**

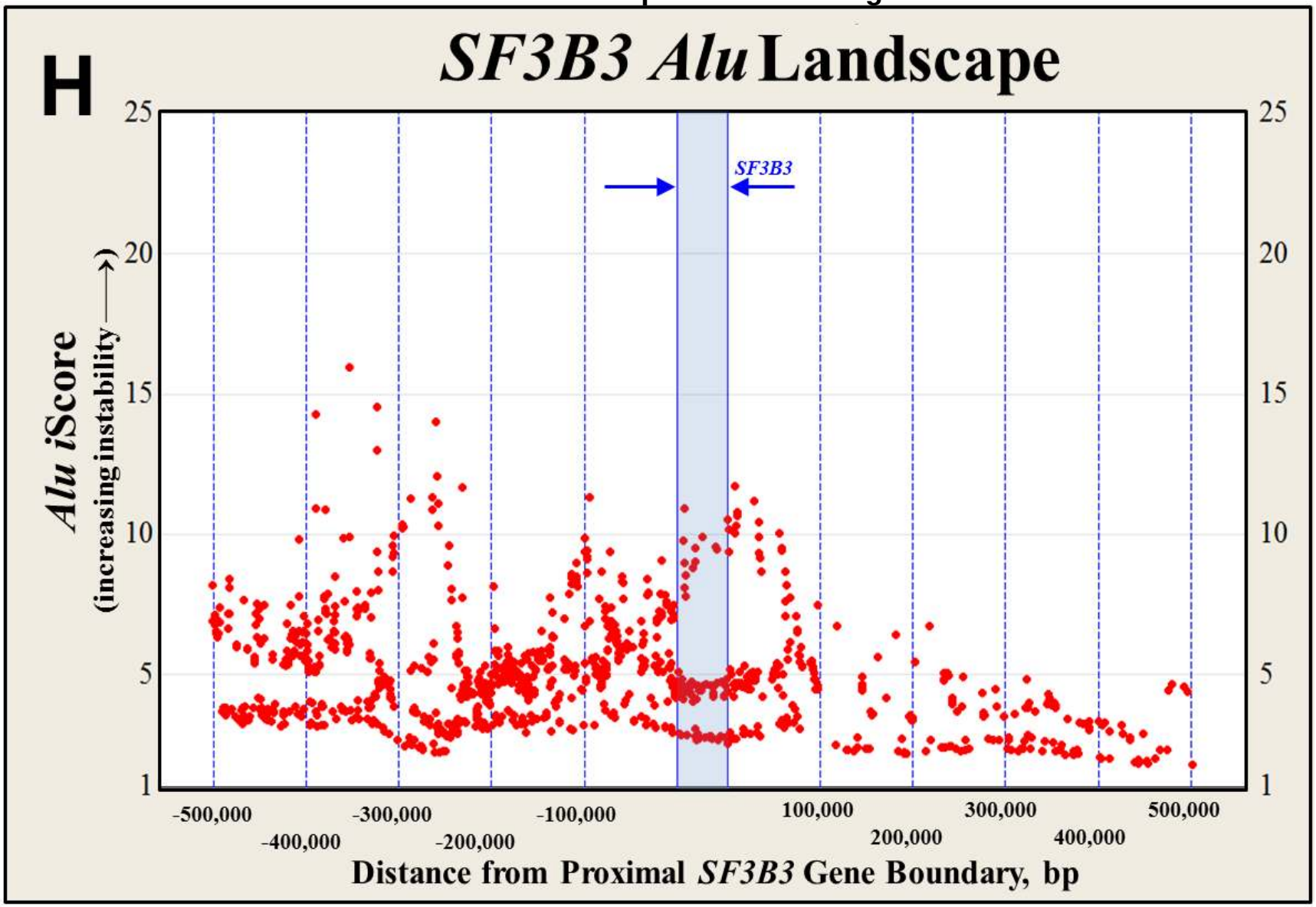


Figure A3.2
Regression fits for 2.5th spacer size percentiles for Type 1, 2 and 3 *Alu* pairs for APSNS 1-115

