HIGH-THROUGHPUT DETECTION OF POTENTIALLY ACTIVE L1 ELEMENTS IN HUMAN GENOMES

AN ABSTRACT

SUBMITTED ON THE 18th DAY OF JULY 2014

TO THE GRADUATE PROGRAM IN BIOMEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF MEDICINE

OF TULANE UNIVERSITY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

VINCENT A. STREVA

APPROVED:

PRESCHOOL DEININGER, Ph.D.

Chairman

VICTORIA BELANCIO, Ph.D.

WILLIAM WIMLEY, Ph.D.

ZACHARY PURSELL, Ph.D.

CHARLES MILLER, Ph.D.
The active human retrotransposon L1 is the most prevalent human retroelement, constituting 17% of the mass of the human genome and contributing significantly to mutagenesis. L1 mutagenizes human genomes in a number of ways including insertional mutagenesis of itself and other retrotransposons, creating of DNA double strand breaks, and induction of non-allelic homologous recombination. Through these processes, the activity of L1 is responsible for approximately 0.5% of all new genetic diseases.

All L1-derived mutagenesis stems from the activity of a small number of intact full-length L1 loci that remain capable of mobilization. A smaller subset of these active L1s are called hot L1s and are responsible for the vast majority of all L1 activity. Hot L1s are polymorphic in the population and represent evolutionarily recent L1 insertion events.

Here, we show that potentially active full length L1 elements are more prevalent in individual genomes than previously believed. We find that the typical individual likely harbors approximately 60 active and 50 hot L1s. However, we also find that there is significant variation between individuals in numbers of potentially active L1s. As a result, the mutagenic burden associated with L1 likely varies between individuals.
HIGH-THROUGHPUT DETECTION OF POTENTIALLY ACTIVE L1 ELEMENTS IN HUMAN GENOMES

A DISSERTATION

SUBMITTED ON THE 18th DAY OF JULY 2014

TO THE GRADUATE PROGRAM IN BIOMEDICAL SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF THE SCHOOL OF MEDICINE

OF TULANE UNIVERSITY

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

VINCENT A. STREVA

APPROVED:

PRESKOTT DEININGER, Ph.D.

Chairman

VICTORIA BELANCIO, Ph.D.

WILLIAM WIMLEY, Ph.D.

ZACHARY PURSELL, Ph.D.

CHARLES MILLER, Ph.D.
ACKNOWLEDGEMENTS

I would like to express sincere gratitude to my advisor, Prescott Deininger, for his excellent mentorship and professional development. His constant advice and encouragement enabled me to grow as a scientist. Additionally, I would like to thank the members of my dissertation committee, Victoria Belancio, Zachary Pursell, Charles Miller, and William Wimley for their guidance and support during my graduate training. Additionally, while not on my dissertation committee, Astrid Engel has provided valuable input into my work, and for that I am grateful.

I would also like to thank the members of the Deininger lab past and present for constant advice, but most of all for their friendship. Particularly Cecily DeFreece for being the one who most influenced my decision to join the Deininger lab and after I did join, made every day working with her a new adventure. I’m also thankful for Geraldine Servant and Maria Morales for their advice and even more for the entertainment and support they provided me with for the last six years. I will miss you all.

Lastly, thank you to my family who always support me in everything I do. To my parents who have, ever since I was a child, supported my ideas no matter how stupid they may have been, I am forever thankful. To my partner Josh, you are my everything. Thanks for always being there for me. And to the two best dogs in the entire world, Hans and Lily, you guys rock.
### TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures and Tables</td>
<td>vii</td>
</tr>
<tr>
<td><strong>CHAPTER 1: INTRODUCTION</strong></td>
<td></td>
</tr>
<tr>
<td>1.1 Mobile elements in mammalian genomes</td>
<td>1</td>
</tr>
<tr>
<td>1.2 LINE-1 (L1) organization</td>
<td>11</td>
</tr>
<tr>
<td>1.2.A L1 5’ UTR</td>
<td>14</td>
</tr>
<tr>
<td>1.2.B L1 ORF1</td>
<td>16</td>
</tr>
<tr>
<td>1.2.C L1 ORF2</td>
<td>18</td>
</tr>
<tr>
<td>1.2.D L1 3’ UTR</td>
<td>21</td>
</tr>
<tr>
<td>1.3 L1 transcription, translation, and expression</td>
<td>23</td>
</tr>
<tr>
<td>1.3.A L1 Transcription</td>
<td>23</td>
</tr>
<tr>
<td>1.3.B L1 Translation</td>
<td>26</td>
</tr>
<tr>
<td>1.3.C L1 Localization</td>
<td>27</td>
</tr>
<tr>
<td>1.3.D L1 Expression in somatic cells</td>
<td>29</td>
</tr>
<tr>
<td>1.4 L1 replication cycle</td>
<td>31</td>
</tr>
<tr>
<td>1.5 L1 trans-mobilized elements</td>
<td>37</td>
</tr>
<tr>
<td>1.6 L1 toxicity and disease</td>
<td>39</td>
</tr>
<tr>
<td>1.7 Cellular control of L1 retrotransposition</td>
<td>43</td>
</tr>
<tr>
<td>1.8 L1 retrotransposition reporter assay</td>
<td>46</td>
</tr>
<tr>
<td>1.9 L1 high throughput sequencing methods</td>
<td>49</td>
</tr>
</tbody>
</table>
CHAPTER 2: PCR of Anchored L1 Extensions (PALE): A METHOD FOR DETECTION OF POTENTIALLY ACTIVE FULL LENGTH L1 ELEMENTS

2.1 Introduction 84

2.2. PCR of Anchored L1 Extensions (PALE) allows for high throughput detection of virtually all full length L1s in a given genome 90

  2.2.A Development of PALE 90

  2.2.B PALE sensitivity 100

  2.2.C PALE specificity 110

2.3 Discussion 117

2.4 Materials and Methods 119
CHAPTER 3: PALE discovery of novel full length L1 elements

3.1 Introduction 123

3.2. Full length L1 elements are more abundant than previously estimated and numbers vary considerably between individuals 128

3.3 Discussion 157

3.4 Materials and Methods 159

CHAPTER 4: PALE discovery of de novo L1 insertion events

4.1 Introduction 161

4.2 3’-PALE can be used for de novo L1 retrotransposition detection 168

4.2.A Development of 3’-PALE 168

4.2.B 3’-PALE Sensitivity 174

4.2.C 3’-PALE Specificity 179

4.2.D Detection of L1Hs events in two NER-deficient family groups 182

4.3 Discussion 188

4.4 Materials and Methods 191

CHAPTER 5: DISCUSSION

5.1 Genomic impact of L1 retrotransposition 194

5.2 Regulation of L1 expression 196
LIST OF FIGURES AND TABLES

Figures

Figure 1: Sequence organization of the human genome 2
Figure 2: Human transposable elements 9
Figure 3: Schematic of a full length L1 element 12
Figure 4: The L1 replication cycle 32
Figure 5: L1 target primed reverse transcription (TPRT) 35
Figure 6: Mobile elements induce DNA damage by multiple mechanisms 41
Figure 7: L1 retrotransposition reporter assay 47
Figure 8: L1-Seq 51
Figure 9: Transposon-Seq 55
Figure 10: ME-Scan 58
Figure 11: RC-Seq 63
Figure 12: Tea 68
Figure 13: Recovery of de novo L1 insertions using the L1 rescue cassette 72
Figure 14: TIP-ChIP 76
Figure 15: Fosmid-based screening for full length L1s 81
Figure 16: PCR of Anchored L1 Extensions (PALE) 88
Figure 17: PALE library quality control 93
Figure 18: Illumina sequencing reaction 96
Figure 19: Distinguishing PCR duplicates from independent ligation events 98
Figure 20: PALE detects fixed full length L1s with high sensitivity 102
Figure 21: PALE detects polymorphic full length L1s at similar levels as fixed L1s 106
Figure 22: PCR validation of 5'-PALE events 111
Figure 23: PALE detects known and novel full length L1 elements 115
Figure 24: Detection of novel full length L1 elements in seven individuals 134
Figure 25: Pairwise comparison showing all polymorphic full length L1 elements shared between any two individuals 141
Figure 26: Pairwise comparison showing all polymorphic full length L1 elements unique between any two individuals 143

Figure 27: Pairwise comparison showing all novel full length L1 elements shared between any two individuals 145
Figure 28: Pairwise comparison showing all novel full length L1 elements unique between any two individuals 147
Figure 29: Chromosomal distribution of polymorphic full length L1 elements across seven individuals 149
Figure 30: Novel full length L1 elements are detected in all seven individuals tested 151
Figure 31: Histogram depicting population study of 40 polymorphic full length L1 elements detected by PALE 153

