Distribution Agreement

Signature:

________________________________ _______________

Rebecca C. Iskow Date

Novel Germline and Somatic Retrotransposon Insertions in Humans

By

Rebecca C. Iskow
Doctor of Philosophy

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

Scott E. Devine, Ph.D.
Advisor

Xiaodong Cheng, Ph.D. Committee Member	Ichiro Matsumura, Ph.D. Committee Member
Paula Vertino, Ph.D. Committee Member	Dr. Michael Zwick, Ph.D. Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Novel Germline and Somatic Retrotransposon Insertions in Humans

By

Rebecca C. Iskow
B.S., University of Maryland, 2003

Advisor: Scott E. Devine, Ph.D.

An Abstract of
a dissertation submitted to the Faculty of the
James T. Laney Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences
Genetics and Molecular Biology
2009

Abstract

Novel Germline and Somatic Retrotransposon Insertions in Humans

By Rebecca C. Iskow

Human genetic variation can cause phenotypic differences as well as provide the substrate for evolutionary forces. Single nucleotide polymorphisms (SNPs) are the most studied form of genetic variation—however, insertions/deletions, copy number variants, and transposable elements are under increasing scrutiny. Retrotransposons (a class of transposable elements) act as endogenous mutagens, and dozens of insertions are responsible for genetic disorders. The majority of human retrotransposons are evolutionary relics, however, a subset of retrotransposons remain active. New retrotransposon insertions are numerous in human populations. We sought to understand the role that young retrotransposons have played in altering human genomes since the divergence from chimpanzee, within human populations, and during an individual's lifetime. First, we developed a computational pipeline to identify lineage-specific insertions in humans and chimpanzees. We identified 11,000 transposable elements that were differentially present between these two species. We also selectively sequenced L1 insertion junctions in diverse humans. We found nearly 1,000 previously unknown retrotransposon polymorphisms. We also found evidence of ongoing and frequent L1 mutagenesis in the germline. We showed that retrotransposition occurs in human lung tumors. L1 mutagenesis is a mechanism of mutation in human tumors and likely contributes to genomic instability. We also identified a potential methylation signature that distinguishes lung tumors that support retrotransposition from those that do not. Altogether, our data show that L1 has been jumping frequently since the divergence of human and chimp and continues to jump in the human germline and soma. Our studies examine human genetic variation through the lens of L1 mutagenesis.

Novel Germline and Somatic Retrotransposon Insertions in Humans

By

Rebecca C. Iskow
B.S., University of Maryland, 2003

Advisor: Scott E. Devine, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Graduate Division of Biological and Biomedical Sciences
Genetics and Molecular Biology
2009

Acknowledgements

My thanks to Scott Devine, Michael Zwick, Andy Bennett, Ryan Mills, Michael McCabe and my committee members.

Thank you to my brother, Scott Iskow, for editorial assistance.

Special thanks for the support of my family and Spencer Torene.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Transposons in the human genome

Transposable elements are segments of DNA capable of moving from one chromosomal location to another, either directly or through an RNA intermediate. In the process, they often increase in copy number without obvious benefit to the host genome. Thus, the scientific community once branded transposable elements as “selfish” or “parasitic” DNA [1, 2]. Barbara McClintock initially discovered transposons in maize [3], but since then, researchers have found evidence of transposons in most genomes. Perhaps the most surprising information to come from the Human Genome Project (HGP) was the vast amount of repetitive sequence in the human genome [4]. Approximately 45% of the human genome is composed of transposable elements [4, 5]. While the majority of transposable elements are incapable of transposition, a handful of elements remain active [6]. In most cases, potentially active elements are kept in check by several cellular mechanisms. When active, transposable elements create novel insertions in populations [5].

Transposable elements fall into two classes: Class I (retrotransposons) and Class II (DNA transposons). DNA transposons have a simpler transposition mechanism, which likely allows for their pervasive existence in diverse hosts. Most DNA transposons contain a single open reading frame (ORF). The ORF encodes a transposase enzyme and is flanked by inverted repeats (IRs). This enzyme catalyzes the excision of the transposon from its donor site [7]. This DNA/protein complex then goes to a second site in the genome where the transposase catalyzes the creation of DNA breaks. Altogether, the transposition process allows the DNA transposon to integrate into a new genomic context. This “cut-and-paste” mechanism does not necessarily increase the copy number of the transposon unless transposition occurs during host DNA replication. If transposition occurs after replication of the donor site, but before replication of the second site, then the transposon can increase in copy number [8]. Remarkably, DNA transposons have increased in copy number throughout evolutionary history and now exist in a multitude of host genomes from bacteria to vertebrates [3, 4, 8-12].

DNA transposons appear to have lost the ability to transpose in the majority of mammals. The last primate-specific transposition occurred nearly 40 million years ago (MYA) [13]. The one mammal that is a glaring exception is the bat. Bat genomes have polymorphic DNA transposons of low sequence divergence, implying recent transposition in this lineage [11].

Unlike DNA transposons, retrotransposons move via a “copy-and-paste” mechanism. Retrotransposable elements belong to one of three classes: (1) long-terminal repeat (LTR), (2) autonomous non-LTR, and (3) nonautonomous non-LTR (Figure 1-1). LTR retrotransposons, including human endogenous retroviruses (HERVs) and yeast Ty1 elements, have a genomic structure and life cycle similar to those of retroviruses [14]. HERV elements and solo HERV LTRs make up 8% of the human genome [4]. Solo HERV LTRs are a result of intra-element recombination between the terminal repeats, which excises the protein-coding region of the element [15]. Only 8 known HERV elements are polymorphic [16]. Thierry Heidmann’s lab recently engineered a retrotranspositionally competent HERV element, but no known naturally occurring elements appear capable of retrotransposition [17].

The autonomous non-LTR class of retrotransposons includes the remarkably successful long interspersed nucleotide element-1 (L1) family. Active L1 elements are 6 kb long and encode their own promoters and proteins necessary for retrotransposition (Figure 1-1). L1 elements contain 5' and 3' untranslated regions (UTRs), two ORFs, a spacer region between the ORFs, and a variable length poly-A tail. The first ORF encodes a protein with nucleic acid chaperone activity [18] while the second ORF

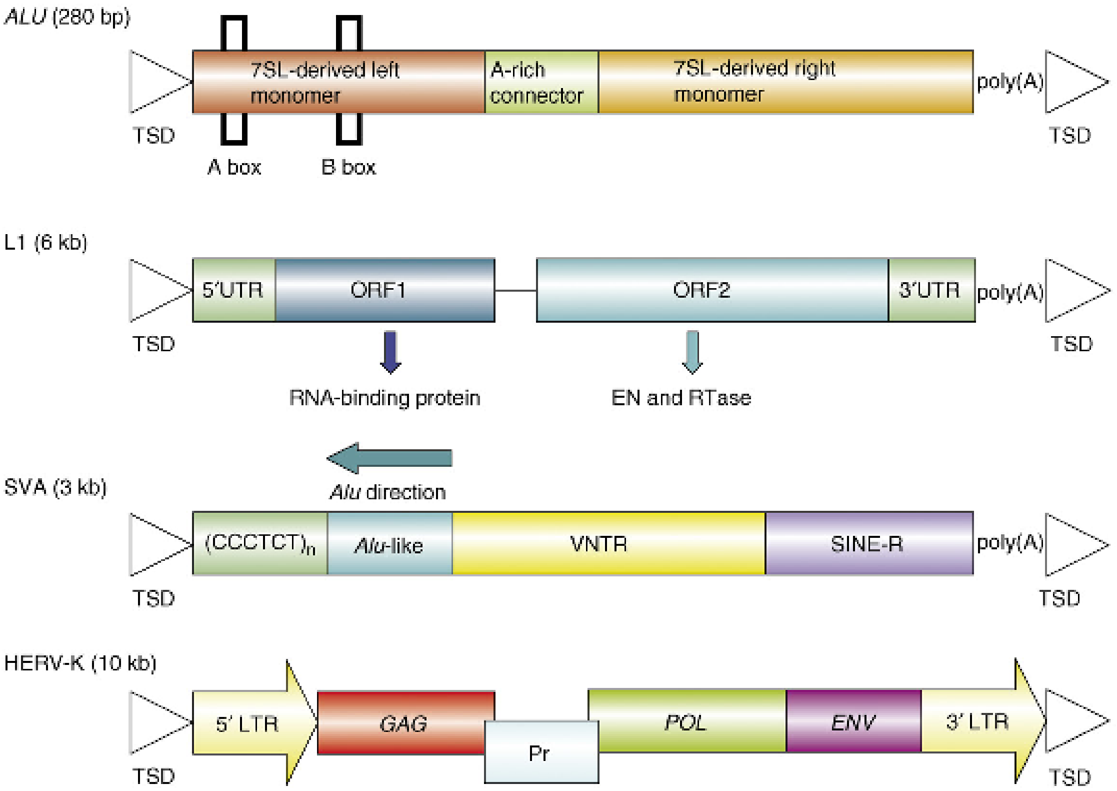


Figure 1-1

Structures of polymorphic human transposable elements. (Adapted from [5].) The figure depicts the structures of the four polymorphic human transposon classes. Definitions of acronyms and abbreviations are as follows: LTR—long-terminal repeat; ORF—open reading frame; Pr—protease; TSD—target site duplication; UTR—untranslated region; VNTR—variable number of tandem repeats.

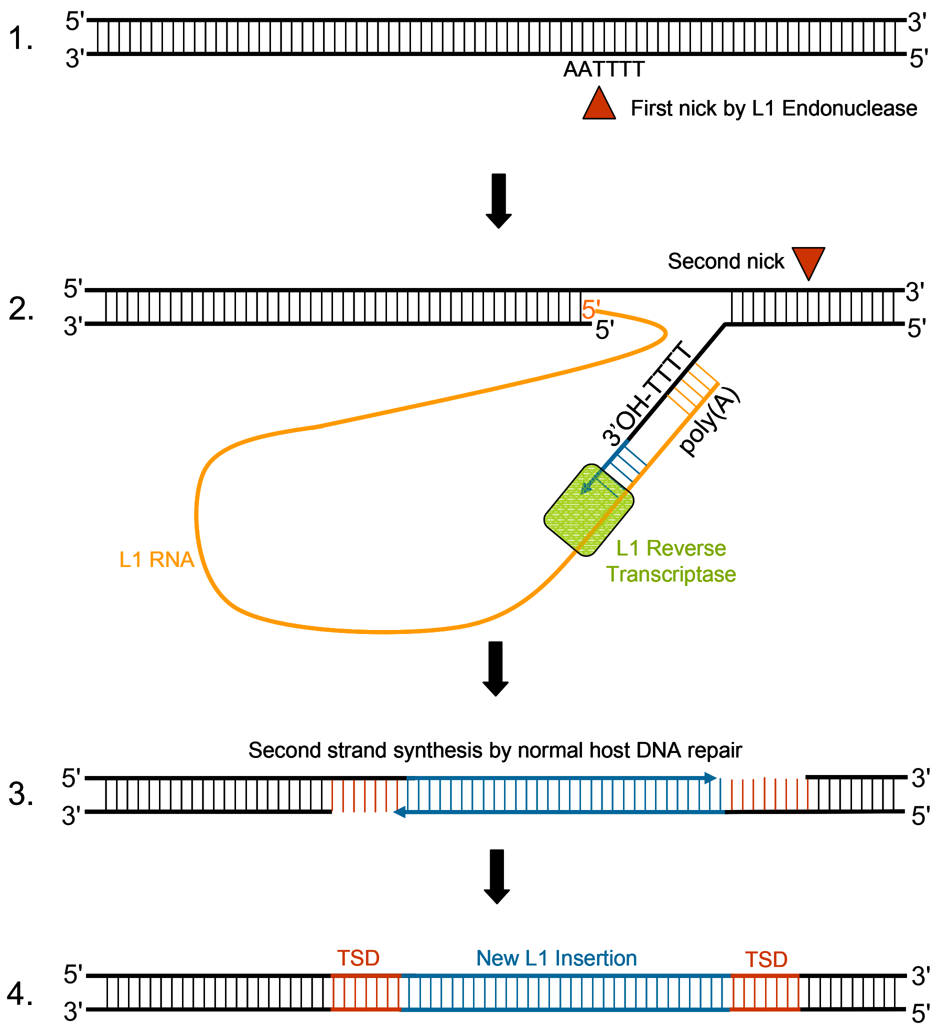


Figure 1-2

Schematic of endonuclease-dependent target-primed reverse transcription. (1) L1 endonuclease creates a nick at a loosely conserved recognition sequence in the genome. (2) The free 3' hydroxyl group created by the nick primes cDNA synthesis of the L1 RNA by its own ORF2p reverse transcriptase. The 5' end of the L1 RNA is thought to interact with the genome at the site of the nick. A second nick is made on the opposite strand downstream of the recognition site. (3) Second-strand synthesis occurs by some yet unknown mechanism, but likely involves normal host DNA repair systems. (4) A new L1 insertion is flanked by target site duplications due to the staggered nicks on opposite strands of host DNA.

encodes a protein with both endonuclease and reverse transcriptase activity [19, 20]. Active L1 elements retrotranspose primarily by target-primed reverse transcription (TPRT, Figure 1-2) [21, 22]. During TPRT, L1 proteins bind to their coding mRNA at the ribosome [23]. This ribonuclear protein complex then catalyzes the creation of a nick in the genome. The nick results in a 3'OH, which primes reverse transcription of the L1 RNA [21]. The endonuclease catalyzes the formation of a second nick on the opposite strand of DNA downstream of the insertion site (Figure 1-2). The reverse transcriptase catalyzes the polymerization of a cDNA copy of the L1 RNA. Afterward, normal host DNA repair systems likely synthesize the complementary strand of DNA. The staggered nicks cause a small duplication of host DNA at the target site, known as target site duplications (TSD) [24]. This “copy-and-paste” process has enabled L1 to reach a copy number of approximately 500,000 in the human genome. The L1 reverse transcriptase, however, is not a processive enzyme. As a result, a majority of L1 elements in the human genome are truncated at their 5' ends [25]. Most L1s have accumulated inactivating mutations, yet each person has 80-100 potentially active L1 elements in their genome [6, 26].

L1-encoded proteins tend to retrotranspose their own encoding RNA (a phenomenon called “cis preference”) [23]. Occasionally, the reverse transcriptase and endonuclease are co-opted by nonautonomous non-LTR retrotransposons and cellular RNAs. Retrotransposition of cellular RNAs creates processed pseudogenes and multi-copy noncoding RNAs [23, 27]. As the name suggests, nonautonomous non-LTR retrotransposons do not appear to encode any proteins. They rely on L1 proteins for retrotransposition [24]. Nonautonomous non-LTR retrotransposons include the primate-specific Alu and Sine-VNTR-Alu (SVA) and the murine-specific B1 families. Alu and B1 members are short interspersed nucleotide elements (SINEs), while SVA members are composite elements that are often greater than 3 kb in length (Figure 1-1) [24, 28, 29]. Alu is the most successful retrotransposon family in primates. Humans have accumulated more than 1.1 million copies over 65 million years. Currently, several different subfamilies of Alu contain potentially active elements [30, 31]. On the other hand, SVA appears to be a relatively young family of retrotransposons. There are only approximately 3,000 copies of SVA in the human genome. Of all the retrotransposon families mentioned thus far, the SVA family has the largest fraction of polymorphic elements [27, 29]. The SVA family appears to be in an upswing of activity.

Retrotransposition and human disease

The vast majority of human retrotransposable elements are relics of evolution [1, 2]. Most retrotransposable elements have accumulated inactivating mutations and are fixed between humans. Only in the past 20 years have scientists realized that some human retrotransposable elements are active [26, 31, 32]. These active elements can cause insertions that alter phenotypes and cause disease [33]. Furthermore, if these new insertions occur in the germline, they can pass on to the next generation, increase in frequency in the larger population, and possibly become fixed.

Both inherited and *de novo* retrotransposon insertions can alter phenotypes. For example, patients with Fukuyama muscular dystrophy inherit two recessive alleles from their parents. The most common causal mutation is an inherited SVA insertion in the 3' UTR of the FCMD gene [34]. Geneticists estimate that this retrotransposon insertion occurred around 100 generations ago [35]. Since then, the element has reached intermediate frequencies in Japanese populations and is inherited faithfully from parent to child.

In contrast, *de novo* insertions of L1 elements caused Hemophilia A in at least two patients. Their parents had no insertion [32, 36]. Since this discovery in 1988, geneticists have found more than 60 additional disease-causing retrotransposon insertions[reviewed in 5 and 33]. Recombination-mediated rearrangements between retrotransposons is likely responsible for even more incidents of genetic disease [37-40]. Retrotransposon insertions account for less than 0.1% of known disease-causing mutations in the Human Genome Mutation Database (HGMD) [41].

Retrotransposons can alter gene function either directly or by post-integration mechanisms. First, retrotransposons can disrupt exons and lead to nonfunctional gene products [e.g. 32, 34, 36, 42, 43]. Second, the presence of a retrotransposon in an intron can alter the stability of premature mRNA, leading to a reduced amount of transcript [44, 45]. Retrotransposons contain cis-acting regulatory sequences, such as promoters, splice sites, and polyadenylation signals which can interfere with nearby genes by altering which sequence is included in mature mRNA[46-50]. Third, retrotransposons can be targets of epigenetic silencing, which can affect the expression of nearby genes. SVA elements, in particular, are “mobile CpG islands” and likely alter the epigenetic backdrop where they transpose [27, 28]. Finally—and perhaps the most dramatic way in which retrotransposons interfere with gene function—retrotransposons can mediate ectopic recombination. Unequal recombination between highly similar interspersed repeats causes large-scale genomic rearrangements such as deletions, duplications, and inversions [51]. Since humans and chimps last shared a common ancestor, the human genome has lost nearly 415 kb of DNA due to recombination between retrotransposable elements [52, 53].

Mechanisms to maintain transposons in populations

One could imagine that the gene-disrupting mechanisms described above could reduce the fitness of a host. Deleterious transposable elements are unlikely to reach fixation because purifying selection should eliminate them. If transposition rates are high and each new insertion kills less than half the individuals who inherit it, a transposon family can increase in copy number [54]. Thus, slightly (and perhaps moderately) deleterious transposon families persist in a population, and individual elements could even reach fixation.

For lower transposition rates (similar to rates for active human retrotransposons), a transposon family can expand if negative selection on most elements is exceedingly low [55]. Otherwise, purifying selection would eliminate transposon insertions faster than new insertions would occur. In this case, the elements that eventually reach fixation are likely neutral or only slightly deleterious. The majority of Alu and L1 insertions in the human genome are, in fact, neutral [56, 57]. Thus, if the majority of retrotransposable elements are neutral, they can reach high frequencies by genetic drift [58].

Alternatively, transposons may achieve fixation through genetic hitchhiking [59, 60]. Positive selection may act upon an allele that is genetically linked to a differentially present transposon. When a beneficial allele increases in frequency, other variants within its haplotype also increase in frequency. Thus, nearby neutral or even deleterious sequences can increase to intermediate and high frequencies within a population [60, 61]. For example, a variant of a pesticide resistance gene was under positive selection in *Drosophila* [59]. Flies with this variant survived exposure to a particular pesticide which caused the variant to increase in frequency. A nearby transposon was genetically linked to the pesticide resistance variant and also increased in frequency until it eventually reached fixation. In theory, genetic hitchhiking may also be responsible for high frequency and fixed retrotransposable elements in humans.

Finally, retrotransposable elements may persist in human populations because they have become beneficial. Approximately 120 retrotransposon insertions are now functional genes in humans [62]. For example, a DNA transposon once inserted into a chromatin remodeling gene [63]. The element's DNA-binding domain is intact and utilized by the gene product [64]. In addition, Alus are often present in mature mRNA. Alus contain cryptic splice sites, which drive their exaptation [50, 65, 66]. The majority of exapted Alus are within 5' UTRs, but many end up in protein-coding RNA [66]. Alus are also a target of RNA editing [67]. Inverted Alu repeats in RNA form a secondary structure, which is targeted by adenosine deaminases that act on RNA (ADARs). ADARs change adenosines to inosines, which cellular splicing and translational machinery read as guanines (Figure 1-3). Thus, retrotransposons, especially Alus, greatly increase the human transcriptome through the addition of exons and the alteration of RNA sequence [68]. Altogether, these mechanisms add a level of complexity to human genes.

Retrotransposition may occur in cancer

For a long time the genetics community thought retrotransposition primarily occurred in the human germline [69]. In fact, retrotransposition occurs most often in early embryogenesis, creating individuals who are mosaic [70-73]. The bulk of these embryonic insertions are in somatic tissues and, therefore, are not heritable. The genome that individuals start with (as fertilized zygotes) is not necessarily the same genome in their adult tissues.

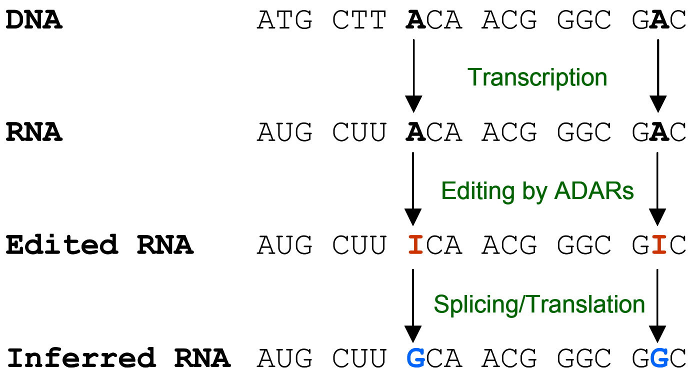

DNA ATG CTT ACA ACG GGC GAC
Transcription
RNA AUG CUU ACA ACG GGC GAC
Editing by ADARs
Edited RNA AUG CUU ICA ACG GGC GIC
Splicing/Translation
Inferred RNA AUG CUU GCA ACG GGC GGC

Figure 1-3

RNA editing alters coding potential of transcripts. Inverted repeats, including Alus, cause the formation of double-stranded RNA, which is the substrate for ADARs. ADARs deaminate adenosines into inosines, which are read by cellular machinery as guanines.

Retrotransposons may be active in somatic cells after embryogenesis and may account for some of the acquired changes in cancer. Methylation appears to be the primary way to silence retrotransposons. In fact, Dnmt3L (a *de novo* methyltransferase) knockout mice are unable to complete spermatogenesis presumably because retrotransposons become unmethylated and create new insertions [74]. Retrotransposons are methylated and transcribed only at low levels in the majority of differentiated tissues [75, 76]. Cancer tissues, on the other hand, often have global loss of retrotransposon methylation [77]. In some studies, global hypomethylation of retrotransposons correlates with genomic instability, tumor progression, and poor prognosis [e.g. 78-81]. For example, researchers found that loss of methylation of Alus and L1s in lung tumors correlated with genomic instability at 11 microsatellite markers [79]. Furthermore, researchers at Harvard University found that colon cancer patients whose L1 methylation decreased by 30% had significantly higher rates of colon cancer-related mortality [80]. The observations above indicate that retrotransposons may be active in cancers. Normally quiescent retrotransposons may suddenly be in a permissive environment when their methylation is lost. In addition, inhibition of reverse transcriptase slows tumor progression [82, 83]. This phenomenon is indirect evidence that retrotransposons may be mobilized in cancers.

Only three groups have identified acquired retrotransposon insertions in human tumors (Figure 1-4) [42, 84, 85]. After careful inspection, however, only one is a *de novo* classical retrotransposition event [42]. The other two are complex rearrangements that involve retrotransposon sequences but lack the hallmarks of a *de novo* insertion, such as target site duplications and a poly-A tail (Figure 1-4). To date only one legitimate example of an acquired retrotransposition event in human tumors exists. This event may be an isolated incident or an example of a more common source of mutation in human tumors.

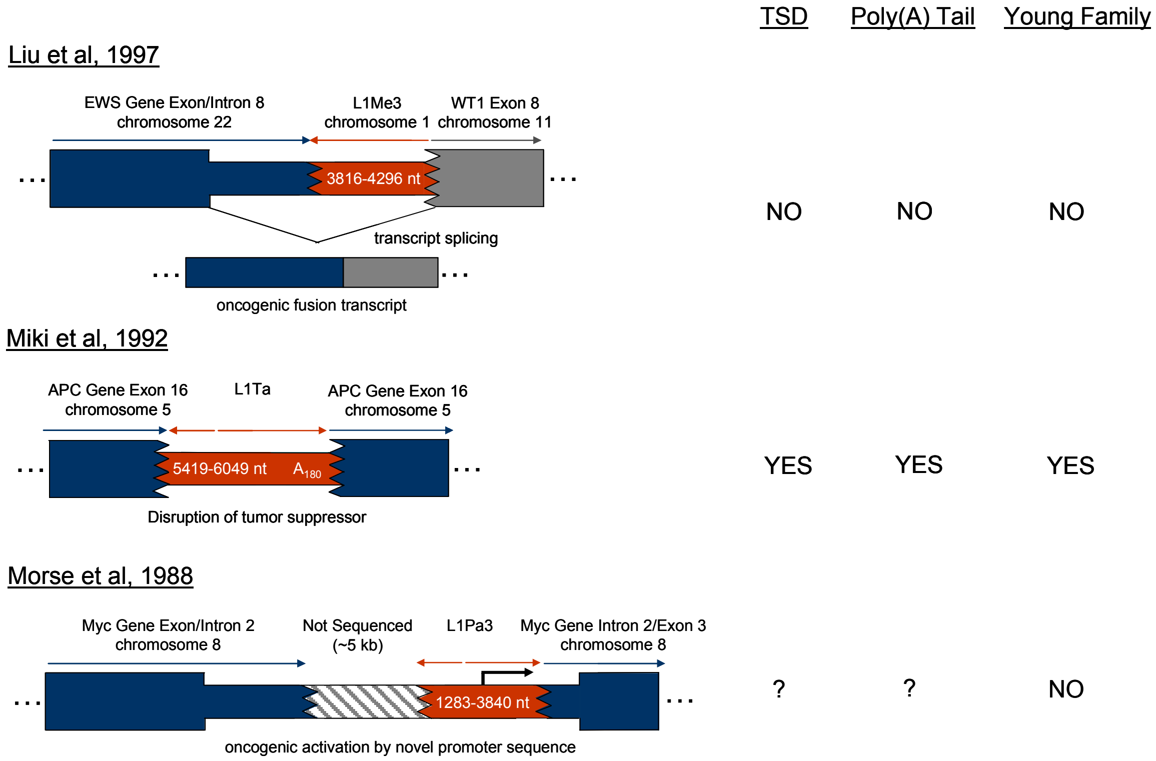


Figure 1-4

Insertion sites of potential somatic retrotranspositions. The insertion described by Liu et al. is a chromosomal translocation between the Ewing sarcoma gene on chromosome 8 and the Wilms' tumor gene on chromosome 11 [84]. A fragment of an L1Me3 (an old L1 subfamily) from chromosome 1 was also included in this rearrangement. The resulting transcript from this hybrid gene splices together exons 8 of the EWS and WT1 genes and creates an oncogenic gene product. The insertion described by Miki et al. is a member of a young L1 subfamily, L1Ta [42]. The element inserted into a coding exon of the anaphase promoting complex gene, a known tumor suppressor. The insertion described by Morse et al. is 7-8 kb long based on Southern blot analysis and in intron 2 of the Myc gene [85]. Only the downstream portion of the insertion was cloned and sequenced.

Retrotransposons can instigate genomic instability at the time of insertion and can mediate recombination after insertion [38-40, 53]. Acquired insertions could possibly accelerate genomic instability in tumors [38, 51, 86]. Therefore, one can hypothesize a possible model for tumor progression. For example, an initiating event causes global epigenetic changes. This event leads to a loss of restraint on retrotransposons, which causes the altered activity of tumor suppressors and oncogenes.

Inherited retrotransposon insertions can also contribute to cancer. One study screened for breast cancer predisposing mutations in BRCA1 and BRCA2 in families with a history of breast cancer. The researchers found that 2 of 50 inherited mutations in the BRCA genes resulted from independent retrotransposition events [87].

Retrotransposition can be an acquired mutation in the murine lineage. For example, the endogenous retrovirus—intracisternal type-A particle (IAP)—is active in somatic cells in a Dnmt1 (a maintenance methyltransferase) knockout mouse [88]. These mice are prone to thymic lymphomas, and 7 out of 16 lymphomas had independent insertions of an IAP element in the Notch gene. The presence of the IAP element created an oncogenic spliceoform of Notch, which likely caused the tumors. Retrotransposons are responsible for some acquired mutations in mouse and could potentially also be responsible for acquired mutations in human.

Estimates of retrotransposition rate

The bulk of retrotransposable elements in the human genome are passive residents. A handful of elements, however, are still capable of creating new insertions. Recently, several groups estimated the retrotransposition rate (RR) for Alu and L1 in humans (1) by comparing the number of disease-causing *de novo* insertions to nucleotide mutations in HGMD [89-91], (2) by comparing the number of lineage-specific insertions in the human and chimp genomes [89], and (3) by extrapolating from the RR of a well-characterized retrotransposon [6]. These methods provide RRs ranging from 1 in 2 to 1 in 100 live human births [6, 89-91].

If the majority of these new insertions are neutral or deleterious, then only a fraction should reach intermediate frequencies and then fixation by genetic drift alone [92]. Thus, polymorphic retrotransposons should have an allelic frequency spectrum skewed toward rare insertions, and the majority of polymorphic retrotransposons would not be in the public human reference sequence. To date, only 530 retrotransposons outside of the reference sequence are in the Database for Retrotransposon Insertion Polymorphism (dbRIP)—the retrotransposon equivalent of dbSNP [93]. The majority of these insertions are likely inherited and not *de novo*. dbRIP is not representative of human populations since no one has conducted a comprehensive search for polymorphic retrotransposons. The bulk of human genetic diversity caused by retrotransposons is presently unknown.

Polymorphic retrotransposons are excellent genetic markers for association studies and population genetics analyses. Unlike single nucleotide polymorphisms (SNPs), retrotransposons are virtually free of homoplasy. Also, the precise deletion of a retrotransposon after integration is exceedingly rare [55]. The lack of a retrotransposon is the ancestral state. Polymorphic retrotransposon insertions are important to discover because (1) they are stable genetic markers and (2) they contribute to human genetic variation and cause disease.

Efforts to find polymorphic retrotransposons

PCR-based methods

Untold numbers of L1 polymorphisms exist, but their locations and potential effects on human health are largely unknown. Recently, several labs joined the hunt for polymorphic retrotransposons (Table 1-1). Initially, these labs discovered many polymorphic retrotransposons by “candidate screening,” a process using PCR primers flanking a known retrotransposon to amplify the preintegration site in other individuals [e.g. 29, 93-96]. This method, while fruitful, cannot find *de novo* insertions and skews dbRIP’s dataset toward common alleles.

Anchored PCR is a less biased way to find polymorphic retrotransposons. For anchored PCR, one primer anneals within the retrotransposable element while a second primer is arbitrary or anneals to a linker. Anchored PCR allows for the amplification of retrotransposon insertion junctions without *a priori* knowledge of their locations. Even so, PCR primers are not usually specific enough to efficiently amplify polymorphic elements amid the background of fixed elements. Thus, many labs also developed enrichment steps or screens that follow PCR.

Several labs identified polymorphic L1s by PCR followed by screening. The labs used a primer specific to the youngest L1 subfamily, L1Ta, and a short, arbitrary primer [97, 98]. They screened the PCR products by Southern blot with a probe to the 3' end of L1. These labs then cut out, reamplified, cloned, and sequenced bands that varied among diverse individuals. These labs identified 42 polymorphic L1s, half of which were not in the human reference sequence [93, 97, 98]. Additional studies used linker-mediated PCR and then identified variable bands without Southern blotting [99] or cloned the PCR products and screened for L1s by filter lifts and dot blots [100]. These studies identified an additional 144 polymorphic L1s.

Method Used	No. of Polymorphic L1s Found	No. of Polymorphic L1s Outside of Hg18	Deposited in dbRIP?
Anchored PCR	292	158	yes
Candidate Screening	32	1	yes
Database Search	55	7	yes
Disease-Causing	11	11	yes
Whole Genome Comparison	202	22	yes
Whole Genome Sequencing	42	0	no
Paired-end sequencing	152	20	no

Table 1-1

Methods to discover polymorphic L1 elements in humans.

Post-PCR screening can be labor intensive, and nearly half of the sequenced retrotransposon insertion junctions are from fixed elements. Thus, several labs used subtractive hybridization to enrich for polymorphic retrotransposons before sequencing. They generated PCR products by anchored PCR and then hybridized their products to those from a chimp [101, 102] or from a "driver" human [103]. PCR products that do not hybridize likely contain a retrotransposon insertion junction, that is absent from the chimp or driver. Several labs used subtractive hybridization to study HERVs [101], Alus [103], and L1s [102], and they efficiently identified previously unknown polymorphic elements.

Although subtractive hybridization is efficient, several lab groups cloned and sequenced their PCR products without additional enrichments or screenings [104-106]. As DNA sequencing becomes cheaper, the hunt for polymorphic retrotransposons occurs at the computer more frequently than the lab bench. Sorting through retrotransposon sequences derived from anchored PCR with a high background of fixed elements is sometimes easier than spending several days in the lab performing enrichments and screens.

Of course, sequencing is not the only method for finding retrotransposons. Recently, Jef Boeke's lab identified the locations of yeast Ty1 elements by anchored PCR followed by hybridization to tiling arrays [107]. The lab argues that similar assays can be done for human retrotransposons.

Anchored PCR-based methods allow for the discovery of retrotransposon polymorphisms, including *de novo* insertions. Although these methods have less bias than candidate screening methods, they still have biases. The researcher must shear or digest the genome before performing linker-mediated PCR. The choice of restriction enzymes or average sheared fragment size will undoubtedly skew the representation of the genome. Also, using gel-based screening methods makes finding common polymorphic retrotransposons, unlikely.

Whole genome sequencing to find retrotransposon polymorphisms

Candidate screening and anchored PCR are successful methods for finding retrotransposon polymorphisms. Even so, the nearly 2,100 known retrotransposon polymorphisms are mostly common insertions, while numerous rare insertions remain undetected. Detecting single polymorphisms amid the background of fixed elements in the human genome requires an extensive effort.

One less biased way to find retrotransposon polymorphisms is analyzing sequence data from whole genome and shotgun sequencing projects. One lab used trace sequence data generated from 24 diverse individuals to identify novel retrotransposon polymorphisms [27]. They aligned the trace sequences to the reference sequence (the publicly available human genome sequence that resulted from the Human Genome Project) and looked for gaps in the alignment. They then compared the sequences within the gap with a library of known repetitive elements. They found 600 polymorphic retrotransposons, 79 of which were not in the reference genome. Their approach used a stringent bioinformatics pipeline that required the trace to encompass the entire insertion. Most trace sequences were less than 1 kb in length while L1s, HERVs, and SVAs are often several kilobases long. Mostly, this lab found Alus in the trace sequences. This study demonstrated the utility of publicly available sequence data for the identification of polymorphic retrotransposons.

Recently, some lab groups have used whole genome comparisons to discover polymorphic retrotransposons. One study identified all the sequenced Alus in the reference sequence as well as the Celera genome [108]. The researchers then aligned Alus plus their flanking sequence to the other genome. If the Alu plus flanks aligned to the other genome, both genomes had the same Alu insertion. This method led to the discovery of 800 polymorphic Alu insertions, 534 of which were not previously known to be polymorphic. A similar study by many of the same researchers found 148 L1 polymorphisms, 73 of which were not previously known to be polymorphic [109]. While these efforts greatly contributed to knowledge of polymorphic retrotransposons, they are limited by the individuals who have been sequenced thus far. Since these papers were published, 5 more individual genomes were sequenced for a total of 7. These genomes likely contain dozens, if not hundreds, of previously unknown retrotransposons [110-116].

To find young and possibly *de novo* insertions with high-throughput methods, scientists need to sort through those that are present in every human’s genome. One promising avenue for retrotransposon discovery is next-gen sequencing technologies such as Roche 454, ABI SOLiD, Illumina, and Helicos. These technologies bypass cloning steps and allow for greater sequencing coverage. As sequencing becomes cheaper, the ability to sequence to greater depth in the population—and thus find rare alleles—becomes more of a reality.

Next-gen technologies produce short reads (about 30 nt for ABI SOLiD, 70 nt for Illumina, and 250 nt for Roche 454 FLX), which make genome assembly computationally difficult. Thus, these technologies are most useful when there is a reference sequence to align to. Even so, short reads that contain retrotransposon sequences often map ambiguously and end up dropped from subsequent analyses. Thus, young retrotransposon subfamily members are underrepresented in human genomes sequenced by next-gen technologies [110, 116]. The length of detectable insertions is limited by read length. Researchers found insertions as large as 208 bp with Roche 454 FLX sequencing, but 208 bp is smaller than the shortest active retrotransposon, Alu (approximately 280 bp) [116]. Researchers can detect larger insertions when they assemble unmapped reads into contigs [115]. Polymorphic retrotransposons are often similar in sequence and, thus, difficult to unambiguously assemble.

So far, researchers using next-gen sequencing have mostly found polymorphic retrotransposable elements that are present in the reference genome and absent from the sample genome. Since the precise removal of a retrotransposon is exceedingly rare, the absence of a retrotransposable element is most likely due to a retrotransposition event that occurred in the reference genome [55]. Sequencing across a pre-insertion locus is much easier than detecting novel insertions with short reads. Whole genomes sequenced with next-gen technology, have a steady drop in the number of deletions detected as deletion size increases, except for an excess of deletions around 300 bp and 6 kb in size [110, 116]. This excess of deletions is primarily due to the absence of Alus and L1s relative to the reference sequence [110, 116].

Paired-end sequencing can discover retrotransposable elements as it has discovered other structural variants (SVs) [110, 112, 115, 117-119]. For paired-end sequencing, researchers shear the genomic DNA, perform size-selection, and clone the fragments into fosmid or BAC libraries. The researchers then sequence the ends of the insert and align the ends to a reference genome [112, 118]. If the distance between the paired-ends in the reference sequence is larger than the size selected, the sample has a deletion relative to the reference sequence. If the distance is smaller, then the sample genome has an insertion relative to the reference sequence. Using this method, Evan Eichler's lab found 40 polymorphic retrotransposon insertions, 20 of which were not in the reference sequence [118]. They were able to find polymorphic L1s, HERVs, and large SVAs, but their resolution of SV detection was not enough to pick up polymorphic Alus. Other researchers applied paired-end sequencing to next-gen technologies in order to find large insertions. As with traditional paired-end sequencing, the insertions identified are limited by the library fragment size. Thus far, next-gen paired-end sequencing has only found the absence of retrotransposons in sample genomes relative to the reference genome [115, 117, 119].