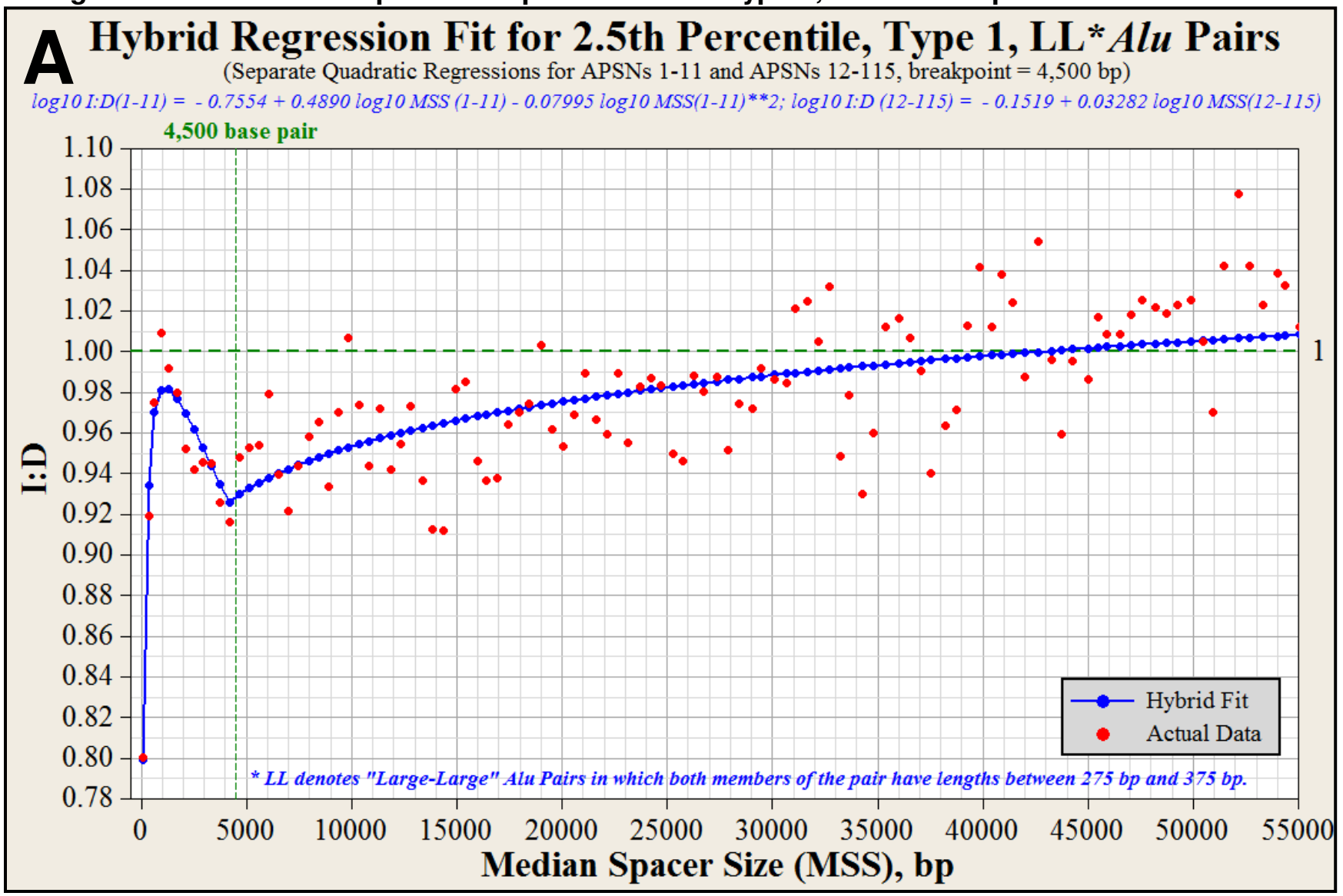


Figure A3.2, continued
Regression fits for 2.5th spacer size percentiles for Type 1, 2 and 3 *Alu* pairs for APSNS 1-115