Figure 32: The nucleotide excision repair (NER) pathway 166
Figure 33: Modification of PALE to detect human-specific L1 elements at the 3' end (3'-PALE) 170

Figure 34: 3'-PALE selectively detects only the youngest, most active L1 subfamilies 172
Figure 35: 3'-PALE detects L1Hs elements from human genomes 175
Figure 36: 3'-PALE detects L1Hs elements with high sensitivity 177
Figure 37: PCR validation of 3'-PALE events 180
Figure 38: XPA-deficient family groups used in this study 184
Figure 39: Model for differing degrees of L1-associated damage in different individuals in the population 202
Figure 40: Model for hot L1s in tumorigenesis 205
Tables

Table 1: PALE-detection of previously described polymorphic full length L1s 108

Table 2: Validation of PALE by 5’ L1 junction PCR 113

Table 3: List of cell lines used in this study 126

Table 4: Novel full length L1 elements identified in this study 130

Table 5: Validation of PALE-detected full length L1 elements by 3’ junction PCR 132

Table 6: Number of polymorphic full length L1 elements detected in each of seven individuals by PALE 139

Table 7: Novel full length L1s detected in each of the seven individuals 155

Table 8: L1Hs elements detected in two XPA-deficient individuals by 3’-PALE 186
CHAPTER 1:

INTRODUCTION

1.1 Mobile elements in mammalian genomes

Only a very small fraction (~1-2%) of the human genome corresponds to gene coding sequence. Of the remaining 98%, the majority is represented by repetitive DNA sequences. The initial sequencing of the human genome estimates that as much as 45% of total human genetic content is repeat element-derived (Lander, Linton et al. 2001) (Figure 1). This estimate, however, is likely quite conservative, with recent groups suggesting as much as two-thirds of human genetic content being repeat-derived (de Koning, Gu et al. 2011). Human repetitive DNA falls into five classes: (1) tandemly repeated sequences (i.e., telomeres, centromeres, and rDNA clusters), (2) segmental duplications (also called copy number variations) which are blocks of typically 10-300kb that have been duplicated from one genomic region to another, (3) simple sequence repeats (e.g., (A)$_n$, (CA)$_n$, (CGG)$_n$), (4) processed pseudogenes, which are inactive retrotransposed copies of functional genes, and (5) transposon-derived repeats (Lander, Linton et al. 2001).
Figure 1

DNA Transposons

- LTR Retrotransposons
- Other nonLTR retrotransposons

non-LTR

53.9%

2.8%

8.3%

6.0%

10.6%

16.9%
Figure 1 Sequence organization of the human genome. The human genome is largely composed of repetitive DNA sequence. Nearly one half of total genomic content consists of repetitive DNA. The three major classes of human repeats are 1) DNA transposons which make up ~3% of genome content, 2) LTR retrotransposons which make up ~8% of genome content and 3) non-LTR retrotransposons which make up ~34% of genome content. The only currently active transposable elements in humans are the non-LTR retrotransposons, which consist of LINEs (L1), SINEs (Alu), and SVA.
Recognizable transposon-derived repeats make up as much as 50% of the human genome. However there are likely many more regions of DNA that belong to this class that have, through millions of years of evolution, diverged too far from the consensus repeat element to be recognized as repeat-derived. All transposable elements fall into one of three types: (1) DNA transposons, (2) long terminal repeat (LTR) retrotransposons, and (3) non-long terminal repeat (non-LTR) retrotransposons (Lander, Linton et al. 2001) (Figure 2).

DNA transposons resemble bacterial transposons, encode a transposase and have terminal inverted repeats near which the transposase binds, mediating mobility of the transposon by a “cut-and-paste” mechanism (Figure 2). The human genome contains at least seven classes of DNA transposons in approximately 300,000 copies, making up about 3% of the mass of the genome (Figure 1). None of these DNA transposons, however, are actively mobilized in humans. This inactivation is believed to be the result of the nature in which the DNA transposons mobilize, with the accumulation of inactive copies decreasing the effectiveness of future transposition (Lander, Linton et al. 2001, Munoz-Lopez and Garcia-Perez 2010).

LTR retrotransposons are viral-like elements that encode all of the necessary machinery (gag and pol) for their mobilization through an RNA intermediate, flanked by long terminal repeats (Figure 2). Transposition of LTR
retrotransposons occurs through a viral-like mechanism involving the reverse transcription of the element primed by cellular tRNA. While numerous LTR retroelement subfamilies make up the approximately 450,000 copies present in the human genome (~8% of genome mass), only the ERV family shows evidence of ever being active in mammalian genomes (Figure 1). Like DNA transposons, LTR retrotransposons are inactive in humans (Lander, Linton et al. 2001, Havecker, Gao et al. 2004).

The only active mobile elements in humans form the third class of repetitive DNA: the non-LTR retrotransposons (Figure 2). Non-LTR retrotransposons retrotranspose through an RNA intermediate by a “copy-and-paste” mechanism, and are divided into two groups on the basis of their ability to encode the necessary machinery to self-mobilize (Daniels and Deininger 1985, Daniels and Deininger 1985, Daniels and Deininger 1991, Curcio and Garfinkel 1994, Feng, Moran et al. 1996, Jurka 1997, Dewannieux, Esnault et al. 2003, Hancks, Goodier et al. 2011, Raiz, Damert et al. 2012). Elements encoding the necessary mobilization machinery are deemed autonomous, while the remaining elements are called non-autonomous. Long Interspersed Elements (LINEs) are the predominant retrotransposon family in humans, accounting for ~21% of human genetic content (Lander, Linton et al. 2001). The predominant LINE is the LINE1 (L1) element, which exists in ~520,000 genomic copies and accounts for ~17% of the human genome (Figure 1). LINE1 is the sole autonomous human
mobile element. Short Interspersed Elements (SINEs) are non-autonomous elements that make up ~13% of the human genome. The actively mobilizing human SINE, Alu, is the most abundant human mobile element, existing in ~1.1 million copies and constituting ~11% of genome mass (Figure 1). The active SINE-VNTR-Alu element (SVA) is a non-autonomous composite element that exists in ~3000 copies. While only representing a modest fraction of genome mass (<0.5%), SVA elements have been implicated in a number of diseases, indicating they are quite active in genomes today (likely a result of their evolutionary youngness). Together, L1, Alu, and SVA are responsible for nearly all human retrotransposition (Skowronski and Singer 1985, Scott, Schmeckpeper et al. 1987, Kazazian, Wong et al. 1988, Skowronski, Fanning et al. 1988, Dombroski, Mathias et al. 1991, Dombroski, Scott et al. 1993, Holmes, Dombroski et al. 1994, Moran, Holmes et al. 1996, Sassaman, Dombroski et al. 1997, Deininger and Batzer 1999, Boissinot, Chevret et al. 2000, Boissinot and Furano 2001, Lander, Linton et al. 2001, Myers, Vincent et al. 2002, Ovchinnikov, Rubin et al. 2002, Salem, Myers et al. 2003, Belancio, Hedges et al. 2008, Belancio, Deininger et al. 2009, Beck, Garcia-Perez et al. 2011, Burns and Boeke 2012).

Distribution of L1 is not random across human chromosomes. There is a four-fold enrichment of LINEs in AT-rich DNA, and a significant depletion of LINEs in gene-coding regions. This phenomenon is the inverse for Alu elements.
which show a five-fold enrichment in GC-rich DNA. This is surprising given the shared insertion mechanism for Alu and L1. In fact, Alu elements are depleted from AT-rich regions of DNA despite their preference for inserting into these regions, indicating a selective force acting to either remove Alu elements from AT-rich regions or prevent their removal from GC-rich regions. Additionally, if one compares only the youngest, most recently inserted L1 and Alu elements, this distribution is much more uniform, supporting an evolutionary selection bias model. There is evidence that Alu distribution bias is not the result of a bias for GC rich DNA, but rather a bias for gene-rich regions, with Alu density on gene-rich chromosome 19 being over-represented when compared to GC content and Alu density on gene-poor chromosome Y being under-represented (Lander, Linton et al. 2001).

While there are over half a million genomic copies of L1 in the human genome, the vast majority of these do not represent full length, 6kb L1 elements. The mean length of genomic L1 copies is 900bp, and the mean length of only the most newly inserted L1Hs subfamily (which makes up all active L1 loci) is not significantly longer (1070bp). There are, however, approximately 4000 copies of full length L1 elements in the human genome. The vast majority of these, however, have been rendered inactive by point mutations in their open reading frames, resulting in only ~100 active elements capable of retrotransposition within the reference genome. Of these active L1 loci, only a small handful (~6) is
considered highly active and account for the bulk of L1 retrotransposition (Lander, Linton et al. 2001, Allen, Horvath et al. 2003, Brouha, Schustak et al. 2003, Beck, Collier et al. 2010).