For one genome sequenced with next-gen technologies, the researchers made an effort to find insertions larger than the fragment size of their paired-end libraries [115]. They looked for regions of the genome that had an excess of "bridge pairs"—paired-end reads with one end anchored to the reference genome and another end that could not be mapped. If the bridge pairs for a single region all had their second end in the same kind of retrotransposon, they considered it a "candidate region" for an insertion of these sequences. Using this analysis, the researchers found 692 potential Alu insertions and

1,601 potential L1 insertions. These numbers appear extraordinarily high when one considers estimates of retrotransposition rate. The majority of these candidate regions are likely insertions that contain retrotransposons and not the result of true retrotransposition events. Further details, such as the presence of target site duplications and the subfamilies of these elements, would help determine if these insertions are the result of retrotransposition. The authors of this study overcame the limitations of library fragment size to potentially find novel retrotransposons.

As high-throughput technologies become cheaper and more accessible, those studying genetic variation will surely benefit, though it is necessary to study genetic variation beyond the single nucleotide level. Retrotransposition is an ongoing source of structural variation. Novel retrotransposon insertions are useful genetic markers and can potentially cause disease. Thus, it is important for the scientific community to identify polymorphic retrotransposable elements.

Scope of Dissertation

The goals of this dissertation are to determine how L1 mutagenesis contributes to human genetic variation. Only 13 years have elapsed since L1 activity was demonstrated in tissue culture, and the impact of L1 mutagenesis on human health is mostly unknown. Recent studies of L1 polymorphisms, are not comprehensive. L1 retrotransposition is a widespread process that goes largely undetected. We took innovative approaches to find novel insertions among the vast majority of fixed insertions. We hoped to further scientific knowledge regarding L1 mutagenesis between species (chapter 2), within human populations (chapter 3), and during the course of an individual's lifetime (chapter 4).

First, we wanted to know whether L1 retrotransposons are responsible for some of the genetic differences between humans and their closest living relative, chimpanzees. We used a bioinformatics pipeline to compare the human and chimp reference genomes in order to uncover lineage-specific retrotransposition events. We followed up with an ORF-trapping experiment to estimate how many potentially retrotransposition-competent L1 elements there are in both species (chapter 2).

Next, we selectively sequenced young L1 insertion junctions in diverse humans. We quadrupled the number of currently known L1 elements that are not in the public human reference sequence [4, 93]. We also found that the allelic spectra of these novel elements indicate common retrotransposition in human populations. Our data strongly suggest that dbRIP has a dearth of rare retrotransposons, probably as a consequence of limited population sampling (chapter 3).

Lastly, we wanted to know whether L1 elements are active in human tumors. We used massively parallel sequencing to determine the polymorphic L1 "transposome" of tumor and matched normal DNA. In doing so, we identified 9 somatic insertions and 1 additional somatic candidate. These data, along with a previously found somatic insertion [42], are the only proof to date that retrotransposons are active in human tumors (chapter 4).

Altogether, these studies help us better understand divergence between species, genetic variation within populations, and tumorigenesis through the lens of L1 mutagenesis.

Chapter 2

Recently mobilized transposons in the human and chimpanzee genomes

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

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

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

Copyright. 2006 by The American Society of Human Genetics. DOI:10.1086/501028

Introduction

Transposable genetic elements collectively occupy ~44% of the human genome [4]. Although most of these transposons lost the ability to transpose long ago, some copies have transposed in relatively recent human history [22, 26, 27, 36, 37, 43]. These recently mobilized transposons are of great interest for a number of reasons. First, recent insertions within or near genes may cause phenotypic changes in humans, including diseases [22, 36, 37, 43]. Several dozen transposon insertions have been identified to date that cause human diseases, and human populations are likely to harbor additional transposon insertions that influence phenotypes as well. Some of these recently mobilized transposons also remain actively mobile today and continue to generate new transposition events elsewhere in the genome [22, 26, 31]. Active retrotransposons in particular have been observed to be the most potent endogenous mutagens in humans, and these elements continue to generate mutations and genetic variation in human populations [22]. In some cases, transposon insertions also may go on to create genomic rearrangements by recombining with other transposon copies [22]. Thus, recently mobilized transposons continue to restructure the human genome through a variety of mechanisms.

Methods/Results

The completion of a draft chimpanzee genome sequence provided an opportunity to identify these recently mobilized transposons in both humans and chimpanzees [9]. Transposons that inserted into either of these genomes during the past ~6 million years (i.e., since the existence of the most recent common ancestor of humans and chimpanzees) would be expected to be present in only one of the two genomes. We used a comparative genomics approach to identify these recently inserted transposon copies (Figure 2-1). We began by aligning the sequences of the human and chimpanzee genomes to identify all insertions and deletions (indels).

We screened indels for the presence of transposable elements by comparing each indel to a library of known transposons (RepBase v. 10.02) [120]. Using this approach, we initially identified a total of 14,783 transposon copies that were differentially present in the two genomes. Many of these copies appeared to be recently mobilized transposon insertions, whereas others were simply transposon copies that happened to be located within larger genomic duplications or deletions in the two genomes.

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

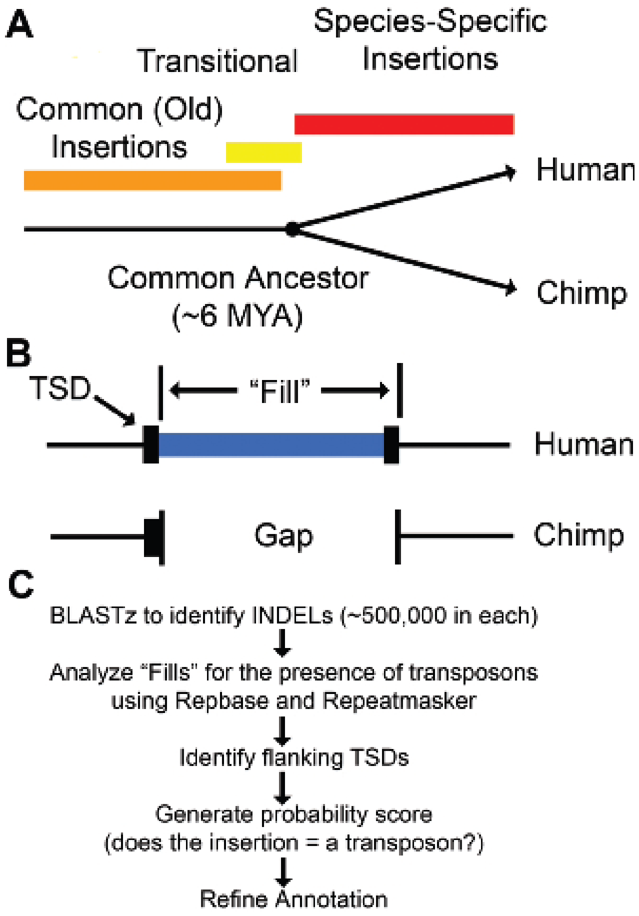
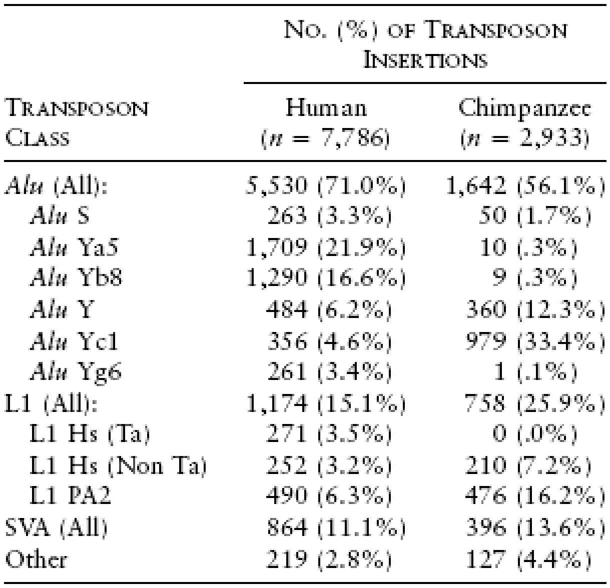

A
Species-Specific Insertions
Transitional
Common (Old) Insertions
Human
Chimp
Common Ancestor (~6 MYA)
B
TSD
"Fill"
Human
Gap
Chimp
C
BLASTz to identify INDELs (~500,000 in each)
Analyze "Fills" for the presence of transposons using Repbase and Repeatmasker
Identify flanking TSDs
Generate probability score (does the insertion = a transposon?)
Refine Annotation

Figure 2-1

Overview of our transposon insertion–discovery pipeline. A. The time line for speciation of humans and chimpanzees is compared with the generation of transposon insertions. Common insertions occurred a very long time ago and are fixed in both species. "Species-specific" insertions are differentially present in the two species and occurred mostly during the past ~6 million years. B. Our strategy for identifying new transposon insertions in humans and chimpanzees. Recently mobilized transposons are flanked by TSDs and are precisely absent from one of the two genomes. Thus, the transposon plus one TSD copy equals the "fill." C. The five sequential steps of our computational pipeline for discovering species-specific transposon insertions are depicted. The draft chimpanzee-genome (build panTro1) and human-genome (build hg17) sequences were obtained from the University of California Santa Cruz browser [129].



	No. (%) of Transposon Insertions	
Transposon Class	Human (n = 7,786)	Chimpanzee (n = 2,933)
Alu (All):	5,530 (71.0%)	1,642 (56.1%)
Alu S	263 (3.3%)	50 (1.7%)
Alu Ya5	1,709 (21.9%)	10 (.3%)
Alu Yb8	1,290 (16.6%)	9 (.3%)
Alu Y	484 (6.2%)	360 (12.3%)
Alu Yc1	356 (4.6%)	979 (33.4%)
Alu Yg6	261 (3.4%)	1 (.1%)
L1 (All):	1,174 (15.1%)	758 (25.9%)
L1 Hs (Ta)	271 (3.5%)	0 (.0%)
L1 Hs (Non Ta)	252 (3.2%)	210 (7.2%)
L1 PA2	490 (6.3%)	476 (16.2%)
SVA (All)	864 (11.1%)	396 (13.6%)
Other	219 (2.8%)	127 (4.4%)

Table 2-1

Summary of transposon insertions.

The most abundant classes of new transposon insertions in both chimpanzees and humans were Alu*,* L1, and SVA element insertions, and these three classes collectively accounted for >95% of the recently mobilized transposons in both species (Table 2-1 and Figure 2-2). However, the relative abundance of these elements and their subfamilies differed between the two species (Table 2-1 and Figure 2-2). Other, less-abundant classes of transposon insertions also were identified in our study. For example, long terminal repeat (LTR) retroelement insertions were observed in both species, including insertions of human endogenous retroviruses (HERVs) and solo LTRs of these elements. Solo LTR insertions have been shown to influence the expression of nearby genes, which makes these insertions of particular interest [121]. Also identified were five full-length HERV-K insertions with relatively long ORFs (up to several thousand amino acids in length) that could remain capable of retrotransposition. Insertions of chimpanzee endogenous retroviruses (CERVs) also were identified [122]. Finally, mammalian interspersed repetitive elements, copies of satellite DNA flanked by unusual TSDs, and small numbers of other interesting transposable elements were identified in the two species.

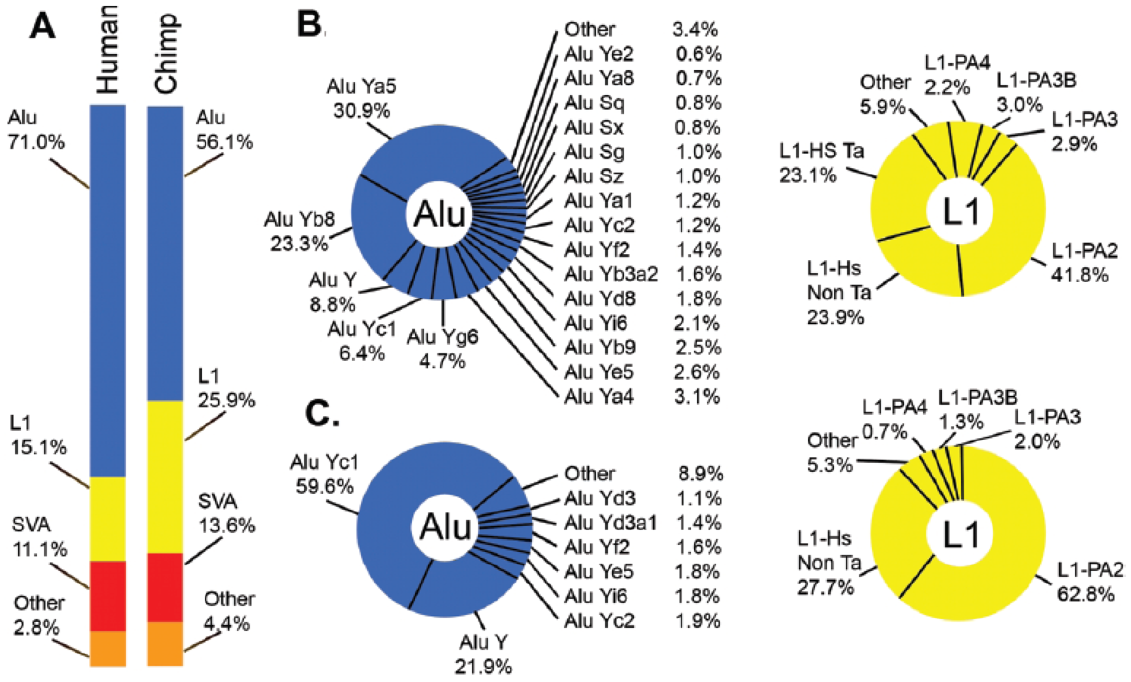


Figure 2-2

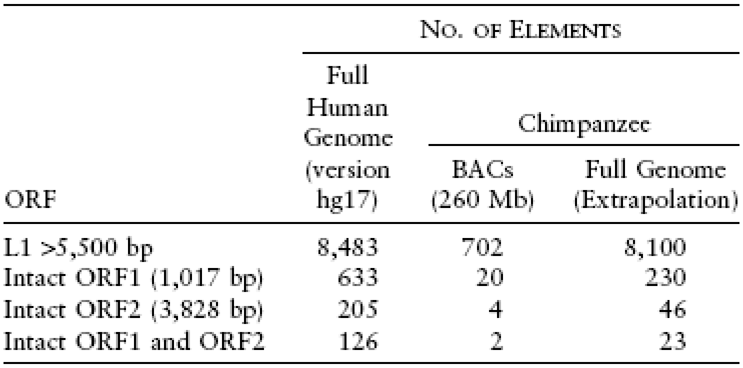
Classes of species-specific transposons in humans and chimpanzees. A. The overall composition of species-specific insertions in humans and chimpanzees. Note that 97.2% of all insertions in humans and 95.6% of all insertions in chimpanzees are Alu, L1, and SVA insertions. B. The distributions of Alu and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

Alu insertions were by far the most abundant class of transposon insertions in both humans and chimpanzees, and these insertions collectively accounted for the bulk of transposons in our study (Table 2-1 and Figure 2-2). The number of *Alu* insertions in humans (5,530) was 3.4-fold higher than the number observed in chimpanzees (1,642). The distributions of these elements among various *Alu* subfamilies also differed between the two organisms (Table 2-1 and Figure 2-2). For example, *Alu* Ya5, *Alu* Yb8, *Alu* Y, and *Alu* Yc1 were highly abundant in humans, whereas only *Alu* Yc1 and *Alu* Y were highly abundant in chimpanzees. Our data indicate that *Alu* S elements, which have been presumed to have been inactive for the past 35 million years [123], apparently have been active in humans and less active in chimpanzees during the past ~6 million years Table 2-1). It is possible that some of these older *Alu* S "insertions" were caused by the precise deletion of *Alu* S elements from one of the two genomes [124] or by gene-conversion events [37]. However, these results also are in agreement with recent data from our laboratory, which indicates that a small number of younger *Alu* S elements are polymorphic in humans and appear to have transposed more recently than the bulk of *Alu* S elements [27]. Overall, our results indicate that the human genome has supported higher levels of *Alu* retrotransposition and has amplified a different set of *Alu* elements than has the chimpanzee genome (Table 2-1 and Figure 2-2). These results confirm and extend previous classifications of *Alu* elements of chimpanzee chromosome 22 [125, 126].

L1 insertions also were abundant in both organisms. In humans, almost 1,200 recently mobilized L1 insertions with TSDs were identified that precisely accounted for gaps in the chimpanzees genome (Table 2-1). These human L1 elements predominantly included members of the L1-Hs and L1-PA2 families (Table 2-1 and Figure 2-2) [6, 127]. The human L1-Hs elements included members of the pre-Ta, Ta0, and Ta1 subfamilies (grouped together as “L1-Hs Ta” in Table 2-1 and Figure 2-2), which are known to be highly active in humans [6]. Also identified in humans were additional L1-Hs and L1-PA2 subfamilies that had unique base combinations at the nine key positions described elsewhere (grouped together as “L1-Hs non-Ta” or “L1-PA2” in Table 2-1 and Figure 2-2) [6, 127]. These novel subfamilies contained 3–65 copies. The remaining L1 insertions in humans belonged to older L1-PA2, L1-PA3, and L1-PA4 groups (Figure 2-2).

The L1 insertions identified in chimpanzees, in contrast, were notably different from those outlined above for humans (Table 2-1 and Figure 2-2). For example, fewer recently mobilized L1 insertions were identified in chimpanzees than in humans (758 in chimpanzees vs. 1,174 in humans). Only 4 of the chimpanzee L1 insertions were full-length (compared with >200 new full-length insertions in humans), and none of the chimpanzee L1 insertions had intact ORFs. The initial draft sequence of the chimpanzee genome is likely to contain assembly errors that may account for at least some of these observed differences. However, we also observed differences in the L1 subfamilies of these organisms that are unrelated to genome assembly issues. For example, proportionally more L1-PA2 insertions and fewer L1-Hs insertions were observed in chimpanzees than in humans (Figure 2-2). Initially, we were surprised to find L1-Hs elements in chimpanzees at all, since these elements were expected to be found only in humans. However, further analysis revealed that most of the L1-Hs elements in chimpanzees actually were “intermediate” elements that matched L1-Hs overall but had ORF1 sequences that were more similar to L1-PA2 elements. Therefore, the L1-Hs family of elements includes subfamilies that are truly human specific as well as other L1-Hs–like elements that are not human specific. We also aligned and analyzed all of our chimpanzee L1 insertions, using ClustalW and PAUP, to determine whether any new L1 subfamilies (equivalent to L1-Ta elements in humans) were present in chimpanzees. In addition, we classified all of our chimpanzee L1 insertions, using the nine key positions that have been used elsewhere to classify human L1 elements [6, 127]. In both cases, we failed to identify any new extended subfamilies of L1 elements within our collection of chimpanzee insertions. Therefore, a dominant class of new offspring elements analogous to L1-Ta elements in humans does not appear to have been produced in recent chimpanzee history.

We next examined all of the existing L1 ORFs in the human and chimpanzee genomes to further characterize possible differences between the L1 elements of these species. We screened the human and chimpanzee genomes for ORFs in all nearly full-length elements (>5,500 bp) and identified 633 L1 elements with intact ORF1 sequences in the human genome but only 39 elements with intact ORF1 sequences in the draft chimpanzee sequence. Moreover, we identified 205 L1 elements with intact ORF2 sequences in the human genome (Table 2-2) but failed to detect intact ORF2 sequences in the draft chimpanzee sequence. These results suggested that functional L1 elements were likely to be rare in chimpanzees. As outlined above, however, it also was possible that the quality of the chimpanzee draft sequence affected our ability to detect ORFs accurately. We determined that the sequence quality of the >5,500 bp L1 elements in chimpanzees had average scores that generally were high (>40 Phred scores) [128, 129]. However, single bases of low quality (<10 Phred scores) [128] also were distributed throughout the draft sequence at sporadic intervals [129]. These single bases, although rare, often resulted in frameshifts. Therefore, it was possible that these sporadic low-quality bases were interfering with our ability to detect ORFs accurately.



	No. of Elements		
	Full Human Genome (version hg17)	Chimpanzee	
ORF		BACs (260 Mb)	Full Genome (Extrapolation)
L1 >5,500 bp	8,483	702	8,100
Intact ORF1 (1,017 bp)	633	20	230
Intact ORF2 (3,828 bp)	205	4	46
Intact ORF1 and ORF2	126	2	23

Table 2-2
Analysis of L1 ORFs.

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

We next examined the L1 ORF1 and ORF2 coding regions from chimpanzees to determine whether the encoded proteins are likely to remain active today. Brouha et al. [6] showed elsewhere that human L1 elements that differ by an average of only 21 nucleotide changes from an active human L1 consensus were inactive. Thus, even elements that were >99% identical to this consensus could be inactive, and no human elements that were <99% identical were found to be active. We determined that the human genome contains at least 119 elements with >99% nucleotide identity to the active human L1 consensus [6] within the regions encoding ORF1 and ORF2. In contrast, no L1 ORFs in the chimpanzee genome or BACs had >99% identity to this active human L1 consensus. We cannot rule out the possibility that some of the chimpanzee L1 elements that are <99% identical to the human L1 consensus could remain active. Other “hot L1” elements might have evolved separately in chimpanzees that are >1% variant from the most-active human elements. However, since we failed to detect extended subfamilies of L1-PA2 or L1-Hs elements within our collection of chimpanzee insertions (analogous to the L1-Ta subfamily in humans), such elements generally would be present at low copy numbers in the chimpanzee genome. Thus, the landscape of potentially functional L1 elements in chimpanzees appears to be quite different from the landscape of active L1 elements in humans.

To verify these ORF results, we used an ORF1-trapping method to recover full-length ORF1 sequences from L1 elements in humans and chimpanzees [130]. We recovered and sequenced 41 intact ORF1 sequences from humans and 51 intact ORF1 sequences from chimpanzees and observed results that were very similar to those obtained through the computational methods described above. None of the intact chimpanzee ORF1 sequences recovered through ORF1 trapping were >99% identical to the active human L1 consensus, whereas 17 (41%) of the 41 intact human ORF1 sequences were >99% identical to the active human L1 consensus. Almost all of the intact ORF1 sequences trapped from chimpanzees (48/51; 94%) were L1-PA2 ORF1 sequences. In contrast, only 21 (51%) of the 41 intact ORF1 sequences recovered from humans were L1-PA2 ORF1 sequences and 18 (44%) were L1-Hs sequences. Thus, our ORF1-trapping experiments confirmed that the most recently active elements in chimpanzees (i.e., those with intact ORFs) contained ORF1 sequences that were divergent from the active L1 consensus in humans [6].

Our method for trapping ORF1 in humans and chimpanzees employed the pβFUS plasmid [130]. Briefly, full-length ORF1 sequences were amplified from human and chimpanzee genomic DNA (NA1MR91 and NA03448A, respectively [Coriell Cell Repository]) using PCR. The PCR primers were identical for humans and chimpanzees, and had the following sequences 5′-CCTGATCTGCGGCCGCATGGGGAAAAAACAGAACAGAAAAACTGG-3′ and 5′-CGTCCGAACGATATCCATTTTGGCATGATTTTGCAGCGGCTGG-3′. We used a combination of human and chimpanzee ORF1 sequences to design these primers. The ORF1 sequences identified in our chimpanzee BAC experiments were aligned to generate a consensus sequence using ClustalW. This sequence was compared to the human L1 consensus, and we determined that the primers chosen were conserved in human L1 sequences. Finally, we compared the candidate primer regions with L1-PA2, L1-PA3, and L1-PA4 elements and determined that the primer sequences also were completely conserved in these elements. Thus, the primers chosen were capable of amplifying a wide spectrum of ORF1 sequences in both humans and chimpanzees (including L1-Hs, L1-PA2, L1-PA3, and L1-PA4 elements). The *Not*I and *Eco*RV restriction sites that were introduced for cloning purposes are underlined. PCR products were cut with these enzymes and ligated to *Not*I/*Sma*I-digested pßFUS, such that the complete ORF1 sequence would be in frame with the AUG-less *lacZ* in the plasmid. Recombinants were identified on LB medium containing X-gal (recombinants with ORFs were blue, whereas those without ORFs and the empty vector alone were white). DNA was prepared and sequenced at Agencourt Biosciences using the primers 5′-CCAGTCACGTTGTAAAACGAC-3′ and 5′-CTAGGCCTGTACGGAAGTGTTAC-3′. High-quality sequences were analyzed and assembled using Sequencher version 4.1.2.

In addition to *Alu* and L1 insertions, we also found that SVA elements have been highly active in humans and chimpanzees (Table 2-1). In fact, SVA insertions were almost as abundant as L1 insertions in humans during the past ~6 million years (Table 2-1). SVA is an unusual composite element that contains four components: (1) a tandem repeat of TCTCCC(n), (2) an unusual *Alu* element in reverse orientation, (3) a central variable-number-of-tandem-repeat (VNTR) region that is rich in CpG sequences, and (4) a SINE-R sequence that was derived from an LTR element [131]. SVA ends with a poly (A) tail and is flanked by TSDs that closely resemble the TSDs of *Alu* and L1 elements [27, 28]. SVA recently was found to be highly polymorphic among humans [27], and a few instances have been reported of SVA insertions causing diseases [28, 34]. Our study now provides further evidence that SVA has been actively mobile in relatively recent primate history and may remain active today.

The ORF1 and ORF2 proteins of L1 elements perform a specific retrotransposition mechanism known as “target-primed reverse transcription” (TPRT) [21], in which L1 mRNAs are copied into cDNAs and integrated into the genome [22]. *Alu* RNAs (and other cellular RNAs) can compete for the L1 machinery during the TPRT process, which leads to the retrotransposition of these alternative RNAs instead of the normal L1 mRNAs [23, 31, 132]. This “trans” mechanism of retrotransposition is thought to have led to the massive expansion of *Alu* [31] and SVA [27, 28] elements in the human genome. Therefore, if L1 elements are indeed less functional in chimpanzees, as predicted above (Table 2-2), we likewise might expect to see fewer *Alu* and SVA insertions in the chimpanzee genome. Table 2-1 and Figure 2-3 show that this is, in fact, the case. Since other factors also influence the amplification rates of *Alu* (and probably SVA) elements, these differences may not be totally caused by lower levels of L1 activity in chimpanzees. It is possible, for example, that humans had a larger number of potentially active *Alu* and SVA source elements than did chimpanzees in recent history.

However, when combined with our other data demonstrating that chimpanzees (1) have fewer full-length L1 insertions than humans, (2) have fewer L1 elements with intact ORFs than humans, (3) have ORF sequences that are divergent from active human L1

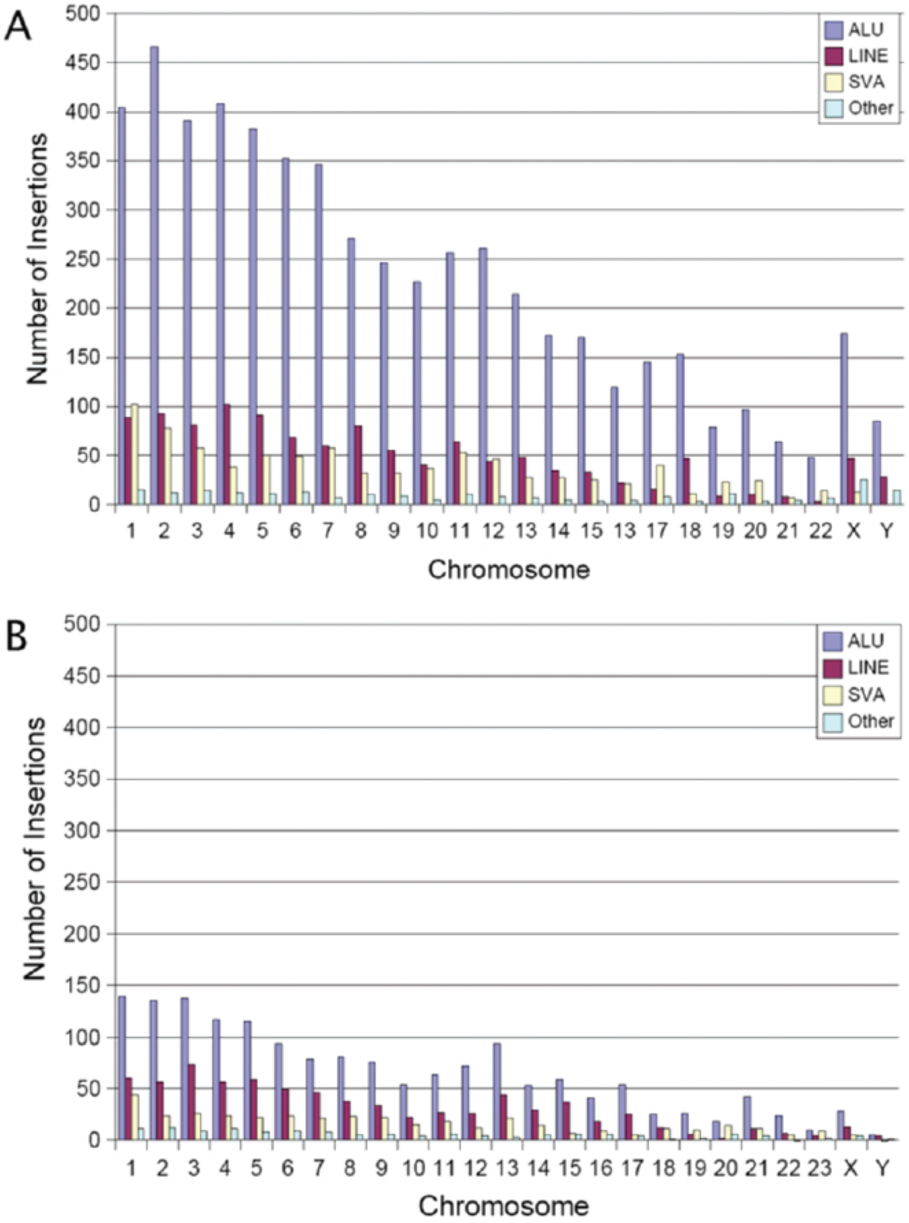
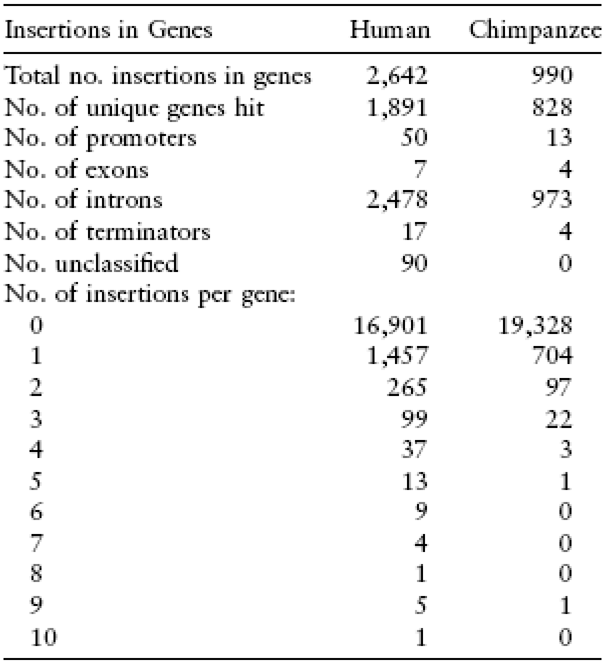


Figure 2-3

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



Insertions in Genes	Human	Chimpanzee
Total no. insertions in genes	2,642	990
No. of unique genes hit	1,891	828
No. of promoters	50	13
No. of exons	7	4
No. of introns	2,478	973
No. of terminators	17	4
No. unclassified	90	0
No. of insertions per gene:		
0	16,901	19,328
1	1,457	704
2	265	97
3	99	22
4	37	3
5	13	1
6	9	0
7	4	0
8	1	0
9	5	1
10	1	0

Table 2-3

Transposon insertions within genes.

elements, and (4) lack extended subfamilies of new insertions, these data collectively indicate that chimpanzees are likely to have supported lower levels of L1 activity in recent history compared with humans.

We next examined the genomic distributions of our recently mobilized transposon insertions. In both humans and chimpanzees, these insertions generally were distributed according to the amount of DNA that was present on each chromosome (Figure 2-3). We also examined the distributions of new insertions relative to genes (Table 2-3). Approximately 34% of the new insertions in both genomes were located within known genes (defined as 3 kb upstream to 0.5 kb downstream of a RefSeq gene) (Table 2-3).

Using the same criteria, we determined that genes occupy ~34% of the human and chimpanzee genomes (33.5% and 34.8%, respectively). Therefore, the fraction of insertions in genes was very close to that expected if integration (and mechanisms that subsequently remove insertions) had occurred randomly during the past ~6 million years.

However, further analysis of these patterns revealed that they were not, in fact, random. Although we identified insertions in only ~14% of all human genes, many of these genes had more than one insertion (Table 2-3). Overall, about a third of the human genes with insertions contained multiple insertions. Similar results were observed in the chimpanzee genome (16.5% of the genes with insertions had multiple insertions). We performed one-sample *Z* tests with our insertions and determined that the observed patterns were not consistent with a random integration model. For example, we observed 16,901 human genes that lacked new transposon insertions from our collections (Table 2-3). The chance of observing this many human genes without such insertions is zero (*P*=0) with a random integration model. Similar results were observed with *Z* tests for the remaining integration classes listed in Table 2-3 (data not shown). Therefore, our statistical tests allowed us to reject the hypothesis of random integration with a very high degree of confidence. On the basis of this analysis, it appears that a large fraction of the new transposon insertions in humans and chimpanzees (the majority of which were *Alu,* L1, and SVA elements) were targeted preferentially to specific genes. It is also possible that negative selection eliminated insertions from a larger initial collection over time, and this led to the appearance of nonrandom integration. Although targeted integration of L1 has not been observed previously in biochemical or cell culture experiments, previous studies indicate that transposons are eliminated through negative selection [56]. Thus, negative selection is likely to have played a role in dictating the final patterns of transposons observed. Our data also may reflect an integration targeting mechanism that is not functional in cell-culture systems but is active in the germline of whole organisms, where all of our insertions occurred.

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

Discussion

In conclusion, we have determined that the original set of transposons in the common ancestor of humans and chimpanzees behaved differently during the subsequent evolution of these organisms. More than 95% of the new transposon insertions in both organisms were *Alu,* L1, and SVA insertions. However, our data indicate that humans and chimpanzees have amplified very different subfamilies of these elements. Our combined data also indicate that chimpanzees have supported lower levels of L1 activity than have humans during the past several million years, and this has led to decreased levels of *Alu,* L1, and SVA transposition in chimpanzees. Other factors, such as differences in population sizes and differences in population bottlenecks, also are likely to have influenced the final patterns of transposon insertions observed in these organisms. In some cases, apparent “insertions” may have been caused by the precise deletion of transposon copies through homologous recombination at the TSDs flanking these elements [124]. A fraction of our insertions also may have been older polymorphisms that were subject to lineage sorting. Thus, the final patterns of transposons in these genomes are likely to have been shaped not only by integration and excision mechanisms but also by the population dynamics of these organisms during the past several million years

Chapter 3

L1 Retrotransposition Occurs Frequently in the Human Germline

Introduction

Polymorphic retrotransposon insertions are a major source of human genetic variation, however, geneticists still no little about where these insertions might be in human genomes [27]. Also, the lack of known polymorphic insertions in public databases indicates that no one has yet conducted a comprehensive search for such polymorphisms [93]. We wanted to know how many previously unknown retrotransposon polymorphisms exist in individual genomes and whether the allelic frequencies of these polymorphisms are consistent with those of currently known polymorphisms.

Long Interspersed Nucleotide Element-1 (L1) is a retrotransposon family that is ubiquitous among mammalian genomes. L1 is the only autonomous, active retrotransposon family in the human genome [26]. L1 elements are essentially endogenous mutagens. New insertions can alter gene function by a variety of mechanisms including insertional mutagenesis, altering transcription start sites, reducing the stability of RNA, promoting aberrant splicing, and recruiting epigenetic silencing marks [e.g. 40, 45, 48, 51, 52, 133, 134]. Dozens of known insertions cause disease [33]. Active L1 elements are 6 kb long and contain two open reading frames (ORFs) that encode the proteins necessary for retrotransposition [18-20]. L1 elements retrotranspose by a process known as target primed reverse transcription (TPRT) [21]. During this process, L1 proteins bind to their coding L1 RNA and create a free 3’OH in the genome that acts as a primer for reverse transcription of the L1 RNA [21, 22]. This “copy & paste” mechanism enabled L1 to reach a copy number of around 500,000 in the human genome [4]. The majority of these elements are truncated at their 5' end and have accumulated inactivating mutations. Each person, however, on average, has 80-100 active L1 elements in their genome [6].

In contrast, Alu is a class of nonautonomous retrotransposon that relies upon L1-encoded proteins in order to retrotranspose [31]. Although L1 proteins tend to retrotranspose their own encoding mRNA (a phenomenon known as "cis-preference") [23], Alu elements have proven successful at commandeering the L1 machinery and have reached copy numbers topping one million in human and other primate genomes [31].

Active retroelements, including L1 and Alu, are copied during retrotransposition so that sequence changes within the element are passed on to offspring elements. Thus, shared sequence changes relative to a consensus sequence have helped define different families of these elements and their ages [Reviewed in 5]. L1Ta is the youngest L1 family and is defined by the sequence ACA at position 5930-2 (relative to L1.2 GenBank Acc # M80343) [32, 135]. Older families of L1 contain the sequence GAG, while a relatively young family, known as pre-TA, contains an ACG. The L1Ta family arose around 2 million years ago, after the divergence of human and chimp and is, thus, a human-specific retrotransposon family [95]. Forty-five percent (45%) of the L1Ta elements in the human reference sequence are polymorphic [95].

Several subfamilies of Alu have active elements and we estimate that each person carries thousands of potentially active Alu elements [30, 123]. In particular, the Alu Ya5 subfamily produces new insertions, including some that have caused disease [33]. About 20% of the members of this subfamily are polymorphic [94].

The majority of retrotransposons are likely neutral [56, 57]. The neutral theory argues that the majority of mutations are neutral, and thus genetic drift, and not natural selection, is the primary force altering allelic frequencies from one generation to the next [92].The probability of a new mutation reaching fixation is 1/2N and the average time to fixation is 4N generations in a population at equilibrium (where N is the effective population size) [92]. Current estimates of retrotransposition rate for Alu and L1 range from 1 in 2.4 to 1 in 100 live human births [6, 89-91]. Therefore, the majority of polymorphic retrotransposable elements are likely rare alleles (presumably resulting from recent retrotransposition events) with a minority reaching higher frequencies and fixation. Furthermore, the recent demographic expansion of the global human population should act to increase the overall bias towards the numbers of rare alleles.