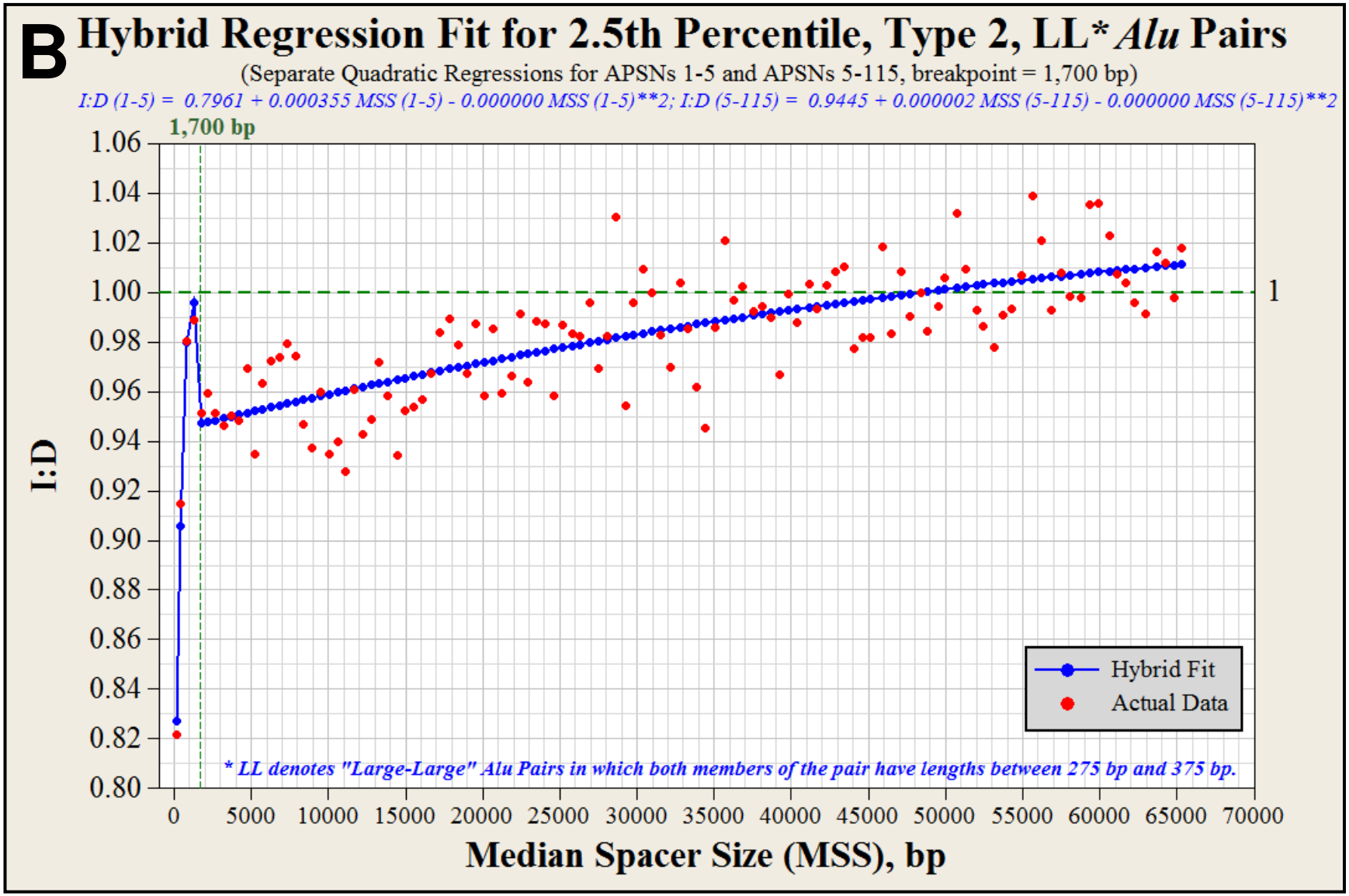


Figure A3.2, continued
Regression fits for 2.5th spacer size percentiles for Type 1, 2 and 3 *Alu* pairs for APSNS 1-115