In addition to these fixed, reference L1 elements, human genomes contain a number of elements that are present in only subsets of individuals. These polymorphic L1 elements are elements that have more recently integrated via retrotransposition into the human genome and are differentially present or absent in subsets of individuals depending on the presence of a common ancestor in which the retrotransposon-insertion even occurred. A majority of the polymorphic full length L1 elements discovered in various genomes are capable of active retrotransposition (Boissinot, Chevret et al. 2000, Myers, Vincent et al. 2002, Ovchinnikov, Rubin et al. 2002, Brouha, Schustak et al. 2003, Salem, Myers et al. 2003, Farley, Luning Prak et al. 2004, Khan, Smit et al. 2006, Seleme, Vetter et al. 2006, Beck, Collier et al. 2010, Ewing and Kazazian 2010, Iskow, McCabe et al. 2010, Rouchka, Montoya-Durango et al. 2010, Beck, Garcia-Perez et al. 2011, Ewing and Kazazian 2011, Burns and Boeke 2012, Hancks and Kazazian 2012, Lee, Iskow et al. 2012, Solyom, Ewing et al. 2012, Shukla, Upton et al. 2013, Helman, Lawrence et al. 2014, Pitkanen, Cajuso et al. 2014).
**Figure 2 Human transposable elements.** Depicted are schematics of the major classes of mobile elements present in human genomes (not to scale). L1 is the only autonomous transposable element in humans and utilizes an internal RNA polymerase II promoter to transcribe a 6kb full length RNA that encodes two proteins, ORF1 and ORF2. L1 ends with a weak polyadenylation signal. Alu is a non-autonomous SINE that uses an internal RNA polymerase III promoter to create a small 300bp RNA that does not code for protein. SVA is a non-autonomous composite element of highly variable length that uses the L1 protein machinery to mobilize. All non-LTR retrotransposons are flanked by target side duplications (TSDs) that are the result of their mechanism of insertion (black triangles). LTR retrotransposons are inactive in present day humans and utilize their long terminal repeats (LTRs) to make RNA expressing two proteins involved in mobilization (gag and pol). DNA transposons are also inactive in humans and transpose via a cut-and-paste mechanism involving inverted terminal repeats (ITRs) and the enzyme transposase.
1.2 **LINE-1 (L1) organization**

The full length human L1 element is ~6kb in length (Figure 3). Each full length L1 element contains both 5’ and 3’ untranslated regions (UTRs), and encode two non-overlapping open reading frames (ORFs). The 5’ UTR contains an internal RNA polymerase II promoter to initiate transcription of the L1 element (Swergold 1990, Ostertag and Kazazian 2001). Each ORF encodes a protein (ORF1p and ORF2p, respectively) that is essential for retrotransposition of the L1 element (Mathias, Scott et al. 1991, Holmes, Singer et al. 1992, Feng, Moran et al. 1996, Hohjoh and Singer 1996, Hohjoh and Singer 1997). The two ORFs are separated by a short region consisting of several stop codons. The 3’ UTR contains a weak polyadenylation signal (pA) followed by a small run of adenine residues (referred to as a polyA tail) that likely originated from reverse transcription of the A-tail of the L1 mRNA upon insertion of the element (Dombroski, Mathias et al. 1991, Alisch, Garcia-Perez et al. 2006).
Figure 3 Schematic of a full length L1 element. A full length L1 element is 6kb long and uses an internal RNA polymerase II promoter to generate a transcript that codes for two proteins. The 40kDa ORF1 protein has nucleic acid binding and chaperone activity, while the 150kDa ORF2 protein has endonuclease and reverse transcriptase catalytic functions. L1 also contains an antisense promoter that can drive transcription of genomic fragments lying upstream of the L1 element.
1.2. A L1 5’ UTR

The GC-rich L1 5’ UTR contains an internal, downstream promoter, which initiates transcription by RNA polymerase II upstream of the ORF region. Additionally, the ~900bp 5’ UTR contains binding sites of a number of cellular factors, some of which are critical for L1 transcription: including YY1, Sp1, and Runx. The L1 5’ UTR YY1 binding site plays a critical role in mRNA transcription at G-rich sequences (Swergold 1990, Minakami, Kurose et al. 1992, Severynse, Hutchison et al. 1992, Becker, Swergold et al. 1993, Yang, Zhang et al. 2003). Additionally, the first ~500 bases of the L1 5’ UTR play a crucial role in the proper and efficient transcription of L1 mRNA (Minakami, Kurose et al. 1992).

In addition to the L1 sense promoter, the L1 5’ UTR also contains an antisense promoter (ASP) located in the 400-600bp region of the 5’ UTR (Speek 2001) (Figure 3). L1 antisense promoter activity can be stronger than that of the actual L1 promoter, however the role of the ASP in the L1 replication cycle has not been elucidated. However, it has been speculated that transcription off the L1 ASP may interfere with sense transcription of the L1 element, acting as a means of cellular restriction on L1-induced host damage (Matlik, Redik et al. 2006).

While the role of the L1 ASP remains unclear, there is evidence that it can function to drive transcription of cellular genes (Matlik, Redik et al. 2006). Expressed sequence tag (EST) analysis has shown increased L1 ASP-driven
transcription of cellular genes in tumor cell lines. The L1 ASP can provide a site of alternate transcription start for cellular genes, resulting in chimeric transcripts (eg, the proto-oncopogene c-MET) (Speek 2001, Birchmeier, Birchmeier et al. 2003).

A large part of the cellular restriction of the activity of L1 elements in the human genome is thought to be attributed to sequence characteristics of the L1 5' UTR. L1 5' UTR promoter CpG methylation has been shown to limit the ability of L1 to retrotranspose, and in turn limit the burden of L1-associated damage (Hata and Sakaki 1997, Woodcock, Lawler et al. 1997). Given the high AT rich nature of the L1 element and the depletion of GC rich regions in the human genome, it is speculated that the presence of a high number of CpG residues in the L1 5' UTR is the result of a cellular defense mechanism that evolved to repress the deleterious effects of L1 mobilization by limiting L1 expression. Loss of L1 5' UTR promoter methylation seems to play a role in L1 expression in various cancers as L1 promoter hypomethylation has been documented in a large number of cancers (Florl, Lower et al. 1999, Santourlidis, Florl et al. 1999).
1.2.B L1 ORF1

The human L1 element mRNA is a bicistronic mRNA that encodes two non-overlapping proteins, ORF1p and ORF2p, both of which are required for L1 retrotransposition (Moran, Holmes et al. 1996) (Figure 3). ORF1p is a 42kD protein with multiple functions. The N-terminus of the human ORF1p consists of a coiled-coil domain that is critical for ORF1p/ORF1p homotrimerization (Martin, Li et al. 2000, Goodier, Ostertag et al. 2001). The C-terminal domain of ORF1p has nucleic acid binding and chaperone activities, allowing it to bind L1 mRNA (Kolosha and Martin 1995, Kolosha and Martin 1997, Martin and Bushman 2001, Kolosha and Martin 2003, Martin, Cruceanu et al. 2005). While the role of ORF1p binding to L1 mRNA has not been elucidated, a number of possibilities for this interaction exist. L1 mRNA-bound ORF1p might protect the nascent L1 mRNA transcript from degradation. Alternatively, L1 ORF1p might in some way interact with cellular factors during insertion of a new L1 element, promoting successful retrotransposition. Regardless of the mechanism, L1 ribonucleoprotein (L1 RNP) formation is required for L1 retrotransposition.

The properties of L1 ORF1p binding help to explain the cis-preference observed with L1 proteins (Wei, Gilbert et al. 2001). In the cis-preference model of retrotransposition, the L1-encoded ORF1 and ORF2 proteins interact with strong preference on the L1 mRNA from which they originated. While the
reasons for the observed cis-preference are not understood, it is possibly the result of either unstable naked L1 mRNA or unstable unbound L1 proteins. This hypothesis is strengthened by the observation that nearly all the observable L1 ORF1p is contained within L1 RNPs. The evolutionary basis for cis-preference could potentially stem from two different areas. First, cis-preference limits the rate at which (non-L1) cellular mRNAs are retrotransposed in cells to below $10^{-7}$ per cell, limiting damage suffered by the host (Werle-Schneider, von Brevern et al. 1999). Second, cis-preference may have been selected to give an evolutionary advantage to the most active L1 elements. By limiting the interaction of functional L1 proteins with nonfunctional, defective L1 mRNAs, cis-preference preferentially selects for the insertion of elements that have the potential for further retrotransposition, while preventing the retrotransposition of L1 elements that would be dead-on-arrival.
1.2.C L1 ORF2

The second L1 open reading frame encodes the considerably larger ~150kD ORF2p. L1 ORF2p encodes three functional domains, all of which are required for L1 retrotransposition. From N-terminal to C-terminal, these domains are the endonuclease domain (EN), the reverse transcriptase domain (RT) and the cysteine-rich domain (CYS) (Mathias, Scott et al. 1991, Feng, Moran et al. 1996, Moran, Holmes et al. 1996, Dhillin, Maestre et al. 1997, Clements and Singer 1998, Cost and Boeke 1998) (Figure 3).