Although retrotransposons are likely neutral (or deleterious) to their human hosts, the Database for Retrotransposon Insertion Polymorphisms (dbRIP) shows an allelic frequency spectrum that is flat with a peak of very common (>0.9) alleles [93]. Researchers discovered many of the retrotransposons in dbRIP by “candidate screening”: using PCR primers flanking a young retrotransposon in the human reference sequence and looking for amplification of the preintegration site in other individuals [e.g. 29, 93-96]. The lack of rare alleles in dbRIP is surprising and suggests that the majority of retrotransposon polymorphisms remain to be discovered.

In order to better survey the complete frequency spectrum of polymorphic retrotransposons in the human genome, an assay capable of discovering both common and rare insertions was developed. A more comprehensive search for polymorphic elements should further the scientific community's understanding of these mutagens. For this assay, the human genome was digested with a restriction enzyme and linkers were then ligated onto the digested DNA. Retrotransposon-specific and linker-specific primers were then used to amplify insertion junctions. The unique sequence directly downstream of the retrotransposon was used to map back to the human reference sequence (Hg18). This assay allowed for the discovery of retrotransposable elements without *a priori* knowledge of their genomic locations. Initially, standard ABI capillary sequencing was used. To increase throughput and genomic coverage, Roche 454 FLX pyrosequencing was later used. These less biased assays should allow for hundreds of previously unknown retrotransposon insertions to be found. The allelic frequency spectrum of polymorphic retrotransposons should show an excess of rare alleles. The population sample consisted of 38 healthy, diverse individuals, 8 tumor-derived cell lines, 20 lung tumor samples, and 10 brain tumor samples. The sequencing of 4600 different clones and 185,348 pyrosequencing reads from PCR libraries and the subsequent analysis of nearly 1,000 novel L1 and Alu insertions in humans is described.

Results

ABI Capillary Sequencing

4600 clones from 46 individuals were sequenced. From this, 3795 (82%) potential L1 insertions were unambiguously mapped. Those that could not be mapped failed to sequence (4.9%), were empty vector (2.3%), had too little genomic sequence to map using BLAT (<20 nts, 2.5%), fell within repeats (6.5%), or had poor sequence quality after the poly-A tail of the L1 (1.5%). As expected, the majority (3343 reads, 88%) of the mapped L1 insertion junctions were from L1s that are present in the human reference sequence (Hg18). In contrast, 12% of the unambiguously mapped reads represent L1 elements that are not in the reference sequence.

The specificity of the assay for amplifying L1Ta subfamily members was determined. L1Ta and preTa are the youngest L1 subfamilies and a large fraction of these subfamily members in Hg18 are polymorphic (45% and 14% respectively). Primers were designed to specifically amplify L1Ta and preTa subfamily members to increase the odds of finding polymorphic elements. Of the 3343 reads belonging to L1s in Hg18, 601 (18%) were L1Ta or preTa elements. Considering that L1Ta and preTa elements make up less than 1% of all L1 elements per genome, there was 88-fold enrichment for these elements.

	L1s in ABI Sequencing	L1s in Pyrosequencing	Alus in Pyrosequencing
Sequenced	4600	150326	35022
Mapped	3795	43676	22338
Nonredundant	783	1101	3799
In Hg18 / Not in dbRIP	500	402	3150
In Hg18 / In dbRIP	100	147	184
Not in Hg18 / In dbRIP	31	60	62
Not in Hg18 / Not in dbRIP	152	492	403

Table 3-1.

Summary of sequencing statistics. Sequenced reads were mapped to the May 2006 build of the human reference sequence (Hg18) using BLAT. Mapped numbers include redundant reads for the same L1 elements. Independent insertions are the number of nonredundant L1 insertions found. The independent insertions were scored for whether they were present in Hg18 and the Database for Retrotransposon Insertion Polymorphisms (dbRIP)

Individual L1 elements were sometimes represented by multiple mapped reads. Many of the 3795 mapped reads were from the same L1 elements found within or between individuals. Altogether, the 3795 mapped L1s represented 783 independent insertions (Table 3-1). Since a large fraction of L1Ta and preTa members are polymorphic, we predicted that some of the L1s that we mapped would also be polymorphic. Each of the 783 independent insertions were categorized into one of the following: (1) those in the human reference sequence (Hg18) but not in the Database for Retrotransposon Insertion Polymorphisms (dbRIP, presumably fixed elements), (2) those in Hg18 and also in dbRIP (presumably common polymorphisms), (3) those outside of Hg18, but in dbRIP (also presumably common polymorphisms), and (4) those neither in Hg18 nor in dbRIP (which we considered "novel" elements). Five hundred L1s fell within the first category (Table 3-1). An additional 183 L1s found in this experiment were not present in the human reference sequence (Table 3-1). Thirty-one of these elements were found in previous studies and were present in dbRIP (Table 3-1). The remaining 152 L1s represent novel, polymorphic insertions (Figure 3-1, genomic coordinates are available in Appendix I). Sixty-six of these elements were randomly chosen for PCR validation and 64 (97%) were confirmed.

Whether these novel L1 insertions were common polymorphisms that had not yet been observed or were more recent insertions was determined. Forty-nine of the 152 novel L1s were randomly chosen to genotype in the same 46 diverse individuals from which they were identified (a subset of the novel L1s for which there was unique sequence nearby for PCR primers to anneal were chosen). Novel L1 elements that could

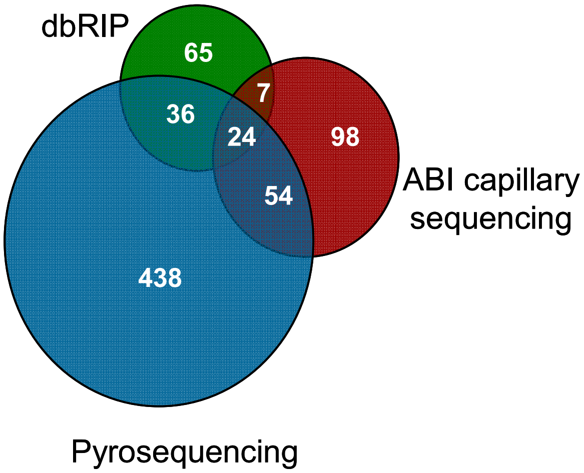


Figure 3-1

L1s outside of the human reference sequence. The number of L1s beyond Hg18 are plotted.

only be verified in the initial tumor-derived cell line from which they were sequenced were omitted from the allelic frequency analysis in case they were somatic insertions and, thus, not heritable. The genotyped novel L1s show an allelic frequency spectrum skewed toward rare insertions (Figure 3-2). This skew is unlike the pattern seen for L1s currently in dbRIP and suggests that dbRIP is biased towards common elements. This pattern is different from the U-shaped allelic frequency pattern observed for single nucleotide polymorphisms (SNPs) under the neutral expectation. Every biallelic segregating site has a minor and major allele whose frequencies add up to 1, thus creating a spectrum from 0 to 0.5 that is mirrored from 0.5 to 1.0. While retrotransposon polymorphisms are also normally biallelic, the preinsertion allelic frequencies are not shown here. Therefore, the pattern of allelic frequencies presented here is consistent with the neutral hypothesis despite appearing different from the pattern observed for neutral SNPs.

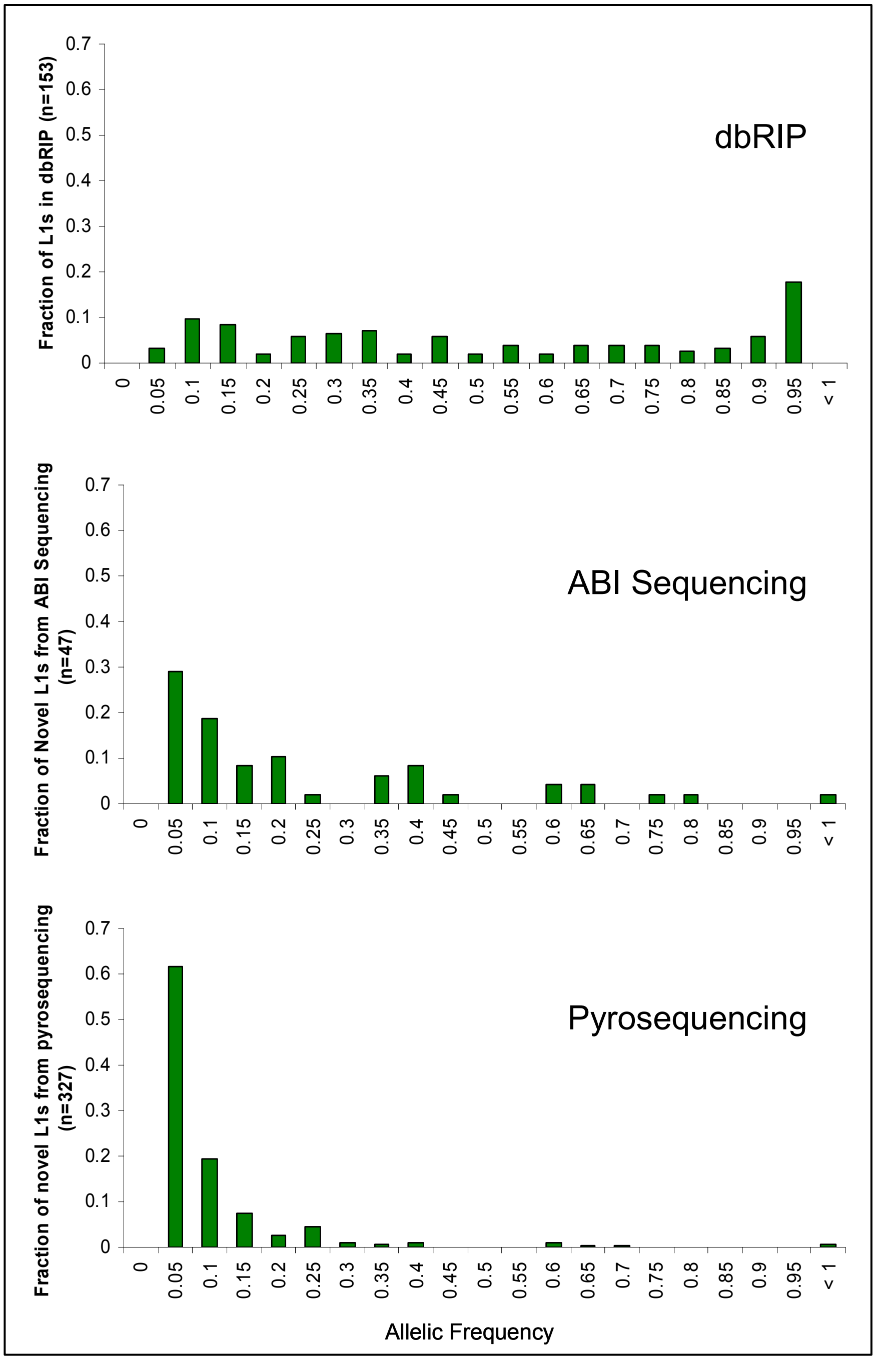


Figure 3-2

Histogram of allelic frequencies for polymorphic L1s. Allelic frequency data was extracted from dbRIP (n = 153). For Dideoxy Sequencing, allelic frequencies were determined by PCR in a panel of 46 diverse individuals (n = 47). For Pyrosequencing, allelic frequencies are estimates based on sequence data and using Hardy-Weinberg (n = 327).

In addition, 8 novel L1 insertions were found only once in heterozygous individuals and were considered "singletons". If we assume that the 47 L1s genotyped by PCR are representative of all the 152 novel L1s (150 when we omit 2 potential somatic insertions), we would expect that 1 in 2 individuals has a singleton. If we assume these singletons are the result of *de novo* insertion, our results are in good agreement with published estimates of retrotransposition rates. However, these singletons are quite possibly inherited and simply not present in the remainder of the panel of 46 individuals.

Pyrosequencing

We were encouraged by our discovery rate of novel L1s in the ABI capillary sequencing experiment, however, the low sampling depth (100 reads per person, with an estimated 520 L1Tas per genome is around 0.2-fold coverage) lead us to believe that many L1s went undiscovered in the samples. Thus, we could not be sure whether the 152 novel L1s comprehensively sampled the frequency spectrum of L1 elements, nor could we get a good idea of the rate of singletons. A follow-up experiment was performed with different restriction sites and primers to have better coverage of the genome. The PCR products were applied to Roche 454 FLX pyrosequencing to sequence the libraries to greater depth. Genomic DNA from 20 lung tumor tissues and adjacent normal lung tissues from the same 20 patients were used. Genomic DNA from 10 brain tumor tissues and blood from the same 10 patients were also used. For this experiment, 5' Alu insertion junctions were also amplified in addition to 3' L1 junctions. L1 insertion junctions for the lung samples and both L1 and Alu insertion junctions for the brain and blood samples were sequenced.

185,348 reads were sequenced collectively for Alu and L1 insertion junctions (Table 3-1). 66,014 (36%) of the reads were mapped to Hg18. The vast majority of reads that could not be mapped were short PCR products that did not have enough unique genomic sequence downstream of the L1 to unambiguously map it using BLAT (<20 nucleotides). Altogether, 1101 independent L1 insertions and 3799 independent Alu insertions were identified and used for subsequent analyses.

Independent insertions were split into 4 categories as before based on their presence in Hg18 and dbRIP (Table 3-1). Four hundred and two (402) L1 elements are likely fixed based on their presence in Hg18 and absence from dbRIP. An additional 552 L1 elements were not found in Hg18. Sixty of these were in dbRIP and the remaining 492 L1 elements were considered novel. Fifty-four of these novel L1s were among the 152 novel L1s in the ABI sequencing assay. These were excluded from subsequent analyses for a total of 438 novel L1 elements and 403 novel Alu elements (genomic coordinates are available in Appendices II and III). Four hundred thirty-eight novel L1s is more than triple the number of L1s in dbRIP that are outside of Hg18 (Figure 3-1). This demonstrates the efficiency of the assay as well as the current lack of comprehensive sampling of L1s.

We wanted to know the frequency of these elements to determine whether our data more closely resembles dbRIP or the ABI sequencing data. First, the estimated allelic frequencies of known polymorphic L1s were compared to known allelic frequencies to gauge how accurate the estimates were. Second, PCR was used to determine whether L1s that were sequenced rarely were, in fact, rare. Finally, allelic frequencies for novel L1s were estimated to determine whether the pattern resembled the expectation of the neutral theory better than dbRIP.

Estimated allelic frequencies are highly correlated with actual frequencies

To determine whether accurate allelic frequencies could be estimated from the pyrosequencing data, L1s of known allelic frequencies were compared to the sequencing-based estimates. For 63 L1s of previously determined allelic frequencies (found in dbRIP), their allelic frequencies were estimated in the 30 samples. Since homozygotes and heterozygotes with an L1 insertion could not be distinguished, the Hardy-Weinberg equation was used to infer allelic frequencies. For each of the 63 L1s, the fraction of individuals in the experiment without the L1 was determined based solely on sequencing data. This fraction was used as the "q^2"value so that "p" (the allelic frequency for each L1) could be derived (Figure 3-3). Estimated and actual allelic frequencies were well-correlated with a high Pearson's correlation coefficient (0.827). Even so, estimated allelic frequencies were generally lower than actual allelic frequencies. Thus, from the sequencing data, allelic frequencies could be crudely estimated, but follow-up with PCR or greater sequencing coverage would be needed to determine more precise allelic frequencies.

Singletons are abundant in human populations

Since the ability to identify rare L1 polymorphisms was limited by low sequencing coverage (Figure 3-6), we determined whether apparent singletons (insertions sequenced

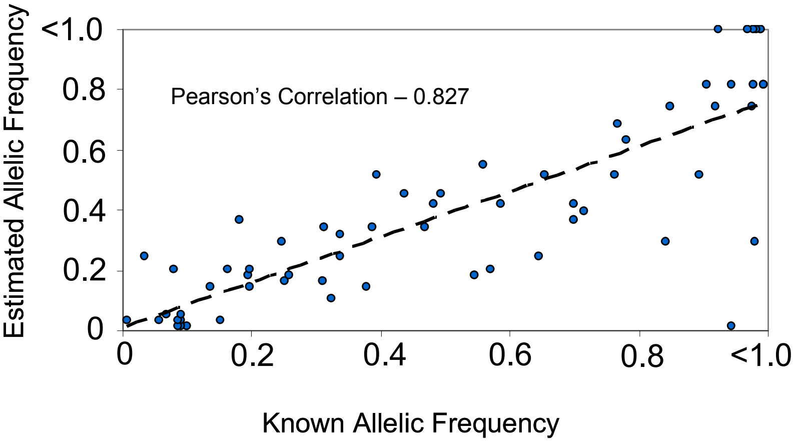


Figure 3-3

Correlation of estimated and previously determined allelic frequencies. For the pyrosequencing experiment, allelic frequencies were estimated using sequence data and Hardy-Weinberg and were compared to known allelic frequencies of polymorphic L1 elements (n = 63).

from only a single individual) were truly present in only one individual. Of 120 apparent singletons, 22 were randomly chosen to validate by PCR (Figure 3-4). The presence of the L1 insertion in the DNA from which it was initially identified was determined. Also, whether the L1 was present in 3 pools each of 15 diverse human genomes was established. Thirteen (59%) singletons were confirmed only in the initial genomes from which they were sequenced while 9 (41%) singletons were present in their initial genomes as well as other genomes. Thus, 41% of singletons were more common than the sequencing data alone indicated. If these 22 singletons are representative of all the 120 singletons identified, then 71 are estimated to be singletons while 49 are likely common. Therefore, each individual in this pyrosequencing study has, on average, 2.3 rare L1s. This is 4 times more than the 1 singleton for every 2 people estimate from the

ABI capillary sequencing project. Also, the low sequencing coverage means there were likely more singletons in these samples, however, singletons may have been inherited and not truly *de novo* insertions. While these singletons indicated that retrotransposition is common in humans, it did not yield a definitive retrotransposition rate.

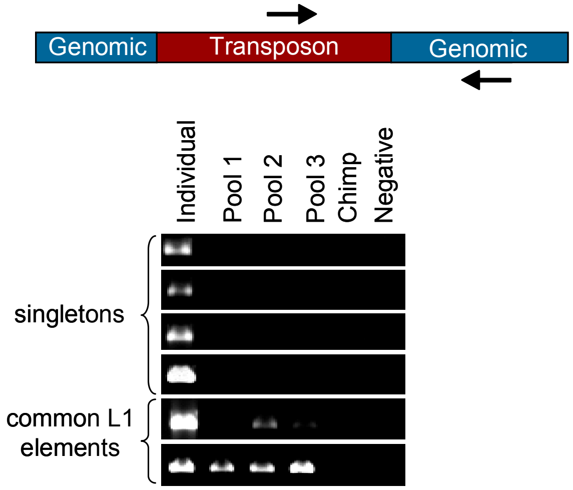


Figure 3-4

Singletons verified by PCR. A subset of L1 elements that appeared to be "singletons" based on their presence in a single individual and absence from Hg18 and dbRIP were verified by PCR in pools of diverse human DNA. The *Individual* lane is the individual from whom the L1 was initially sequenced. *Pool* lanes contain DNA from 15 diverse humans. *Chimp* is Coriell #NA03448A and *Negative* is a control PCR with no template.

Estimated allelic frequencies show a sharp skew toward rare insertions

Although the 30 samples from cancer patients used in the pyrosequencing study were not representative of human populations, allelic frequencies from the sequencing data were estimated (Figure 3-2) as they were previously done for the 63 known L1s. L1s that were sequenced in only a single tumor tissue were omitted from this analysis in case they were absent from the germline and, thus, not heritable. Once these were eliminated, there were a total of 327 novel L1s.

Sixty percent (60%) of the novel L1s, had an estimated allelic frequency of <0.1 (Figure 3-2). This skew toward rare insertions was much more pronounced than expected and was likely due to the lack of sequencing coverage (Figure 3-6). Since sequencing was done to only an average of 1.8-fold per person, it was likely that many of these "rare" L1s were actually more common and just happened to not be sequenced in every genome where it was present. The analysis of L1s of known allelic frequencies (Figure 3-3) and apparent singletons (Figure 3-4) supported this prediction in that many of these "rare" L1s were more common than they appeared. Even so, it seemed likely that L1 mutagenesis occurred frequently enough in this "population" to skew the allelic frequency spectrum toward rare insertions.

The allelic frequencies of novel L1s found by both ABI capillary sequencing and pyrosequencing appeared different from those of L1s in dbRIP (Figure 3-2). Kolmogorov-Smirnov tests (KS-tests) were performed to determine whether L1 elements in dbRIP, and the novel L1s found by both methods differed significantly. Each pair-wise comparison of data sets gave a significant ($P<0.001$) result, thus rejecting the null hypothesis of the datasets being similar. The smallest difference between datasets was between the novel L1s found by ABI sequencing and those found by pyrosequencing (D=0.4552). As expected, frequencies of L1s in dbRIP differed most significantly from the L1s found by pyrosequencing (D=0.7354). Therefore, we concluded that dbRIP's L1 dataset is significantly different from the dataset found by our assay and is enriched for common polymorphisms.

Discussion

Many novel L1s are rare and their allelic frequencies differ from those in dbRIP

Current estimates of L1 retrotransposition rate (RR) range from 1 in 2.4 to 1 in 100 live human births [90, 91]. The neutral hypothesis states that the probability of a neutral mutation reaching fixation is 1/2N where N is the effective population size. Thus, if these estimates of RR are accurate and the majority of polymorphic L1 insertions are neutral or deleterious, polymorphic L1s should be at low frequencies in the population and, thus less likely to be observed. L1s in the Database of Retroviral Insertion Polymorphisms (dbRIP) have an allelic frequency spectrum that is relatively uniform across most frequencies with an excess of very common (>0.9) insertions (Figure 3-2). While the HapMap project mined genomes for single nucleotide polymorphisms (SNPs), no such thorough effort has been taken to identify L1 polymorphisms. Thus, the excess of common insertions in dbRIP is a result of a lack of extensive sampling as well as ascertainment bias.

In this study, L1 insertions were sequenced without *a priori* knowledge of their genomic locations which allowed for the discovery of rare L1s with less bias than previous methods. In fact, although the ABI capillary and pyrosequencing experiments differed greatly, their allelic frequency spectra for novel L1s showed an excess of rare alleles (Figure 3-2). The skew toward rare insertions demonstrated the sensitivity of both assays, and differed greatly from dbRIP.

In fact, the average individual in the pyrosequencing experiment had an estimated 2.3 L1s that no other individual in the experiment had. This number was higher than current estimates of retrotransposition rate [6, 89-91]. Some explanations that may account for this are: (1) some of these singletons were likely inherited and not *de novo*, (2) the 22 L1s chosen for validation were not representative of the remaining singletons, or (3) retrotransposition is more frequent than current estimates. More of the L1s may be common than estimated from the pyrosequencing experiment (Figures 3-3 and 3-4). Also, additional singletons were likely present in the samples. Perhaps the best way to estimate L1 retrotransposition rate would be to apply the assay to parent-child trios with higher sequencing coverage.

Selective resequencing is an effective method to study retrotransposon polymorphisms

Retrotransposition is an ongoing source of polymorphisms in human populations. These insertions have the potential to alter phenotypes and cause disease. In this study, we demonstrated the utility of selective sequencing in identifying novel retrotransposon insertion junctions. 993 novel Alu and L1 insertions were identified from 76 individuals (Table 3-1). The efficiency of this assay in finding previously unknown retrotransposon insertions in humans is a testament to how understudied these elements are.

While both technologies differed greatly in their throughput, chemistry, and sample preparation, they were both used to discover novel retrotransposon insertions, including singletons. As shown in Figure 3-6, greater sequencing depth allowed for greater discovery in the sample. Thus, while this data was the most comprehensive collection of human retrotransposon polymorphisms to date, it indicated that many more L1s, even within the genomes that were screened, are yet to be found. Rare L1 insertions are just that: rare. In order to find these young, and possibly *de novo* insertions with high-throughput methods, scientists need to sequence through the hundreds known to be present in every genome. In order to do so, applying high throughput sequencing technologies such as Helicos, ABI SOLiD, Roche 454 FLX/Titanium, and Illumina Genome Analyzer, are the likely future for retrotransposon discovery

Methods

ABI Capillary Sequencing

DNA preparation

Linker-mediated PCR methods were adapted from a previous study [136]. Genomic DNA samples were pooled. 165 ng of each pool was digested with HpyCH4IV (NEB) or MspI (NEB) in 20 uL reactions for 4 hours and then heat inactivated at 65C for 20 min. Water was digested and carried through the assay as a negative control. The top and bottom strands of the linker were annealed by combining them in equal molar ratios and using the following cycling conditions: 95C 3 min, 95 C 40 sec, -1C / cycle for 90 cycles. When annealed, the linker is partially double-stranded with a 5’ CG overhang that is compatible with HpyCH4IV and MspI digested DNA. Oligo sequences are available in Table 3-2. The entire digest was ligated to 40-fold molar excess of prepared linker in 50 uL reactions with T4 DNA ligase (NEB). The reactions were incubated overnight at 15C in a thermocycler. The ligations were heat inactivated at 65C for 20 min. Excess linker was removed with the QIAquick PCR Purification Kit. Samples were eluted in 50 uL water.

Linker-mediated PCR

3 uL of the purified ligation was used as a template in PCR. DNA was amplified with 1 unit of Platinum Taq (Invitrogen), 1X of the manufacturer's provided reaction buffer, 1.5 mM $MgCl_2$, 0.2 mM dNTPs (Invitrogen), 0.4 uM each of L1Ta primer and Linker Out primer (Table 3-2) in a total reaction volume of 50 uL. To prevent random amplification from the linker-specific primer, the bottom strand of the linker is modified to have a 3'-amine group (Figure 3-5). The following cycling conditions were used: 96C 2 min initial denaturation, followed by 12 cycles of 96C 30 sec, 60C 1.5 min, 72C 1.5 min, and then a final extension of 72C for 3 min. 1-5 uL of the above reactions was used as a template in a second, nested PCR. L1 Nested and Linker In primers (Table 3-2) were used in 50 uL total reaction volume. The following cycling conditions were used: 96C 2 min initial denaturation, followed by 30 cycles of 96C 30 sec, 60C 30 sec, and 72C 1 min, and then a final extension of 72C for 5 min. In order to have ample colonies to pick for sequencing, each nested PCR reaction was done in triplicate and like reactions were combined. For quality control, 10 uL of each PCR product was separated on an agarose gel. All samples had diffuse smearing. There was less amplification from chimp. And the negative control lanes were empty (data not shown).

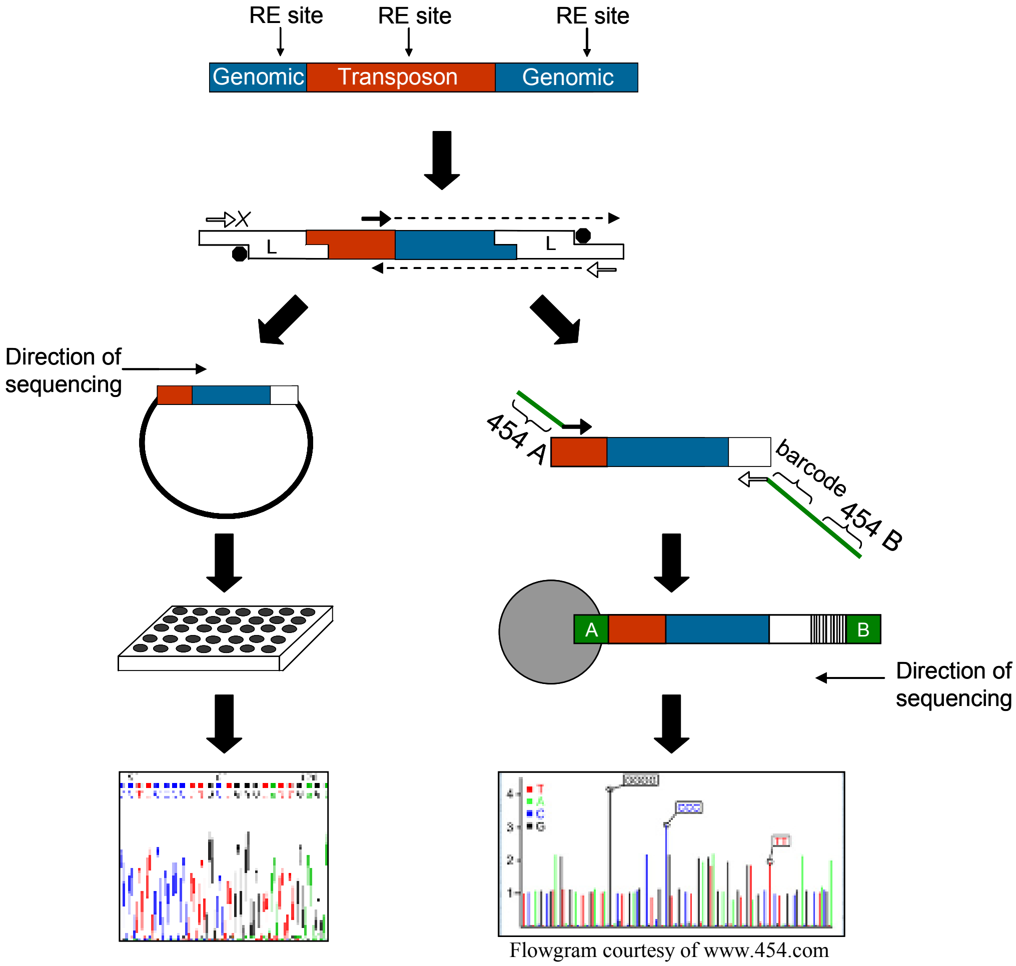
Primer Name	**Sequence**	**Special Features**
Linker Top	GGAAGCTTGACATTCTGGATCGATCG**CTGCAG**GGTATACGCGTCGACAAC	PstI site
Linker Bottom	CGGTTGTCGACG	5' phosphate, 3' amine
L1Ta	ATACCTAATGCTAGATGACACA	L1Ta "ACA" at 3' end
Linker Out	AGCTTGACATTCTGGATCGATC	none
L1 nested	ACCAA**GCGGCCGC**CATGGCACATGTATACATATGTAACTAACCTGCACAATGTG	NotI site
Linker In	GCAGGGTATACGCGTCGACAAC	none
TA Linker Bottom	TAGTTGTCGACG	5' phosphate, 3' amine
L1 nested 454A	**GCCTCCCTCGCGCCATCAG**CATATGTAACTAACCTGCACAATGTG	454A adapter
Linker In B4	**GCCTTGCCAGCCCGCTCAG****AACCAACC**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B5	**GCCTTGCCAGCCCGCTCAG****AACCGCAT**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B6	**GCCTTGCCAGCCCGCTCAG****AACGAACG**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B7	**GCCTTGCCAGCCCGCTCAG****AACGAAGC**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B8	**GCCTTGCCAGCCCGCTCAG****AACGGCTT**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B9	**GCCTTGCCAGCCCGCTCAG****AACGTTGC**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B10	**GCCTTGCCAGCCCGCTCAG****AATTCCGG**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B11	**GCCTTGCCAGCCCGCTCAG****ACACGACT**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B12	**GCCTTGCCAGCCCGCTCAG****ACACTCAG**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B13	**GCCTTGCCAGCCCGCTCAG****ACAGGACA**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B14	**GCCTTGCCAGCCCGCTCAG****ACCAACCA**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B15	**GCCTTGCCAGCCCGCTCAG****ACCACTAG**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B16	**GCCTTGCCAGCCCGCTCAG****ACCATCGA**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B17	**GCCTTGCCAGCCCGCTCAG****ACCTCATC**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B18	**GCCTTGCCAGCCCGCTCAG****ACCTCTTG**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B19	**GCCTTGCCAGCCCGCTCAG****ACCTTGCT**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B20	**GCCTTGCCAGCCCGCTCAG****ACGACTAC**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B21	**GCCTTGCCAGCCCGCTCAG****ACGAGATG**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B22	**GCCTTGCCAGCCCGCTCAG****ACGTAGGA**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
Linker In B23	**GCCTTGCCAGCCCGCTCAG****ACGTTCCT**GCAGGGTATACGCGTCGACAAC	454B adapter, 8 bp barcode
T3	AATTAACCCTCACTAAAGGG	Universal Primer
T7	GTAATACGACTCACTATAGGGC	Universal Primer

Table 3-2

Oligos used in ligation-mediated PCR. Oligos used in this experiment. All oligos, except for Linker Bottom, were purchased from Invitrogen, desalted, and kept in 20 uM stocks. Linker Bottom was purchased from Operon.

Cloning of L1Ta libraries

The second round PCR products described above were purified with the QIAquick PCR Purification Kit. Samples were eluted in 50 uL water. The entire purified PCR product was digested with NotI and PstI (NEB) in 60 uL reaction volumes for 2 hours. The reactions were heat-inactivated at 80C for 20 min. The DNA was cleaned by



Flowgram courtesy of www.454.com

Figure 3-5

Strategy to sequence retrotransposon insertion junctions. Human genomic DNA was digested and ligated to a linker. The linker is partially double-stranded with a 3' amine group on the short strand. This prevents amplification of random genomic DNA from the linker primers. Amplification only occurs if there is extension from the transposon-specific primer. This completes the double-stranded linker and creates the sequence for the linker-specific primer to anneal to. Left side: After amplification, there was a second round of PCR with a nested primers. Second round PCR products were cleaned and cloned using a restriction site in the nested retrotransposon primer and another restriction site embedded in the linker. Colonies were used to inoculate cultures in 96-well plates which were then subjected to standard dideoxy sequencing. Right side: After the initial PCR amplification, there was a second PCR with nested primers. The retrotransposon primer contained the “A” adapter sequence for 454 FLX sequencing while the linker-specific primer contained an 8 bp unique barcode for each sample and the “B” adapter sequence. Samples were then pooled in equal molar ratios for emulsion PCR with beads binding only the “A” adapter and were thus sequenced from the "B" end.

phenol:chloroform extraction and ethanol precipitation. The resulting DNA pellets were resuspended in a 25 uL ligation reaction with T4 DNA Ligase from NEB. The reaction was incubated at room temperature for 1 hour. The vector for the ligation was pBluescript that had been digested with NotI and PstI and then purified by gel extraction. A vector-only ligation was used as a negative control. The ligation reactions were then purified with the QIAquick PCR Purification Kit. Samples were eluted in 30 uL water. 10 uL of the cleaned ligation reaction was electroporated with 5 uL DH10B (Invitrogen) electrocompetent *E. coli*. The cells were then rescued in 200 uL SOC media and incubated at 37C for 30 min. 50-100 uL of these cells were then spread on LB Amp plates that had been treated with 100 uL 10 mg/mL Xgal.

Colony Picking and Sequencing

Colony PCR was performed using several colonies from each plate to ensure that a variety of L1 insertions were represented in the libraries. A homemade Taq prep was used with 10X PCR buffer and $MgCl_2$ from Applied Biosystems. 0.2 mM dNTPs and 0.8 uM of T3 and T7 primers were used in a 25 uL reaction at the following cycling conditions: 94C 2 min initial denaturation, then 35 cycles of 94C 1 min, 58C 1 min, 72C 2 min, and a final extension of 72C 5 min. Colonies from the vector only control ligation plate were used as a negative control. Colonies were used to inoculate 200 uL LB Amp + 15% glycerol in each well of 96-well plates. The plates were loosely wrapped in cellophane to prevent condensation from collecting on the lids and incubated at 37C overnight without shaking. They were then sealed with foil lids and frozen at -80C. The frozen plates were shipped on dry ice overnight for sequencing. In total, 4600 clones from human libraries and 96 clones from a chimp library were sent for sequencing.

PCR Validations

Sixty-six of the 152 novel L1 insertions were chosen for validation. Primers were designed to flank the putative insertion site. The majority of PCRs were performed in a 25 uL reaction with 15 ng template, 1X PCR buffer II (Applied Biosystems), 1 mM $MgCl_2$ (Applied Biosystems), 0.2 mM dNTPs, 1.2 uM each of the L1Ta primer and downstream primer, and 0.1 uL of a homemade Taq prep. The following cycling conditions were used: 94C 2 min initial denaturation, then 35 cycles of 94C 1 min, 62C for 1 min, 72C 1 min, and a final extension of 72C 10 min. If these conditions did not yield conclusive results, the annealing temperature was adjusted, Invitrogen Taq and its provided buffers were used, and, in a few cases, new primers were designed. PCR results and primers can be found in Appendix I.

Roche 454 FLX pyrosequencing

Samples analyzed

Glioblastoma and medulloblastoma specimens and their matching B-lymphocyte specimens were obtained from the Emory University School of Medicine Tissue Procurement and Banking Service. Samples were frozen by cooling them in a container with isopentane in liquid nitrogen. The samples were then transferred to cryovials and stored in liquid nitrogen. Genomic DNA was extracted from 25 mg of tissue using the DNeasy Tissue Kit (Qiagen). Twenty snap-frozen non-small cell lung cancer specimens

(16 adenocarcinoma, 4 squamous cell carcinoma) and paired adjacent normal tissue were also obtained from the Emory University School of Medicine Tissue Procurement and Banking Service. Genomic DNA was extracted from 15 mg of snap-frozen lung tissues using the DNeasy Tissue Kit.

Linker-mediated PCR

DNA prep and linker-mediated PCR were the same as for the ABI capillary sequencing experiment, but with the following modifications: DNA samples were treated individually and were not pooled as before. 100 ng of DNA was digested with MseI (NEB) in 30 uL reactions to create 5' TA overhangs. The bottom strand of the linker was changed to have a compatible overhang (Table 3-2). Purified ligation products were eluted in 30 uL of water. 8 uL (~25 ng) of the purified ligation was used as template for PCR with 2 units of Platinum Taq polymerase. Cycling and primers for the first PCR are the same as for the ABI capillary sequencing experiment. 2 uL of PCR product was used as a template for the second PCR. Each sample used the L1 nested 454A primer, and different Linker In B primers (Table 3-2). The L1 nested 454A primer contains the A adapter sequence required for emulsion PCR and Roche 454 FLX sequencing. Each Linker In B primer has the B adapter sequence at the 5' end followed by a unique 8 bp barcode and then linker-specific sequence at the 3' end. The barcodes were chosen based on a previous study [137]. For Alu sequences, linker primers were the same as for L1, but Alu-specific primers were used. There were 80 different samples (20 lung tumors with matched normals for L1, 10 brain tumors with matched normals for L1, and 10 brain tumor and matched normals analyzed for Alu insertions) but only 20 different barcode primers so samples were split into groups of 20, each with a unique barcode. Samples in each group were pooled together in equal molar ratios and gel purified to remove small, uninformative PCR products. An aliquot of each pool was cloned into the pCR4-TOPO vector for quality control. Colonies were used for PCR to look for a variety of inserts as before. Each pool was sent to the University of Florida Interdisciplinary Center for Biotechnology Research for emulsion PCR and sequenced in different regions of a Roche 454 FLX plate.