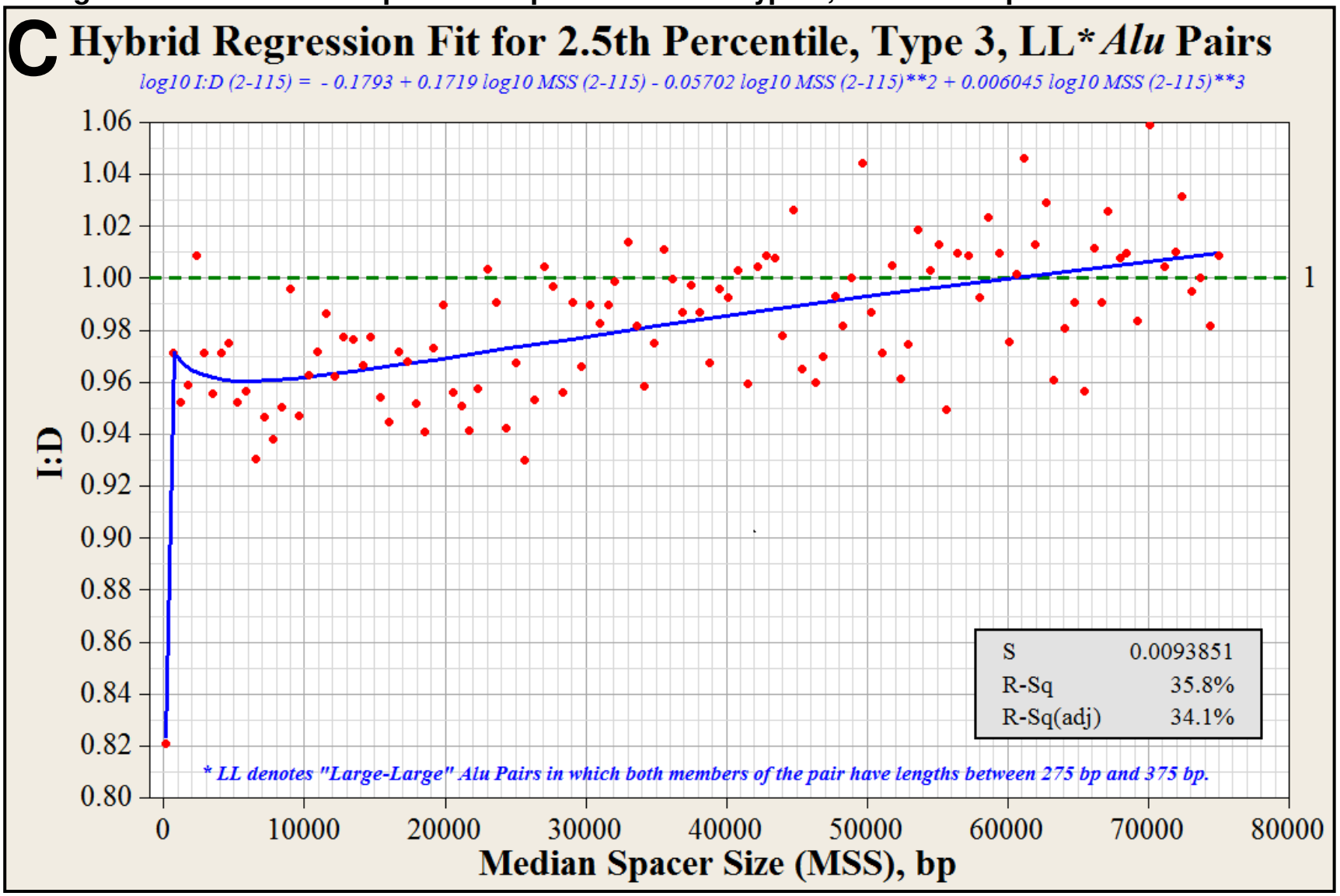
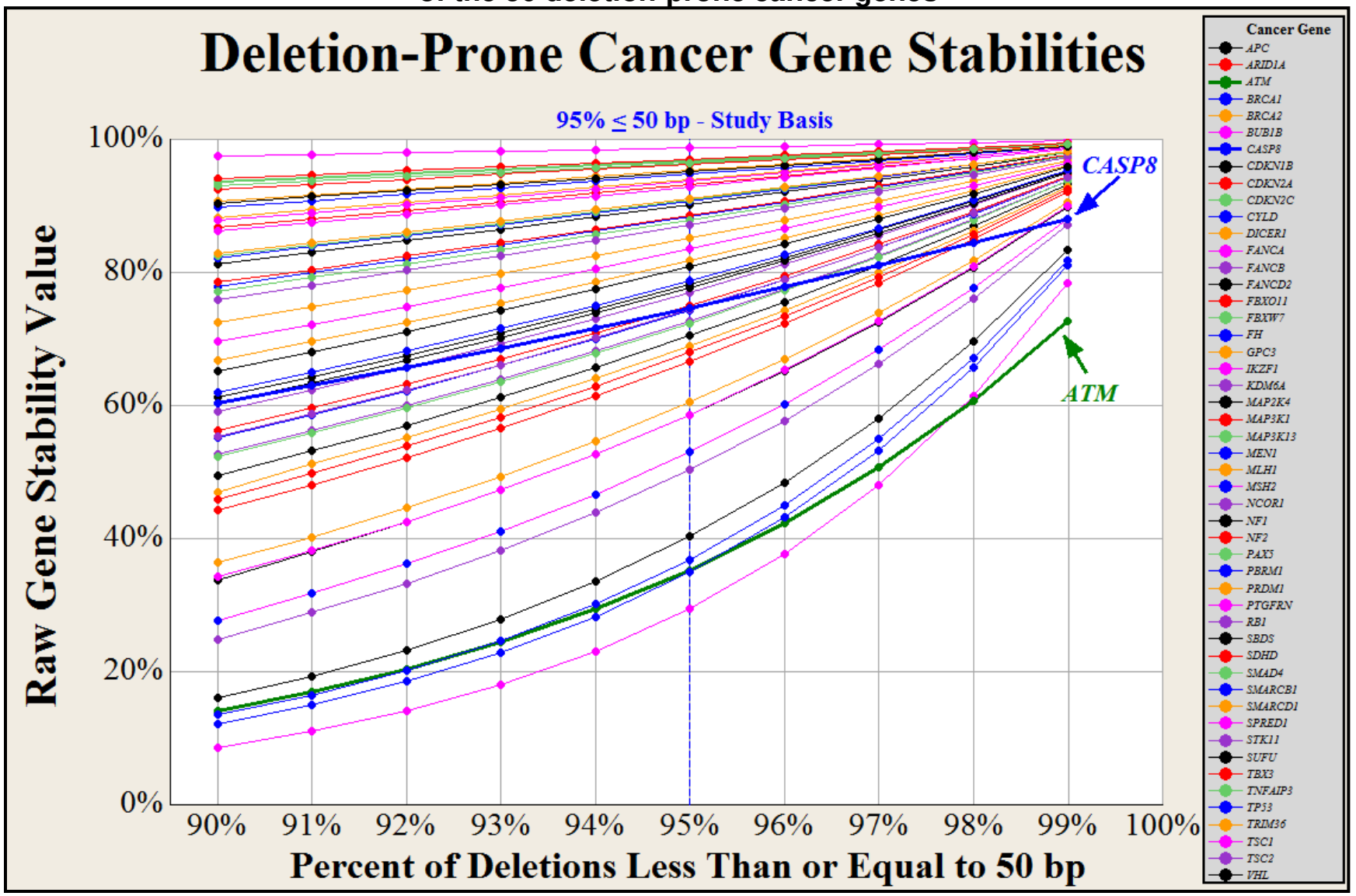