The L1 ORF2p EN domain consists of the first ~26kD of the ORF2p. ORF2p EN is an endonuclease of the apurinic/apyrimidinic (AP) type and has high sequence conservation in the active site to ExoIII, a major AP endonuclease in *Escherichia coli* (Feng, Moran et al. 1996). Despite significant active site similarities, the L1 endonuclease shows no cleavage preference for apurinic/apyrimidinic sites, but rather cleaves at a rough 5’-TTAAAA-3’ consensus sequence, allowing for base modifications within the consensus, particularly A→G and T→C. Cleavage nearly always occurs on the opposite (antisense) strand of the above consensus between the 3’-AA and TTTT-5’. Cleavage of the sense strand occurs at a variable distance from the initial EN cut site and either occurs by L1 EN or another cellular endonuclease (possibly one of the endonucleases in the nucleotide excision repair (NER) pathway), resulting in a small target site duplication (TSD) at the site of L1 insertion (Feng, Moran et al. 1996, Jurka 1997, Cost and Boeke 1998, Gilbert, Lutz-Prigge et al. 2002, Gilbert,
Lutz et al. 2005). The expression of L1 EN is required for successful integration of new L1 copies into the genome.

The L1 ORF2p RT domain is located in the middle portion of the second ORF following the EN domain. L1 RT is not sequence-specific and is capable of reverse transcribing any mRNA template (Dhellin, Maestre et al. 1997). It is this promiscuity that allows L1 to act as a retrotransposition driver for other parasite retroelements like Alu and SVA (Moran, Holmes et al. 1996, Dewannieux, Esnault et al. 2003, Hancks, Goodier et al. 2011, Raiz, Damert et al. 2012). Like L1 EN, L1 RT is required for generation of new L1 inserts into the genome. Work from the Boeke lab at Johns Hopkins indicates that L1 RT is susceptible to inhibition by nucleoside analog RT inhibitors (NRTIs) but not to inhibition by non-nucleoside inhibitors (NNRTIs) (Dai, Huang et al. 2011). Measurements of the activity of the L1 RT conducted using a technique developed in the Moran lab called L1 element amplification protocol (LEAP) has shown that the L1 RT is capable of extending from mismatched templates, can function in the absence of the L1 EN, preferentially extends L1 RNA sequence (acts in cis), colocalizes in an RNP with ORF1p and L1 mRNA (Kulpa and Moran 2005, Kulpa and Moran 2006).

The L1 CYS domain makes up the C-terminal region of the ORF2p. The functional importance of the CYS domain has not been established, however it has been shown to be required for retrotransposition of L1 (Moran, Holmes et al. 1996). The L1 CYS domain is well conserved among all clades of mammalian L1 elements, as well as in the related L1-like yeast element Zorro, indicating a
conserved CCHC zinc knuckle domain (Wagstaff, Barnerssoi et al. 2011, Piskareva, Ernst et al. 2013, Wagstaff, Kroutter et al. 2013). Such zinc knuckles are found in retroviral reverse transcriptases and facilitate protein binding to single stranded RNA, indicating a role for ORF2p in protein-L1 mRNA interaction (Ostertag and Kazazian 2001).
1.2.D L1 3’ UTR

The L1 element ends in a 3’ UTR with three predominant features. The first, a weak polyadenylation signal, is present to terminate transcription of the L1 mRNA. Because of the weak nature of this polyadenylation signal, transcriptional read-through sometimes occurs until a downstream, cellular genomic polyadenylation signal is reached, resulting in a transduction of the genomic flanking region 3’ of the L1 element (Moran, DeBerardinis et al. 1999, Ejima and Yang 2003, Helman, Lawrence et al. 2014, Pitkanen, Cajuso et al. 2014). Analysis of reference full length L1 elements indicates that ~20% result in 3’ transduction events with the size of the transduction varying from 30-970bp. Such L1 transductions of 3’ flanking sequences account for ~0.5-1% of the human genome and may represent an important mechanism for exon shuffling in the human genome (Moran, DeBerardinis et al. 1999, Lander, Linton et al. 2001, Ejima and Yang 2003).

The second feature of the L1 3’ UTR is the presence of a stretch of adenine residues at the very 3’ end of the element. This stretch of As most likely is the result of reverse transcription of the polyA-tail of the L1 mRNA during the insertion event that led to the element being present in the genome. It has been speculated that this stretch of A residues at the 3’ end might facilitate the function of the L1 EN (Moran, DeBerardinis et al. 1999, Ejima and Yang 2003).
The final feature of the L1 3' UTR is the presence of a nuclear export factor (NXF) binding site, which may facilitate the export of the L1 mRNA from the nucleus (Lindtner, Felber et al. 2002). Because L1 is a large (~6kb) mRNA that lacks any intronic regions to facilitate nuclear export, it is likely that the presence of the NXF binding site may play a role in exporting the large L1 mRNA to the cytoplasm where it can undergo translation.
1.3 L1 transcription, translation, and expression

1.3.A L1 Transcription

L1 transcription makes use of an internal RNA polymerase II promoter to encode a full length (6 kb) bicistronic mRNA. While internal priming is more typical of RNA polymerase III transcription, the full length L1 transcript at 6 kb in length is much longer than any known RNA polymerase III transcripts (Swergold 1990, White 2004). Additionally, the presence of a 5’ methylguanosine cap implicates RNA polymerase II as the transcription machinery. Within the 5’ internal promoter region of L1 are a number of transcription factor binding sites including Sp1, Runx, and YY1. The YY1 transcription factor binding site, which has been shown to be involved in efficient transcription of L1 mRNA (Minakami, Kurose et al. 1992, Severynse, Hutchison et al. 1992, Becker, Swergold et al. 1993, Yang, Zhang et al. 2003).

In addition to the internal L1 promoter, the L1 5’ UTR contains an antisense promoter (ASP) whose role is not well understood despite its measured activity. One hypothesis for the role of the L1 ASP is a means of cellular L1 restriction; ASP promoter activity may interfere with transcription from the L1 sense promoter, resulting in decreased expression of L1 mRNA. Despite lacking knowledge of the role of the L1 ASP, it has potential to impact human
gene expression by driving transcription of cellular genes. L1 ASP has been shown to represent an alternative transcription initiation site for a number of human genes, including the proto-oncogene c-MET (Speek 2001, Birchmeier, Birchmeier et al. 2003).

Transcription of L1 RNA results in the production of two proteins, open reading frame 1 and 2 (ORF1 and ORF2). Both ORF1 and ORF2 are essential for L1 mobilization and ORF2 is essential for the mobilization of all other retrotransposable elements (Moran, Holmes et al. 1996). Full length L1 transcripts encoding two intact open reading frames are functionally active and capable of generating new retroelement inserts by target primed reverse transcription. The majority of genomic L1 elements, however are 5’ truncated or contain inactivating mutations in their ORFs rendering them non-functional (Lander, Linton et al. 2001). As a result, there are only about 90-100 human L1 elements capable of expressing full length L1 transcripts that code for functional ORFs (Lander, Linton et al. 2001, Brouha, Schustak et al. 2003, Beck, Collier et al. 2010).

1.3.A.i L1 splicing

mRNA splicing involves a number of cellular factors that act both in \textit{cis} and in \textit{trans} to generate differentially spliced RNA species (McManus and Graveley 2011). A study from the Deininger lab in 2006 predicted the presence of numerous splice sites within the L1 RNA sequence and were the first to show that L1 elements produce numerous functional splice products (Belancio, Roy-
Engel et al. 2008). The production of these alternative L1 splice products limits the production of full length L1 mRNA and competes with an alternate means of L1 RNA modification (premature polyadenylation, discussed below) to affect L1 mRNA levels in cells. Additionally, Belancio et al. uncovered some commonly used splice variants that lacked ORF1 sequence but generated functional ORF2. The implications of this discovery include the potential presence of multiple L1 elements in the genome that may appear to be nonfunctional (e.g., by containing mutations in ORF1) but might in fact be functional to drive the retrotransposition of *trans*-mobilized elements such as Alu. Additionally, in this study, the authors characterized some human and mouse L1 splice products that made use of alternative splice sites, but contained both functional ORFs that are retrotranspositionally active. One final discovery by Belancio et al., was that L1 splice sites could be utilized to splice into the exons of the gene into which an L1 was inserted in a number of tissues for which sequence information was available (Belancio, Roy-Engel et al. 2008).