Analysis of Sequences

For both L1 and Alu, fastA sequences were parsed so only the unique genomic sequence was aligned back to the human genome (May 2006 build, Hg18). Barcodes that contained sequencing errors were run through a custom Perl script to determine the most likely barcode. Custom Perl scripts were also used to batch BLAT the sequences and choose a best hit. For L1, each nonredundant Hg18 coordinate was manually inspected to see if an L1 was upstream in the reference sequence. For Alu, sequence flanking the insertion site in Hg18 was used as a query in CAlu, a web-based Alu classifier (available at http://clustbu.cc.emory.edu/calu/index.cgi) to determine whether an Alu was already present in the reference sequence. PCR validations were performed as before except with Invitrogen Taq and its buffers. Genomic coordinates and PCR results are available in Appendices II and III.

Calculating sequencing coverage

For the L1 reads, we wanted to know how well our data set represented the true L1Ta composition of our samples. We chose 63 L1Tas that were absent from dbRIP and previously shown to be fixed in a panel of 80 diverse individuals [93, 95]. For each of these L1s, we plotted the percentage of samples from which the L1 was sequenced as a function of the number of total reads representing the L1 (Figure 3-6). Thus, L1s that were sequenced 60 times were sequenced at 1X coverage (30 individuals with two tissue samples each). Those L1s that were sequenced to at least 4X coverage were all found in >90% of the samples. The average fixed L1 was sequenced with 108 reads or 1.8X coverage and was found in 74% of the samples.

Since DNA samples were pooled before pyrosequencing, we wanted to make sure all samples were uniformly represented. We plotted the number of mapped reads for each sample in ascending order for both L1 and Alu (Figure 3-7). We found that, for L1s, the most represented sample had 2.3 times more mapped reads than the least represented sample. For Alu, there were 1.8 times more mapped reads in the most represented sample. Altogether, we found this data reassuring that our samples were, for the most part, equally represented in the pooled sequencing.

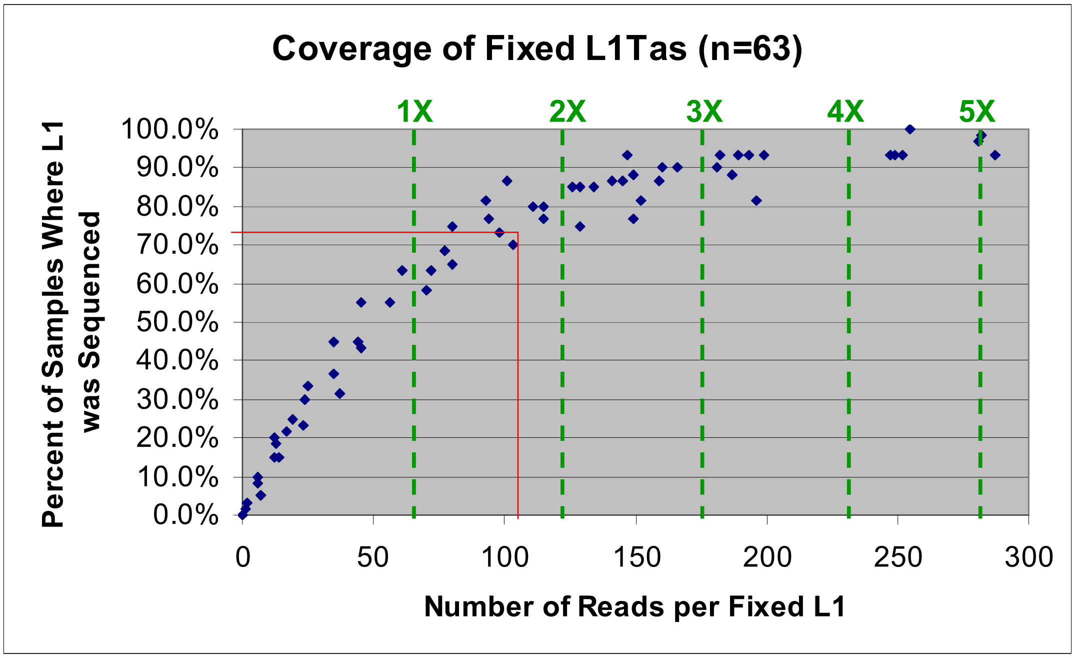


Figure 3-6

Sequencing Coverage for L1. Using a custom Perl script, we predicted which L1Tas in Hg18 we would be able to sequence based on how well they matched our primers and whether specific restriction sites were nearby downstream. Of those we predicted we would sequence, 63 L1Tas were found to be fixed in the population. We used these 63 L1Tas as internal controls to measure our sequencing coverage. Each data point represents one of the 63 fixed L1Tas. We plotted the percentage of the 60 tumor and normal samples in which we sequenced the L1 as a function of the number of sequence reads representing each L1. Since there were 60 samples, 60 reads is 1x, 120 reads is 2x and so on. The average number of reads per fixed L1Ta was 108 and is indicated by a red line. Thus, we discovered ~74% of the L1Tas present in our 60 samples and sequenced to an average of 1.8x coverage.

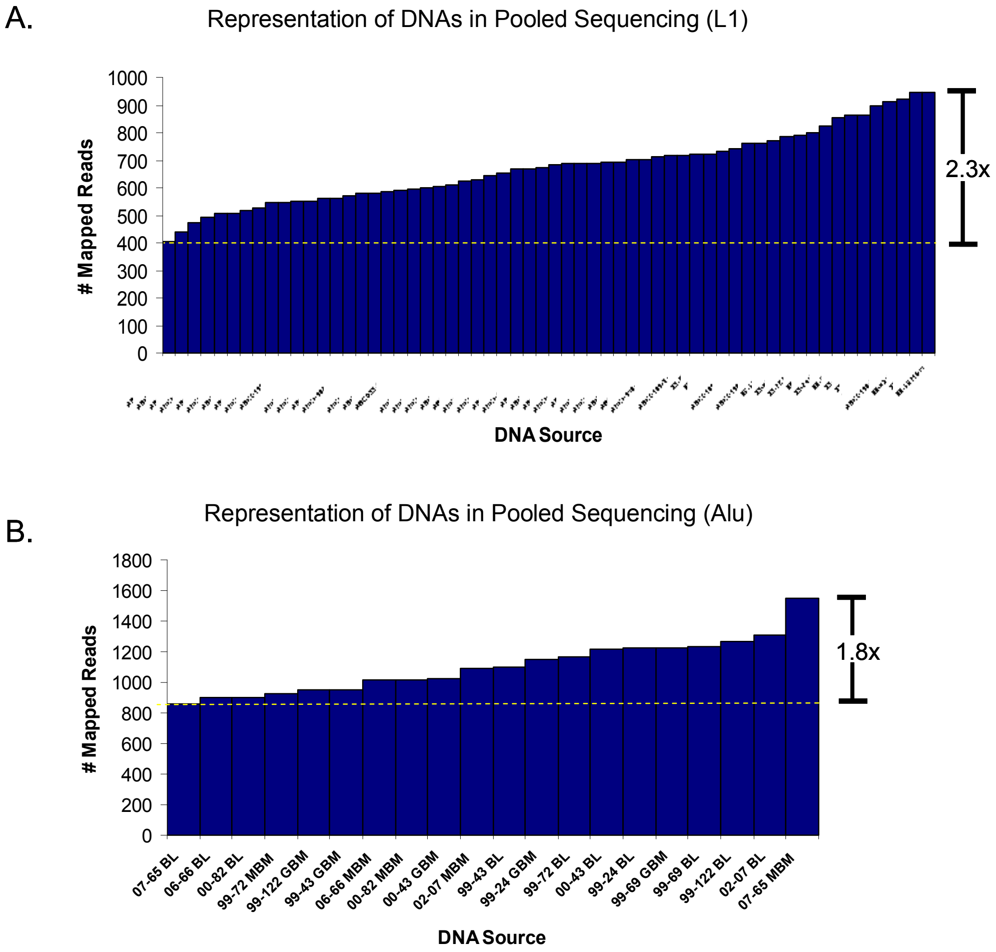


Figure 3-7

Representation of samples. We determined the number of mapped reads from each sample and plotted them is ascending order. For L1, the most represented sample had 2.3 times more mapped reads than the least represented sample. For Alu, the most represented sample had 1.8 times more mapped reads than the least represented sample.

Chapter 4

Frequent Mutagenesis of Lung Tumors by the Human L1 Retrotransposon

Introduction

Only 21 years have passed since the first discovery of an active retrotransposon in humans [36]. Since then, geneticists have identified dozens of disease-causing insertions and hundreds of polymorphisms [Reviewed in 5 and 33]. Thus far, all but a handful of these insertions appear to have occurred in the germline. However, a growing body of literature indicates retrotransposition can occur in other cell types as well [69, 70, 72, 73, 138-140]. Coufal and colleagues recently found that the human brain likely accommodates retrotransposition [138]. The authors found that neural progenitor cells (NPCs) derived from human fetal brains supported retrotransposition of a marked L1. Thus, brain tissue (and potentially other tissues) are likely mosaic for L1 insertions. Recently, Haig Kazazian's laboratory found that the majority of retrotransposon insertions occur during early embryogenesis with only a small fraction entering the germline [70]. Therefore, most retrotransposon insertions are not heritable, but still may be affecting human health. Haig Kazazian's laboratory also demonstrated that the majority of adult human tissues express low levels of L1 [75]. Together, these studies indicate that retrotransposition is likely frequent in somatic tissues.

The recent revelation that human retrotransposons are active in tissues other than the germline leads to the possibility that they may play a role in causing cancer. Retrotransposons are essentially endogenous mutagens that can disrupt gene function directly or indirectly. For example, an acquired insertion disrupted the APC tumor suppressor gene in a patient's colon cancer [42]. The inactivation of this gene likely contributed to the patient's tumorigenesis. Additionally, L1s can induce genomic instability in a cell-based assay [26, 86]. L1s often create deletions of host DNA at insertion sites [134]. Cells that express L1 elements incorporate more H2AX histones into their nucleosomes than cells that do not. H2AX incorporation is a sign of DNA damage and indirectly supports the hypothesis of retrotransposons being mobilized in somatic tissues [141]. Additional young L1s of low sequence divergence can act as instigating sites for nonallelic homologous recombination [142]. These, and other mechanisms, contribute to genomic instability, a hallmark of most types of cancer.

At this point, the idea of active retrotransposons contributing to tumor progression is purely speculative. Even so, implicit evidence supports this possibility. Hypomethylation of retrotransposons occurs in many different tumor types [e.g. 78, 88, 143]. Some of these studies show that loss of methylation is followed by reanimation of transposable elements at the transcriptional level [79, 82, 144]. In fact, one of the earliest studies of human retrotransposons showed expression of potentially retrotransposition-competent elements in tumor cells [135]. Also, expression of retrotransposons in tumors can be an indicator of genomic instability and poor prognosis [32, 79, 82, 83, 88, 144, 145]. All of these observations support the hypothesis that retrotransposons may be responsible for some of the acquired mutations in tumors.

In fact, somatic transposition occurs in non-human model systems. For example, Barbara McClintock discovered transposons (what she called "mutable loci") from the observation of somatic pigment changes in maize [3]. Also, spots found on drosophila wings result from the activation of P-elements in somatic cells [146]. Somatic transposition also occurs endogenously in C. elegans [147] and Arabidopsis [148]. In particular, somatic retrotransposition appears to be common in the murine lineage [69, 72, 74, 149]. A Dnmt1 *de novo* methyltransferase knockout mouse model was prone to spontaneous thymic lymphomas. The mice had independent somatic insertions of a retrotransposon in 7 of 16 tumors [88]. These insertions created an oncogenic spliceoform of the Notch gene which further contributed to tumor progression. These observations increase the likelihood that humans are not immune to somatic retrotransposition.

So far, the somatic insertion in the APC gene described above is the only example of a tumor-specific retrotransposition in humans. Thus, it is unclear whether this retrotransposition event is an isolated incident or an example of a more common mechanism of mutagenesis.

In this study, we sought to directly test whether retrotransposition occurs in human tumors by sequencing transposon insertion junctions in tumor and matched normal tissues. We used linker-mediated PCR to sequence tens of thousands of L1 and Alu insertion junctions. In doing so, we found hundreds of previously unknown retrotransposons some of which were somatic insertions.

Results

Potential somatic insertion in a lung tumor-derived cell line

Initially, 4600 clones were sequenced from L1 insertion junction libraries. These libraries were derived from 38 B-lymphocyte-derived cell lines from healthy individuals and 8 tumor-derived cell lines. The 8 tumor-derived cell line genomic DNAs were split into 3 pools.

Four breast cancer-derived cell lines, three leukemia-derived cell lines and a single lung tumor-derived cell line were screened. Of 152 novel L1s found (those absent from public databases), 20 were observed in only one of the tumor-derived DNA pools (Table 4-1). Thus, candidates for cancer-specific insertions were identified. Fifteen of these insertions were chosen for PCR validation because of nearby unique sequence where primers could anneal. Of these 15, one could not be verified and 12 were present in at least one of the DNAs from the pool it came from, but were also present in healthy diverse human DNAs. Thus, these insertions were common, inherited polymorphisms and not acquired.

Interestingly, one of the L1s that was initially sequenced in a lung tumor-derived cell line appeared to show loss of heterozygosity (LOH). B-lymphocyte-derived DNA from the same individual was heterozygous for the L1 insertion (located at 3p12.1) while the tumor-derived cell line only had the allele with the L1 (Figure 4-1). Previous studies using microsatellite markers and spectral karyotyping with this particular pair of lung tumor and matched normal DNA also showed significant losses of chromosome 3 [150, 151]. LOH of 3p is typical in lung tumors [152]. This L1 insertion had an allelic frequency of 0.378 in a panel of 41 diverse individuals. In theory, this L1 could be a useful marker in determining LOH.

After PCR validation, two L1s appeared to be cancer-specific. One of the two potentially cancer-specific insertions came from a breast cancer-derived cell line. This particular L1 was found only in a single sample (Figure 4-1). We could not be sure if this was a germline or somatic mutation since no matched normal DNA was available. The second insertion was found in a lung tumor-derived cell line. The insertion was not present in any other genomic DNA screened, including B-lymphocyte-derived DNA from the same individual (Figure 4-1). The L1 insertion was in the last intron of the Ras-like

GTPase-encoding RIT2 gene. This L1 element could have retrotransposed during tumorigenesis or during early embryogenesis. Haig Kazazian's laboratory previously demonstrated that the majority of human retrotransposition events occur in early embryogenesis and cause individuals to be mosaic for such insertions [70]. Since we only had B-lymphocyte-derived DNA from the same individual as the lung-tumor-derived DNA, we could not distinguish whether the somatic insertion we identified was embryonic or acquired.

Insertion Site	DNA pool	Found in non-cancer DNAs during validations?	Gene Abbreviation	Where L1 is relative to gene
chr1:187096534-187096535	breast cancer	not validated	FAM5C	1237 kb downstream
chr1:190528883-190528884	breast cancer	yes	RGS21	24 kb upstream
chr3:48361750-48361751	breast cancer	not validated	SPINK8	17 bp upstream
chr10:92720929-92720930	breast cancer	yes	ANKRD1	50 kb upstream
chr11:94317492-94317493	breast cancer	no - found only in single breast cancer	LOC51503	19 kb downstream
chr3:38601070-38601071	leukemia	yes	SCN5A	in intron 13
chr4:46702535-46702536	leukemia	yes	GABRA4	12 kb upstream
chr4:191206809-191206810	leukemia	yes	FRG2	24 kb downstream
chr9:67967174-67967175	leukemia	yes	BC063132	23 kb upstream
chr12:83489268-83489269	leukemia	yes	SLC6A15	288 kb downstream
chr14:62418237-62418238	leukemia	yes	KCNH5	in intron 7
chr16:59636693-59636694	leukemia	could not be validated	CDH8	608 kb downstream
chrX:122492541-122492542	leukemia	not validated	GRIA3	42 kb downstream
chr2:191186952-191186953	lung cancer	not validated	NAB1	35 kb bp upstream
chr3:85659258-85659259	lung cancer	yes	CADM2	in intron 1
chr5:41464296-41464297	lung cancer	yes	PLCXD3	in intron 1
chr11:114400722-114400723	lung cancer	not validated	CADM1	144 kb downstream
chr12:125221187-125221188	lung cancer	yes	AKO57632	187 kb upstream
chr18:38642598-38642599	lung cancer	no - found only in lung cancer	RIT2	in intron 4
chr21:27991047-27991048	lung cancer	yes	BC043580	248 kb upstream

Table 4-1

Potential cancer-specific insertions from ABI capillary seqeuncing. Several novel L1s that were sequenced were found only in cancer DNAs. Many of these were selected for PCR validation in the DNA pools they were originally sequenced as well as all the other DNA pools used in this assay. Insertion site coordinates are relative to Hg18. DNA pool is where the sequenced clone came from. Nearest gene is based on UCSC Gene Prediction tract in the UCSC Genome Browser.

For the potential somatic insertion, the 5’ end of the insertion junction was sequenced for confirmation. The L1 was truncated at the 5' end and 4889 bp long. Interestingly, the L1 element did not have target site duplications flanking it. Therefore, this particular element may have used an atypical mechanism to retrotranspose.

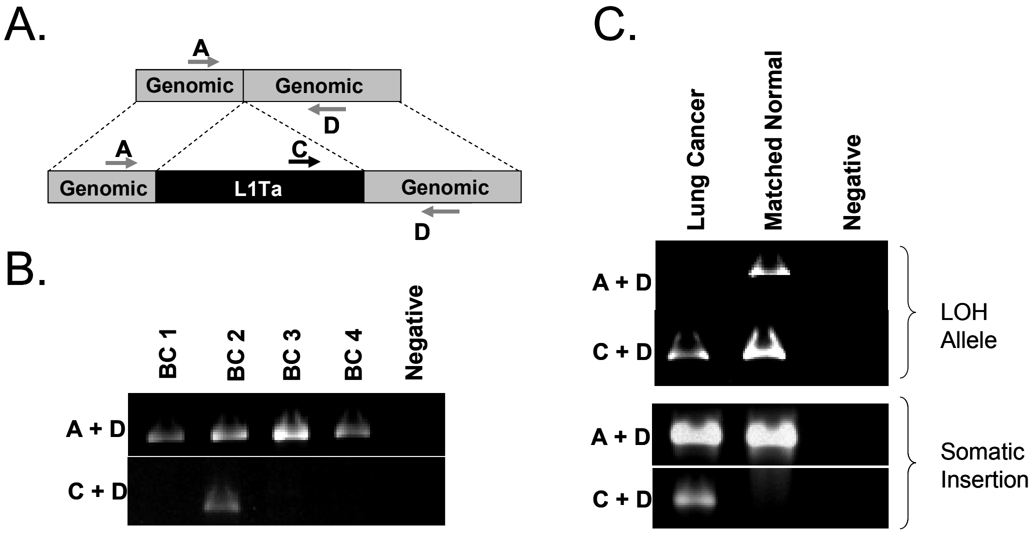


Figure 4-1

Selected validations. A) 64 of the novel L1s were validated by PCR. Primers flanking the putative insertion site (A & D) were used to amplify the preinsertion allele. A primer within the L1 and primers downstream in the genome (C & D) were used to verify the presence of each L1. B) An L1 insertion initially sequenced in pooled breast cancer DNAs was validated in a single breast cancer. C) DNA from a lung tumor-derived cell line, and from the B-lymphocyte derived-cell line from the same individual, was used as templates in PCR.

Somatic Insertions in lung tumor tissues

After sequencing thousands of L1 insertion junctions by standard ABI sequencing, we found a potential somatic insertion in a lung tumor-derived cell line. However, retrotransposition in early embryogenesis or during the culturing of the cell line could not be ruled out. Also, the lack of target site duplications made us question whether this was an endonuclease-independent retrotransposition event or some other form of chromosomal rearrangement. To determine whether retrotransposition occurs in tumors, sequencing to greater depth was performed using Roche 454 FLX pyrosequencing. Genomic DNA from 20 lung tumor cases and adjacent normal lung tissue from the same patients were acquired. Also, since the brain appears to be a permissive environment for retrotransposition [138], genomic DNA from 10 brain tumor cases and matching blood were acquired. L1 insertion junctions were sequenced for the lung tumors and both L1 and Alu insertion junctions were sequenced for the brain tumors.

For both L1 and Alu, hundreds of novel elements (those absent from public databases) were found only in a single sample. These were considered as potential somatic insertions. For L1, there were 232 such insertions (Figure 4-2). There were 221 potential somatic Alu insertions. There were more potential normal-specific L1 insertions in the brain tumor patients than tumor-specific insertions. Lung tumor patients collectively had more potential tumor-specific insertions. Due to low sequencing coverage (1.8-fold, see Chapter 3 for how sequencing coverage was calculated) the majority of these insertions are likely present in both tissues, but happened to only be observed in a single tissue. Thus, 109 potential somatic L1 insertions were chosen for validation by PCR (79 tumor-specific and 30 normal-specific).

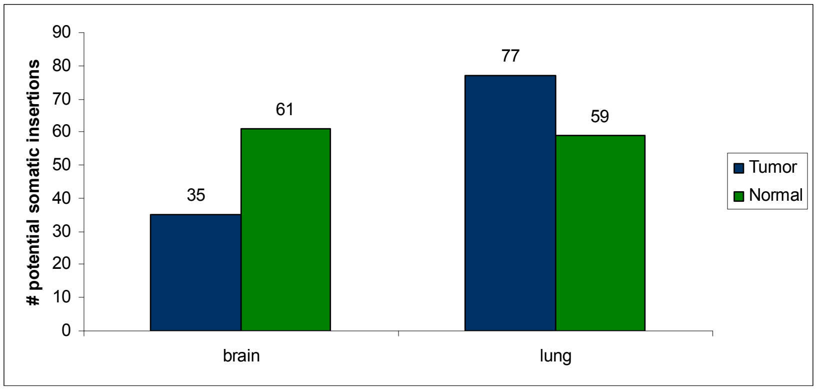


Figure 4-2

Potential somatic L1 insertions from pyrosequencing.

Of the 109 potential somatic L1 insertions, 85 were amplified in both tissues of the individual, 15 could not be verified by PCR and 9 amplified only in a single tissue (Figure 4-3A). For Alus, 56 that were potentially tumor-specific were chosen for PCR validation. Three could not be verified and the remainder were present in the normal tissue or other individuals as well. The 9 L1s that could only be verified in a single tissue were all found in lung tumors. Their presence in tumor tissue and precise absence from adjacent matched normal tissue indicated that these retrotransposon insertions were acquired mutations. No somatic insertions were found in the brain tumors, normal tissues, or with Alus.

We confirmed the somatic insertions by repeating the PCRs several times and sequencing the amplified product. One of the somatic insertions (found in patient 106), could not be confirmed. We suspect this is because we ran out of the original DNA and

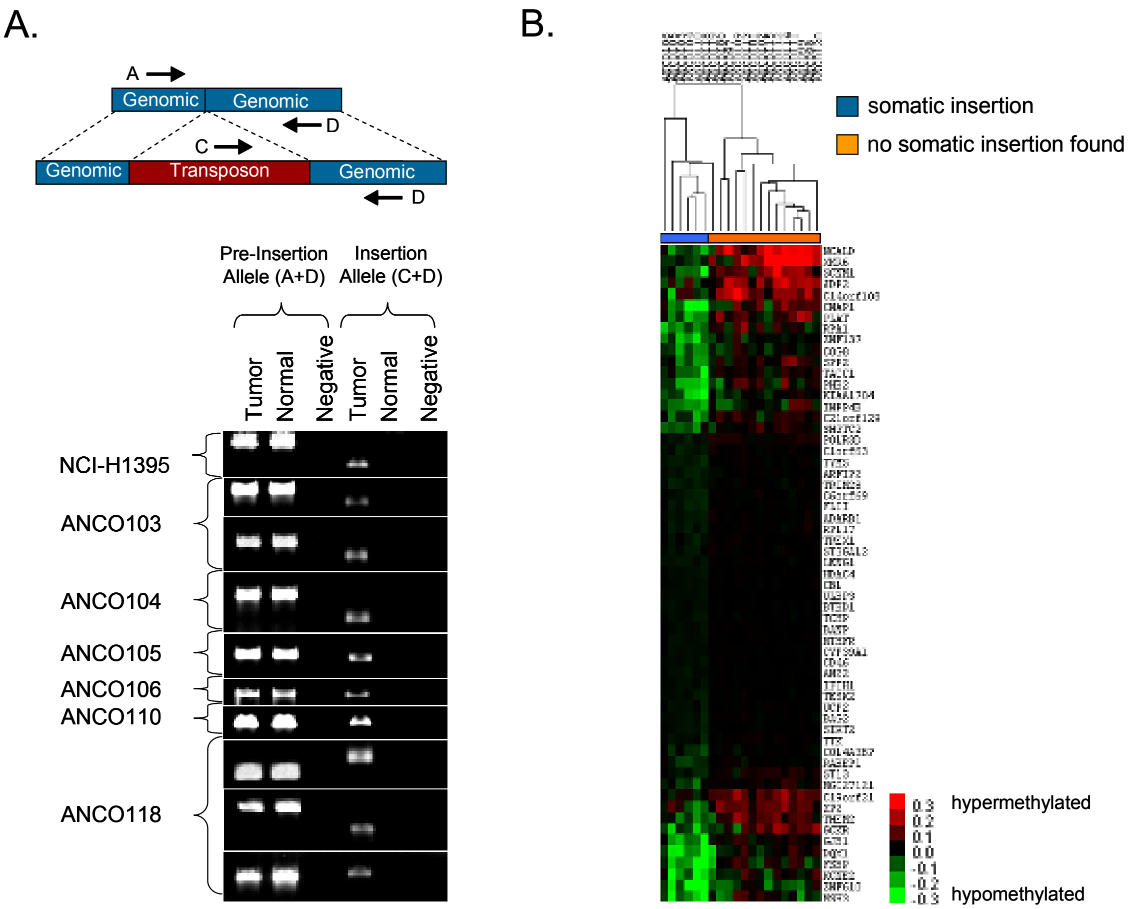


Figure 4-3

PCR validation of somatic insertions & correlation of somatic methylation changes. A) L1 retrotransposons that were somatic candidates based on sequence data were further characterized by PCR validation. Shown here are those verified as somatic insertions by their presence in the tumor tissue and absence from adjacent normal tissue. Negative lanes are control PCRs with no template. Primers flanking the putative insertion site were used to amplify the pre-insertion allele. A primer within the 3' end of an L1 consensus sequence and a primer downstream of the putative insertion sites were used to verify the presence of each L1. Anonymous patient identifiers are on the left. NCI-H1395 is DNA from a tumor-derived cell line and its normal is B-lymphocyte-derived purchased from Coriell. B) The 20 lung tumors used in the pyrosequencing experiment were also analyzed by methylation-specific microarray analysis. Shown here is an unsupervised hierarchical clustering of the samples based on 59 probes whose differential methylation between tumor and normal pairs was correlated with the presence of a somatic retrotransposition event. This revealed a dendrogram consisting of two distinct sample groups.

used a new prep from a different part of the tumor. It may be that this particular lung tumor was mosaic for the L1 insertion.

Three of the 9 insertions were within introns of genes while the other insertions ranged from 33 kb to 859 kb away from the nearest gene. None of the insertions had a predictable effect on gene function or tumorigenesis. One somatic insertion, however, was in intron 27 of the Werner syndrome, RecQ helicase-like gene (WRN). This gene is involved in DNA repair and, when mutated, can lead to a rare form of adult progeria. The individual with this insertion also had 2 other somatic insertions. This lead us to speculate that the WRN gene function was altered by the somatic insertion which, in turn, allowed for further genomic instability.

The 9 somatic insertions occurred in 6 different lung tumors (4 squamous cell carcinomas and 2 adenocarcinomas). We wanted to know whether methylation played a role in driving somatic retrotransposition in our lung tumor samples. Since genomic DNA methylation is altered in lung tumors and is thought to play a role in suppressing retrotransposon mobilization, the methylation status of >14,000 gene promoters was examined in the lung tumors and their paired adjacent normal tissues using the Illumina Infinium platform. Dr. Michael McCabe analyzed the methylation data and performed the analyses described below.

In order to study the relationship between DNA methylation and somatic retrotransposition, patient samples were divided into those in which somatic insertions were identified in the tumor sample and those in which no tumor-specific insertions were identified. Based on this classification, 59 probes were identified whose methylation status significantly correlated with somatic retrotransposition potential (randomly permuted datasets averaged only 1.5 significantly correlated probes). Unsupervised hierarchical clustering of the samples based on these 59 probes revealed a dendrogram consisting of two distinct sample groups (Figure 4-3). As expected all six patient samples exhibiting somatic insertions were clustered within one branch. However, only 13 of 14 samples with no detected insertions clustered within the other branch. While no insertions were detected in patient 119, this sample was more related to the insertion group based upon the methylation status of this subset of probes. This observation suggests that an existing insertion may either have not been detected due to: (1) the modest sequencing coverage, (2) the L1(s) fell within a region of the genome that could not be unambiguously mapped or, (3) this patient's tumor exhibited a permissive environment which had not yet fully mobilized an L1 element. Interestingly, all of the correlated probes were hypomethylated to varying degrees in the tumors exhibiting insertions relative to their normal tissue. These data reveal a methylation signature that distinguishes L1-permissive lung tumors from non-L1 permissive tumors. One possible interpretation of these data is that the hypomethylation signature associated with L1-permissive cells represents a broader genomic methylation state that releases L1 elements from mobilization constraint.

Discussion

L1s May Contribute to Genomic Instability in Tumors

We identified 10 somatic insertions in lung tumors (9 in tumor tissue and 1 in a tumor-derived cell line). These insertions are evidence of a common mechanism of mutagenesis in cancer genomes. Retrotransposons, particularly those that are young, can cause deletions and rearrangements upon integration [86], can alter gene function by influencing the sequence that is included in mature mRNA [Reviewed in 51], and can instigate ectopic recombination events [38]. Thus, these elements have the potential to play a role in tumor-specific genomic instability even after they have retrotransposed.

One of the L1s was found in a single lung tumor-derived cell line and not in any other DNA tested, including matched normal-derived DNA (B lymphocyte) from the same individual. This is likely an example of retrotransposition as an acquired mutation in a tumor, however, retrotransposition in early embryogenesis cannot be ruled out. Ideally, we would have used adjacent matched normal tissue instead of B lymphocyte matched normal to demonstrate that the insertion was acquired, however, none was available. Also, it cannot be ruled out whether retrotransposition occurred during culturing of the cell line. Even so, a previous study that included the same lung cancer cell line and its primary tumor when studying LOH concluded that the cell line faithfully represented the tumor [150].

The lack of target site duplications indicates that this retrotransposon used an endonuclease-independent pathway [153]. Recently, Mark Batzer's laboratory showed that retrotransposons, on occasion, use double strand breaks (DSBs) in the genome as insertion sites without the action of their endonuclease [153, 154] Often times, these insertions use microhomologies at the DSB to prime reverse transcription [153, 154]. In this case, the microhomology may have been a string of four thymines precisely at the insertion site. The thymines likely annealed to the L1 element’s poly-A tail, priming reverse transcription from the 3’ end of the L1 RNA. While this insertion demonstrates that L1s are expressed and capable of retrotransposition in somatic tissues, it is not an example of classical retrotransposition.

Of course, a single retrotransposition event in a single lung tumor-derived cell line is supportive, but not demonstrative, of retrotransposition being a common source of mutation in cancer. Therefore, we set out to find more such somatic retrotransposition events in 20 lung and 10 brain tumors.

Of the 20 lung tumors sampled by Roche 454 FLX pyrosequencing, 6 (30%) had at least 1 somatic insertion. Only 1 other published incident of an acquired retrotransposition in a human tumor exists [42]. Until this point, it was unclear whether this was a lone occurrence or an example of a more common form of mutation in tumors. Now, with this current analysis, we can say that tumor-specific retrotransposition does, in fact, occur frequently at least in lung tumors.

We did not detect any somatic retrotranspositions in the brain tumors. It is possible that our low sequencing coverage prevented the detection of such events. A recent study demonstrated that L1 elements are likely active in the developing human brain [138]. Since we had B-lymphocyte matched normal for the brain tumor patients, we would not have been able to distinguish between brain tumor-specific insertions and those that occurred during fetal brain development. Our inability to observe developmental retrotransposition events in the brain tissues is likely due to their rarity (8-12 events per 100,000 cells) [138]. Even so, we hope to follow up with deeper sequencing in order to confirm the suspicions of L1 jumping in developing brains.

Interesting, we did not observe any somatic retrotransposition events caused by Alu insertions. Since there are far more AluYa5 and AluYa8 subfamily members per genome than L1Ta (2500 compared with 560), there was even less sequencing coverage for Alus. Based on the number of mapped reads alone, there was 0.4 fold coverage per haploid genome. It is no surprise then that all of the 53 "singleton" Alus that could be validated were also observed in additional samples. The possibility of Alus jumping in human tumors cannot be excluded by these data.

Since retrotransposons are hypomethylated in many tumors, we hypothesized that specific changes in methylation could be associated with retrotransposition in lung tumor samples. Using data from the Illumina Infinium platform to assess methylation changes within genes, 59 probes whose methylation status could be correlated with whether the tumor had a somatic retrotransposition were identified. After performing hierarchical clustering based the methylation changes of these 59 probes, there was 1 lung tumor for which no somatic retrotransposition event was observed, but nevertheless, clustered with more those lung tumors with a somatic retrotransposition event. These data suggest that this tumor may have had a permissive environment for retrotransposition, but low sequencing coverage prevented its discovery or no such event had yet occurred. Also, if the retrotransposition event could not be mapped to the human reference sequence, it would have been omitted from subsequent analyses. When this tumor is grouped in with the 6 tumors that supported retrotranspositions, more probes (139) become correlated. Altogether, this analysis identified a potential methylation signature for lung tumors that support retrotransposition.

Two of the correlated genes were possible regulators of retrotransposition. RPA1 is a single-stranded DNA binding protein and is involved with the replication of some retroviruses [155]. TREX1 is also a single-stranded DNA binding protein.

Overexpression of TREX1 has been shown to prevent the integration of endogenous retroviruses in mouse [156], however, knockdown of TREX1 inhibits integration of HIV [157]. Therefore, TREX1 both inhibits and promotes retroviral integration. Its complex relationship with retroviruses may be contingent upon the family of retrovirus and the level of viral expression. This leads us to speculate that the loss of methylation of both RPA and TREX1 could potentially regulate L1 retrotransposition in lung tumors, however, since only one probe for each gene was correlated from our analysis, it seems more likely that hypomethylation of individual L1 elements is primarily responsible for their reanimation.

Retrotransposons were long suspected to play a role in tumorigenesis, but whether they were responsible for some of the acquired mutations in human tumors was unclear. Now, we can say for certain that retrotransposons are capable of jumping in lung tumors. Even so, we cannot tease apart whether retrotransposition is a byproduct of or a contributor to genomic instability. The acquired retrotransposition events identified in this study have no obvious consequences on gene functions, but may further genomic instability by a variety of mechanisms. Whether retrotransposition contributes to tumorigenesis remains unknown, but it is a mechanism of mutation in lung tumors.

Methods

Linker-mediated PCR, DNA sequencing, sequence analysis and PCR validations were performed as described in Chapter 3.

Methylation analysis

Genomic DNA was extracted from 15 mg of snap-frozen lung tissues using the DNeasy Tissue Kit (Qiagen). Genomic DNA (1 ug) was bisulfite-modified with the EZ DNA Methylation-Gold kit (Zymo Research) according to manufacturer's recommendations. Genome-wide DNA methylation profiling was then performed by the Emory University Biomarker Service Center using the Illumina Infinium HumanMethylation27 v1.0 platform. This platform assesses the methylation status of 27,578 CpG dinucleotides located within 14,475 RefSeq genes.

Probes whose change in methylation status correlated with somatic retrotransposition potential were identified using the Quantitative Response feature within the Significance Analysis of Microarrays (SAM) software package. This approach calculates a score based on the linear regression coefficient of each probe on the qualitative transposition potential score and calculates the significance and false discovery rate based on 100 random permutations of the dataset. After calculating the score of each probe within the dataset, the delta statistic was adjusted to obtain a false discovery rate of ~1%. Data for significantly correlated genes was extracted, hierarchically clustered using Cluster 3.0, and visualized with Java Treeview.

Chapter 5

Discussion

Discussion

The human genome is surprisingly malleable. Genetic variation helps shape individuals, giving them unique phenotypes and affecting their health. Major international efforts such as the HapMap Project seek to understand the consequences of human genetic variation [158]. These efforts, however, are primarily focused on variation at the single nucleotide level. Recently, the genomics community has come to realize the importance of other, larger forms of genetic variation including indels [159], structural variants [118, 119], and—the focus of this work—transposons [5]. Transposons were long considered "selfish" or "junk" DNA [1, 2], but geneticists now know they play a major role in altering the landscape of human genomes.

We sought to understand the role of retrotransposons in human and chimpanzee divergence (Chapter 2). We identified nearly 11,000 transposable elements that were differentially present between these two species (Table 2-1). Human has more lineage-specific transposons than chimpanzee and that this is likely due to the lack of active L1s in chimp (Table 2-2). We also found that the nonrandom genomic distribution of these elements indicates that some regions of the genome are under purifying selection and exclude retrotransposon insertions. Over one-third of differentially present transposons fell within genes (Table 2-3) and some of these insertions are bound to affect the function of those genes. Some of the phenotypic differences between human and chimp are likely due to retrotransposon insertions.

We also wanted to know whether ample rare retrotransposon alleles exist in human populations as published estimates of retrotransposition rate would predict (Chapter 3). L1 insertion junctions were selectively sequenced and nearly 1,000 novel

Alu and L1 elements were identified in 76 individuals. We more than quadrupled the number of known L1s that are outside of the human reference sequence. Novel L1s are enriched for low frequency (<5%) alleles and this is unlike L1s in the Database for Retrotransposon Insertion Polymorphisms (dbRIP).