Figure A3.3
Sensitivity of the shape of the human deletion size frequency distribution on the relative stabilities of the 50 deletion-prone cancer genes



APPENDIX B:
LETTERS OF REQUEST AND PERMISSION

From: George Cook [mailto:gcook2@tigers.lsu.edu]
Sent: Sunday, November 18, 2012 4:59 PM
To: Mobile DNA Editorial
Subject: Request for Written Permission to Publish Mobile DNA Article in Dissertation

Dear *Mobile DNA* Editorial Staff,

I am requesting your written permission to publish my first author paper, "*Alu* pair exclusions in the human genome" in my Ph.D. dissertation. This article was published in *Mobile DNA* on September 23, 2011.

Please know that I have carefully read the BioMed Central copyright and license agreement. BioMed Central makes very clear that authors of *Mobile DNA* articles are free to reproduce their work. Nonetheless, Louisiana State University requires that written permission be obtained from the journal of a published paper before it can be included in a dissertation.

I would be most appreciative if you could provide me with an email granting me this permission.

Thanks so much for your help.

Best Regards,

George Cook
Ph.D. Student (for Mark A. Batzer)
Louisiana State University
Baton Rouge, Louisiana 70803

From: Mobile DNA Editorial <editorial@mobilednajournal.com>
Sent: Monday, November 19, 2012 4:55 AM
To: George Cook
Subject: RE: Request for Written Permission to Publish Mobile DNA Article in Dissertation

Dear Mr Cook,

Thank you for your email.

I can confirm that you are an author for"*Alu* pair exclusions in the human genome" which was published in Mobile DNA in September 2011. Mobile DNA is part of BioMed Central and as this is an Open Access publishing model, you are free to reproduce your work. Perhaps you could provide a link to the published version and acknowledge it has been published in Mobile DNA in your dissertation however this is entirely up to you as an author.

I hope this helps but if you have any further questions please do not hesitate to get in touch.

Best wishes,

Arianna Vaccaro

Editorial Assistant

On behalf of
Mobile DNA

BioMed Central
Floor 6, 236 Gray's Inn Road
London,
WC1X 8HL

VITA

George Wyndham Cook, Jr. is the son of George Windham Cook, Sr. and Jo Nell Cook. He was born in Nashville, Tennessee in 1952. George graduated from the University of Arkansas with a BS in Chemical Engineering in 1975. After graduation, George accepted a job with Ethyl Corporation which was renamed Albemarle Corporation in 1992. With the exception of a three year stint at Exxon Chemical, George has worked his entire 37 year career for the same company.

Following the announcement of the draft human genome sequence in 2001, George enrolled in a freshman biology class at San Jacinto Junior College in Pasadena, Texas. He then attended various life science classes at the University of Houston in Clear Lake, Texas from 2002-2004. In the summer of 2004 George was transferred by his company to Baton Rouge, Louisiana where he enrolled in the Department of Biological Sciences at Louisiana State University. In the spring semester of 2008, he joined Mark Batzer's lab and began his doctoral studies as a part-time graduate student. George is scheduled to graduate with the degree of Doctor of Philosophy in May 2013.