1.3.A.ii L1 premature polyadenylation

An additional means by which L1 is regulated at the cellular level is by premature polyadenylation of the L1 RNA. Polyadenylation of mammalian mRNAs occurs during transcription and occurs downstream of a consensus polyadenylation sequence of either AATAAA or ATTAAA. L1 contains a weakly active polyadenylation signal in its 3' UTR which is used for polyadenylation of full length L1 transcripts. For many years, it was thought that the only L1 mRNAs created were full length, however in 2003, the Deininger lab reported the
presence of multiple strong polyadenylation signals within the L1 sequence and showed that many of these alternative polyadenylation signals were active (Perepelitsa-Belancio and Deininger 2003). In fact, many of the alternative L1 polyadenylation signals acted much more strongly than the canonical polyadenylation site in the L1 3’ UTR, generating an array of prematurely terminated L1 mRNAs (Perepelitsa-Belancio and Deininger 2003).

Additionally, when mutations were made to abolish premature polyadenylation sites in L1, the rate of L1 retrotransposition increased, indicating that premature polyadenylation may represent a means for the cell to control the mobility of L1. Further, the presence of strong polyadenylation sites in the sequence of L1 may result in an increase in the genomic burden of L1 insertion events into genes, resulting in the premature polyadenylation of the gene’s transcript (Perepelitsa-Belancio and Deininger 2003).

1.3.B L1 Translation

L1 translation occurs through a unique and poorly understood mechanism owing to the bicistronic nature of the L1 mRNA. Because bicistronic mRNA transcripts are not normally encountered by mammalian translational machinery, much of what we know about L1 translation is speculative. Bicistronic mRNAs normally contain an internal ribosomal entry site (IRES), which helps mediate proper ribosomal initiation leading to accurate translation of the second reading frame.
The translation of L1 ORF1p is generally accepted to occur via a conventional 5’ methyl-guanosine cap-dependent mechanism, however the translation of L1 ORF2p is less understood. When expressed from the same mRNA, the levels of L1ORF1p produced are significantly greater than the levels of L1ORF2p produced off the same mRNA, leading to two models of ORF2p translation (Goodier, Ostertag et al. 2004, Li, Li et al. 2006).

In one model, L1 ORF2p translation initiation occurs at an IRES located at the end of the L1 ORF1 domain. In this model, there is also evidence that translation of L1 ORF1p occurs from a separate IRES located upstream of the ORF1 coding region, accounting for the differing levels of protein production as the result of two IRES with differing efficiencies (Li, Li et al. 2006).

In the second model, L1 ORF2p translation occurs through an unconventional mechanism independent of any identifiable IRES. This model involves translation termination by the ribosome after ORF1p translation and reinitiation of translation for ORF2p. In this model, the differential protein levels are explained by inefficient translational reinitiation following ORF1p translation termination (Alisch, Garcia-Perez et al. 2006).

1.3.C L1 Localization

Because of its mechanism of insertion, we know that L1 proteins must localize to both cellular compartments: nuclear and cytoplasmic. Translation of L1 obviously occurs in the cytoplasm, however it is also apparent that in order to successfully retrotranspose to a new location, the L1 RNP must somehow enter
the nucleus and gain access to the genomic template. A number of groups have reported differing findings on the subcellular localization of L1 proteins.

In 2010, the Kazazian lab used two systems to detect L1 RNA and protein expression in living cells (Goodier, Mandal et al. 2010). The first made use of an MS2-EGFP fusion construct designed to be cotransfected with an episome expressing a tagged L1 RNA, allowing the MS2-EGFP to become tethered to the stem loops and mark the L1 RNA in the cell. Using this system, Goodier et al. showed that L1 RNA concentrates in cytoplasmic foci. To visualize L1 protein localization, the authors used immunohistochemistry (IHC) to detect L1 ORF1 protein in cells. In doing so, they confirmed that L1 ORF1 protein co-localizes with L1 RNA in cytoplasmic foci, which the authors go on to demonstrate represent stress granules (Goodier, Zhang et al. 2007, Goodier, Mandal et al. 2010). Other groups have confirmed these observations in similar manners (Harris, Normart et al. 2010, An, Dai et al. 2011). Unfortunately, however, there are some concerns with the methodology of these experiments. Because they were performed with tagged L1 proteins, these experiments were looking at a non-native subset of L1 proteins; ie, those that are generated as fusion proteins with their tag. The presence of a tag on a protein may in some way affect its subcellular localization. Additionally other experiments that use IHC to visualize L1 proteins lack the ability to differentiate a true L1 protein signal from background noise generated by non-specific binding products from the antibody used for the IHC. In fact, most currently available L1 antibodies present numerous off-target bands on Western blot, which complicates IHC analysis.
One study to use Western blots to determine the subcellular localization of L1 ORF1 protein looked at both untagged versions of ORF1 as well as two different tagged versions (Sokolowski, deHaro et al. 2013). In this study, the authors report untagged L1 ORF1 protein localization predominantly to the nuclear compartment with no detectable ORF1 protein levels in the cytoplasmic fraction. Intriguingly, the addition of a VP16 tag to L1 ORF1 protein substantially altered the subcellular localization and resulted in almost equivalent levels of nuclear versus cytoplasmic localization (Sokolowski, deHaro et al. 2013). Taken together, these differing results merit further analysis to tease out the details of L1 subcellular localization.

1.3.D L1 Expression in somatic cells

Understanding the subcellular localization of L1 proteins is important given recent evidence correlating L1 protein subcellular localization with cancer prognosis (Harris, Normart et al. 2010). Harris et al. looked at L1 ORF1 expression by immunohistochemistry in a number of normal human cell lines, human cancer cell lines, human tumors, and matched normal human tissue. They found a strong increase of L1 ORF1 expression in tumor cells types when compared to normal tissues, with a particularly strong effect seen for breast tumors where 437/441 (99%) of breast tumors studied tested positive for L1 ORF1 protein expression. In this instance, the authors also noted a significant correlation between nuclear localization of L1 ORF1 protein and poor prognosis, indicating that increased L1 nuclear localization may be a contributing factor to tumorigenesis (Harris, Normart et al. 2010). More recently, another group
reported increased levels of L1 protein expression correlate with an advanced stage of tumor (Rodic, Sharma et al. 2014). Admittedly, the findings in these studies must also be interpreted with caution as they make use of immunohistochemistry which can suffer from off-target effects as discussed above.

Expression of L1 is not uniform across cell or tissue types. Many cell lines do not support expression of L1 mRNA or proteins, while others seem to support L1 expression extremely well (Perepelitsa-Belancio and Deininger 2003, Belancio, Roy-Engel et al. 2010). Northern blot analysis of L1 expression in esophagus, prostate, stomach, and heart tissues showed that these tissues support L1 RNA expression levels at about 80, 50, 150, and 200% of the levels expressed by HeLa cells, respectively, while no expression of L1 was detected in adrenal, spleen, kidney, cervix, or several other tissues (Belancio, Roy-Engel et al. 2010). Additionally, L1 RNA expression levels in testis were found to be extremely high, however these RNAs were not full length L1 RNAs but rather L1 RNAs that were severely restricted by processing through either premature polyadenylation or alternative splicing or some other mechanism (Belancio, Roy-Engel et al. 2010). It is important to note, however, that it remains possible for truncated L1 mRNAs to produce an ORF2 protein that can function to drive Alu retrotransposition and create DNA double strand breaks (Gasior, Wakeman et al. 2006, Belancio, Roy-Engel et al. 2010).
1.4 L1 replication cycle

The replication cycle of an L1 element begins with the RNA polymerase II mediated transcription of a full length genomic L1 element to yield a full length L1 mRNA, which is processed as a normal cellular mRNA would be (Figure 4). Transcription of L1 is initiated by the internal 5' UTR L1 promoter and relies on the presence of a YY1 transcription factor binding site within the 5' UTR (Swergold 1990).