Finally, we wanted to know whether L1 mutagenesis occurs in tumors (Chapter 4). Anecdotal evidence suggests that L1s are active in tumors, however, only one example has been published to date [42]. Massively parallel pyrosequencing of tumor and matched normal tissues was used to find candidates for somatic insertions. PCR was then used to verify their presence in tumors and precise absence from matched normal tissues. Ten somatic L1 insertions (nine from lung tumor tissues and one from a lung tumor-derived cell line) were identified. Six out of twenty (30%) lung tumors tested had at least one somatic insertion. Thus, retrotransposition is a common occurrence in lung tumors and likely contributes to genomic instability and tumor progression. We also identified a methylation signature for lung tumors that support retrotransposition.

Over time, the human genome has accumulated approximately 500,000 L1 elements. These successful parasites continue to mutagenize human genomes and alter their evolutionary potential. It is the author's hope that this work contributes to an understanding of the human L1 retrotransposon so that scientists can better predict, prevent and treat human diseases caused by retrotransposons as well as better understand the evolutionary history of humans.

Active L1s contribute to species divergence

We wanted to know what role transposons have in human evolution. In order to understand the genetic changes that led to uniquely human traits, we needed to know the differences between the human genome and that of its closest living relative, the chimpanzee. This would provide a snapshot of divergence that has occurred since human and chimp last shared a common ancestor, around 6 million years ago (MYA).

When the chimp genome was published, scientists realized how similar these two species are at the sequence level [9]. Nucleotide substitutions occur at a rate of 1.23% between human and chimp. Therefore, changes in protein-coding regions of genes are unlikely to be solely responsible for phenotypic differences between humans and chimps. Other differences such as epigenetic marks, gross chromosomal rearrangements and indels likely play a role in creating species-specific traits. We wanted to identify transposons that are differentially present between humans and chimps to understand more clearly what makes one human.

We used available sequence data from the Human Genome Project and Chimpanzee Sequencing and Analysis Consortium to perform whole genome alignments. A computational pipeline was then used to filter for lineage-specific transpositions. Fewer chimp-specific L1s were found than human-specific L1s, but on the whole, these numbers were comparable (Table 2-1). Non-autonomous retrotransposons, on the other hand, have been much more successful in the human lineage. This is in agreement with other studies [9, 160].

Active retrotransposons are near extinction in chimps. Alus and SVAs rely upon L1-encoded proteins in order to retrotranspose. Thus, a decline in active L1 copy number affects Alu and SVA retrotransposition potential as well. The chimp-specific L1 elements identified in our study were far less likely to be full length than their human counterparts. In fact, only four chimp-specific L1 elements were identified that were full length. Using high quality BAC sequence data for chimps as well as an ORF-trapping genetic screen, it was found that chimps have very few L1s with intact ORFs (Table 2-2). Even those with intact ORFs have diverged substantially from active elements and are unlikely to be active. Furthermore, many of the human-specific L1s fall into novel subfamilies while no predominant subfamily is active in chimp (Figure 2-2). Thus, chimp specific L1 activity occurs primarily from the remnants of once-active subfamilies. Without a new subfamily to spawn offspring, retrotransposition will continue to decline as chimp evolution continues.

The lack of potentially active L1s in chimps highlights the difference in evolutionary history between humans and chimps. Chimps may be under different selective pressures that require clearing out full length L1 elements before they reach fixation. Or perhaps, chimp L1-encoded reverse transcriptase is less processive than that of humans. Another possibility is that chimp host factors are more efficient at preventing integration of full length L1 elements. Demographic factors also may have led to the differences in the number of species-specific insertions. Even so, the lack of intact L1s supports the hypothesis that retrotransposition is on the decline in chimp. Whatever the case may be, retrotransposons continue to thrive in humans while they struggle for a foothold in chimpanzees. Retrotransposons could once again thrive in chimp genomes if a new subfamily could produce offspring that evade repressive host factors or are more efficient at retrotransposition. Such waves of retrotransposition activity were recurrent during primate evolution [161]. As one retrotransposon subfamily died out, another subfamily would take over. Perhaps chimps L1s are currently in a lull until a new subfamily can emerge.

Retrotransposition has the potential for altering genes and causing species-specific phenotypes. As described in Chapter 1, retrotransposons can alter phenotypes by directly disrupting genes or by post integration mechanisms. In our study, 3,632 differentially present retrotransposons within genes were identified. Some of these insertions were within exons and may alter the function of those genes (Table 2-3). A non-random distribution of lineage-specific retrotransposons relative to genes was also found. The majority of genes had no insertions while some genes had as many as ten. This implies that some genes likely exclude retrotransposons through purifying selection.

We wanted to know if the types of genes with lineage-specific insertions were different between humans and chimps. Although, a similar fraction of lineage-specific insertions were within genes (~34%), they could be in different sets of genes and may play a role in creating phenotypic differences. We wanted to determine whether genes with specific functions were enriched or depleted for retrotransposon insertions (Table 5-1) [162]. Despite the fact that these retrotransposons have been jumping independently in these two species for the past 6 million years, several gene functions are enriched or depleted for insertions in both species (Table 5-1). Thus, for these gene functions, similar evolutionary pressures likely took place in both human and chimp to either tolerate or reject retrotransposon insertions. Some gene functions were enriched or depleted only in one of the species. For these gene functions, different evolutionary pressures may have influenced which lineage-specific retrotransposons were allowed to reach fixation.

Chimp	Human	Both
Cell proliferation and differentiation	Cell motility	Cation transport
	Chemosensory perception	Cell adhesion
	General vesicle transport	Cell adhesion-mediated signaling
	Homeostasis	Cell communication
	Intracellular protein traffic	Cell structure and motility
	Lipid, fatty acid and steroid metabolism	Developmental processes
	Mesoderm development	Ectoderm development
	Metabolism of cyclic nucleotides	Intracellular signaling cascade
	Neurotransmitter release	Ion transport
	Olfaction	Nerve-nerve synaptic transmission
	Phosphate metabolism	Neurogenesis
	Protein metabolism and modification	Neuronal activities
		Other intracellular signaling cascade
		Other neuronal activity
		Protein biosynthesis
		Protein modification
		Protein phosphorylation
		Receptor protein tyrosine kinase signaling pathway
		Signal transduction
		Synaptic transmission
		Transport

Table 5-1

Enrichment of Gene Functions with Lineage-Specific Retrotransposon Insertions. We wanted to know whether lineage-specific retrotransposons were within different types of genes in humans and chimps. We used the PANTHER classification system [162]. The program uses Gene Ontology classifications to determine the number of genes of a particular function one would expect to have when given a list of genes. This table has gene functions that were shown to be enriched or depleted with a P-value of 0.05 or less. Gene functions in black are enriched for lineage-specific retrotransposon insertions. Gene functions in red are depleted.

We speculate that differentially present retrotransposons have the potential to help drive speciation. As described in Chapter 1, retrotransposons can mediate nonallelic homologous recombination (NAHR) which leads to gross chromosomal rearrangements. Speciation can be driven by reproductive isolation caused by differences in chromosomal structure. During meiotic crossing over, regions of homology that are within chromosomal inversions can recombine and lead to nonviable gametes. In fact, the majority of chromosomal differences between primate species are inversions [163]. Since human and chimp divergence began, retrotransposons have been responsible for the creation of forty-nine lineage-specific inversions [142]. L1s and Alus of younger subfamilies have higher sequence identity to each other and are more likely to be involved in chromosomal rearrangements [142]. Thus, differences in the number of lineage-specific retrotransposons creates differences in evolutionary potential and may contribute to reproductive isolation. Retrotransposons continue to be a driving force in human genome evolution while their influence in chimpanzee is less potent.

L1 retrotransposition is a major source of genetic variation in human populations.
Young retrotransposons are not as well-characterized as single nucleotide polymorphisms (SNPs). The HapMap project's goal was to create a set of well-correlated SNPs that could be used in genome-wide association studies (GWAS) [158]. The second generation HapMap contained over three million SNPs. No similar grand international effort has been dedicated to identifying retrotransposon polymorphisms. The HapMap shed light on the frequency and evolution of SNPs. Outside of ENCODE regions, known SNPs had uniform allelic frequencies. SNPs within ENCODE regions were deeply sampled and had

an excess of rare alleles, 10% of which were singletons [158]. Thus, deep sampling allowed for the discovery of rare variants. No such deep sampling has been done on retrotransposons. In fact, the allelic frequency spectrum of L1s in dbRIP is mostly uniform with a peak of high frequency (>0.9) alleles (Figure 3-2) [93]. This skew toward common alleles is a reflection of ascertainment bias and not the population dynamics of retrotransposons.

We sought to further our knowledge of L1 biology by selectively sequencing the insertion junctions of L1Ta elements. L1Ta is the youngest and most active subfamily of LINEs [6, 95]. L1Ta elements are defined by a diagnostic sequence at their 3' end [32, 135]. We took advantage of this sequence to selectively amplify, clone and sequence potentially polymorphic elements with a subfamily-specific primer. The PCR products were later applied directly to highthroughput pyrosequencing. In doing so, nearly 1600 different L1s were sequenced, 590 of which had never been observed before. These novel L1s more than quadrupled the number of L1s outside of Hg18 that were in dbRIP as of October 2009.

This new collection of L1s allowed us to confirm estimates of retrotransposition rate (RR) in human populations. Current estimates of L1 RR rely upon disease-causing mutations or cell culture assays [6, 90, 91]. For the former, the number of L1 insertions causing disease was compared to the number of SNPs known to cause disease. Since the SNP mutation rate is known, one can extrapolate to get a mutation rate for L1s. An argument against this estimation of RR is that L1 insertions are more likely to be deleterious than SNPs. For example, a SNP within an exon can be silent or synonymous, while an L1 insertion within an exon is almost certainly deleterious. Thus, using nucleotide substitution rates as a standard to determine L1 insertion rates will likely lead to an overestimation of RR. The estimate of RR from cell culture assays is also imperfect. For this estimate, L1s that were active in a cell culture assay were compared to a disease-causing "hot" L1 whose RR rates in HeLa cells and mouse germ cells were known [6]. By adding the relative activities of all the L1s predicted to be in each person's genome, the authors of the study estimated that 1 in 2 to 1 in 33 human beings have a new L1 insertion. This method has certain assumptions that may not hold true *in vivo*. For instance, L1 activity was based solely on core sequence and not genomic context. Thus, L1s that may be active in cell culture may be in heterochromatic regions of the genome where they are not normally expressed. Also, activity in HeLa cells may not necessarily translate to activity in germ or embryonic cells.

These estimates of L1 activity are good starting points, but not reliable retrotransposition rates. Even so, both methods estimate comparable RRs of 1 in 2 to 1 in 100 live human births. The neutral theory states that the majority of new mutations are neutral (or nearly neutral) and that genetic drift is the primary influence on their frequencies in the population. Therefore, if these RRs are accurate, we would expect an unbiased subset of polymorphic L1Tas should have an allelic frequency skewed toward rare insertions. We used polymorphism data from novel L1s identified in our study to determine whether rare L1s are collectively abundant in human populations (Figure 3-2). We found dozens of rare (<0.1) L1 insertions and this pattern differs greatly from L1s in dbRIP. We conclude that dbRIP is biased toward common alleles and that retrotransposon polymorphisms are a major source of genetic variation that has gone largely unstudied. If we were to follow up our study with a larger sample size and greater sequencing depth, we would be able to use allelic frequencies to better estimate RR. Or we could apply our assay to parent-child trios and more directly determine how often insertions are introduced into the population.

The studies described in Chapters 3 of this dissertation collectively identified nearly 1,000 novel retrotransposons. This is the most comprehensive and least biased study of polymorphic retrotransposons to date. The elements identified in these studies can be used as markers for GWAS and population analyses. They can also be used to study the biology of young insertions. Furthermore, these novel insertions may serve as sources of future retrotransposition events or sites of genomic instability. Thus, these elements have the potential to contribute to human genetic variation long after they have retrotransposed.

L1 elements are active in tumors

One of the hallmarks of cancer is genomic instability [Reviewed in 164]. Cancer genomes are altered at the nucleotide, epigenetic, and chromosomal levels. A complex series of events must occur to turn a healthy cell into a cancerous one [165]. One event that appears to be universal across many cancer types is the loss of transcriptional control over retrotransposons. The expression of retrotransposons in tumors is associated with genomic instability and poor prognosis [78-80, 82, 88, 135, 144, 145]. This is tantalizing, yet anecdotal, evidence that retrotransposons are active in tumors and their activity promotes tumor progression.

An acquired insertion of an L1 retrotransposon disrupted a tumor suppressor gene in a colon cancer patient [42]. Seventeen years have passed since this discovery and no other tumor-specific insertion in humans has been found (although several inherited insertions in known tumors suppressor genes have been found). There have been several efforts to identify mutations in cancer genomes and none have found retrotranspositions.

It is the author's opinion that the recent focus on SNPs in cancer genomes distracted from the discovery of retrotransposons. The acquired insertion described above was found by screening Southern blots of tumor suppressor genes. Since current cancer genome projects tend to use PCR screening and sequencing instead of Southerns, retrotransposons will likely be overlooked [113, 117, 166]. For example, a massive effort to sequence all the exons of 22 tumor genomes omitted larger-than-expected PCR products from subsequent analyses [166]. Since their target regions were around 300 bp long, they would likely miss retrotransposon insertions, even ones as small as Alus. Other cancer sequencing projects use next-gen sequencing technologies [113, 117]. As described in Chapter 1, these technologies have difficulty handling repetitive sequence, thus retrotransposons are often disregarded. Despite the finding of a somatic retrotransposition 17 years ago, there has been little follow-up and the role retrotransposons play in tumor progression is largely unknown.

We sought to directly address this problem by sequencing L1 insertion junctions in tumor and matched normal tissues. We chose to work with lung tumors because we initially found a somatic insertion in a lung tumor-derived cell line (Figure 4-1). We also chose to work with glioblastomas and medulloblastomas because L1 has been shown to be active in neural cell lines and murine brains [72]. Also, since conducting our study, a recent study demonstrated L1 activity in human fetal brains [138]. Since we had matching blood DNA for our brain tumor tissues, we would not be able to distinguish between brain tumor-specific insertions and those occurring during development. Our predictions were the following: i) The majority of L1s found would be inherited and not tumor-specific, ii) A handful of L1s will be found in a single tumor tissue and not in matched normal, iii) The brain tumors will appear to have more somatic insertions than the lung tumors because of developmental insertions.

Using massively parallel pyrosequencing (Roche 454 FLX), we identified 232 L1 and 221 Alu insertions that were sequenced in only a single tissue and were considered somatic candidates (Figure 4-2). Since we were only able to sequence to 1.8X coverage (Figure 3-6), the majority of these somatic candidates were in fact present in both tumor and normal tissue, but happened to only be sequenced in one or the other. Of 109 somatic candidates tested by PCR, 9 turned out to be somatic insertions in lung tumors (Figure 4-3). These insertions, along with one identified in a lung tumor-derived cell line (Figure 4-1) indicate that retrotransposition occurs commonly in lung tumors.

We predicted we would observe more somatic insertions in the brain. Since we had B-lymphocyte normal tissue for our brain tumor patients, we would not be able to tease apart somatic from embryonic insertions in the brain. And, although we sequenced to slightly higher coverage in the brain tumors than the lung tumors (Figure 3-7), we did not observe any brain-specific insertions. This was surprising to us. It is possible that our low sequencing coverage prevented us from observing brain-specific insertions. Or, perhaps, L1 activity in murine brains and fetal brain-derived cell lines is not demonstrative of endogenous activity in the human brain [72, 138]. Also, different neural cell types and regions of the brains seem to have different tolerances for retrotransposition [72, 138]. It may be that we did not have the appropriate samples to observe retrotransposition in the brain.

We also identified 59 genes whose hypomethylation in lung tumors was correlated with the presence of a somatic L1 insertion. Randomly permutated sets of lung tumors on average only had 1.5 correlated genes. Thus, we are confident we have identified a methylation signature that can be used to predict L1 retrotransposition potential. Among these 59 genes were 2 genes which are known regulators of retroviruses. These genes are potential host factors that may regulate L1 retrotransposition.

We knew that expression of retrotransposons was associated with genomic instability, but we did not know whether genomic instability could possibly result from acquired retrotransposition events. Now, we know that retrotransposition does, in fact, occur in lung tumors. Even so, we still cannot tease apart cause and effect. None of the insertions found in our study have an obvious effect on nearby genes unlike the previously described somatic insertion [42]. Thus, whether retrotransposons contribute to or are merely consequences of genomic instability remains to be determined. Even so, it is likely that the expression of L1 plays some role in tumor progression. The L1 endonuclease has been shown to induce double-strand breaks throughout the genome [134]. Increasing the copy number of elements may instigate more ectopic recombinations. It is unclear whether the acquired L1 insertions found in our study play an active role in tumorigenesis or are "passenger" mutations resulting from global genomic destabilization.

The knowledge that retrotransposons are jumping in tumors and possibly contributing to tumor progression creates the possibility of new therapeutic agents. Reverse transcriptase inhibitors can slow tumor progression *in vitro* [82]. While the mechanism is unknown, it is possible that the prevention of somatic retrotransposition helps maintain genome integrity. Reverse transcriptase inhibitors (as well as endonuclease inhibitors) should be considered as novel agents in preventing or slowing tumorigenesis.

Future directions in the hunt for retrotransposons

As sequencing technology gets cheaper, it will likely play a major role in the hunt for novel retrotransposons. Additional innovative platforms, such as nanopore sequencing, offer the potential for a *next*-next-generation of DNA sequencing technology. Meanwhile, the next-gen technologies continue to improve. Roche and other companies already offer platforms with longer reads than before (400-500 nt for Roche 454 GS Titanium). Five-hundred (500) nt reads are long enough to span an Alu with enough sequence left over on either end to anchor them to the human genome.

Scientists should improve upon current alignment algorithms to find novel L1s, SVAs, and HERVs. They should use paired-end sequencing and determine whether the unanchored ends of bridge pairs are from the same retrotransposon subfamily (Figure 5-1). The anchored ends should also include target site duplications. While researchers would not be able to sequence the entire retrotransposon, they would be able to see what the 5' and 3' ends of the insertion look like. Thus, they could have a set of candidate retrotransposition events (Figure 5-1). The ability to detect novel retrotransposons will bottleneck until alignment algorithms become powerful enough to handle larger indels and repetitive sequences.

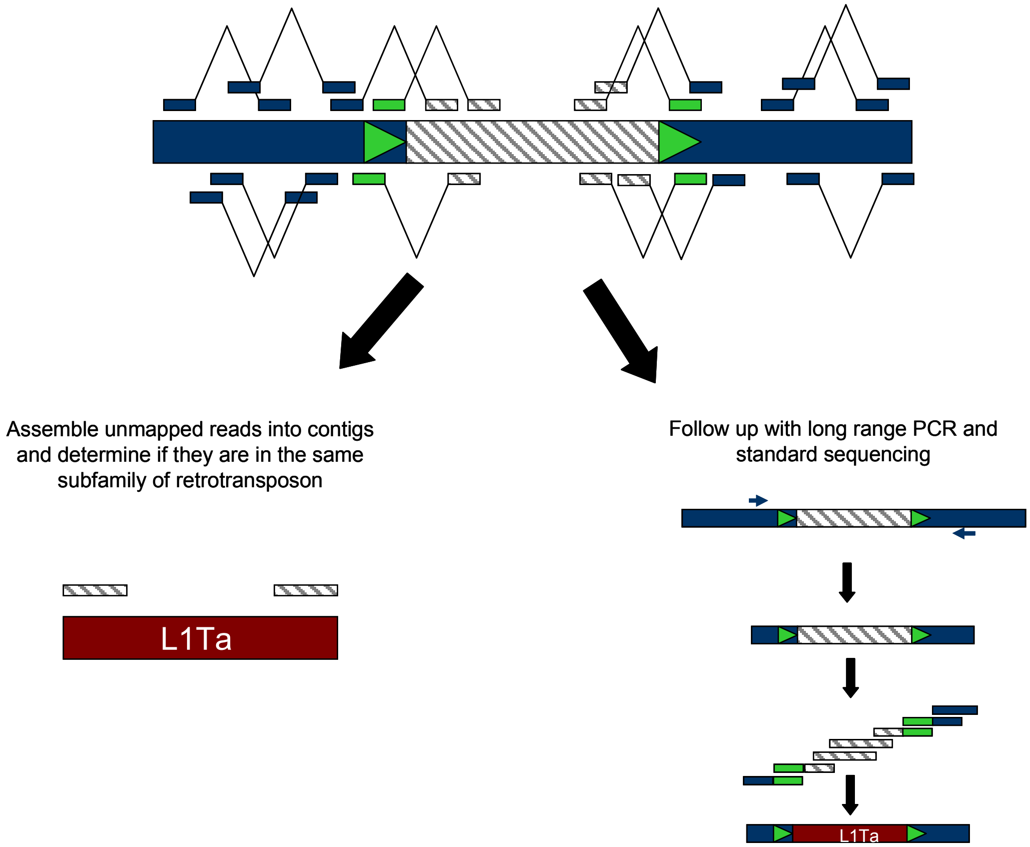

Assemble unmapped reads into contigs and determine if they are in the same subfamily of retrotransposon
Follow up with long range PCR and standard sequencing
L1Ta
L1Ta

Figure 5-1

Paired-end sequencing to identify retrotransposons. Genomic DNA is sheared and size selected. The ends are sequenced and mapped back to a reference genome. Blue regions are sequences present in the reference while the hashed area is unknown sequence. Green triangles represent target site duplications.

Another possibility would be to start with next-gen paired-end sequencing and follow up with traditional sequencing. Researchers using next-gen paired-end sequencing can find large insertions in the population, however the intervening sequence is missing. Perhaps potential insertions should be further characterized with long-range PCR followed by capillary sequencing (Figure 5-1). Researchers could then sequence precise SV junctions and inserted DNA. Then researchers could determine the cause of insertion, whether by retrotransposition or some other mechanism.

Whole genome sequencing with next-gen technology has not yet led to the discovery of novel retrotransposons, but, as described in the body of this thesis, the targeted sequencing of insertion junctions shows great promise. In addition, David Largaespada’s lab combined anchored PCR with next-gen sequencing to find insertion junctions of a synthetic transposon in mice [167, 168]. The researchers used a transposon designed to both disrupt genes from within and overexpress genes from upstream. The researchers induced the transposon to mobilize in specific tissues in mice and then isolated tumors from those tissues. By mapping the synthetic transposon insertions in these tumors, the researchers hoped to identify novel tumor suppressors and oncogenes. To isolate their transposon insertion junctions, they used linker-mediated PCR followed by Roche 454 sequencing. Altogether, they were able to find nearly 25,000 nonredundant insertions. The work from David Largaespada's lab in addition to assays performed by our lab demonstrate that selectively sequencing transposon insertion junctions is a high throughput method for identifying novel transposons.

References

1. Doolittle WF, Sapienza C., *Selfish genes, the phenotype paradigm and genome evolution.* Nature, 1980. **284**(5757): p. 601-3.

2. Orgel LE, Crick FH., *Selfish DNA: the ultimate parasite.* Nature, 1980. **284**(5757): p. 604-7.

3. McClintock, Barbara, *The origin and behavior of mutable loci in maize.* Proc Natl Acad Sci U S A., 1950. **36**(6): p. 344-55.

4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T,

Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.

5. Mills RE, Bennett EA, Iskow RC, Devine SE, *Which transposable elements are active in the human genome?* Trends Genet. , 2007. **23**(4): p. 183-91.

6. Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr., *Hot L1s account for the bulk of retrotransposition in the human population.* Proc Natl Acad Sci U S A. , 2003. **100**(9): p. 5280-5.

7. Vos JC, van Luenen HG, Plasterk RH, *Characterization of the Caenorhabditis elegans Tc1 transposase in vivo and in vitro.* Genes Dev, 1993. **7**(7A): p. 1244-53.

8. Feschotte C, Pritham EJ, *DNA transposons and the evolution of eukaryotic genomes.* Annu Rev Genet, 2007. **41**: p. 331-68.

9. Chimpanzee Sequencing and Analysis Consortium, *Initial sequence of the chimpanzee genome and comparison with the human genome.* Nature, 2005. **437**(7055): p. 69-87.

10. Kempken F, Kück U, *Transposons in filamentous fungi--facts and perspectives.* Bioessays, 1998. **20**(8): p. 652-9.

11. Ray DA, Pagan HJ, Thompson ML, Stevens RD, *Bats with hATs: evidence for recent DNA transposon activity in genus Myotis.* Mol Biol Evol, 2007. **24**(3): p. 632-9.

12. Zillig W, Prangishvilli D, Schleper C, Elferink M, Holz I, Albers S, Janekovic D, Götz D, *Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea.* FEMS Microbiol Rev, 1996. **18**(2-3): p. 225-36.

13. Pace JK 2nd, Feschotte C, *The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage.* Genome Res, 2007. **17**(4): p. 422-32.

14. Bayev AA Jr, Lyubomirskaya NV, Dzhumagaliev EB, Ananiev EV, Amiantova IG, Ilyin YV, *Structural organization of transposable element mdg4 from Drosophila melanogaster and a nucleotide sequence of its long terminal repeats.* Nucleic Acids Res, 1984. **12**(8): p. 3707-23.

15. Leib-Mösch C, Haltmeier M, Werner T, Geigl EM, Brack-Werner R, Francke U, Erfle V, Hehlmann R, *Genomic distribution and transcription of solitary HERV-K LTRs.* Genomics, 1993. **18**(2): p. 261-9.

16. Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M, *Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity.* J Virol., 2005. **79**(19): p. 12507-14.

17. Dewannieux M, Harper F, Richaud A, Letzelter C, Ribet D, Pierron G, Heidmann T, *Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements.* Genome Res, 2006. **16**(12): p. 1548-56.

18. Martin SL, Bushman FD, *Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon.* Mol Cell Biol., 2001. **21**(2): p. 467-75.

19. Feng Q, Moran JV, Kazazian HH Jr, Boeke JD, *Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition.* Cell, 1996. **87**(5): p. 905-16.

20. Mathias SL, Scott AF, Kazazian HH Jr, Boeke JD, Gabriel A, *Reverse transcriptase encoded by a human transposable element.* Science, 1991. **254**(5039): p. 1808-10.

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

22. Ostertag EM, Kazazian HH Jr., *Biology of mammalian L1 retrotransposons.* Annu Rev Genet., 2001(35): p. 501-38.

23. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV, *Human L1 retrotransposition: cis preference versus trans complementation.* Mol Cell Biol., 2001. **21**(4): p. 1429-39.

24. Schmid CW, Jelinek WR, *The Alu family of dispersed repetitive sequences.* Science, 1982. **216**(4550): p. 1065-70.

25. Voliva CF, Martin SL, Hutchison CA 3rd, Edgell MH, *Dispersal process associated with the L1 family of interspersed repetitive DNA sequences.* J Mol Biol, 1984. **178**(4): p. 795-813.

26. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH Jr., *High frequency retrotransposition in cultured mammalian cells.* Cell, 1996. **87**(5): p. 917-27.

27. Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE, *Natural genetic variation caused by transposable elements in humans.* Genetics. , 2004. **168**(2): p. 933-51.

28. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH Jr, *SVA elements are nonautonomous retrotransposons that cause disease in humans.* Am J Hum Genet., 2003. **73**(6): p. 1444-51.

29. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA., *SVA elements: a hominid-specific retroposon family.* J Mol Biol, 2005. **354**(4): p. 994-1007.

30. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE, *Active Alu retrotransposons in the human genome.* Genome Res, 2008. **18**(12): p. 1875-83.

31. Dewannieux M, Esnault C, Heidmann T, *LINE-mediated retrotransposition of marked Alu sequences.* Nat Genet., 2003. **35**(1): p. 41-8.

32. Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH Jr., *Isolation of an active human transposable element.* Science, 1991. **254**(5039): p. 1805-8.

33. Chen JM, Férec C, Cooper DN, *LINE-1 Endonuclease-Dependent Retrotranspositional Events Causing Human Genetic Disease: Mutation Detection Bias and Multiple Mechanisms of Target Gene Disruption.* J Biomed Biotechnol., 2006. **2006**(1): p. 56182.

34. Kobayashi K, Nakahori Y, Miyake M, Matsumura K, Kondo-Iida E, Nomura Y, Segawa M, Yoshioka M, Saito K, Osawa M, Hamano K, Sakakihara Y, Nonaka I, Nakagome Y, Kanazawa I, Nakamura Y, Tokunaga K, Toda T, *An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy.* Nature, 1998. **394**(6691): p. 388-92.

35. Colombo R, Bignamini AA, Carobene A, Sasaki J, Tachikawa M, Kobayashi K, Toda T, *Age and origin of the FCMD 3'-untranslated-region retrotransposal insertion mutation causing Fukuyama-type congenital muscular dystrophy in the Japanese population.* Hum Genet., 2000. **107**(6): p. 559-67.

36. Kazazian HH Jr., Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE., *Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man.* Nature, 1988. **332**(6160): p. 164-6.

37. Batzer MA, Deininger PL, *Alu repeats and human genomic diversity.* Nat Rev Genet., 2002. **3**(5): p. 370-9.

38. Burwinkel B, Kilimann MW, *Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease.* J Mol Biol, 1998. **277**(3): p. 513-7.

39. Segal Y, Peissel B, Renieri A, de Marchi M, Ballabio A, Pei Y, Zhou J, *LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis.* Am J Hum Genet., 1999. **64**(1): p. 62-9.

40. Temtamy SA, Aglan MS, Valencia M, Cocchi G, Pacheco M, Ashour AM, Amr KS, Helmy SM, El-Gammal MA, Wright M, Lapunzina P, Goodship JA, Ruiz-Perez VL, *Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in Ellis-van Creveld syndrome with borderline intelligence.* Hum Mutat., 2008. **29**(7): p. 931-8.

41. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN, *The Human Gene Mutation Database: 2008 update.* Genome Med, 2009. **1**(1): p. 13.

42. Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y., *Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer.* Cancer Res, 1992. **52**(3): p. 643-5.

43. Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS, *A de novo Alu insertion results in neurofibromatosis type 1.* Nature, 1991. **353**(6347): p. 864-6.

44. Lebedev YB, Amosova AL, Mamedov IZ, Fisunov GY, Sverdlov ED., *Most recent AluY insertions in human gene introns reduce the content of the primary transcripts in a cell type specific manner.* Gene, 2007. **390**(1-2): p. 122-9.

45. Ustyugova SV, Lebedev YB, Sverdlov ED., *Long L1 insertions in human gene introns specifically reduce the content of corresponding primary transcripts* Genetica, 2006. **128**(1-3): p. 261-72.

46. Lev-Maor G, Sorek R, Shomron N, Ast G, *The birth of an alternatively spliced exon: 3' splice-site selection in Alu exons.* Science, 2003. **300**(5623): p. 1288-91.

47. Lin L, Shen S, Tye A, Cai JJ, Jiang P, Davidson BL, Xing Y., *Diverse splicing patterns of exonized Alu elements in human tissues.* Plos Genet, 2008. **4**(10): p. e1000225.

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

49. Perepelitsa-Belancio V, Deininger P., *RNA truncation by premature polyadenylation attenuates human mobile element activity`.* Nat Genet., 2003. **35**(4): p. 363-6.

50. Sorek R, Ast G, Graur D, *Alu-containing exons are alternatively spliced.* Genome Res, 2002. **12**(7): p. 1060-7.

51. Kazazian HH Jr, Goodier JL, *LINE drive. retrotransposition and genome instability.* Cell, 2002. **110**(3): p. 277-80.

52. Han K, Sen SK, Wang J, Callinan PA, Lee J, Cordaux R, Liang P, Batzer MA, *Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages.* Nucleic Acids Res, 2005. **33**(13): p. 4040-52.

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

54. Hickey DA, *Selfish DNA: a sexually-transmitted nuclear parasite.* Genetics, 1982. **101**(3-4): p. 519-31.

55. Charlesworth B, Langley CH, *The population genetics of Drosophila transposable elements.* Annu Rev Genet, 1989. **23**: p. 251-87.

56. Boissinot S, Entezam A, Furano AV., *Selection against deleterious LINE-1-containing loci in the human lineage.* Mol Biol Evol, 2001. **18**(6): p. 926-35.

57. Cordaux R, Lee J, Dinoso L, Batzer MA, *Recently integrated Alu retrotransposons are essentially neutral residents of the human genome.* Gene, 2006. **373**: p. 138-44.

58. Sved JA, *The stability of linked systems of loci with a small population size.* Genetics, 1968. **59**(4): p. 543-63.

59. Schlenke TA, Begun DJ, *Strong selective sweep associated with a transposon insertion in Drosophila simulans.* Proc Natl Acad Sci U S A., 2004. **101**(6): p. 1626-31.

60. Smith JM, Haigh J, *The hitch-hiking effect of a favourable gene.* Genet Res, 2007. **89**(5-6): p. 391-403.

61. Stephan W, Song YS, Langley CH, *The hitchhiking effect on linkage disequilibrium between linked neutral loci.* Genetics, 2006. **172**(4): p. 2647-63.

62. Vinckenbosch N, Dupanloup I, Kaessmann H, *Evolutionary fate of retroposed gene copies in the human genome.* Proc Natl Acad Sci U S A., 2006. **103**(9): p. 3220-5.

63. Cordaux R, Udit S, Batzer MA, Feschotte C, *Birth of a chimeric primate gene by capture of the transposase gene from a mobile element.* Proc Natl Acad Sci U S A., 2006. **103**(21): p. 8101-6.

64. Liu D, Bischerour J, Siddique A, Buisine N, Bigot Y, Chalmers R, *The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase.* Mol Cell Biol., 2007. **27**(3): p. 1125-32.

65. Krull M, Brosius J, Schmitz J, *Alu-SINE exonization: en route to protein-coding function.* Mol Biol Evol, 2005. **22**(8): p. 1702-11.

66. Zhang XH, Chasin LA, *Comparison of multiple vertebrate genomes reveals the birth and evolution of human exons.* Proc Natl Acad Sci U S A., 2006. **103**(36): p. 13427-32.

67. Athanasiadis A, Rich A, Maas S, *Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome.* PLoS Biol, 2004. **2**(12): p. e391.

68. Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, Ast G, *RNA-editing-mediated exon evolution.* Genome Biol, 2007. **8**(2): p. R29.

69. Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, Gerton GL, Kazazian HH Jr, *A mouse model of human L1 retrotransposition.* Nat Genet., 2002. **32**(4): p. 655-60.

70. Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, Kazazian HH Jr, *L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism.* Genes Dev, 2009. **23**(11): p. 1303-12.

71. Kubo S, Seleme MC, Soifer HS, Perez JL, Moran JV, Kazazian HH Jr, Kasahara N, *L1 retrotransposition in nondividing and primary human somatic cells.* Proc Natl Acad Sci U S A., 2006. **103**(21): p. 8036-41.

72. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH, *Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition.* Nature, 2005. **435**(7044): p. 903-10.

73. van den Hurk JA, Meij IC, Seleme MC, Kano H, Nikopoulos K, Hoefsloot LH, Sistermans EA, de Wijs IJ, Mukhopadhyay A, Plomp AS, de Jong PT, Kazazian HH, Cremers FP, *L1 retrotransposition can occur early in human embryonic development.* Hum Mol Genet. , 2007. **16**(13): p. 1587-92.

74. Bourc'his D, Bestor TH., *Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L.* Nature, 2004. **431**(7004): p. 96-9.

75. Rangwala SH, Zhang L, Kazazian HH Jr, *Many LINE1 elements contribute to the transcriptome of human somatic cells.* Genome Biol, 2009. **10**(9): p. R100.

76. Yoder JA, Walsh CP, Bestor TH, *Cytosine methylation and the ecology of intragenomic parasites.* Trends Genet., 1997. **13**(8): p. 335-40.

77. Szpakowski S, Sun X, Lage JM, Dyer A, Rubinstein J, Kowalski D, Sasaki C, Costa J, Lizardi PM, *Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements.* Gene, 2009. **448**(2): p. 151-67.

78. Alves G, Tatro A, Fanning T., *Differential methylation of human LINE-1 retrotransposons in malignant cells.* Gene, 1996. **176**(1-2): p. 39-44.

79. Daskalos A, Nikolaidis G, Xinarianos G, Savvari P, Cassidy A, Zakopoulou R, Kotsinas A, Gorgoulis V, Field JK, Liloglou T, *Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer.* Int J Cancer, 2009. **124**(1): p. 81-7.

80. Ogino S, Nosho K, Kirkner GJ, Kawasaki T, Chan AT, Schernhammer ES, Giovannucci EL, Fuchs CS, *A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer.* J Natl Cancer Inst, 2008. **100**(23): p. 1734-8.

81. Wiemels JL, Hofmann J, Kang M, Selzer R, Green R, Zhou M, Zhong S, Zhang L, Smith MT, Marsit C, Loh M, Buffler P, Yeh RF, *Chromosome 12p deletions in TEL-AML1 childhood acute lymphoblastic leukemia are associated with retrotransposon elements and occur postnatally.* Cancer Res, 2008. **68**(23): p. 9935-44.

82. Oricchio E, Sciamanna I, Beraldi R, Tolstonog GV, Schumann GG, Spadafora C, *Distinct roles for LINE-1 and HERV-K retroelements in cell proliferation, differentiation and tumor progression.* Oncogene, 2007. **26**(29): p. 4226-33.

83. Sciamanna I, Landriscina M, Pittoggi C, Quirino M, Mearelli C, Beraldi R, Mattei E, Serafino A, Cassano A, Sinibaldi-Vallebona P, Garaci E, Barone C, Spadafora C, *Inhibition of endogenous reverse transcriptase antagonizes human tumor growth.* Oncogene, 2005. **24**(24): p. 3923-31.

84. Liu J, Nau MM, Zucman-Rossi J, Powell JI, Allegra CJ, Wright JJ., *LINE-I element insertion at the t(11;22) translocation breakpoint of a desmoplastic small round cell tumor.* Genes Chromosomes Cancer, 1997. **18**(3): p. 232-9.

85. Morse B, Rotherg PG, South VJ, Spandorfer JM, Astrin SM., *Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma.* Nature, 1988. **333**(6168): p. 87-90.

86. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, Parmigiani G, Boeke JD, *Human l1 retrotransposition is associated with genetic instability in vivo.* Cell, 2002. **110**(3): p. 327-38.

87. Teugels E, De Brakeleer S, Goelen G, Lissens W, Sermijn E, De Grève J, *De novo Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes.* Hum Mutat., 2005. **26**(3): p. 284.

88. Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A., *Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice.* Oncogene, 2008. **27**(3): p. 404-8.

89. Cordaux R, Hedges DJ, Herke SW, Batzer MA, *Estimating the retrotransposition rate of human Alu elements.* Gene, 2006. **373**: p. 134-7.

90. Kazazian HH Jr., *An estimated frequency of endogenous insertional mutations in humans.* Nat Genet. , 1999. **22**(2): p. 130.

91. Li X, Scaringe WA, Hill KA, Roberts S, Mengos A, Careri D, Pinto MT, Kasper CK, Sommer SS., *Frequency of recent retrotransposition events in the human factor IX gene.* Hum Mutat., 2001. **17**(6): p. 511-9.

92. Kimura M, Ota T, *Protein polymorphism as a phase of molecular evolution.* Nature, 1971. **229**(5285): p. 467-9.

93. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P, *dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans.* Hum Mutat., 2006. **27**(4): p. 323-9.

94. Carroll ML, Roy-Engel AM, Nguyen SV, Salem AH, Vogel E, Vincent B, Myers J, Ahmad Z, Nguyen L, Sammarco M, Watkins WS, Henke J, Makalowski W, Jorde LB, Deininger PL, Batzer MA., *Large-scale analysis of the Alu Ya5 and Yb8 subfamilies and their contribution to human genomic diversity.* J Mol Biol, 2001. **311**(1): p. 17-40.

95. Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, Jorde LB, Batzer MA., *A comprehensive analysis of recently integrated human Ta L1 elements.* Am J Hum Genet., 2002. **71**(2): p. 312-26.

96. Salem AH, Myers JS, Otieno AC, Watkins WS, Jorde LB, Batzer MA, *LINE-1 preTa elements in the human genome.* J Mol Biol, 2003. **326**(4): p. 1127-46.

97. Ovchinnikov I, Troxel AB, Swergold GD, *Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion.* Genome Res, 2001. **11**(12): p. 2050-8.

98. Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, Batzer MA, Swergold GD, *Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition.* Genome Res, 2000. **10**(10): p. 1496-508.

99. Badge RM, Alisch RS, Moran JV, *ATLAS: a system to selectively identify human-specific L1 insertions.* Am J Hum Genet., 2003. **72**(4): p. 823-38.

100. Boissinot S, Entezam A, Young L, Munson PJ, Furano AV, *The insertional history of an active family of L1 retrotransposons in humans.* Genome Res, 2004. **14**(7): p. 1221-31.

101. Buzdin A, Khodosevich K, Mamedov I, Vinogradova T, Lebedev Y, Hunsmann G, Sverdlov E, *A technique for genome-wide identification of differences in the interspersed repeats integrations between closely related genomes and its application to detection of human-specific integrations of HERV-K LTRs.* Genomics, 2002. **79**(3): p. 413-22.

102. Buzdin A, Ustyugova S, Gogvadze E, Lebedev Y, Hunsmann G, Sverdlov E, *Genome-wide targeted search for human specific and polymorphic L1 integrations.* Hum Genet., 2003. **112**(5-6): p. 527-33.

103. Mamedov IZ, Arzumanyan ES, Amosova AL, Lebedev YB, Sverdlov ED., *Whole-genome experimental identification of insertion/deletion polymorphisms of interspersed repeats by a new general approach.* Nucleic Acids Res, 2005. **33**(2): p. e16.

104. Cordaux R, Srikanta D, Lee J, Stoneking M, Batzer MA., *In search of polymorphic Alu insertions with restricted geographic distributions.* Genomics, 2007. **90**(1): p. 154-8.

105. Hollies CR, Monckton DG, Jeffreys AJ, *Attempts to detect retrotransposition and de novo deletion of Alus and other dispersed repeats at specific loci in the human genome.* Eur J Hum Genet., 2001. **9**(2): p. 143-6.

106. Roy AM, Carroll ML, Kass DH, Nguyen SV, Salem AH, Batzer MA, Deininger PL., *Recently integrated human Alu repeats: finding needles in the haystack.* Genetica, 1999. **107**(1-3): p. 149-61.

107. Wheelan SJ, Scheifele LZ, Martínez-Murillo F, Irizarry RA, Boeke JD., *Transposon insertion site profiling chip (TIP-chip).* PNAS, 2006. **103**(47): p. 17632-7.

108. Wang J, Song L, Gonder MK, Azrak S, Ray DA, Batzer MA, Tishkoff SA, Liang P, *Whole genome computational comparative genomics: A fruitful approach for ascertaining Alu insertion polymorphisms.* Gene, 2006. **365**: p. 11-20.

109. Konkel MK, Wang J, Liang P, Batzer MA, *Identification and characterization of novel polymorphic LINE-1 insertions through comparison of two human genome sequence assemblies.* Gene, 2007. **390**(1-2): p. 28-38.

110. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson ML, Pratt MR, Rasolonjatovo IM, Reed MT, Rigatti R, Rodighiero C, Ross MT, Sabot A, Sankar SV, Scally A, Schroth GP, Smith ME, Smith VP, Spiridou A, Torrance PE, Tzonev SS, Vermaas EH, Walter K, Wu X, Zhang L, Alam MD, Anastasi C, Aniebo IC, Bailey DM, Bancarz IR, Banerjee S, Barbour SG, Baybayan PA, Benoit VA, Benson KF, Bevis C, Black PJ, Boodhun A, Brennan JS, Bridgham JA, Brown RC, Brown AA, Buermann DH, Bundu AA, Burrows JC, Carter NP, Castillo N, Chiara E Catenazzi M, Chang S, Neil Cooley R, Crake NR, Dada OO, Diakoumakos KD, Dominguez-Fernandez B, Earnshaw DJ, Egbujor UC, Elmore DW, Etchin SS, Ewan MR, Fedurco M, Fraser LJ, Fuentes Fajardo KV, Scott Furey W, George D, Gietzen KJ, Goddard CP, Golda GS, Granieri PA, Green DE, Gustafson DL, Hansen NF, Harnish K, Haudenschild CD, Heyer NI, Hims MM, Ho JT, Horgan AM,

Hoschler K, Hurwitz S, Ivanov DV, Johnson MQ, James T, Huw Jones TA, Kang GD, Kerelska TH, Kersey AD, Khrebtukova I, Kindwall AP, Kingsbury Z, Kokko-Gonzales PI, Kumar A, Laurent MA, Lawley CT, Lee SE, Lee X, Liao AK, Loch JA, Lok M, Luo S, Mammen RM, Martin JW, McCauley PG, McNitt P, Mehta P, Moon KW, Mullens JW, Newington T, Ning Z, Ling Ng B, Novo SM, O'Neill MJ, Osborne MA, Osnowski A, Ostadan O, Paraschos LL, Pickering L, Pike AC, Pike AC, Chris Pinkard D, Pliskin DP, Podhasky J, Quijano VJ, Raczy C, Rae VH, Rawlings SR, Chiva Rodriguez A, Roe PM, Rogers J, Rogert Bacigalupo MC, Romanov N, Romieu A, Roth RK, Rourke NJ, Ruediger ST, Rusman E, Sanches-Kuiper RM, Schenker MR, Seoane JM, Shaw RJ, Shiver MK, Short SW, Sizto NL, Sluis JP, Smith MA, Ernest Sohna Sohna J, Spence EJ, Stevens K, Sutton N, Szajkowski L, Tregidgo CL, Turcatti G, Vandevondele S, Verhovsky Y, Virk SM, Wakelin S, Walcott GC, Wang J, Worsley GJ, Yan J, Yau L, Zuerlein M, Rogers J, Mullikin JC, Hurles ME, McCooke NJ, West JS, Oaks FL, Lundberg PL, Klenerman D, Durbin R, Smith AJ, *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-9.

111. Kim JI, Ju YS, Park H, Kim S, Lee S, Yi JH, Mudge J, Miller NA, Hong D, Bell CJ, Kim HS, Chung IS, Lee WC, Lee JS, Seo SH, Yun JY, Woo HN, Lee H, Suh D, Lee S, Kim HJ, Yavartanoo M, Kwak M, Zheng Y, Lee MK, Park H, Kim JY, Gokcumen O, Mills RE, Zaranek AW, Thakuria J, Wu X, Kim RW, Huntley JJ, Luo S, Schroth GP, Wu TD, Kim H, Yang KS, Park WY, Kim H, Church GM,

Lee C, Kingsmore SF, Seo JS, *A highly annotated whole-genome sequence of a Korean individual.* Nature, 2009. **460**(7258): p. 1011-5.

112. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC., *The diploid genome sequence of an individual human.* PLoS Biol, 2007. **5**(10): p. e254.

113. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, Chen K, Dooling D, Dunford-Shore BH, McGrath S, Hickenbotham M, Cook L, Abbott R, Larson DE, Koboldt DC, Pohl C, Smith S, Hawkins A, Abbott S, Locke D, Hillier LW, Miner T, Fulton L, Magrini V, Wylie T, Glasscock J, Conyers J, Sander N, Shi X, Osborne JR, Minx P, Gordon D, Chinwalla A, Zhao Y, Ries RE, Payton JE, Westervelt P, Tomasson MH, Watson M, Baty J, Ivanovich J, Heath S, Shannon WD, Nagarajan R, Walter MJ, Link DC, Graubert TA, DiPersio JF, Wilson RK, *DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome.* Nat Biotechnol, 2008. **456**(7218): p. 66-72.

114. Pushkarev D, Neff NF, Quake SR, *Single-molecule sequencing of an individual human genome.* Nat Biotechnol, 2009. **27**(9): p. 847-52.

115. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Zhang J, Guo Y, Feng B, Li H, Lu Y, Fang X, Liang H, Du Z, Li D, Zhao Y, Hu Y, Yang Z, Zheng H, Hellmann I, Inouye M, Pool J, Yi X, Zhao J, Duan J, Zhou Y, Qin J, Ma L, Li G, Yang Z, Zhang G, Yang B, Yu C, Liang F, Li W, Li S, Li D, Ni P, Ruan J, Li Q, Zhu H, Liu D, Lu Z, Li N, Guo G, Zhang J, Ye J, Fang L, Hao Q, Chen Q, Liang Y, Su Y, San A, Ping C, Yang S, Chen F, Li L, Zhou K, Zheng H, Ren Y, Yang L, Gao Y, Yang G, Li Z, Feng X, Kristiansen K, Wong GK, Nielsen R, Durbin R, Bolund L, Zhang X, Li S, Yang H, Wang J, *The diploid genome sequence of an Asian individual.* Nature, 2008. **456**(7218): p. 60-5.

116. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM, *The complete genome of an individual by massively parallel DNA sequencing.* Nature, 2008. **452**(7189): p. 872-6.

117. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, Stebbings LA, Leroy C, Edkins S, Hardy C, Teague JW, Menzies A, Goodhead I, Turner DJ, Clee CM, Quail MA, Cox A, Brown C, Durbin R, Hurles ME, Edwards PA, Bignell GR, Stratton MR, Futreal PA., *Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing.* Nat Genet., 2008. **40**(6): p. 722-9.

118. Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E, Zerr T, Yamada NA, Tsang P, Newman TL, Tüzün E, Cheng Z, Ebling HM, Tusneem N, David R, Gillett W, Phelps KA, Weaver M, Saranga D, Brand A, Tao W, Gustafson E, McKernan K, Chen L, Malig M, Smith JD, Korn JM, McCarroll SA, Altshuler DA, Peiffer DA, Dorschner M, Stamatoyannopoulos J, Schwartz D, Nickerson DA, Mullikin JC, Wilson RK, Bruhn L, Olson MV, Kaul R, Smith DR, Eichler EE, *Mapping and sequencing of structural variation from eight human genomes.* Nature, 2008. **453**(7191): p. 56-64.

119. Korbel JO, Urban AE, Affourtit JP, Godwin B, Grubert F, Simons JF, Kim PM, Palejev D, Carriero NJ, Du L, Taillon BE, Chen Z, Tanzer A, Saunders AC, Chi J, Yang F, Carter NP, Hurles ME, Weissman SM, Harkins TT, Gerstein MB, Egholm M, Snyder M, *Paired-end mapping reveals extensive structural variation in the human genome.* Science, 2007. **318**(5849): p. 420-6.

120. Jurka J, *Repbase update: a database and an electronic journal of repetitive elements.* Trends Genet., 2000. **16**(9): p. 418-20.

121. Landry JR, Mager DL, *Functional analysis of the endogenous retroviral promoter of the human endothelin B receptor gene.* J Virol., 2003. **77**(13): p. 7459-66.

122. Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, Eichler MY, McPherson JD, Zhao S, Pääbo S, Eichler EE, *Lineage-specific expansions of*

retroviral insertions within the genomes of African great apes but not humans and orangutans. PLoS Biol, 2005. **3**(4): p. e110.

123. Johanning K, Stevenson CA, Oyeniran OO, Gozal YM, Roy-Engel AM, Jurka J, Deininger PL, *Potential for retroposition by old Alu subfamilies.* J Mol Evol, 2003. **56**(6): p. 658-64.

124. van de Lagemaat LN, Gagnier L, Medstrand P, Mager DL, *Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates.* Genome Res, 2005. **15**(9): p. 1243-9.

125. Hedges DJ, Callinan PA, Cordaux R, Xing J, Barnes E, Batzer MA, *Differential alu mobilization and polymorphism among the human and chimpanzee lineages.* Genome Res, 2004. **1068-75**(14): p. 6.

126. Watanabe H, Fujiyama A, Hattori M, Taylor TD, Toyoda A, Kuroki Y, Noguchi H, BenKahla A, Lehrach H, Sudbrak R, Kube M, Taenzer S, Galgoczy P, Platzer M, Scharfe M, Nordsiek G, Blöcker H, Hellmann I, Khaitovich P, Pääbo S, Reinhardt R, Zheng HJ, Zhang XL, Zhu GF, Wang BF, Fu G, Ren SX, Zhao GP, Chen Z, Lee YS, Cheong JE, Choi SH, Wu KM, Liu TT, Hsiao KJ, Tsai SF, Kim CG, OOta S, Kitano T, Kohara Y, Saitou N, Park HS, Wang SY, Yaspo ML, Sakaki Y, *DNA sequence and comparative analysis of chimpanzee chromosome 22.* Nature, 2004. **429**(6990): p. 382-8.

127. Boissinot S, Chevret P, Furano AV, *L1 (LINE-1) retrotransposon evolution and amplification in recent human history.* Mol Biol Evol, 2000. **17**(6): p. 915-28.

128. Ewing B, Hillier L, Wendl MC, Green P, *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome Res, 1998. **8**(3): p. 175-85.

129. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D, *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

130. Ivics Z, Izsvák Z, *Family of plasmid vectors for the expression of beta-galactosidase fusion proteins in eukaryotic cells.* Biotechniques, 1997. **22**(2): p. 254-6, 258.

131. Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY, *Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication.* J Biol Chem, 1994. **269**(11): p. 8466-76.

132. Esnault C, Maestre J, Heidmann T, *Human LINE retrotransposons generate processed pseudogenes.* Nat Genet., 2000. **24**(4): p. 363-7.

133. Belancio VP, Hedges DJ, Deininger P., *LINE-1 RNA splicing and influences on mammalian gene expression.* Nucleic Acids Res, 2006. **34**(5): p. 1512-21.

134. Gilbert N, Lutz-Prigge S, Moran JV, *Genomic deletions created upon LINE-1 retrotransposition.* Cell, 2002. **110**(3): p. 315-25.

135. Skowronski J, Fanning TG, Singer MF., *Unit-length line-1 transcripts in human teratocarcinoma cells.* Mol Cell Biol., 1988. **8**(4): p. 1385-97.

136. Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA, *An improved PCR method for walking in uncloned genomic DNA.* Nucleic Acids Res, 1995. **23**(6): p. 1087-8.

137. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R, *Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex.* Nat Methods, 2008. **5**(3): p. 235-7.

138. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, Morell M, O'Shea KS, Moran JV, Gage FH, *L1 retrotransposition in human neural progenitor cells.* Nature, 2009. **460**(7295): p. 1127-31.

139. Garcia-Perez JL, Marchetto MC, Muotri AR, Coufal NG, Gage FH, O'Shea KS, Moran JV, *LINE-1 retrotransposition in human embryonic stem cells.* Hum Mol Genet, 2007. **16**(13): p. 1569-77.

140. Prak ET, Dodson AW, Farkash EA, Kazazian HH Jr., *Tracking an embryonic L1 retrotransposition event.* Proc Natl Acad Sci U S A., 2003. **100**(4): p. 1832-7.

141. Gasior SL, Wakeman TP, Xu B, Deininger PL, *The human LINE-1 retrotransposon creates DNA double-strand breaks.* J Mol Biol, 2006. **357**(5): p. 1383-93.

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

143. Feltus FA, Lee EK, Costello JF, Plass C, Vertino PM, *DNA motifs associated with aberrant CpG island methylation.* Genomics, 2006. **87**(5): p. 572-9.

144. Golan M, Hizi A, Resau JH, Yaal-Hahoshen N, Reichman H, Keydar I, Tsarfaty I, *Human endogenous retrovirus (HERV-K) reverse transcriptase as a breast cancer prognostic marker.* Neoplasia, 2008. **10**(6): p. 521-33.

145. Ishihara H, Tanaka I., *Detection and cloning of unique integration sites of retrotransposon, intracisternal A-particle element in the genome of acute myeloid leukemia cells in mice.* FEBS Letters, 1997. **418**: p. 205-209.

146. Getz C, van Schaik N., *Somatic mutation in the wings of Drosophila melanogaster females dysgenic due to P elements when reared at 29 degrees C.* Mutat Res, 1991. **248**(1): p. 187-94.

147. Eide D, Anderson P, *Insertion and excision of Caenorhabditis elegans transposable element Tc1.* Mol Cell Biol., 1988. **8**(2): p. 737-46.

148. Liu D, Wang R, Galli M, Crawford NM, *Somatic and germinal excision activities of the Arabidopsis transposon Tag1 are controlled by distinct regulatory sequences within Tag1.* Plant Cell, 2001. **13**(8): p. 1851-63.

149. An W, Han JS, Wheelan SJ, Davis ES, Coombes CE, Ye P, Triplett C, Boeke JD, *Active retrotransposition by a synthetic L1 element in mice.* Proc Natl Acad Sci U S A., 2006. **103**(49): p. 18662-7.

150. Girard L, Zöchbauer-Müller S, Virmani AK, Gazdar AF, Minna JD, *Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering.* Cancer Res, 2000. **60**(17): p. 4894-906.

151. Grigorova M, Lyman RC, Caldas C, Edwards PA., *Chromosome abnormalities in 10 lung cancer cell lines of the NCI-H series analyzed with spectral karyotyping.* Cancer Genet Cytogenet., 2005. **162**(1): p. 1-9.

152. Whang-Peng J, Kao-Shan CS, Lee EC, Bunn PA, Carney DN, Gazdar AF, Minna JD, *Specific chromosome defect associated with human small-cell lung cancer; deletion 3p(14-23).* Science, 1982. **215**(4529): p. 181-2.

153. Sen SK, Huang CT, Han K, Batzer MA, *Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome.* Nucleic Acids Res, 2007. **35**(11): p. 3741-51.

154. Srikanta D, Sen SK, Conlin EM, Batzer MA, *Internal priming: An opportunistic pathway for L1 and Alu retrotransposition in hominins.* Gene, 2009. **448**233-41

155. Amacker M, Hottiger M, Mossi R, Hübscher U, *HIV-1 nucleocapsid protein and replication protein A influence the strand displacement DNA synthesis of lentiviral reverse transcriptase.* AIDS, 1991. **11**(4): p. 534-6.

156. Stetson DB, Ko JS, Heidmann T, Medzhitov R., *Trex1 prevents cell-intrinsic initiation of autoimmunity.* Cell, 2008. **134**(4): p. 587-98.

157. Yan N, Cherepanov P, Daigle JE, Engelman A, Lieberman J, *The SET complex acts as a barrier to autointegration of HIV-1.* PLoS Pathog, 2009. **5**(3): p. e1000327.

158. The International HapMap Consortium, *A second generation human haplotype map of over 3.1 million SNPs.* Nature, 2007. **449**(7164): p. 851-61.

159. Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS, Devine SE, *An initial map of insertion and deletion (INDEL) variation in the human genome.* Genome Res, 2006. **16**(9): p. Sep.

160. Lee J, Cordaux R, Han K, Wang J, Hedges DJ, Liang P, Batzer MA., *Different evolutionary fates of recently integrated human and chimpanzee LINE-1 retrotransposons.* Gene, 2007. **390**(1-2): p. 18-27.

161. Boissinot S, Furano AV, *Adaptive evolution in LINE-1 retrotransposons.* Mol Biol Evol, 2001. **18**(12): p. 2186-94.

162. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A, *PANTHER: a library of protein families and subfamilies indexed by function.* Genome Res, 2003(13): p. 2129-41.

163. Locke DP, Archidiacono N, Misceo D, Cardone MF, Deschamps S, Roe B, Rocchi M, Eichler EE, *Refinement of a chimpanzee pericentric inversion breakpoint to a segmental duplication cluster.* Genome Biol, 2003. **4**(8): p. R50.

164. Murga M, Fernández-Capetillo O, *Genomic instability: on the birth and death of cancer.* Clin Transl Oncol., 2007. **9**(4): p. 216-20.

165. Hanahan D, Weinberg RA, *The Hallmarks of Cancer.* Cell, 2000. **100**: p. 57-70.

166. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE., *The consensus coding sequences of human breast and colorectal cancers.* Science, 2006. **314**(5797): p. 268-74.

167. Keng VW, Villanueva A, Chiang DY, Dupuy AJ, Ryan BJ, Matise I, Silverstein KA, Sarver A, Starr TK, Akagi K, Tessarollo L, Collier LS, Powers S, Lowe SW, Jenkins NA, Copeland NG, Llovet JM, Largaespada DA, *A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma.* Nat Biotechnol, 2009. **27**(3):264-74.

168. Starr TK, Allaei R, Silverstein KA, Staggs RA, Sarver AL, Bergemann TL, Gupta M, O'Sullivan MG, Matise I, Dupuy AJ, Collier LS, Powers S, Oberg AL, Asmann YW, Thibodeau SN, Tessarollo L, Copeland NG, Jenkins NA, Cormier RT, Largaespada DA, *A Transposon-Based Genetic Screen in Mice Identifies Genes Altered in Colorectal Cancer.* Science, 2009. **323(**5922): 1747-50.

Appendix I – Novel L1 insertions identified by ABI Capillary sequencing

Name	Hg18 coordinates	Allelic Frequency	PCR Validated	Validation Primer A	Validation Primer D
20-4-31_A06	chr1:119354875-119355311	0.391	VAL	CCCTGTAGATGGGTAATAGTGGTCC	AAGTCATTCTAGTTGGAAAGCCATAAC
20-4-32_F02	chr1:147483426-147483679	NT	VAL	ACAGTCCCCTGAAGCAAAGTG	ATCAACCTGGTCCCTTGCTCTAAC
20-4-23_B01	chr1:177841660-177841983	NT	NT		
20-4-7_H04	chr1:187096120-187096534	NT	NT		
20-2-4_D03	chr1:189239391-189239636	NT	NT		
20-4-7_G04	chr1:190528633-190528883	0.250	VAL	CGGAATGAAAACACTAGAACTATGGAG	GAAATCACACAATCACCTTGCTTAGAA
20-4-7_F06	chr1:247158091-247158253	NT	NT		
20-4-34_H10	chr1:45472027-45472170	NT	NT		
20-4-9_H06	chr1:56177519-56177983	0.087	VAL	AAGCCATAGCAAACAAATACAGTTAGTG	CCCCGCCACAAGGTAATAAGC
20-4-10_G12	chr1:81451261-81451598	0.011	VAL	TTTTGGCTGCCTAGTTTATCTGC	AATGTAAAGTGCTGTTGCCTCAGTG
20-2-8_H08	chr1:83382073-83382326	NT	NT		
20-2-10_F06	chr1:88404553-88404596	NT	VAL	GAAGAGAAGCCTGAAGGAACTGC	GCAGCATTTGGTATTTGTGGG
20-4-34_G07	chr1:95574103-95574443	NT	NT		
20-4-4_A07	chr10:124445207-124445515	0.182	VAL	GTGAAGGTGCTGGCTGGACTC	GTCAGTCCAATCTCCTTCCCTCC
20-4-19_A01	chr10:24589623-24590090	0.011	VAL	GTGTCTTGGCTTAGTTGCTGTGG	AATGCCTGTTTACAAGAGACCCAG
20-4-24_E01	chr10:25616538-25616952	NT	NT		
20-4-9_A11	chr10:4624464-4624600	NT	NT		
20-4-3_F11	chr10:52732822-52733103	0.038	VAL	CCATCATGCCCAGTAAGACGAG	TCCAACCAGGGAAAGAAGGG
20-4-8_E07	chr10:67848162-67848290	NT	NT		
20-4-31_F09	chr10:68619323-68619583	NT	VAL	CCTGGGTGACAGAGCGAGAC	ATAGATTCATCCCACATTGTTCACTC
20-4-34_B07	chr10:91706903-91707616	0.182	VAL	TGTTCCAGACTGACTGAGAGGGC	GTGTGGAATCAGAGCAAGGGG
20-4-7_A04	chr10:92720700-92720929	0.023	VAL	GCCCTAAAGCATTCGGTCACCTAG	GTGTAGTCACTGCTAGAATGGTTAGTCAGAAC
20-4-34_B01	chr11:109882845-109883099	0.158	VAL	GGGCTCAGTGGTCACAAGGG	GCAGGGATGATGAGTGTGGC
20-4-29_B06	chr11:114400723-114401050	NT	NT		
20-4-5_G08	chr11:127872768-127873040	0.100	VAL	CCCGTTAGCAGATTGGAGGATG	TAACACATTATTACCACTGAAGTCCTGAG

20-4-12_E03	chr11:133810106-133810850	NT	VAL	TTGTCTGGGCAAGGCGAGG	GGGATGAGGAGATGGAGTGGG
20-4-22_H05	chr11:83329650-83330012	0.091	VAL	GCTTCTACAAATGCTGTGACAATGAC	CCAGCCCCATCACCTTGC
20-4-7_G11	chr11:94316929-94317492	0.013	VAL	GTCACTAAGAAGGTCTCAAATCTCTATGG	GATGAGTGGGTGAGGCGTGAG
20-2-4_B08	chr11:95499833-95499948	0.022	VAL	GCACACTTCAACAAATCACCCATCCAG	GCAGTCATCTCAGGTGCGGTTCCTACAG
20-4-26_F03	chr11:96938909-96939516	NT	NT		
20-4-32_G05	chr12:100696857-100697382	NT	NT		
20-4-17_C09	chr12:125221188-125221249	0.063	VAL	GAGATAGACCCAACTTACTAACATACCTGAG	GCATAGCCTCAGCTCATGTGG
20-4-35_B06	chr12:38432990-38433454	0.013	VAL	ACCTGAATAACGGACACCTGAATAG	TAATGTCACGGTTTTTGTTTGAATC
20-4-35_B03	chr12:69306354-69306688	NT	NT		
20-4-38_G05	chr12:7996882-7997043	NT	NT		
20-2-1_E12	chr12:83489269-83489390	0.315	VAL	GAAATGAGGGGAGCAGACGC	TGCTGGTTCTGTGGTTTCTACTGAC
20-4-19_G01	chr13:103093852-103094234	NT	NT		
20-2-8_H06	chr13:21387796-21388346	0.341	VAL	CACTTTGGCATCCCTTCCTTCTG	TGGATAGATAGGTAAACAGGTTAGTGTAGTAGC
20-4-10_G11	chr13:26386112-26386789	0.321	VAL	GCCTACCATCCCCAGAACTTACC	CTCTGCCACATTCCTGTTTCCC
20-4-24_C03	chr13:37640100-37640201	NT	NT		
20-4-11_E10	chr13:72094397-72094630	NT	NT		
20-4-5_E12	chr13:77038245-77038608	NT	NT		
20-4-4_D12	chr13:78157303-78157794	0.955	VAL	AACCAGTGCCAATAGGTGTCAGG	GTAGTGGAACCTTGATAACAGATGGG
20-4-11_F02	chr13:85281099-85281712	0.011	VAL	GAATAGAAATTGACAGCGTGGGAG	GGGCAGGATTGAAATAGGAGAGC
20-4-35_G06	chr14:35744498-35744702	0.420	VAL	GGTATTAGGTGACTTGGTATGTTAGCG	TACAAATGCTGGGGCTCTTAGG
20-4-2_D12	chr14:58230644-58230666	0.074	VAL	ATGGGGTGGGTGTTACAGGAGAG	AGATGTCACCGAAGCACAATGG
20-4-36_C12	chr14:62303031-62303497	NT	NT		
20-4-27_B01	chr14:62418238-62418672	0.015	VAL	CCACATAACAGACCACATTTTTCTACTC	TGAGTGTCATACAATAAGCATAAGTCTGG
20-4-20_A12	chr14:66489458-66489629	NT	NT		
20-4-19_B11	chr14:71403036-71403468	0.091	VAL	ACCTGCTGAAAGTGTTGCCTCTC	GAAAAACAGACAAAAGTTCATAGAGGC
20-4-34_F11	chr14:85451132-85451558	NT	NT		
20-2-9_B04	chr15:25848804-25849392	NT	NT		
20-2-3_C08	chr15:54038431-54038787	NT	NT		
20-4-3_D02	chr16:17662715-17662807	NT	NT		
20-1-1_A09	chr16:59636694-59636784	NT	NO	CACAAAGCATAGGGATCAATCTCAG	CACATGATTAAGTCCTTCTGTTCTGCTG
20-4-32_G12	chr17:30158894-30159533	NT	NT		
20-4-30_F08	chr17:31296264-31296614	NT	NT		

20-4-32_H03	chr18:1751470-1751534	NT	NT		
20-4-17_G03	chr18:38642599-38642848	0.000	SOM	GCATTCTGCTAAAACACATCAAGAGG	GTTTTGCCTATCCTATCTCAACAGTG
20-4-32_C12	chr18:57730954-57731237	NT	NT		
20-4-32_A05	chr18:67853774-67854040	NT	NT		
20-4-37_D07	chr18:72643827-72644259	NT	NT		
20-4-38_A06	chr19:32560430-32560622	NT	NT		
20-4-9_C03	chr2:105529597-105529750	NT	NT		
20-4-23_D08	chr2:105582843-105583108	0.091	VAL	CGAGCATCAGCAGTGAGTTTAGG	GGACCAGCAAATGTTTCAGCC
20-4-19_A03	chr2:126689817-126689941	NT	NT		
20-4-28_E12	chr2:13100639-13101295	NT	NT		
20-4-21_C11	chr2:152258683-152258706	NT	NT		
20-4-4_A06	chr2:156236106-156236198	NT	NT		
20-4-36_E02	chr2:168230202-168230570	NT	NT		
20-4-11_F12	chr2:1853986-1854058	0.011	VAL	TGGGTCTGAGAGTCGGTCTAATCC	ACAGAGTCCCCTCCATCAGCG
20-4-35_A12	chr2:188639561-188640185	NT	VAL	AGGGGACATTTGAAGAACTATGACAG	GTCAGTCAGTTCAGCCCAATGTG
20-4-17_C01	chr2:191186719-191186952	NT	NT		
20-4-34_A12	chr2:201346315-201346839	0.091	VAL	CCTTCTGAAACGCCTTGAGCC	CCTCCTAACCTACCCAAAACATCG
20-4-22_E08	chr2:212415432-212415830	NT	VAL	AGTCAAAAGTGATGGCAAAACAGC	AGTGTTTGTGTCAGTGGCGGG
20-4-18_B09	chr2:41904911-41905304	NT	NT		
20-4-36_C07	chr2:59207759-59208008	NT	NT		
20-4-10_E06	chr2:83041849-83042317	0.121	VAL	ACATTATTCCCTGTGTATTCCTGGTG	TCAGACACACAAGCAAATCAATGG
20-4-7_C09	chr20:17808931-17808964	0.136	VAL	GCAAAGGGCAGCAGGGTGACTGGGT	CGTTCAGAAGCCCCGAGCAGTGG
20-2-9_D09	chr20:22649936-22650080	NT	NT		
20-4-38_D09	chr20:41385103-41385518	NT	VAL	TTGCTCTGGCTGCTCTGCG	TAGAAGGGCACGGTAGTCAGAGG
20-2-9_A08	chr20:59565973-59566222	NT	VAL	TGAGTCAGCCTAAAATGAGTCCACC	TTTCGGCGGGAGAGGGAAC
20-4-17_G01	chr21:27991048-27991145	0.645	VAL	TTGAGGAGATTGTAGAGAAGTGAGGAAC	GATGTCTCCCTGCCTAACTCCTTC
20-2-1_B10	chr22:20816180-20816371	NT	VAL	TAGTTAATGCTAATGTTGCCACAAGTG	TTGAGATTTGAAAGAATGTTGAGGCAGC
20-4-33_C07	chr22:42655407-42655933	0.136	VAL	GCCACTGTAGAAGGGGATGAAGC	AAGTCCGAGGGTGTATGTTAGTTCC
20-4-12_C12	chr3:107447880-107448380	0.011	VAL	GGTCTTTACTACCCTACTACCGATGTG	GGTTATCTTTCGTTACTTTAGGTTTGGG
20-4-8_C09	chr3:152630884-152631199	0.561	VAL	ACACCAGCACTACAGGCAAGATTAC	CTTGAATGCTTCCTTTCCTGGG
20-4-10_H12	chr3:153170683-153170785	NT	NT		
20-4-2_D02	chr3:154184011-154184561	NT	NT		

20-4-11_A11	chr3:30394212-30394397	NT	NT		
20-2-1_E04	chr3:38601071-38601173	0.387	VAL	GCACCTGCCCAAGGACAGAC	ATCAGTAGATTCTTCCCCAGTTACCA
20-4-7_B02	chr3:48361751-48362006	NT	NT		
20-4-33_B11	chr3:75183913-75184071	NT	NT		
20-4-17_G02	chr3:85659259-85659928	0.378	VAL	TTCACATATCCTTCACAGACACTTGC	TGATAATGACAACAGGAAGGGGAC
20-4-30_G01	chr4:103562770-103563039	NT	NT		
20-4-36_E10	chr4:118551527-118551856	NT	NT		
20-4-10_C05	chr4:140866852-140867240	NT	NT		
20-4-29_E05	chr4:17771203-17771783	NT	NT		
20-2-3_E05	chr4:190896697-190896991	0.196	VAL	GCTATCCATAAAAGCACAACCTCTG	TATTCTCTACTAAACACCTGCTCATATCCTG
20-1-1_B09	chr4:191206810-191207159	NT	VAL	GGTTTCAAATGACCCCAAGTGCTCC	CCTATCCTTCTCAGGTTCCTCTTTCAGTTG
20-4-35_G07	chr4:24251813-24251903	NT	NT		
20-4-30_A09	chr4:28930949-28931263	NT	NT		
20-1-1_C06	chr4:46702477-46702535	0.587	VAL	TGTCGGAAATGATCTAATGGAAGGGATTC	TGTCCTTTTGTCAATCACAGAGAAC
20-4-3_D03	chr4:53323238-53323941	NT	NT		
20-4-34_H11	chr5:109595429-109595493	NT	NT		
20-4-2_G01	chr5:11444269-11445017	NT	VAL	ACATTCAAAGGAAGGCACAGGC	CGCCTTGTCTGTCTTTGTGTGC
20-4-15_F11	chr5:155552443-155553106	0.364	VAL	GATTTATCAGTGACATTTTAGCGTGC	AAGAGGAGAGTAAGCAGGGTCGG
20-4-21_D05	chr5:159283598-159283769	NT	NT		
20-4-14_E01	chr5:33668478-33669097	0.762	VAL	TAGGATGACTGCTTTCAAACATAGGC	CAGAGAGAAACGGACAGATTCACTAAG
20-4-29_G10	chr5:41464297-41464481	NT	VAL	none	CCAGTTCCTCCAAGTCAGTCCG
20-4-8_G06	chr5:41637789-41638057	NT	NT		
20-4-2_E02	chr5:84274784-84275360	NT	NT		
20-4-26_A08	chr6:13299016-13299425	0.182	VAL	CAGCAAAGAAAGGAAACTGAGGC	ATGGAGGGTTGGGAATGTGG
20-4-35_B01	chr6:157888209-157888698	0.738	VAL	ATGCCTGCTATGTTGACCCTGC	GGAAGTTTTGACCCACTGATTTGAC
20-4-14_H12	chr6:19901102-19901343	0.022	VAL	ATTGCCCTGGTTCATACCTTCC	CGCATCTCTTAGCACTTACTCCATAGC
20-2-11_F10	chr6:69255220-69255461	NT	VAL	CCGTGTTCCTTTGCTTTCTTGTAG	ATCTCTATCTCTAATCCTGACTTCCTGC
20-4-36_B11	chr6:92678449-92678651	NT	NT		
20-4-23_A06	chr7:111020046-111020168	0.136	VAL	GAACCTGTGCTTTGGACTGCC	AGTTTTACCCAGTTGGCACTTATCC
20-4-31_B05	chr7:127717664-127718096	NT	NT		
20-4-4_G04	chr7:142528781-142529368	NT	NT		
20-2-10_B03	chr7:155694089-155694150	0.011	VAL	GGAAAACCTTCACAAGCCTCAGC	GCCAAGCCGTGGACCATACC

20-4-38_B01	chr7:61461081-61461622	0.024	VAL	GTTTTGGCTTTCTCCCTGACTCC	TTGGAAGGACAGTTTGCTCGG
20-4-20_B08	chr7:62130365-62130596	NT	NT		
20-4-14_G11	chr7:69826300-69826625	NT	NT		
20-4-18_E12	chr7:84738998-84739656	0.609	VAL	CCTATGTTTAGTCTGTGAAGGCTTATTGC	CCTGTTTACTCAAATCATTGGTTCTTCTGCTC
20-4-24_C01	chr7:89782355-89782563	NT	NT		
20-4-2_F07	chr7:97212581-97212861	NT	NT		
20-4-36_A02	chr8:100927834-100928460	NT	NT		
20-4-36_C09	chr8:140214044-140214423	NT	NT		
20-2-8_A02	chr8:24981914-24982456	NT	NT		
20-2-3_A07	chr8:25559452-25559550	0.065	VAL	GACTTACCCACTACATATTCTTGACCATCCCTG	GTTACCCTGCTTGGCAACAGAATC
20-4-30_C08	chr8:29112007-29112456	NT	NT		
20-4-30_B02	chr8:67544199-67544422	NT	VAL	AGTTGAATGGCACAGTAATGCTTTTCTC	TTGGGGTGTGGAGAACTAAGAGAAAATG
20-4-2_C10	chr8:72561556-72561782	NT	NT		
20-4-16_D08	chr8:84736526-84736731	NT	NT		
20-4-36_C02	chr9:116967805-116968008	NT	NT		
20-4-30_B05	chr9:117123580-117123699	NT	NT		
20-4-31_F02	chr9:30664336-30664458	NT	NT		
20-4-18_F10	chr9:5481411-5481590	NT	NT		
20-4-27_A10	chr9:67967080-67967174	NT	VAL	AGTTTCAGGAATAAGCACAGGACAGC	CTTCTGCTGCTTTTGCTTGATGC
20-4-14_C12	chr9:74094556-74094736	NT	NT		
20-4-28_A09	chr9:81272413-81272608	NT	NT		
20-4-34_B06	chrX:101325923-101325980	NT	NT		
20-4-13_D06	chrX:102249841-102250273	NT	NT		
20-4-9_G05	chrX:111444126-111444410	NT	NT		
20-4-27_G01	chrX:122492159-122492541	NT	NT		
20-4-15_B03	chrX:123057684-123057794	NT	NO	ATGAGGTAGGAGAGGGCAAGGTG	AATCCTGGCTGACTGGGACTTG
20-4-19_F02	chrX:123439835-123439951	NT	NT		
20-4-3_H03	chrX:129352772-129352957	NT	NT		
20-4-11_D01	chrX:71229151-71229344	NT	NT		
20-4-26_D07	chrX:75866020-75866307	NT	NT		
20-4-2_E12	chrX:76328251-76328522	NT	NT		