The nascent L1 mRNA receives a 5' 7-methylguanylate cap and is polyadenylated before being exported from the nucleus (Dmitriev, Andreev et al. 2007). Nuclear export is potentially assisted by binding of NXF1 to the NBE located in the L1 3' UTR (Lindtner, Felber et al. 2002). In the cytoplasm, the bicistronic L1 mRNA is translated to make L1 ORF1p and ORF2p. These proteins act with a strong cis preference and preferentially bind to the L1 mRNA that encoded them, resulting in the formation of an L1 ribonucleoprotein (L1 RNP) (Wei, Gilbert et al. 2001). The L1 RNP reenters the nucleus and initiates insertion of a new L1 element into the genome via a process called target primed reverse transcription (TPRT) (Luan, Korman et al. 1993).
Figure 4
**Figure 4 The L1 replication cycle.** Replication of L1 begins with transcription of a genomic full length L1 in the nucleus (1). The L1 mRNA (red) is exported to the cytoplasm (2) where it is translated into ORF1 and ORF2 proteins (3). The newly generated ORF1 and ORF2 proteins act *in cis* to form a ribonucleoprotein (RNP) complex with the L1 mRNA from which they were generated (4). The L1 RNP complex is then imported back into the nucleus (5) where L1 integration occurs via target-primed reverse transcription (6) (see Figure 5).
The first step in the L1 TPRT process is the creation of a DNA nick by the L1 ORF2p-encoded endonuclease (EN) at a rough genomic consensus 5’-TTTT|AA-3’ (Figure 5). Following cleavage by the L1 EN, the poly-A tail of the L1 mRNA base pairs with the 3’-T-rich sequence at the site of the genomic break. Using the genomic DNA as a primer and the L1 mRNA as a template, the L1 ORF2p reverse transcriptase (RT) creates an L1 cDNA. Because L1 RT lacks sufficient processivity, the majority (~90%) of new L1 insertion events are 5’-truncated, with the average genomic element being around 1kb in length (Luan, Korman et al. 1993, Lander, Linton et al. 2001).

The generation of the second strand nick and resolution of the L1 element insertion event remain poorly understood. A second genomic nick must occur in order for the L1 cDNA to complete insertion. The nuclease responsible for the generation of the second strand nick has not been definitively shown, however there is speculation that either ORF2 endonuclease or the nucleotide excision repair (NER) flap endonuclease XPG (ERCC5) may be responsible for generation of the second strand nick in the TPRT process (Gasier, Roy-Engel et al. 2008). Resolution of the insertion event is also poorly understood, but results in the new L1 insert flanked on either end with target site duplications (TSDs), direct repeats usually ~20nt in length, but sometimes kilobases in length (Gilbert, Lutz-Prigge et al. 2002, Symer, Connelly et al. 2002, Gilbert, Lutz et al. 2005, Giasi, Preston et al. 2007, Wagstaff, Hedges et al. 2012).
Figure 5
Figure 5 L1 target primed reverse transcription (TPRT). The endonuclease domain of L1 ORF2 protein recognizes a rough genomic consensus sequence and creates a nick in the genomic DNA (1). The ORF2-induced nick creates a 3’ polyT overhang that anneals to the polyA tail of the L1 mRNA and primes cDNA reverse transcription by L1 ORF2 reverse transcriptase (2). Following reverse transcription, a second strand nick is generated and second strand DNA synthesis occurs via unknown mechanisms (3), resulting in the integration of a new L1 copy at a new genomic locus (4). Most often, this new genomic L1 copy is highly truncated at the 5’ end.
1.5 L1 trans-mobilized elements


Unlike L1 elements which require both L1 ORF1p and L1 ORF2p for their effective retrotransposition, the trans-mobilization of Alu elements by the L1
machinery requires only the L1 ORF2p (Dewannieux, Esnault et al. 2003). Although not strictly required for Alu mobilization, L1 ORF1p has been shown to enhance Alu retrotransposition, possibly through an interaction between the ORF1p and the Alu RNA (Wallace, Wagstaff et al. 2008). Interestingly, an artificial RNA polymerase II-expressed Alu element requires L1 ORF1p for its mobilization, indicating that the polymerase (which is RNA pol III for Alu elements) may somehow dictate the requirement for L1 ORF1p. Interestingly, a second non-autonomous retroelement, SINE/VNTR/Alu (SVA), is transcribed by RNA polymerase II and shows varying degrees of reliance on L1 ORF1p (Hancks, Goodier et al. 2011, Raiz, Damert et al. 2012). As a result, the exact details of L1 protein requirements in trans remain to be fully elucidated.
1.6 L1 toxicity and disease

The ability of L1 to cause disease is predominantly a consequence of L1 endonuclease function (Feng, Moran et al. 1996). Through the use of a functional endonuclease, L1 is responsible for DNA double strand break formation, insertional mutagenesis by not only L1 but also Alu, SVA, and processed pseudogenes, and Alu/Alu recombination (Batzer and Deininger 2002, Gasior, Wakeman et al. 2006, Hedges and Deininger 2007, Belancio, Hedges et al. 2008, Belancio, Deininger et al. 2009, Belancio, Roy-Engel et al. 2010) (Figure 6).

Because of their ability to actively mobilize throughout the human genome, retroelements have been implicated in a number of human diseases {reviewed in (Deininger and Batzer 1999, Hedges and Deininger 2007, Belancio, Hedges et al. 2008, Belancio, Deininger et al. 2009, Belancio, Roy-Engel et al. 2010, Beck, Garcia-Perez et al. 2011, Burns and Boeke 2012, Hancks and Kazazian 2012)}. As of now, 96 single-gene human diseases have been the result of a retrotransposition event of L1, Alu, SVA or pseudogenes. Diseases caused by retrotransposition are varied and range from hemophilia to muscular dystrophy to HNPCC colon cancer. Interestingly, there are six independent genomic locations at which at least two independent retroelement insertions have occurred,
indicating the potential for regions of the genome that are particularly prone to insertion of retroelements (Figure 6A).

Even more prevalent than direct insertional mutagenesis by retroelements are mutations that are the result of non-allelic homologous recombination (NAHR) between non-allelic Alu elements (Figure 6B). Such recombination events result in deletions, duplications, and translocations, and have been shown to result in as much as 0.5% of new human genetic diseases {reviewed in (Batzer and Deininger 2002, Hedges and Deininger 2007)}. Intrachromosomal NAHR between Alu elements resulting in deletions occurs when a DNA double strand break (DSB) occurs between two Alu repeats, resulting in either single strand annealing (SSA) or non-homologous end joining (NHEJ) repair of the break, resulting in an inter-Alu deletion (Elliott, Richardson et al. 2005). One major source of genomic DSBs is from the function of L1-encoded endonuclease, which has been shown to result in the formation of DSBs in vivo (Gasior, Wakeman et al. 2006) (Figure 6C). Additionally, L1 endonuclease has been shown to initiate Alu-mediated deletions in a tissue culture based reporter system (Deininger lab, unpublished data).

In addition to the diseases discussed above, a number of recent groups have reported somatic mobilization of L1 in a number of human cancers (discussed in Section 1.9). While these associations are not proven to be causative, there is a growing body of evidence suggesting some role for L1 in a growing number of human diseases.
Figure 6

A. Retroelement insertion

B. L1-induced DSBs

C. Unequal recombination

- Deletion
- Duplication
Figure 6 Mobile elements induce DNA damage by multiple mechanisms. (A) Retrotransposons insert themselves in new genomic loci resulting in insertional mutagenesis. (B) L1 expression generates DNA double strand breaks (DSBs) which are a highly mutagenic form of DNA damage. (C) Ectopic recombination between repetitive sequences results in non-allelic homologous recombination causing deletions and duplications of genomic sequence.
1.7 Cellular control of L1 retrotransposition

Because their expression and activity is deleterious to cells, retrotransposons are often down-regulated by a number of varied cellular processes including DNA repair, RNA interference, promoter methylation, and action by ABOBECs. Through its ability to cause double strand breaks, insertional mutagenesis, and Alu-mediated deletions (all discussed in Section 1.6), L1 exacts a significant toll on the human genetic landscape.

One of the powerful cellular processes controlling the activity of L1 is restriction of L1 expression by promoter methylation. The L1 promoter located in the 5' UTR of the L1 element contains a CpG island whose methylation controls the activity of L1 (Swergold 1990, Hata and Sakaki 1997). In malignant or transformed cells, genome wide hypomethylation may release L1 promoter regions from negative regulation and result in the upregulation of L1 expression in these cells (Weisenberger, Campan et al. 2005).

A second cellular pathway influencing L1 expression and activity is the presence of tissue-specific transcription factors. The L1 internal promoter region has been shown to contain binding sites for the runt-domain transcription factor 3 (RUNX3). Abolishing the RUNX3 binding site results in a decrease in L1 transcription and retrotransposition (Yang, Zhang et al. 2003). Providing further evidence for a role for RUNX3 in expression of L1, RUNX3 expression has been
found to be elevated in testis, corroborating the elevated expression of L1 in germ cells.

Additionally, RNA interference is a mechanism used by cells to limit the activity of L1. In 2006, the Kazazian lab showed that L1 has the ability to self-regulate itself via RNAi (Yang and Kazazian 2006). L1 antisense promoter transcripts were shown to generate L1-specific siRNAs, which in turn caused the degradation of the 5' UTR-derived L1 transcript and resulted in the inhibition of L1 retrotransposition both *in vitro* and *in vivo*.