Appendix II – Novel L1 insertions identified by Roche 454 pyrosequencing

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	99122BL	99122GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr1:101023271-101023295	no	NT																															Y																													
chr1:101759562-101759612	no	VAL	Y																																																											
chr1:102122550-102122589	no	SOM																																																		Y										
chr1:114543977-114544002	no	VAL																														Y																														
chr1:116782333-116782357	no	NT	Y	Y	Y																															Y															Y	Y												
chr1:119039998-119040227	no	NT																				Y																																								
chr1:121060021-121060056	no	NT																						Y						Y																							Y									
chr1:121186240-121186396	no	NT													Y																																															
chr1:121186719-121186832	no	NT											Y																																																	
chr1:121186830-121186925	no	NT																																			Y																									
chr1:12798482-12798699	no	NT					Y			Y							Y							Y			Y				Y	Y															Y						Y					Y				
chr1:148892291-148892314	no	NT																			Y																																									
chr1:155999574-155999632	no	NT									Y	Y	Y						Y	Y																Y				Y				Y																		
chr1:160282466-160282521	no	NT																																										Y	Y																	
chr1:160877364-160877485	no	VAL												Y																																																
chr1:163819567-163819771	no	NT												Y				Y																																				Y								Y
chr1:173233297-173233356	no	VAL																																																	Y											
chr1:177841926-177841956	yes	NT													Y	Y	Y	Y	Y	Y						Y																												Y		Y	Y					
chr1:186787743-186787798	no	VAL	Y	Y																																																										

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-12BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr1:189239517-189239585	yes	NT																			Y	Y																																								
chr1:189497188-189497318	no	NO																			Y																																									
chr1:190528821-190528858	yes	NT											Y	Y						Y					Y			Y			Y	Y	Y	Y		Y					Y	Y			Y	Y																
chr1:215281531-215281601	no	VAL																												Y																																
chr1:215615527-215615576	no	NT											Y	Y								Y																																								
chr1:220645719-220645747	no	NT		Y									Y																										Y	Y											Y	Y										
chr1:23029620-23029808	no	NT											Y	Y																			Y	Y																												
chr1:247158088-247158289	yes	NT		Y																																		Y											Y													
chr1:40974280-40974300	no	NT			Y														Y	Y			Y						Y	Y			Y	Y	Y	Y	Y	Y					Y	Y					Y	Y	Y								Y			
chr1:41275305-41275358	no	VAL																																			Y																									
chr1:42256550-42256602	no	NT	Y	Y	Y						Y	Y							Y	Y	Y	Y	Y									Y						Y	Y				Y		Y																	
chr1:56177832-56177985	yes	NT			Y																																																						Y	Y		
chr1:69574391-69574425	no	NT			Y																																	Y												Y												
chr1:74965080-74965257	no	NT	Y			Y		Y	Y						Y			Y						Y																	Y									Y						Y			Y		Y	
chr1:78134550-78134602	no	VAL																				Y																																								
chr1:89556114-89556284	no	VAL																										Y																																		
chr1:94746951-94747016	no	NT																													Y																															
chr1:95574423-95574443	yes	NT			Y															Y																					Y	Y																				
chr1:98136175-98136314	no	VAL				Y																																																								
chr2:102989107-102989139	no	NT																																																		Y									
chr2:105529637-105529841	yes	NT				Y									Y	Y					Y																														Y	Y			Y				Y	Y		
chr2:105583022-105583111	yes	NT																																																Y										Y		
chr2:117497922-117497955	no	VAL																																																									Y	Y		
chr2:11847423-11847498	no	NT		Y					Y	Y			Y	Y							Y	Y							Y								Y	Y					Y	Y	Y	Y			Y	Y	Y				Y	Y			Y	Y		

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr2:125944398-125944417	no	NT																																																				Y											
chr2:13100639-13100674	yes	NT									Y	Y													Y	Y	Y	Y													Y	Y	Y	Y																				
chr2:132679307-132679358	no	NT																																																													Y	
chr2:134521056-134521123	no	NT												Y																												Y																						
chr2:136872693-136872847	no	NT											Y	Y											Y				Y														Y	Y																				
chr2:150415784-150415808	no	NT	Y									Y																																																				
chr2:156236121-156236324	yes	NT																																																		Y								Y					
chr2:168443409-168443616	no	VAL																													Y																Y															
chr2:17885165-17885195	no	VAL						Y																																																						
chr2:182231112-182231170	no	VAL		Y																																																										
chr2:187026550-187026580	no	VAL																																		Y																										
chr2:190631456-190631641	no	NT																			Y																																									
chr2:191186912-191186943	yes	NT				Y	Y	Y							Y		Y						Y				Y	Y	Y	Y			Y	Y									Y	Y									Y	Y			Y	Y	Y	Y	Y	Y
chr2:201346369-201346483	yes	NT				Y							Y																								Y	Y																								
chr2:203854152-203854177	no	NO																																																					Y											
chr2:206415768-206415791	no	NT																																																							Y								
chr2:212415428-212415557	yes	NT				Y			Y	Y																	Y	Y													Y	Y								Y							Y					
chr2:221211764-221211974	no	NT																			Y	Y											Y																						Y	Y				Y		
chr2:23336125-23336185	no	NT			Y																																																									
chr2:235184534-235184697	no	NT				Y			Y																					Y																	Y															
chr2:35732862-35732913	no	VAL																		Y																												Y														
chr2:41904908-41905014	yes	NT				Y	Y	Y	Y	Y	Y		Y	Y		Y				Y	Y	Y							Y	Y			Y	Y	Y				Y	Y	Y		Y	Y		Y	Y	Y							Y				Y	Y		
chr2:43794085-43794125	no	VAL																																	Y	Y																										
chr2:5216404-5216425	no	NT						Y																																																						

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	99122BL	99122GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr2:53313722-53313958	no	NT					Y	Y											Y	Y																																						Y			
chr2:69022388-69022407	no	NT																															Y	Y				Y					Y																			
chr2:75579677-75579741	no	NT																															Y																													
chr2:77816748-77816809	no	VAL																																Y																							Y			
chr2:79783012-79783228	no	VAL																																															Y												
chr2:81951539-81951593	no	NT		Y			Y		Y	Y				Y		Y												Y	Y												Y		Y			Y			Y			Y		Y	Y		Y				
chr2:83041838-83042058	yes	NT			Y							Y					Y	Y																			Y	Y		Y														Y						Y	Y		
chr2:91545160-91545218	no	NT			Y								Y	Y			Y					Y																															Y									
chr3:112074560-112074853	no	NT																Y																																												
chr3:114098759-114098801	no	VAL																	Y	Y																																										
chr3:122246361-122246386	no	NT					Y				Y										Y																																		Y						
chr3:125073403-125073463	no	VAL			Y	Y	Y	Y	Y		Y						Y	Y								Y	Y	Y	Y	Y	Y	Y	Y	Y					Y	Y		Y			Y	Y	Y	Y	Y	Y			Y			Y		Y	Y			
chr3:126404222-126404286	no	NT																																														Y			Y										
chr3:126542191-126542278	no	NT																																																Y										
chr3:132123776-132123923	no	NT																																							Y																		
chr3:132279201-132279308	no	NT											Y																																															
chr3:134071664-134071800	no	VAL																											Y																															
chr3:140527311-140527520	no	VAL															Y				Y																																									
chr3:140617549-140617762	no	NT							Y																																																					
chr3:154184011-154184185	yes	NT																						Y					Y	Y												Y			Y	Y					Y	Y					Y					
chr3:167841181-167841202	no	VAL														Y																																														
chr3:170752651-170752759	no	VAL							Y	Y										Y	Y	Y													Y	Y																									Y	Y		Y
chr3:174231785-174231821	no	VAL							Y	Y			Y	Y	Y	Y	Y	Y					Y	Y					Y	Y	Y	Y		Y										Y														Y	Y		Y	Y
chr3:177160627-177160649	no	NT							Y	Y	Y	Y															Y	Y	Y	Y												Y	Y	Y	Y		Y	Y										Y			Y	Y

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	99122BL	99122GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr3:177574017-177574155	no	VAL	Y		Y	Y	Y		Y	Y			Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y	Y			Y	Y	Y	Y	Y	Y			Y			Y	Y	Y		Y
chr3:177811562-177811593	no	NT			Y																																																									
chr3:182648493-182648513	no	NT																	Y																																									
chr3:182675599-182675830	no	NT							Y	Y		Y	Y	Y	Y	Y	Y						Y								Y	Y		Y																											Y			
chr3:189717114-189717323	no	VAL			Y																																											Y	Y												
chr3:197738304-197738520	no	VAL																																				Y																								
chr3:20723771-20723816	no	VAL					Y		Y	Y			Y		Y	Y																																														
chr3:21209923-21210120	no	NT																																	Y	Y	Y		Y							Y				Y	Y											
chr3:21775878-21775956	no	VAL																																										Y																
chr3:23081661-23081867	no	NT	Y																																																										
chr3:24936873-24936907	no	VAL																		Y																																								
chr3:30394182-30394396	yes	NT				Y																																					Y	Y																		
chr3:37865011-37865144	no	NT																															Y																											
chr3:43237680-43237758	no	VAL																																																				Y						
chr3:46081731-46081750	no	NT																																																Y										
chr3:58759169-58759370	no	VAL																											Y																															
chr3:631407-631444	no	VAL	Y	Y		Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
chr3:80672816-80672847	no	VAL		Y								Y	Y	Y		Y					Y	Y										Y																									Y	Y		Y		
chr3:81351638-81351738	no	NT	Y																																																										
chr3:82073623-82073691	no	NT					Y	Y																																																	Y				
chr3:83696685-83696735	no	NT																																																	Y									
chr3:84516074-84516104	no	NO			Y		Y	Y											Y	Y							Y	Y	Y	Y																											Y					
chr3:85659243-85659389	yes	NT		Y					Y								Y		Y	Y		Y										Y																				Y			Y							
chr3:85745923-85746053	no	NT																																					Y											Y											

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr3:86347615-86347635	no	VAL																																	Y				Y	Y					Y					Y												
chr3:89104976-89105036	no	NT	Y						Y	Y	Y				Y	Y										Y				Y					Y		Y					Y			Y	Y		Y			Y	Y			Y	Y	Y	Y	Y			
chr3:96281050-96281071	no	NT													Y																																													
chr4:101807126-101807182	no	NT											Y																																															
chr4:110467753-110467772	no	NT			Y				Y														Y	Y																			Y																	
chr4:114129384-114129433	no	VAL																																		Y																								
chr4:114982851-114982960	no	NT									Y																																						Y											
chr4:115352153-115352365	no	VAL							Y	Y					Y	Y				Y			Y			Y	Y	Y	Y	Y	Y											Y					Y								Y	Y			Y	Y
chr4:118984709-118984752	no	VAL							Y	Y									Y	Y					Y	Y	Y				Y	Y																									Y	Y		
chr4:127815709-127815944	no	NT					Y	Y					Y				Y	Y					Y			Y	Y	Y	Y		Y						Y			Y					Y								Y	Y						Y
chr4:129184820-129184884	no	NT											Y	Y								Y																												Y										
chr4:130581083-130581116	no	NT															Y		Y	Y																						Y																		
chr4:132401041-132401107	no	NT																	Y											Y		Y																												
chr4:133744070-133744358	no	NT			Y																																																							
chr4:134815876-134815981	no	VAL						Y							Y	Y	Y	Y	Y			Y	Y					Y														Y									Y				Y			Y		Y
chr4:138396777-138396798	no	NT					Y								Y			Y																											Y				Y						Y		Y			
chr4:14458257-14458496	no	NT											Y	Y													Y																																	
chr4:147825260-147825347	no	NT																																										Y																
chr4:149614954-149615049	no	VAL																																				Y																						
chr4:150497664-150497734	no	NT																																										Y																
chr4:15480895-15481107	no	VAL																										Y																																
chr4:154814724-154814944	no	NT																																				Y																						
chr4:159138560-159138636	no	VAL																																						Y																				
chr4:170518425-170518493	no	NT	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y					Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y					Y	Y	Y	Y	Y		Y		Y	Y	Y	Y	Y		Y	Y						Y	Y	Y	Y	Y	Y

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr4:174193819-174193998	no	NT					Y																																																							
chr4:175921998-175922080	no	VAL																																						Y						Y			Y													
chr4:17771649-17771787	yes	NT											Y																	Y	Y																															
chr4:190896692-190896740	yes	NT	Y		Y						Y		Y							Y				Y																		Y									Y							Y				Y	
chr4:22531092-22531222	no	NT																																						Y																						
chr4:27493740-27493824	no	SOM																															Y																													
chr4:28930949-28930973	yes	NT				Y	Y	Y			Y	Y	Y	Y			Y	Y	Y	Y																					Y	Y					Y	Y						Y			Y	Y	Y	Y		
chr4:29089680-29089811	no	VAL																											Y				Y																													
chr4:29503814-29503860	no	NT																																			Y																									
chr4:32381031-32381053	no	VAL																																										Y	Y																	
chr4:37628485-37628504	no	NT															Y			Y																	Y											Y			Y												
chr4:46662763-46662888	no	NT									Y																																																			
chr4:46702417-46702539	yes	NT	Y	Y			Y		Y	Y	Y				Y	Y	Y	Y	Y	Y			Y		Y	Y		Y		Y				Y	Y						Y	Y		Y		Y										Y	Y			Y	Y	
chr4:53785689-53785800	no	NT																																											Y																
chr4:54703480-54703506	no	NT																																											Y			Y						Y		Y					
chr4:5718723-5718810	no	NT																																								Y			Y	Y														
chr4:57191791-57191821	no	VAL										Y																																												Y					
chr4:59887131-59887156	no	NT											Y																											Y	Y																				
chr4:69908523-69908669	no	NT	Y																																																										
chr4:71127432-71127937	no	NT					Y					Y				Y					Y	Y				Y							Y				Y																					Y	Y	Y	Y
chr4:73134152-73134331	no	NT																																													Y													
chr4:85536442-85536463	no	NT																								Y								Y		Y	Y							Y																	
chr4:86518471-86518564	no	NT																		Y																																									
chr4:86826980-86827140	no	NT								Y	Y	Y								Y	Y																															Y									

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr4:98322709-98322838	no	VAL			Y																																		Y	Y										Y	Y							Y	Y		
chr5:105963264-105963422	no	VAL																Y						Y																						Y							Y									
chr5:106137545-106137920	no	NT																					Y																																							
chr5:109508117-109508157	no	VAL		Y			Y	Y		Y	Y	Y									Y	Y			Y								Y						Y	Y	Y	Y		Y	Y								Y	Y								
chr5:109595429-109595577	yes	NT		Y	Y																	Y														Y	Y	Y	Y												Y						Y					
chr5:117037426-117037522	no	VAL																					Y																																							
chr5:119494067-119494227	no	NT	Y	Y																																	Y	Y													Y	Y											
chr5:120957604-120957693	no	NT		Y																																																										
chr5:124182992-124183040	no	VAL	Y	Y																																																										
chr5:125067940-125068094	no	NT													Y	Y				Y		Y			Y					Y											Y													Y								Y	
chr5:130285280-130285325	no	VAL																																														Y																
chr5:13284757-13284903	no	VAL																	Y	Y																																										
chr5:137158036-137158057	no	NT																			Y	Y					Y	Y																																		
chr5:140466580-140466970	no	NT																																																							Y			
chr5:144989548-144989575	no	NT																																							Y	Y																						
chr5:155266250-155266442	no	NT																		Y																	Y							Y																				
chr5:155552441-155552579	yes	NT	Y	Y	Y		Y	Y			Y	Y						Y	Y	Y			Y	Y				Y	Y								Y		Y				Y			Y		Y			Y					Y				Y		
chr5:155831416-155831530	no	VAL																																			Y																									
chr5:157799629-157799867	no	NT					Y						Y		Y		Y	Y													Y	Y									Y						Y		Y				Y			Y						
chr5:159283740-159283770	yes	NT			Y	Y		Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y									Y	Y					Y	Y			Y	Y	Y	Y	Y				Y	Y
chr5:16935683-16935903	no	NT																																													Y													
chr5:23059115-23059142	no	VAL																																									Y																	
chr5:23106031-23106067	no	VAL																																																										Y
chr5:23447062-23447101	no	NT			Y																																																							

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr5:30729512-30729536	no	VAL							Y	Y																																																			
chr5:31011736-31011917	no	NT																																																							Y			
chr5:31018443-31018643	no	NT																																																							Y			
chr5:33669054-33669102	yes	NT	Y	Y		Y	Y	Y		Y		Y	Y	Y		Y	Y				Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y			Y	Y					Y	Y					Y	Y	Y				Y		Y	Y		Y	Y		Y	Y
chr5:33872919-33872986	no	VAL																																																								Y		
chr5:39961866-39962065	no	NT		Y																	Y																							Y													Y				
chr5:57315273-57315368	no	VAL			Y													Y																								Y										Y										
chr5:67621052-67621100	no	NT			Y																																																										
chr5:72241454-72241531	no	VAL	Y																																																											
chr5:86963690-86963732	no	NT	Y																																																											
chr5:88386287-88386317	no	NT																																											Y			Y													
chr6:100095441-100095466	no	NT			Y																																																									
chr6:102647736-102647792	no	NT																																											Y																
chr6:115457417-115457507	no	NT								Y																																Y			Y															
chr6:121229783-121229876	no	NO																																						Y																				
chr6:125733748-125733786	no	VAL																				Y												Y																												
chr6:129510682-129510752	no	NO																				Y																																								
chr6:13298995-13299019	yes	NT				Y			Y	Y	Y	Y		Y		Y		Y	Y		Y	Y			Y	Y	Y	Y	Y	Y		Y						Y					Y				Y					Y					Y	Y			Y	Y
chr6:13795536-13795585	no	NO																												Y																																
chr6:142164195-142164214	no	NT								Y							Y													Y	Y																															
chr6:145542512-145542570	no	NT							Y	Y				Y	Y		Y	Y								Y	Y	Y	Y	Y																				Y		Y			Y		Y	Y				Y
chr6:145659993-145660073	no	VAL												Y																																														
chr6:148205263-148205300	no	NT				Y		Y	Y		Y	Y			Y					Y		Y																				Y	Y				Y					Y									
chr6:149518735-149518830	no	VAL			Y																																Y																								

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr6:152539226-152539433	no	NT									Y																																																			
chr6:154626089-154626272	no	NT																															Y																						Y			Y					
chr6:157441855-157441882	no	NT			Y																																																									
chr6:160760946-160760979	no	NT	Y																																																											
chr6:162209303-162209522	no	VAL																																															Y												
chr6:163517682-163517904	no	VAL						Y																																																						
chr6:34994940-34995044	no	VAL	Y	Y																																																										
chr6:45433804-45433838	no	NT																																				Y	Y																					Y			
chr6:56837069-56837268	no	VAL																																				Y																								
chr6:63718540-63718606	no	NT																																																									Y		
chr6:64669375-64669605	no	NT																										Y																																	
chr6:6542087-6542230	no	VAL													Y	Y																																													
chr6:69255426-69255464	yes	NT										Y																Y																Y			Y														Y	Y
chr6:77436629-77436653	no	NT									Y	Y																																				Y													
chr6:78706363-78706406	no	NO						Y																																																						
chr6:83734729-83734764	no	NT																																																							Y	Y		
chr6:84122338-84122364	no	NT											Y																																															
chr6:87409145-87409167	no	NT																																											Y					Y											
chr6:91726192-91726242	no	NT			Y		Y	Y	Y	Y		Y				Y			Y	Y	Y	Y	Y		Y	Y	Y			Y	Y	Y	Y	Y		Y		Y	Y	Y		Y	Y	Y		Y	Y	Y	Y				Y	Y	Y	Y	Y	Y			Y	Y
chr6:91824376-91824530	no	NT																																		Y																									
chr6_random:396637-396713	no	NT				Y																																																								
chr7:111020136-111020166	yes	NT									Y	Y									Y					Y																																				
chr7:111656690-111656918	no	NT													Y																																														
chr7:112576705-112576853	no	NT													Y	Y					Y	Y																			Y	Y	Y	Y																		

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	99122BL	99122GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr7:117150490-117150516	no	NT													Y																																													
chr7:119461494-119461572	no	VAL																Y																																												
chr7:124238514-124238549	no	VAL							Y																																						Y															
chr7:12691792-12691995	no	NT															Y												Y	Y																	Y															
chr7:127718046-127718066	yes	NT	Y	Y			Y						Y	Y		Y					Y	Y			Y	Y	Y			Y					Y									Y							Y	Y										
chr7:136334944-136335159	no	VAL																																				Y																								
chr7:144539765-144539800	no	NT																																	Y	Y																										
chr7:19192160-19192376	no	NT													Y																																															
chr7:24244271-24244366	no	NT																																																				Y														
chr7:24378775-24378860	no	VAL																																						Y																							
chr7:26761433-26761519	no	VAL		Y																																		Y																								
chr7:30780700-30780733	no	NT																																															Y													
chr7:38583812-38584081	no	NT											Y																																																
chr7:46316627-46316861	no	VAL								Y																Y						Y										Y	Y					Y	Y													Y	Y
chr7:47831356-47831569	no	VAL																																																						Y				
chr7:49162166-49162233	no	NT																												Y																															
chr7:51553145-51553249	no	VAL																																															Y											
chr7:51695173-51695418	no	NT											Y																																																
chr7:51946310-51946362	no	VAL																																																							Y			
chr7:63723735-63723960	no	NT																																Y																										
chr7:69826575-69826620	yes	NT	Y	Y												Y		Y	Y	Y					Y									Y											Y						Y											
chr7:7985539-7985683	no	NT											Y																																																
chr7:83519927-83520037	no	NT																					Y																		Y		Y	Y				Y		Y				Y						
chr7:84739367-84739590	yes	NT		Y			Y				Y	Y						Y	Y		Y		Y		Y		Y	Y	Y					Y			Y				Y			Y											Y	Y	Y	Y	Y		

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	99122BL	99122GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr7:84882678-84882703	no	VAL								Y	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y				Y			Y										Y	Y		Y									Y	Y										Y
chr7:9258874-9259117	no	NT																			Y																																									
chr7:92695457-92695479	no	NT	Y	Y					Y	Y		Y															Y																																Y		
chr7:92984392-92984549	no	NT	Y	Y																															Y	Y			Y	Y					Y	Y				Y	Y	Y										
chr7:9476039-9476267	no	NT											Y																																																	
chr7:9660696-9660913	no	VAL																																				Y						Y							Y											
chr8:112468405-112468606	no	NT					Y																																																							
chr8:115842301-115842471	no	VAL																																																		Y										
chr8:115857721-115857746	no	NT								Y														Y				Y					Y				Y						Y	Y										Y					Y			
chr8:119790736-119790810	no	VAL							Y	Y			Y	Y																																					Y		Y									
chr8:24982081-24982301	yes	NT																																																								Y				
chr8:25378970-25379017	no	NT																																			Y	Y											Y	Y	Y											
chr8:25559480-25559554	yes	NT																																					Y	Y									Y													
chr8:31119927-31119974	no	SOM																																																		Y										
chr8:39695616-39695711	no	NT																												Y																																
chr8:58510494-58510518	no	NT	Y	Y	Y																														Y	Y									Y	Y			Y	Y												
chr8:62277701-62277763	no	NT									Y																																																			
chr8:62844225-62844348	no	NT			Y																																																									
chr8:75550377-75550604	no	NT																															Y																													
chr8:75886282-75886332	no	VAL	Y	Y							Y	Y											Y																										Y	Y					Y							
chr8:84736577-84736810	yes	NT																															Y	Y																												
chr8:85591163-85591186	no	NT																				Y				Y	Y	Y																											Y	Y						
chr9:100434506-100434526	no	VAL																															Y																													
chr9:113366231-113366373	no	VAL																																													Y															

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	9912BL	9912GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr9:120434331-120434422	no	NT																																						Y	Y																									
chr9:134301794-134301871	no	NT			Y																																																											
chr9:17183067-17183271	no	NT																																																							Y										
chr9:29082640-29082672	no	VAL																				Y																																																	
chr9:29250391-29250553	no	NT			Y																																																											
chr9:30485866-30485884	no	NT																																	Y	Y		Y	Y	Y																				Y	Y							Y	Y		
chr9:30664377-30664435	yes	NT	Y	Y													Y	Y																									Y			Y	Y																			Y							Y			
chr9:32453763-32453886	no	NT		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y				Y	Y	Y		Y	Y	Y	Y				Y	Y	Y	Y	Y	Y		Y				Y		Y		Y	Y	Y	Y	Y		
chr9:32551984-32552030	no	VAL																												Y																																		
chr9:32747350-32747392	no	NT			Y																																																											
chr9:5013608-5013844	no	NT																																							Y																			
chr9:5481411-5481438	yes	NT																	Y							Y	Y																													Y										
chr9:6280622-6280846	no	VAL							Y	Y			Y							Y						Y		Y				Y		Y								Y					Y	Y					Y		Y						Y	Y
chr9:67907361-67907442	no	NT																																					Y																									
chr9:71245063-71245086	no	NT																											Y																																			
chr9:7416394-7416601	no	NT																																Y																												
chr9:89081242-89081468	no	VAL										Y																																																				
chr10:108547069-108547148	no	NO													Y																																																	
chr10:110189959-110190006	no	VAL								Y																																																						
chr10:117310075-117310215	no	NT							Y	Y																																																					
chr10:120990265-120990468	no	NT			Y																																																											
chr10:126973605-126973629	no	NT			Y																														Y		Y																									
chr10:128974913-128974932	no	NT																			Y																																											
chr10:131358226-131358398	no	NT																																																												Y	Y

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	9912BL	9912GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr10:14350646-14350848	no	VAL																															Y																													
chr10:19256658-19256717	no	NT	Y														Y	Y															Y							Y														Y					Y	Y
chr10:21252682-21252900	no	VAL				Y																																																								
chr10:23882003-23882062	no	NT				Y									Y								Y	Y													Y								Y										Y					
chr10:25747601-25747701	no	VAL				Y																				Y	Y														Y					Y	Y											Y		
chr10:27815768-27815833	no	NO				Y	Y	Y					Y	Y															Y	Y				Y							Y	Y	Y	Y					Y				Y	Y		Y		Y				
chr10:39081744-39081806	no	NT																																						Y																				
chr10:39091250-39091311	no	VAL									Y																																																	
chr10:42279718-42279747	no	VAL			Y																																							Y																
chr10:52862278-52862468	no	VAL																															Y	Y																										
chr10:54542518-54542866	no	NT																											Y																															
chr10:85532615-85532682	no	NT																															Y																											
chr10:94282442-94282466	no	VAL																																		Y																								
chr11:104061320-104061340	no	NT										Y																																		Y														
chr11:108214028-108214106	no	VAL	Y																																																									
chr11:116370732-116370833	no	VAL																														Y																												
chr11:127872850-127872894	yes	NT													Y	Y																																												
chr11:15336822-15336865	no	NT												Y																		Y																													Y	Y
chr11:29141007-29141144	no	VAL																																		Y	Y	Y																						
chr11:31559603-31559794	no	NO																																												Y														
chr11:33626699-33626762	no	VAL	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y		Y	Y	Y	Y						Y	Y		Y	Y	Y	Y	Y	Y		Y							Y	Y	Y	Y	Y		Y	Y	Y	Y			Y		Y		Y	Y		Y		Y
chr11:39181705-39181730	no	NT																																	Y	Y																								
chr11:41788252-41788383	no	NO																								Y																																		
chr11:41826441-41826495	no	NT			Y																																Y										Y													

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	9912BL	9912GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr11:41948200-41948296	no	NT																	Y	Y																																										
chr11:50356447-50356551	no	NT																																															Y													
chr11:54920535-54920641	no	NT													Y		Y	Y							Y																																			
chr11:57496250-57496396	no	VAL																																									Y																	
chr11:82630210-82630442	no	VAL																																									Y																	
chr11:83329907-83330012	yes	NT												Y	Y																																													
chr11:94324068-94324259	no	NT			Y		Y				Y			Y			Y	Y						Y		Y		Y		Y	Y		Y								Y									Y		Y		Y		Y		Y		Y	Y
chr12:102122467-102122608	no	SOM																																						Y																					
chr12:103717906-103718121	no	NT																																								Y	Y																		
chr12:116298777-116298839	no	VAL				Y									Y					Y				Y			Y						Y													Y	Y													Y	Y
chr12:125368868-125368892	no	NT			Y				Y	Y							Y	Y											Y	Y					Y	Y			Y	Y		Y			Y				Y	Y			Y	Y					Y	Y		
chr12:19811429-19811654	no	NT																Y													Y																													
chr12:20036450-20036477	no	VAL																														Y																													
chr12:24759928-24759984	no	VAL																															Y																								Y				
chr12:25666910-25667117	no	VAL																																		Y																									
chr12:36574211-36574301	no	NT																																					Y																					Y		
chr12:56390536-56390697	no	NT												Y																																														
chr12:58091068-58091097	no	VAL																				Y																																						
chr12:63975476-63975519	no	VAL																																																Y												
chr12:7520556-7520612	no	NT				Y		Y																																																						
chr12:77862125-77862163	no	NT		Y																							Y										Y								Y	Y						Y								Y		
chr12:80254951-80255024	no	VAL																																																Y										
chr12:82535354-82535374	no	NT																																							Y	Y																				
chr12:82901370-82901615	no	SOM																																																Y										

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr12:85918162-85918208	no	NT																																	Y	Y															Y	Y										
chr12:95889620-95889674	no	NT																			Y												Y	Y																												
chr13:106234148-106234356	no	NT																																										Y									Y										
chr13:20412688-20412913	no	NT			Y																														Y																											
chr13:32288166-32288373	no	VAL			Y																																																									
chr13:39858824-39858866	no	VAL																																																		Y										
chr13:48419714-48419762	no	VAL																																																	Y											
chr13:51292232-51292262	no	VAL											Y																																																	
chr13:60579764-60579806	no	VAL																																																				Y								
chr13:60632488-60632712	no	VAL	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y	Y	Y	Y	Y	Y						Y	Y				Y					Y		Y				Y						Y		Y	Y	Y	Y										Y
chr13:62587154-62587188	no	VAL						Y																																																						
chr13:70798788-70798875	no	NO																																												Y														
chr13:74894408-74894431	no	NT																							Y																					Y			Y											
chr13:75682497-75682705	no	SOM																												Y																														
chr13:76953334-76953523	no	NT																													Y							Y										Y												
chr13:90424304-90424555	no	NT																			Y				Y	Y					Y									Y																Y				
chr14:24062538-24062754	no	NT																			Y																																							
chr14:24381245-24381447	no	VAL														Y																																												
chr14:25076474-25076499	no	SOM																										Y																																
chr14:25857189-25857321	no	NT			Y																																																							
chr14:27011899-27011949	no	VAL																																										Y																
chr14:29040239-29040350	no	NT																		Y																																								
chr14:29595000-29595039	no	VAL				Y																	Y							Y					Y	Y													Y	Y												
chr14:29731440-29731482	no	NO														Y																																												

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	9912BL	9912GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr14:38944834-38944897	no	VAL			Y				Y	Y			Y	Y	Y	Y		Y	Y	Y					Y	Y					Y		Y									Y			Y			Y									Y	Y	Y	Y		
chr14:43128864-43128894	no	VAL																																						Y																						
chr14:50481665-50481771	no	VAL																																					Y																							
chr14:58592039-58592175	no	VAL							Y																																Y																					
chr14:65223160-65223197	no	SOM																												Y																																
chr14:66489476-66489628	yes	NT		Y																													Y	Y			Y	Y			Y	Y						Y	Y		Y		Y	Y										
chr14:70267533-70267700	no	VAL				Y		Y			Y												Y						Y																		Y	Y											Y	Y				
chr14:70922004-70922023	no	NT																																																	Y													
chr14:71403345-71403472	yes	NT	Y																																																											
chr14:80546868-80546971	no	NT			Y																																																									
chr14:85219341-85219394	no	NT																															Y	Y																									Y			
chr14:85451520-85451555	yes	NT					Y	Y	Y	Y	Y	Y	Y	Y			Y		Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y									Y	Y		Y			Y						Y	Y			Y	Y				
chr14:86022244-86022280	no	VAL																															Y														Y													
chr14:87485244-87485431	no	NT																																	Y																									
chr14:89745839-89745948	no	NT																																											Y															
chr14:91124198-91124221	no	NT										Y	Y														Y																				Y													
chr14:97063375-97063594	no	NT																															Y																											
chr15:18337608-18337710	no	VAL																																																						Y				
chr15:18376666-18376802	no	NT													Y						Y					Y											Y									Y				Y					Y				Y		Y	
chr15:31466165-31466307	no	NT																				Y																																						
chr15:36126408-36126552	no	NT			Y																																																								
chr15:45294469-45294634	no	VAL														Y		Y																		Y			Y	Y					Y						Y		Y									
chr15:49631104-49631149	no	NT																																			Y								Y															
chr15:59116219-59116265	no	NT			Y																																Y																			Y				

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	9912BL	9912GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chr15:60893558-60893708	no	VAL		Y																																																										
chr15:89771542-89771607	no	VAL																				Y																																								
chr15:96000976-96001016	no	NT																																																							Y					
chr16:17662626-17662806	yes	NT		Y									Y	Y																										Y																				Y		
chr16:27333356-27333429	no	NT	Y	Y																																	Y	Y																								
chr16:32011398-32011490	no	VAL																																					Y																							
chr16:5278654-5278779	no	NT			Y																																Y	Y																								
chr16:59636689-59636734	yes	NT		Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y				Y	Y		Y		Y	Y	Y		Y	Y			Y	Y	Y		Y	Y	Y	Y			Y	Y	Y	Y	Y	Y		Y	Y	Y		
chr16:64469011-64469053	no	NT																																																									Y			
chr16:6733832-6733886	no	NT															Y	Y	Y							Y			Y					Y									Y														Y					
chr16:74950652-74950870	no	NT	Y																																																											
chr17:22177186-22177417	no	NT																																												Y																
chr17:49347950-49348173	no	NT		Y																																																										
chr17:57033826-57033945	no	VAL																																											Y																	
chr17:70588305-70588349	no	NO								Y																																																				
chr18:10758675-10758762	no	VAL																																																		Y										
chr18:14880023-14880178	no	VAL																																												Y																
chr18:20157168-20157347	no	NT																													Y	Y																														
chr18:30944185-30944206	no	NT																																																							Y	Y				
chr18:41876107-41876129	no	NO																																									Y																	Y		
chr18:47470187-47470229	no	NT							Y										Y																	Y																			Y							
chr18:48351250-48351421	no	NT					Y						Y																																													Y				
chr18:51626703-51626745	no	NT																																														Y														
chr18:52501441-52501480	no	VAL																																											Y																	

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr18:55721740-55721763	no	NT																																			Y																									
chr18:57617370-57617535	no	NT				Y										Y																		Y							Y	Y																				
chr18:57730944-57731043	yes	NT					Y	Y																																																						Y
chr18:66856768-66856820	no	SOM																																		Y																										
chr18:67953636-67953825	no	NT													Y																																															
chr18:98166-98349	no	NT														Y						Y																					Y																			
chr19:23494217-23494248	no	NO																																																		Y										
chr19:39048243-39048283	no	NT																																															Y													
chr19:43036628-43036872	no	NT																			Y																																									
chr19:58863018-58863157	no	VAL																																														Y														
chr19_random:88931-89113	no	NT																																																Y												
chr20:1178999-1179064	no	NT	Y	Y																																																										
chr20:19175274-19175293	no	NO			Y																															Y																										
chr20:29937984-29938122	no	NT																																																							Y					
chr20:30134102-30134126	no	NT										Y			Y												Y		Y																						Y											
chr20:42840322-42840344	no	NO												Y																																																
chr20:52966549-52966725	no	VAL																		Y									Y																																	
chr20:59405833-59406001	no	VAL				Y	Y																																																							
chr20:7240206-7240332	no	NT												Y	Y	Y		Y																										Y																		
chr20:9425805-9425923	no	VAL	Y	Y																																																										
chr21:15179509-15179631	no	VAL																					Y																																							
chr21:18004002-18004221	no	NT																																																											Y	Y
chr21:19953729-19953975	no	NT					Y																																																							
chr21:27902875-27902900	no	VAL																		Y																																										

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	00-43BL	00-43GBM	00-82BL	00-82MBM	02-07BL	02-07MBM	06-66BL	06-66MBM	07-65BL	07-65MBM	99-122BL	99-122GBM	99-24BL	99-24GBM	99-43BL	99-43GBM	99-69BL	99-69GBM	99-72BL	99-72MBM	ANCO100-Normal	ANCO100-Tumor	ANCO101-Normal	ANCO101-Tumor	ANCO102-Normal	ANCO102-Tumor	ANCO103-Normal	ANCO103-Tumor	ANCO104-Normal	ANCO104-Tumor	ANCO105-Normal	ANCO105-Tumor	ANCO106-Normal	ANCO106-Tumor	ANCO109-Normal	ANCO109-Tumor	ANCO110-Normal	ANCO110-Tumor	ANCO111-Normal	ANCO111-Tumor	ANCO112-Normal	ANCO112-Tumor	ANCO113-Normal	ANCO113-Tumor	ANCO114-Normal	ANCO114-Tumor	ANCO115-Normal	ANCO115-Tumor	ANCO118-Normal	ANCO118-Tumor	ANCO119-Normal	ANCO119-Tumor	ANCO120-Normal	ANCO120-Tumor	ANCO97-Normal	ANCO97-Tumor	ANCO98-Normal	ANCO98-Tumor	ANCO99-Normal	ANCO99-Tumor
chr21:27991045-27991228	yes	NT		Y	Y		Y	Y	Y	Y		Y													Y	Y			Y	Y	Y	Y									Y								Y	Y			Y	Y					Y	Y	Y	Y
chr21:28214703-28215093	no	NT			Y																					Y				Y													Y																			
chr21:29833004-29833062	no	NT																																					Y	Y																						
chr21:38835361-38835387	no	NT					Y																																																							
hr21_random:1312916-13129	no	NO	Y	Y	Y	Y			Y	Y		Y			Y	Y	Y	Y	Y	Y	Y	Y							Y								Y	Y	Y	Y	Y		Y	Y			Y				Y	Y			Y	Y	Y					
chr21_random:435178-43524	no	VAL	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y		Y	Y	Y	Y	Y	Y	Y	Y
chr22:15626058-15626096	no	NT																	Y						Y	Y																Y																			Y	
chr22:15733461-15733522	no	NT																																							Y					Y																
chr22:20816300-20816319	yes	NT			Y	Y	Y	Y	Y	Y									Y		Y	Y										Y	Y		Y	Y	Y		Y	Y	Y	Y	Y	Y							Y	Y	Y	Y								
chr22:21045586-21045636	no	NT	Y	Y		Y	Y		Y	Y	Y	Y					Y	Y	Y	Y			Y			Y	Y	Y	Y												Y	Y	Y	Y				Y			Y	Y			Y	Y	Y	Y		Y	Y	Y
chr22:26785840-26785877	no	NT							Y	Y	Y										Y		Y	Y																				Y																		
chr22:42655911-42655933	yes	NT							Y											Y	Y	Y	Y				Y							Y							Y	Y											Y									
chrX:100841322-100841344	no	NT				Y																											Y																													
chrX:101325923-101325977	yes	NT																																																			Y									
chrX:102249830-102249859	yes	NT																										Y													Y	Y											Y	Y	Y		Y				Y	Y
chrX:111116139-111116181	no	NT										Y	Y	Y			Y		Y	Y													Y	Y																						Y						
chrX:111444196-111444334	yes	NT											Y				Y						Y			Y			Y									Y	Y	Y		Y																				
chrX:115909316-115909350	no	VAL															Y	Y								Y																Y																				
chrX:119776694-119776740	no	VAL																	Y	Y						Y	Y	Y	Y		Y	Y		Y		Y																	Y	Y	Y	Y	Y	Y				Y
chrX:121650485-121650541	no	NT																																						Y																						
chrX:129352843-129352901	yes	NT							Y	Y				Y														Y			Y	Y	Y	Y	Y	Y							Y	Y	Y	Y				Y												
chrX:14061858-14061931	no	VAL																							Y	Y					Y	Y			Y		Y						Y	Y	Y	Y			Y											Y		
chrX:140913804-140914022	no	NT			Y																																																									
chrX:142745608-142745682	no	NT																																							Y																					

Blat Coordinates (Hg18)	In ABI Sequencing Project?	PCR Validation	0043BL	0043GBM	0082BL	0082MBM	0207BL	0207MBM	0666BL	0666MBM	0765BL	0765MBM	9912BL	99122GBM	9924BL	9924GBM	9943BL	9943GBM	9969BL	9969GBM	9972BL	9972MBM	ANCO100Normal	ANCO100Tumor	ANCO101Normal	ANCO101Tumor	ANCO102Normal	ANCO102Tumor	ANCO103Normal	ANCO103Tumor	ANCO104Normal	ANCO104Tumor	ANCO105Normal	ANCO105Tumor	ANCO106Normal	ANCO106Tumor	ANCO109Normal	ANCO109Tumor	ANCO110Normal	ANCO110Tumor	ANCO111Normal	ANCO111Tumor	ANCO112Normal	ANCO112Tumor	ANCO113Normal	ANCO113Tumor	ANCO114Normal	ANCO114Tumor	ANCO115Normal	ANCO115Tumor	ANCO118Normal	ANCO118Tumor	ANCO119Normal	ANCO119Tumor	ANCO120Normal	ANCO120Tumor	ANCO97Normal	ANCO97Tumor	ANCO98Normal	ANCO98Tumor	ANCO99Normal	ANCO99Tumor
chrX:149949996-149950031	no	VAL												Y																																														
chrX:17517204-17517453	no	VAL																																											Y																	
chrX:28536772-28536832	no	NT		Y																	Y	Y			Y										Y	Y				Y				Y																	Y	Y
chrX:54709844-54709875	no	NT			Y																																																									
chrX:57823142-57823169	no	VAL		Y	Y																																Y	Y		Y					Y	Y													Y	Y		
chrX:58070249-58070338	no	NT																																											Y	Y																
chrX:58179599-58179802	no	NT																																																		Y										
chrX:75866019-75866229	yes	NT	Y	Y	Y	Y	Y		Y	Y		Y	Y	Y	Y	Y		Y	Y	Y	Y	Y			Y	Y	Y	Y		Y	Y	Y	Y	Y			Y	Y				Y	Y	Y		Y			Y		Y	Y	Y		Y	Y	Y	Y	Y		Y	Y
chrX:76328254-76328308	yes	NT								Y																Y								Y																												
chrX:84601947-84601987	no	NT																																													Y	Y														
chrX:85230508-85230654	no	NT																										Y																																		
chrX:85668440-85668504	no	NT			Y	Y														Y																																										Y
chrX:91634589-91634656	no	NO																																						Y																						
chrX:98150768-98150952	no	VAL																		Y										Y		Y	Y	Y				Y	Y	Y						Y			Y	Y								Y	Y				
chrY:12820652-12820695	no	NT													Y																																															

Appendix III – Novel Alu insertions sites from pyrosequencing

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr1:111925199-111925371	no	no	C	N/A	NT													Y							Y
chr1:158767536-158767614	no	no	C	N/A	NT	Y																			
chr1:161580944-161581067	no	no	+	N/A	NT										Y						Y				
chr1:162389496-162389533	no	no	+	N/A	NT																			Y	
chr1:177214781-177214808	no	no	+	N/A	NT							Y				Y								Y	
chr1:180153094-180153297	no	no	+	N/A	NO						Y														
chr1:180687464-180687491	no	no	C	N/A	NT				Y	Y			Y			Y	Y			Y	Y				
chr1:185850461-185850505	no	no	+	N/A	NT			Y																	
chr1:187513630-187513662	no	no	+	N/A	NT												Y					Y	Y		
chr1:188149930-188149973	no	no	+	N/A	NT									Y	Y	Y									Y
chr1:189141069-189141182	no	no	C	N/A	NT								Y			Y						Y	Y	Y	
chr1:196735963-196736106	no	no	C	N/A	NT																			Y	
chr1:20129670-20129706	no	no	C	N/A	NT				Y																
chr1:213351902-213352085	no	no	C	N/A	NT													Y				Y			
chr1:213501317-213501531	no	no	+	N/A	NT																			Y	
chr1:215539704-215539924	no	no	C	N/A	NT	Y																			
chr1:216068468-216068734	no	no	C	N/A	NT										Y										
chr1:216248548-216248774	no	no	+	N/A	NT																		Y		
chr1:219020969-219020995	no	no	C	N/A	NT				Y							Y									
chr1:231585701-231585767	no	no	+	N/A	NT		Y					Y								Y					
chr1:232189880-232189985	no	no	C	N/A	NT																Y				
chr1:232206337-232206464	no	no	+	N/A	VAL								Y												
chr1:232872336-232872460	no	no	C	N/A	VAL										Y										
chr1:23853639-23853725	no	no	+	N/A	NT						Y	Y	Y	Y			Y								
chr1:239975198-239975242	no	no	+	N/A	NT											Y			Y						
chr1:244055928-244055956	no	no	C	N/A	NT		Y	Y	Y	Y	Y		Y			Y	Y			Y	Y	Y	Y	Y	

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr1:28618128-28618184	no	no	C	N/A	NT												Y								
chr1:38220248-38220304	no	no	+	N/A	NT											Y	Y								
chr1:42162878-42162905	no	no	+	N/A	NT																	Y		Y	
chr1:43981406-43981511	no	no	+	N/A	NT							Y								Y		Y	Y		
chr1:48745027-48745123	no	no	+	N/A	NT	Y	Y	Y																	
chr1:49634408-49634580	no	no	+	N/A	VAL								Y												
chr1:59364055-59364076	no	no	C	N/A	NT					Y	Y					Y					Y		Y		
chr1:69863645-69863706	no	no	C	N/A	NT						Y			Y	Y					Y					
chr1:79103784-79103853	no	no	C	N/A	NT										Y										
chr1:81296704-81296723	no	no	C	N/A	NT								Y	Y	Y								Y		
chr1:8369215-8369420	no	no	C	N/A	VAL																Y				
chr1:89990070-89990121	no	no	C	N/A	NT									Y										Y	
chr2:11271055-11271158	no	no	+	N/A	NT										Y	Y			Y				Y		
chr2:112718007-112718077	no	no	+	N/A	NT			Y						Y											
chr2:113822659-113822874	no	no	+	N/A	NT							Y													
chr2:121801718-121801748	no	no	C	N/A	NT	Y						Y													
chr2:139097056-139097257	no	no	+	N/A	NT						Y												Y		
chr2:158990821-158990881	no	no	+	N/A	NT										Y							Y	Y		
chr2:173024970-173025005	no	no	C	N/A	NT							Y												Y	
chr2:175444071-175444191	no	no	+	N/A	NT	Y								Y	Y				Y	Y	Y	Y			
chr2:176436482-176436543	no	no	+	N/A	NT	Y	Y	Y			Y				Y										
chr2:185469802-185469891	no	no	C	N/A	VAL																				Y
chr2:192690925-192691026	no	no	C	N/A	VAL		Y																		
chr2:194274764-194274894	no	no	C	N/A	VAL						Y														
chr2:195286489-195286538	no	no	+	N/A	NT	Y	Y												Y						
chr2:205360026-205360170	no	no	+	N/A	VAL		Y																		
chr2:209226532-209226630	no	no	+	N/A	NT											Y									
chr2:212416577-212416601	no	no	+	N/A	NT							Y													
chr2:214959892-214959960	no	no	C	N/A	NT			Y																	

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr2:226678328-226678417	no	no	C	N/A	NT	Y																			
chr2:30523467-30523497	no	no	+	N/A	NT	Y																			
chr2:34108885-34108919	no	no	C	N/A	NT	Y				Y		Y	Y			Y	Y						Y		
chr2:36029852-36029998	no	no	C	N/A	VAL										Y										
chr2:54178549-54178777	no	no	C	N/A	NT			Y																	
chr2:6434807-6434884	no	no	+	N/A	NT	Y								Y											
chr2:70651208-70651314	no	no	C	N/A	NT	Y												Y	Y					Y	
chr2:80888939-80889073	no	no	+	N/A	NT														Y						
chr2:83685232-83685286	no	no	C	N/A	NT				Y				Y					Y							Y
chr2:83978663-83978682	no	no	+	N/A	NT															Y					
chr3:101138075-101138224	no	no	C	N/A	NT	Y					Y														
chr3:109376012-109376112	no	no	C	N/A	VAL																Y				
chr3:11675109-11675207	no	no	C	N/A	NT	Y																			
chr3:123928333-123928437	no	no	+	N/A	NT			Y		Y	Y														
chr3:124609216-124609319	no	no	+	N/A	NT										Y					Y				Y	
chr3:146552104-146552153	no	no	+	N/A	NT															Y					
chr3:150582304-150582523	no	no	C	N/A	VAL		Y																		
chr3:156865540-156865565	no	no	+	N/A	NT															Y					
chr3:168657670-168657919	no	no	C	N/A	NT																			Y	
chr3:169681758-169681806	no	no	C	N/A	NT				Y									Y	Y	Y					
chr3:183781667-183781707	no	no	+	N/A	NT	Y		Y																	
chr3:189644775-189644810	no	no	+	N/A	NT					Y															
chr3:192482145-192482308	no	no	C	N/A	NT			Y		Y		Y	Y		Y					Y		Y			
chr3:22348050-22348298	no	no	C	N/A	NT			Y																	
chr3:23124169-23124316	no	no	+	AluSp	NT														Y						
chr3:25253452-25253587	no	no	C	N/A	NT	Y		Y																	
chr3:26881574-26881729	no	no	C	N/A	VAL																Y				
chr3:29214879-29214903	no	no	+	N/A	NT											Y		Y	Y						
chr3:44315277-44315306	no	no	+	N/A	NT																	Y			

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr3:45517657-45517700	no	no	C	N/A	NT	Y					Y									Y	Y	Y			
chr3:63183911-63184052	no	no	+	N/A	NT						Y								Y						
chr3:6611917-6611967	no	no	C	N/A	NT			Y																	
chr3:83285395-83285493	no	no	+	N/A	VAL																		Y		
chr3:89389479-89389529	no	no	+	N/A	NT																		Y		Y
chr4:110821142-110821177	no	no	C	N/A	NT			Y																	
chr4:111154410-111154474	no	no	C	N/A	NT										Y					Y	Y		Y		
chr4:11475088-11475112	no	no	C	N/A	NT	Y	Y																		
chr4:118008636-118008695	no	no	C	N/A	NT													Y							
chr4:118938842-118939014	no	no	+	N/A	VAL		Y																		
chr4:125878459-125878545	no	no	C	N/A	NT	Y										Y									Y
chr4:130285145-130285367	no	no	+	N/A	NT																	Y			
chr4:135404577-135404639	no	no	+	N/A	VAL																		Y		
chr4:136428009-136428052	no	no	C	N/A	NT	Y																			
chr4:136947023-136947062	no	no	C	N/A	NT			Y																	
chr4:138099998-138100028	no	no	+	N/A	NT															Y					
chr4:145614683-145614758	no	no	+	N/A	NT		Y		Y					Y	Y										Y
chr4:162576013-162576106	no	no	+	N/A	NT							Y	Y												
chr4:165264617-165264728	no	no	+	N/A	NT	Y	Y																		
chr4:166849460-166849543	no	no	+	N/A	VAL																Y				
chr4:169993363-169993389	no	no	+	N/A	NT							Y													
chr4:172683393-172683437	no	no	C	N/A	VAL		Y																		
chr4:180935245-180935477	no	no	+	N/A	NT												Y								
chr4:186598897-186598917	no	no	+	N/A	NT													Y	Y					Y	
chr4:187330281-187330481	no	no	+	N/A	NT										Y		Y		Y						Y
chr4:190906686-190906771	no	no	+	N/A	NT					Y															
chr4:232387-232512	no	no	C	N/A	NT							Y													
chr4:2541598-2541731	no	no	C	N/A	VAL																Y				
chr4:27929822-27929865	no	no	+	N/A	NT			Y			Y								Y					Y	

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr4:33688424-33688547	no	no	C	N/A	NT													Y							
chr4:36445814-36446014	no	no	+	N/A	VAL												Y								
chr4:39111366-39111453	no	no	C	N/A	NT																			Y	
chr4:41615571-41615699	no	no	C	N/A	NT																		Y		
chr4:42473064-42473097	no	no	C	N/A	VAL																Y				
chr4:49207751-49207807	no	no	+	N/A	NT													Y	Y			Y		Y	
chr4:53082770-53082786	no	no	+	N/A	NT															Y					
chr4:59634716-59634843	no	no	+	N/A	NT										Y										
chr4:61535440-61535663	no	no	+	N/A	NT													Y							
chr4:62665660-62665899	no	no	+	N/A	NT																			Y	
chr5:115526777-115526820	no	no	C	N/A	NT				Y									Y			Y				
chr5:118506192-118506405	no	no	+	N/A	NT																			Y	
chr5:119058422-119058448	no	no	C	N/A	NT															Y	Y				
chr5:160479405-160479424	no	no	+	N/A	NT							Y													
chr5:163526760-163526790	no	no	C	N/A	NT													Y							
chr5:16444702-16444788	no	no	+	N/A	NT													Y	Y						
chr5:17745271-17745305	no	no	+	N/A	NT											Y					Y	Y	Y		
chr5:18293212-18293324	no	no	+	N/A	NT														Y				Y		
chr5:4128834-4128919	no	no	C	N/A	NT								Y												
chr5:51454268-51454290	no	no	C	N/A	NT		Y	Y																	
chr5:55547613-55547683	no	no	C	N/A	NT			Y																	
chr5:58552387-58552524	no	no	C	N/A	NT																	Y			
chr5:64647218-64647280	no	no	+	N/A	NT			Y																	
chr5:74359550-74359607	no	no	C	N/A	NT	Y																			
chr5:75310917-75311087	no	no	+	N/A	NT							Y				Y			Y						Y
chr5:7899040-7899069	no	no	C	N/A	NT					Y						Y	Y	Y	Y		Y				
chr5:83580574-83580593	no	no	C	N/A	NT		Y																Y		
chr5:84580146-84580237	no	no	C	N/A	NT						Y		Y		Y				Y		Y		Y		
chr5:89887522-89887618	no	no	+	N/A	NT			Y																	

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr5:98963674-98963773	no	no	C	N/A	NT	Y																			
chr5:99327805-99327931	no	no	C	N/A	NT				Y											Y					
chr6:104597347-104597376	no	no	+	N/A	NT															Y					
chr6:108951859-108951891	no	no	+	N/A	NT					Y															
chr6:110733362-110733447	no	no	+	N/A	NT	Y												Y	Y						
chr6:11159639-11159755	no	no	+	N/A	NT											Y			Y						
chr6:117862578-117862659	no	no	C	N/A	VAL						Y														
chr6:120770383-120770435	no	no	C	N/A	NT	Y																			
chr6:122957215-122957395	no	no	C	N/A	NT												Y								
chr6:123021692-123021800	no	no	C	N/A	NT								Y					Y						Y	Y
chr6:124906913-124907141	no	no	+	N/A	NT																		Y		
chr6:129087530-129087638	no	no	C	N/A	NT			Y																	
chr6:132057788-132057965	no	no	C	N/A	NT	Y																			
chr6:150578674-150578867	no	no	C	N/A	NT			Y																	
chr6:155957440-155957669	no	no	+	N/A	VAL																		Y		
chr6:156663843-156663868	no	no	+	N/A	NT													Y							
chr6:159086239-159086287	no	no	+	N/A	NT																	Y			
chr6:19007222-19007450	no	no	C	N/A	NT																	Y			
chr6:4144519-4144754	no	no	C	N/A	NT		Y																		
chr6:47780291-47780361	no	no	C	N/A	VAL		Y																		
chr6:56495431-56495535	no	no	+	N/A	VAL						Y														
chr6:57375418-57375441	no	no	+	N/A	NT					Y	Y			Y	Y										Y
chr6:5979754-5979775	no	no	+	N/A	NT			Y				Y								Y			Y		
chr6:69367782-69367834	no	no	C	N/A	NT																	Y			
chr6:74561479-74561571	no	no	+	N/A	NT		Y			Y				Y		Y							Y	Y	
chr6:93632767-93632869	no	no	+	N/A	NT					Y			Y	Y	Y			Y		Y	Y				
chr6:93909902-93910004	no	no	C	N/A	NT	Y																			
chr6:96223682-96223795	no	no	+	N/A	NT								Y						Y						
chr6:9809956-9810167	no	no	+	N/A	NT							Y													

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-12 2 BL	99-12 2 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr6:98358050-98358117	no	no	C	N/A	NT						Y					Y	Y						Y		Y
chr7:121134687-121134909	no	no	+	N/A	VAL												Y								
chr7:129162639-129162688	no	no	+	N/A	NT	Y																			
chr7:136020247-136020319	no	no	C	N/A	NT				Y									Y							
chr7:138154944-138154999	no	no	+	N/A	NT				Y												Y				
chr7:153559080-153559128	no	no	C	N/A	VAL								Y												
chr7:156223159-156223348	no	no	C	N/A	NT									Y	Y										
chr7:157141535-157141639	no	no	C	N/A	NT					Y					Y				Y			Y		Y	
chr7:18239601-18239642	no	no	C	N/A	VAL										Y										
chr7:38175725-38175765	no	no	C	N/A	NT								Y								Y				
chr7:43857102-43857337	no	no	C	N/A	NT									Y											
chr7:56465701-56465721	no	no	+	N/A	NT							Y													
chr7:61426325-61426526	no	no	+	N/A	NT																	Y			
chr7:66124034-66124091	no	no	C	N/A	NT	Y																			
chr7:76301636-76301741	no	no	C	N/A	NT													Y							
chr7:76818685-76818713	no	no	+	N/A	NT													Y							
chr7:77984440-77984459	no	no	C	N/A	NT					Y													Y		
chr7:79835225-79835432	no	no	C	N/A	NT															Y					
chr7:82268445-82268542	no	no	C	N/A	VAL														Y						
chr7:88218276-88218302	no	no	C	N/A	NT		Y				Y	Y		Y	Y						Y				
chr7:88322625-88322857	no	no	C	N/A	NT			Y																	Y
chr7:96905459-96905490	no	no	C	N/A	NT																	Y			
chr8:10363679-10363707	no	no	C	N/A	NT												Y						Y		
chr8:105969587-105969659	no	no	+	N/A	NT									Y	Y										
chr8:117935441-117935486	no	no	+	N/A	NT	Y																			
chr8:124842484-124842541	no	no	C	N/A	NT	Y																			
chr8:126893625-126893678	no	no	+	N/A	NT			Y																	
chr8:134236822-134236878	no	no	+	N/A	NT																			Y	
chr8:141430281-141430468	no	no	+	N/A	NT			Y																	

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr8:17380374-17380445	no	no	C	N/A	NT			Y	Y																
chr8:33881194-33881223	no	no	C	N/A	NT					Y	Y		Y			Y						Y			
chr8:35188433-35188579	no	no	+	N/A	NT			Y			Y			Y	Y			Y	Y						
chr8:3614315-3614382	no	no	+	N/A	NT																	Y	Y		
chr8:49775354-49775472	no	no	+	N/A	NT											Y						Y		Y	Y
chr8:49923605-49923819	no	no	+	N/A	NT			Y																	
chr8:50310047-50310173	no	no	+	N/A	NT			Y																	
chr8:52059359-52059397	no	no	C	N/A	NT							Y													
chr8:70510929-70511088	no	no	+	N/A	NT	Y																			
chr8:71937346-71937378	no	no	C	N/A	NT								Y										Y		
chr8:72368435-72368610	no	no	+	N/A	NT																				
chr8:72672759-72672837	no	no	+	N/A	VAL																Y				
chr8:74175694-74175716	no	no	+	N/A	NT									Y	Y										
chr8:74445869-74445937	no	no	+	N/A	NT				Y								Y			Y					
chr8:76114604-76114643	no	no	+	N/A	NT										Y										
chr8:77410914-77410984	no	no	C	N/A	VAL										Y										
chr8:9501060-9501306	no	no	C	N/A	NT																			Y	
chr8:96852679-96852724	no	no	+	N/A	NT											Y				Y					Y
chr9:100130133-100130212	no	no	C	N/A	VAL																Y				
chr9:101741534-101741612	no	no	+	N/A	NT																			Y	
chr9:101889056-101889130	no	no	+	N/A	NT	Y										Y									
chr9:107067063-107067262	no	no	C	N/A	NT							Y													
chr9:10923373-10923518	no	no	C	N/A	NT													Y							
chr9:10944429-10944462	no	no	+	N/A	NT								Y												
chr9:110952081-110952146	no	no	C	N/A	NT															Y	Y				
chr9:111301127-111301214	no	no	+	N/A	NT											Y									
chr9:111947887-111947937	no	no	C	N/A	NT				Y						Y										
chr9:116674675-116674700	no	no	+	N/A	NT																	Y			
chr9:123725494-123725522	no	no	+	N/A	VAL		Y																		

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr9:124398771-124398794	no	no	C	N/A	NT				Y		Y						Y			Y					
chr9:16573595-16573646	no	no	+	N/A	NT																	Y			
chr9:22278708-22278794	no	no	C	N/A	NT				Y		Y	Y	Y		Y			Y	Y	Y	Y	Y	Y		
chr9:22385549-22385721	no	no	+	N/A	NT		Y							Y							Y		Y		
chr9:28185366-28185389	no	no	+	N/A	NT	Y														Y	Y				
chr9:336845-336945	no	no	C	N/A	NT	Y																			
chr9:4981870-4981986	no	no	C	N/A	NO																				Y
chr9:67923618-67923707	no	no	C	N/A	NT				Y	Y		Y		Y	Y	Y			Y						
chr9:78085570-78085612	no	no	+	N/A	NT								Y												
chr9:78141858-78141884	no	no	+	N/A	NT						Y		Y			Y									
chr9:78641185-78641442	no	no	C	N/A	NT																			Y	
chr9:78792178-78792336	no	no	C	N/A	VAL								Y												
chr9:94322156-94322232	no	no	C	N/A	NT	Y																			
chr9:96750520-96750557	no	no	C	N/A	NT	Y		Y																	
chr9:98594596-98594640	no	no	C	N/A	NT	Y																			
chr10:100293448-100293473	no	no	+	N/A	NT						Y														
chr10:103750180-103750354	no	no	+	N/A	VAL														Y						
chr10:10639664-10639890	no	no	C	N/A	NT												Y								
chr10:123548104-123548291	no	no	+	N/A	NT																	Y			
chr10:127817421-127817440	no	no	C	N/A	NT	Y																			
chr10:130464969-130465132	no	no	C	N/A	NT			Y																	
chr10:14436665-14436703	no	no	+	N/A	VAL										Y										
chr10:14524758-14524820	no	no	+	N/A	NT				Y	Y			Y	Y	Y										
chr10:14577727-14577794	no	no	+	N/A	NT		Y						Y												
chr10:31041142-31041285	no	no	+	N/A	VAL						Y														
chr10:35075463-35075566	no	no	+	N/A	VAL		Y																		
chr10:36507298-36507370	no	no	C	N/A	NT		Y	Y		Y															
chr10:56572360-56572439	no	no	C	N/A	VAL												Y								
chr10:60116336-60116494	no	no	C	N/A	NT																				

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr10:63356522-63356571	no	no	C	N/A	NT																	Y			
chr10:66507061-66507222	no	no	C	N/A	NT						Y		Y									Y		Y	
chr10:92455252-92455279	no	no	+	N/A	NT															Y	Y				
chr10:93223592-93223621	no	no	+	N/A	NT	Y																			
chr10:97658293-97658521	no	no	+	N/A	NT	Y																Y		Y	
chr10:98984396-98984444	no	no	+	N/A	NT													Y							
hr11:102018233-10201828	no	no	C	N/A	NT									Y	Y										
hr11:103417318-10341736	no	no	+	N/A	NT	Y																			
hr11:105675600-10567569	no	no	+	N/A	NT			Y					Y	Y		Y			Y	Y	Y	Y	Y	Y	Y
chr11:10689958-10689993	no	no	C	N/A	NT								Y									Y			Y
hr11:121431045-12143107	no	no	+	N/A	VAL		Y																		
hr11:121737729-12173795	no	no	C	N/A	NT			Y																	
hr11:127503037-12750317	no	no	C	N/A	NT		Y																		
hr11:128502624-12850269	no	no	+	N/A	NT															Y	Y				
hr11:134164752-13416487	no	no	C	N/A	VAL																		Y		
chr11:23315057-23315091	no	no	+	N/A	NT					Y						Y	Y	Y		Y		Y			Y
chr11:34353947-34354131	no	no	C	N/A	NT	Y		Y						Y	Y				Y			Y			
chr11:36233651-36233681	no	no	C	N/A	NT											Y				Y					
chr11:37662843-37662871	no	no	+	N/A	NT	Y	Y		Y	Y	Y				Y				Y	Y		Y		Y	Y
chr11:39360327-39360416	no	no	+	N/A	VAL				Y																
chr11:40450505-40450593	no	no	+	N/A	NT	Y	Y	Y						Y					Y	Y	Y		Y		
chr11:40683658-40683718	no	no	C	N/A	NT															Y					
chr11:49612394-49612536	no	no	C	N/A	NT		Y											Y		Y	Y		Y		
chr11:55076999-55077083	no	no	+	N/A	NT							Y	Y												
chr11:83580682-83580740	no	no	C	N/A	NT																			Y	
chr11:86564465-86564526	no	no	+	N/A	NT																	Y	Y		
chr11:88019323-88019421	no	no	+	N/A	NT		Y									Y	Y								
chr11:88548586-88548653	no	no	C	N/A	NT																			Y	
chr11:89673404-89673446	no	no	C	N/A	NT	Y																			

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr11:90840978-90841007	no	no	C	N/A	VAL						Y														
chr11:97848933-97849161	no	no	+	N/A	NT															Y					
hr12:108414443-10841455	no	no	+	N/A	VAL																				Y
hr12:120253358-12025338	no	no	+	N/A	NT			Y																	
chr12:12500910-12501059	no	no	+	N/A	VAL						Y														
hr12:125455817-12545591	no	no	C	N/A	NT					Y										Y	Y				
chr12:14065137-14065162	no	no	C	N/A	NT								Y		Y			Y	Y						Y
chr12:25353887-25354106	no	no	+	N/A	NT				Y					Y											
chr12:25892917-25892955	no	no	C	N/A	NT	Y																			
chr12:39118921-39119063	no	no	+	N/A	NT							Y				Y						Y			
chr12:56369398-56369544	no	no	+	N/A	NT		Y			Y	Y							Y							
chr12:56845502-56845544	no	no	C	N/A	NT					Y															
chr12:67194582-67194751	no	no	C	N/A	NT															Y					
chr12:68408894-68409107	no	no	C	N/A	NT								Y				Y								
chr12:71947562-71947605	no	no	+	N/A	NT													Y							
chr12:73487911-73487942	no	no	C	N/A	NT				Y			Y	Y				Y								
chr12:79839351-79839548	no	no	C	N/A	VAL										Y										
chr13:18156324-18156440	no	no	+	N/A	NT							Y	Y				Y				Y	Y	Y		
chr13:29857471-29857703	no	no	+	N/A	NT													Y							
chr13:33769359-33769450	no	no	C	N/A	NT																	Y			
chr13:35555137-35555170	no	no	+	N/A	NT						Y						Y								
chr13:52331778-52331905	no	no	C	N/A	VAL																		Y		
chr13:58275674-58275908	no	no	C	N/A	NT						Y	Y													
chr13:60265272-60265409	no	no	+	N/A	NT						Y	Y													
chr13:61486960-61487034	no	no	C	N/A	NT	Y	Y	Y	Y						Y					Y	Y				
chr13:63862478-63862535	no	no	+	N/A	NT			Y													Y		Y		
chr13:68580843-68580873	no	no	+	N/A	NT	Y																			
chr13:78169552-78169602	no	no	+	N/A	VAL																Y				
chr13:91623219-91623239	no	no	+	N/A	NT					Y	Y										Y	Y			Y

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
hr14:105884583-10588462	no	no	+	N/A	NT				Y		Y		Y						Y	Y	Y	Y	Y	Y	
chr14:30513760-30513979	no	no	+	N/A	NT											Y				Y					
chr14:39649398-39649435	no	no	+	N/A	NT							Y			Y	Y					Y				
chr14:39952637-39952862	no	no	C	N/A	NT						Y				Y					Y					
chr14:42810267-42810387	no	no	+	N/A	VAL																				Y
chr14:42864209-42864426	no	no	C	N/A	NT																			Y	
chr14:50788694-50788750	no	no	+	N/A	NT		Y			Y				Y	Y	Y	Y						Y		Y
chr14:66004153-66004214	no	no	+	N/A	NT		Y																		
chr14:69282381-69282451	no	no	+	N/A	NT																		Y		
chr14:77113204-77113280	no	no	+	N/A	NT									Y	Y										
chr14:84767255-84767356	no	no	C	N/A	NT			Y																	
chr14:86445242-86445320	no	no	+	N/A	NT			Y																	
chr15:43357205-43357280	no	no	C	N/A	NT										Y	Y	Y		Y						
chr15:43913233-43913294	no	no	+	N/A	NT	Y		Y		Y	Y									Y	Y			Y	Y
chr15:52073635-52073693	no	no	+	N/A	NT		Y										Y								
chr15:60393339-60393362	no	no	C	N/A	NT							Y	Y		Y	Y	Y	Y						Y	Y
chr15:60873041-60873090	no	no	+	N/A	NT							Y					Y		Y						
chr15:79637834-79637874	no	no	C	N/A	NT		Y																		
chr15:92620875-92620905	no	no	+	N/A	NT								Y												
chr15:95571105-95571238	no	no	+	N/A	NT				Y			Y	Y								Y			Y	
chr16:20534298-20534395	no	no	+	N/A	NT									Y											
chr16:23772963-23773184	no	no	C	N/A	NT									Y											
chr16:45040439-45040612	no	no	C	N/A	NT		Y						Y												
chr16:5348325-5348377	no	no	C	N/A	NT			Y																	
chr16:58598649-58598718	no	no	C	N/A	NT	Y			Y				Y				Y				Y		Y	Y	Y
chr16:79254276-79254493	no	no	+	N/A	NT							Y	Y					Y	Y						Y
chr16:8357708-8357916	no	no	+	N/A	NT																			Y	
chr17:11585499-11585540	no	no	+	N/A	NT	Y																			
chr17:12985873-12985934	no	no	C	N/A	NT		Y											Y	Y					Y	

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-12 2 BL	99-12 2 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr17:13740329-13740381	no	no	+	N/A	NT									Y	Y							Y			
chr17:22321507-22321621	no	no	C	N/A	NT	Y		Y		Y	Y		Y		Y			Y			Y	Y		Y	
chr17:22619220-22619412	no	no	C	N/A	NT		Y		Y																
chr17:33243443-33243522	no	no	C	N/A	NT	Y		Y			Y	Y		Y		Y	Y		Y	Y		Y			Y
chr17:36149516-36149611	no	no	C	N/A	NT							Y													
chr17:47404718-47404740	no	no	+	N/A	VAL		Y																		
chr17:47515530-47515561	no	no	+	N/A	NT	Y																			
chr17:47753490-47753673	no	no	C	N/A	NT				Y					Y							Y				
chr17:49311836-49312066	no	no	+	N/A	VAL										Y										
chr17:50663209-50663245	no	no	+	N/A	NT											Y	Y								
chr17:53135420-53135469	no	no	C	N/A	NT			Y																	
chr17:53347521-53347607	no	no	+	N/A	NT					Y						Y									
chr17:59085409-59085597	no	no	C	N/A	NT								Y												
chr17:6479348-6479395	no	no	+	N/A	NT													Y							
chr17:73447550-73447600	no	no	+	N/A	NT	Y												Y		Y	Y				
chr17:75702107-75702217	no	no	+	N/A	VAL																		Y		
chr18:17959149-17959253	no	no	C	N/A	NT	Y		Y																	
chr18:20355888-20355915	no	no	+	N/A	NT					Y					Y	Y					Y				
chr18:2327117-2327158	no	no	C	N/A	NO		Y																		
chr18:25435954-25436101	no	no	+	N/A	NT																				Y
chr18:42895845-42895892	no	no	+	N/A	NT		Y									Y									
chr18:43178721-43178753	no	no	C	N/A	NT							Y	Y												
chr18:65311033-65311063	no	no	C	N/A	NT											Y	Y		Y				Y		
chr18:72666365-72666510	no	no	C	N/A	NT	Y																			
chr18:8917252-8917353	no	no	+	N/A	NT																				
chr19:15788883-15788952	no	no	+	N/A	NT			Y		Y	Y	Y						Y	Y		Y	Y	Y		
chr19:32423986-32424131	no	no	C	N/A	NT															Y	Y				
chr19:35140040-35140133	no	no	C	N/A	NT								Y							Y	Y				
chr19:51628150-51628170	no	no	C	N/A	VAL		Y																		

Blat Coordinates	In Hg18?	In dbRIP?	strand	CAlu Subfamily	PCR Validation	00-43 BL	00-43 GBM	00-82 BL	00-82 MBM	02-07 BL	02-07 MBM	06-66 BL	06-66 MBM	07-65 BL	07-65 MBM	99-122 BL	99-122 GBM	99-24 BL	99-24 GBM	99-43 BL	99-43 GBM	99-69 BL	99-69 GBM	99-72 BL	99-72 MBM
chr20:10325837-10325921	no	no	+	N/A	NT										Y										
chr20:16778483-16778708	no	no	+	N/A	NT	Y																			
chr20:16822742-16822969	no	no	+	N/A	NT																	Y			
chr20:3190389-3190416	no	no	+	N/A	NT	Y														Y				Y	
chr20:5003147-5003244	no	no	+	N/A	NT	Y																			
chr20:52071320-52071356	no	no	+	N/A	VAL																		Y		
chr20:58271529-58271753	no	no	+	N/A	NT																			Y	
chr20:8764434-8764504	no	no	+	N/A	NT								Y								Y				
chr21:13707152-13707187	no	no	+	N/A	NT		Y												Y					Y	Y
chr21:36242954-36242986	no	no	C	N/A	NT													Y	Y						
chr21:9989077-9989101	no	no	C	N/A	NT			Y																	
hr21_random:713843-71399	no	no	C	N/A	NT			Y																	
chr22:30269498-30269649	no	no	+	N/A	NT									Y											
chr22:38925728-38925889	no	no	C	N/A	NT														Y						
chrX:11091785-11091939	no	no	+	N/A	NT																			Y	
chrX:114121008-114121145	no	no	+	N/A	NT	Y												Y							
chrX:119411871-119411912	no	no	C	N/A	NT							Y												Y	
chrX:135853376-135853530	no	no	C	N/A	NT									Y											
chrX:136177487-136177506	no	no	+	N/A	NT				Y		Y	Y		Y	Y	Y	Y	Y							Y
chrX:142326792-142326813	no	no	+	N/A	NT	Y																			
chrX:145346714-145346811	no	no	+	N/A	NT																	Y			
chrX:154502975-154502997	no	no	C	N/A	NT												Y								
chrX:29031527-29031573	no	no	+	N/A	NT	Y		Y																	
chrX:44793070-44793109	no	no	C	N/A	NT			Y																	
chrX:45203816-45203876	no	no	C	N/A	NT					Y															
chrX:5791742-5791772	no	no	C	N/A	NT			Y				Y				Y	Y								
chrX:6766320-6766421	no	no	+	N/A	NT																			Y	
chrX:93428586-93428707	no	no	C	N/A	NT		Y																		
chrY:4437444-4437484	no	no	+	N/A	NT				Y			Y													