In addition to the mechanisms discussed above, there are also various cellular proteins that have been shown to effect L1 mobilization. The viral defense proteins of the apolipoprotein B mRNA editing enzyme (particularly APOBEC3) have been shown to limit the rate of L1 retrotransposition (Stenglein and Harris 2006, Hulme, Bogerd et al. 2007). This limiting action appears to be independent of the deaminase activity of the APOBEC proteins and acts on the L1 RT to reduce the production of L1 cDNA (Horn, Klawitter et al. 2014).

Another set of cellular proteins that appear to modulate the activity of L1 are various DNA repair proteins. Ataxia telangiectasia mutated kinase, a serine/threonine protein kinase recruited to the site of DNA double strand breaks appears to inhibit the rates of L1 retrotransposition in many cell types, as do members of the non-homologous end joining pathway of double strand break repair (Gasior, Wakeman et al. 2006, Suzuki, Yamaguchi et al. 2009, Coufal, Garcia-Perez et al. 2011, Wallace, Gasior et al. 2013). Additional DNA repair
pathways thought to modulate L1 activity include the mismatch repair machinery (Deininger lab, unpublished data) and the nucleotide excision repair pathway (Gasior, Roy-Engel et al. 2008). Particularly, multiple proteins of the nucleotide excision repair pathway (XPA, XPD, XPB, ERCC1/XPF) have been shown to have strong effects on the rate of L1 retrotransposition in a cultured cell model implicating the entire NER pathway in the retrotransposition process.
1.8 L1 retrotransposition reporter assay

The first assay to directly assess the jumping of a retrotransposable element was an assay developed to detect the mobilization of L1 elements in a tissue culture system (Figure 7). This elegant system relies on an antibiotic selection cassette cloned into the 3’ UTR of the sequence of a genomic L1 element known to be active. This selection cassette is interrupted by a gamma-globin intron placed in the antisense orientation and is flanked by splice donor and acceptor sites in the sense direction. This design is such that a functional selectable marker (in this case the neomycin resistance gene driven by an SV40 promoter) is only generated if the L1 transcript is spliced, reverse transcribed, and reintegrated into a new chromosomal location, allowing transcription of the neomycin resistance gene. This system ensures that G418-resistant colonies can only result from successful L1 retrotransposition events. This system has allowed for the quantitative detection of new L1 retrotransposition events in tissue culture, and has been adapted for use in detecting Alu and SVA retrotransposition events (Moran, Holmes et al. 1996, Dewannieux, Esnault et al. 2003, Hancks, Goodier et al. 2011, Raiz, Damert et al. 2012).
Figure 7
Figure 7 L1 retrotransposition reporter assay. Rates of L1 retrotransposition in tissue culture can be measured using a retrotransposition reporter cassette. A full length L1 element is tagged at the 3’ end with a reporter gene in the antisense orientation. The reporter gene is interrupted with an intron in the sense orientation so reporter gene expression is only attained after transcription, splicing and retrotransposition.
1.9 L1 high throughput sequencing methods

To date, a number of high throughput sequencing methods have been used to study repetitive element biology. Because of their repetitive nature, interspersed repeats are often automatically removed from high throughput sequencing datasets in the post-processing phase of data analysis. Because they exist in hundreds of thousands or even millions of copies in the human genome, repetitive DNA must be analyzed by specially designed analysis pipelines that eliminate the trimming that often occurs with sequencing data. To this end, a number of laboratories have developed tools for the analysis of repetitive elements using high throughput sequencing. The tools developed range from targeted enrichment protocols that specifically isolate and identify regions containing a repetitive element to modified analysis pipelines that begin with previously completed whole genome sequencing and use a custom-designed analysis pipeline to extract read pairs containing repeats.

1.9.A L1-Seq

The first laboratory to successfully develop an L1-specific high throughput sequencing approach was the Kazazian lab (Figure 8). Their technique relies on a hemi-specific nested PCR. The initial reaction consists of five cycles of single asymmetric primer extension off of a human-specific L1 primer followed by hemi-
specific PCR using an L1-specific primer paired with one of eight degenerate primers, both tagged at the 5’ end with Illumina sequencing adapters, which allows for sequencing on an Illumina Genome Analyzer. Using this technique, the authors identified 1139 L1 elements in 25 individuals with a 90% sensitivity and 80% specificity (Ewing and Kazazian 2010).

Because only eight degenerate primers were used in the library preparation, this L1-seq method is limited in the number of independent reads that can be generated from each L1 element. Additionally, because of the nature of degenerate priming, some loci will be more prone to amplification during library preparation, resulting in an overrepresentation of these specific elements and a potential loss of overall signal. A further concern is the initial asymmetric primer extension which makes use of a human-specific L1 primer. Because there are >1000 L1 loci in the genome that can act as primer binding sites for the L1-specific primer used in this initial priming step, the potential for L1 elements in opposite orientations creating artifacts is a concern.

The L1-seq experiments described by the Kazazian lab contributed a great deal of information to the knowledge of L1-based sequence variation between individuals in a given population, however this study focused exclusively on detecting 3’ ends of L1. Due to the nature of the sequencing method, no information on the extent of full length L1 elements was obtainable using this technology.
Figure 8

Asymmetric PCR

gDNA

Hemispecific PCR

5mer

5merN5Seq

Adapter 2

Adapter 2

L1Hs 3' UTR

pA

pA

pA

pA

L1Hs 3' UTR

L1Hs 3' UTR

L1Hs 3' UTR

Adapter 1

AC

AC

AC

51
**Figure 8 L1-Seq.** Restriction digested DNA is subjected to five cycles of asymmetric PCR using an L1Hs-specific primer allow for amplification of fragments containing L1 elements. Element-specific and degenerate primers are then used to generate Illumina sequencing libraries, which are sequenced on an Illumina Genome Analyzer.
1.9.B Transposon-seq

A second method for detection of L1 elements was developed by the Devine lab and named transposon-seq (Figure 9). This technique relies on restriction digestion of genomic DNA followed by Roche 454 linker ligation and nested PCR using L1 and linker specific primers, generating a library compatible with Roche 454 sequencers. Using their transposon-seq pipeline, Iskow et al. identified 650 previously unreported L1 insertions in 76 individuals with an 89% sensitivity and 97% specificity (Iskow, McCabe et al. 2010).

One additional finding reported by Iskow et al. was the presence of a somatic L1 insertion event in a lung adenocarcinoma cell line that was not present in the matched normal DNA from the same individual. Given the small number of lung tumors analyzed (eight total) this represents the potential for high levels of L1 mobilization in lung tumors, and was the first report of identification of an L1 insertion event in a tumor using high throughput sequencing.

Similarly to L1-seq, transposon-seq was designed to identify the 3’ end of L1 elements, and as such, yields no direct information about full length L1 elements without individual PCR validation of selected hits. Similarly, there are a number of technical concerns with the transposon-seq method that leave room for improvements. Transposon-seq begins with restriction digested genomic DNA followed by linker ligation. This technique results in DNA ends that are a fixed
distance away from any given L1 elements, preventing visualization of PCR artifacts. Because all libraries generated from a single L1 locus will be cut in the same region of flanking DNA, there is no way to isolate independent linker ligation events from PCR duplicates because all reads from that locus will appear identical. This makes it hard to get a good idea of the ability of differentiate rare L1 insertion events from common ones. Additionally, because linker ligation occurs in a random manner, there exists the potential for ligation artifacts of blunted genomic DNA fragments.
Figure 9
Figure 9 Transposon-Seq. Restriction digested DNA is ligated to a linker sequence on both ends. A retrotransposon and linker specific primer are used to PCR amplify ligation fragments that contain a repeat. Nested PCR is used to add barcodes required for 454 pyrosequencing.
1.9.C ME-Scan

In 2010, Witherspoon et al. developed a pipeline for the detection of Alu insertions in human genomes which they call ME-scan (Figure 10). In their study of four unique individuals, they found 487 non-reference Alu insertions with 95% sensitivity and specificity. ME-scan is based on the standard Illumina library preparation protocol and begins with a random shearing of genomic DNA by sonication followed by end repair and linker ligation to both DNA ends. Nested PCR with L1-specific and linker-specific primers occurs and generated libraries are sequenced on an Illumina genome analyzer using linker and Alu-specific primers to initiate sequencing. Random shearing coupled with a custom sequencing primer allows for the differentiation of PCR duplicates as well as detection of the exact Alu-genomic DNA junction (Witherspoon, Xing et al. 2010).

In a later paper, Witherspoon et al. expanded their ME-scan analysis to 169 independent individuals from 12 diverse human populations. In this experiment, the authors describe ~2500 novel Alu insertions that are not present in known databases of Alu elements or in the human genome reference build (Witherspoon, Zhang et al. 2013). However, while it is the most robust of the high throughput mobile element detection methods discussed, ME-scan is designed to isolate only Alu elements from human genomes, and no attempts were made to assay L1 insertion events.
Figure 10

Genomic  Alu  Genomic

Shear; end repair; ligate

B

L  Alu  L

L  Alu  L

Genomic  Alu  Genomic
**Figure 10 ME-Scan.** Genomic DNA is randomly fragmented by sonication and ligated on both ends to adapters of known sequence. PCR using Alu-specific and adapter-specific primers followed by selection for biotin tags allows for generation of libraries that are sequenced on an Illumina Genome Analyzer.
1.9.D RC-Seq

In an effort to improve the specificity of existing L1 detection techniques, the Faulkner lab in 2011 attempted to use a hybridization capture approach to detect somatic L1 and Alu insertion events in the brains of three independent individuals (Figure 11). In this report, the authors design hybridization probes based on the 5’ and 3’ termini of hundreds of full length L1 and Alu elements in the human genome. Following standard Illumina library preparation, samples are hybridized to the L1 and Alu element capture arrays and then undergo PCR amplification to amplify post-hybridization libraries. Following this, samples are sequenced on an Illumina Genome Analyzer (Baillie, Barnett et al. 2011).

Using this approach, Baillie et al. report an astounding ~7,700 somatic L1 and ~13,700 somatic Alu insertion events in three independent individuals. While these numbers are staggering {note: Witherspoon et al. only find ~2500 germline Alu polymorphisms in 169 different individuals (Witherspoon, Zhang et al. 2013)}, one must be cautious about their implications given some major flaws with the approach used in this study. While the authors report a 100% and 80% validation rate for L1 and Alu, respectively, a closer inspection of the reported sequences reveals major flaws that call into question the validity of the results (Baillie, Barnett et al. 2011).
A major concern with the validation is the use of post-hybridization DNA rather than genomic DNA for the purposes of validation PCRs. This strategy allows any ligation artifacts that may have occurred during sample preparation to be “validated” by post-sequencing validation PCR analysis. Interestingly an extremely high percentage (71%) of “validated” somatic L1 inserts contain 5’ transductions. This rate is about five- to ten-fold higher than what has previously been reported in the literature. Even more questionably, 40% of the validated Alu somatic insertions contain 5’ transductions (a physical impossibility given the mechanism of Alu element transcription).

These inconsistencies warrant a closer inspection of the sequences that Baillie et al. call validated. In doing so, an alarming discovery is made. In three independent instances, the 5’ transduced sequence in the validated product contains Illumina linker sequence. The only way for this sequence to appear in the validated product is via a ligation artifact. This unfortunately introduces a great deal of question into the method presented by Baillie et al. for the detection of repetitive elements from human genomes.

In a paper from 2012, Solyom et al. reported the use of L1-seq coupled with RC-seq in parallel to detect L1 insertion events in 16 colorectal tumors. Using a combination of L1-seq and RC-seq, the authors detected a total of 107 potential tumor-specific L1 insertion events, however only 69 (64%) were able to be validated by PCR. Interestingly in a subset of 11 high-confidence L1-seq insertion events, only four were able to be validated by RC-seq indicating the increased sensitivity of L1-seq over RC-seq. Further, of eight high-confidence
RC-seq insertion events that were not detected by L1-seq, only one was able to be validated by PCR indicating a significant false discovery rate using RC-seq (Solyom, Ewing et al. 2012).

In a later paper, Shukla et al. reported improvements to RC-seq which they used to detect novel L1 insertion events in 19 hepatocellular carcinoma genomes. The updated RC-seq protocol incorporated multiplex liquid-phase sequence capture and an updated probe set as well as a reduced insert size that allowed for overlapping paired-end reads to be used to obtain single-nucleotide resolution of insertion junctions. Using their improved RC-seq protocol, Shukla et al. were able to identify 12 tumor-specific L1 insertion events, all of which they indicate passed validation (Shukla, Upton et al. 2013). Unfortunately, however, no actual data is given in the manuscript showing validation of the twelve tumor-specific insertion events, and given the poor quality of the original RC-seq report, these results must be taken with a bit of skepticism (Baillie, Barnett et al. 2011).
Figure 11

Standard Illumina library

L1

L1 capture probes

L1 Enriched Libraries

L1

L1

L1
Figure 11 RC-Seq. Restriction digested DNA is ligated to known adapter sequences on both ends. Hybridization to retrotransposon-specific probes occurs in liquid phase, and selection for biotin tags generates fragments that are PCR amplified using transposon- and adapter-specific primers to generate libraries that are sequenced on an Illumina Genome Analyzer.
1.9.E Computational approaches

In addition to targeted sequencing approaches to detect L1 insertions, there have been multiple additional efforts to delineate L1 insertion events in tumors and individuals using data mining of readily available whole genome sequencing data sets. In one report, Lee et al. describe the analysis of 43 high-coverage whole genome sequencing data sets comprising five cancer types: colon, ovarian, prostate, multiple myeloma, and glioblastoma using a customized bioinformatics pipeline they termed transposable element analyzer (Tea) (Figure 12). Tea makes use of “repeat-anchored mate” reads (RAMs) in which one read of a read pair maps to a repeat sequence and the other maps somewhere within the human genome. Using Tea, the authors report extensive levels of somatic L1 retrotransposition in tumors of epithelial origin (colon, ovarian, and prostate) but not in tumors of other cell origin. Together, Lee et al. report 183 high-confidence L1 insertion events (28 in four colon cancer samples, four in seven prostate cancer samples, four in eight ovarian cancer samples, and none in either seven multiple myeloma samples or 16 glioblastoma samples), with an overall PCR validation rate of 97% (Lee, Iskow et al. 2012).

Some minor concerns are raised about the data analysis performed by Lee et al. with respect to recovery of L1 insertion target site duplications. Upon retrotransposition, new L1 elements are expected to be flanked by target site...
duplications (TSDs) on the order of ~15 base pairs and end with polyA tails. While a significant proportion of the L1 insertion events recovered by Lee et al. (64%) had TSDs ≥5 base pairs and polyA-tails, a significant fraction (i.e., the remaining 36%) did not exhibit these characteristic hallmarks of L1 insertion events. While this result could be explained by various retrotransposition oddities, this rate of non-canonical L1 insertion events is significantly higher than has been reported.

An additional report using data from the 1000 Genomes Project by Stewart et al. reported L1 insertion polymorphisms in 179 individuals sequenced to low coverage (1-3X) as part of the 1000 Genomes Project (Stewart, Kural et al. 2011). The method employed by this group is similar to that employed by Lee et al. and uses read coverage across the genome to identify regions where one read of a read-pair maps to an L1 reference sequence and the other pair maps to a known region in the reference genome where no known L1 exists (Stewart, Kural et al. 2011, Lee, Iskow et al. 2012).

Using this bioinformatic approach, Stewart et al. identified ~1000 polymorphic L1 elements in the 179 individuals tested with ~96% specificity and ~75% sensitivity. The low sensitivity obtained by this analysis is almost certainly due to the relatively low sequencing depth of the samples used for this analysis. In fact, when Stewart et al. looked at two high coverage (15-40X) family trios, the sensitivity improved to ~90%. Unsurprisingly, the authors found no evidence of de novo L1 insertion events in either offspring of the two trios, in agreement with
estimates of the rate of L1 insertion events at one in every ~200 live births (Brouha, Schustak et al. 2003).

One final recent study out of Matthew Meyerson’s lab used high-depth paired whole genome sequencing data to analyze L1 retrotransposition in 200 tumors and matched normal pairs in a pipeline they call TranspoSeq. Using TranspoSeq, Helman et al. were able to classify 810 somatic retrotransposon insertion events, primarily in lung squamous, head and neck, colorectal, and endometrial carcinomas with a sensitivity of about 83% (Helman, Lawrence et al. 2014).
Figure 12

Repeat anchored mate reads

Inserted TE

mutated genome

Ref genome
Figure 12 Transposable element analyzer (Tea). High-depth paired end sequencing data from tumor and matched normal samples was put through a custom analysis pipeline where they are mapped to both the reference genome and to a custom repeat assembly. This generates two types of supporting reads: repeat-anchored mate (RAM) reads and clipped reads which support the presence of a non-reference retroelement insertion at that location.
1.10 Additional L1 detection methods

In addition to the high throughput methods for L1 detection discussed above, a number of groups have also reported a variety of low-throughput detection strategies. While a number of these rely on tagged L1 elements and lack the ability to discern insertions resulting from endogenous, native L1 elements (L1 rescue, L1 inverse PCR), others have the ability to identify native L1 elements (TIP-chip, fosmid-based L1 detection). These methods are detailed below.

1.10.A L1 rescue

The first low throughput method for the detection of L1 insertion events is the L1 recovery "rescue" system (Figure 13). This system works by transfecting cells with an L1 tagged in its 3’ UTR with a reporter cassette designed to detect retrotransposition events as neomycin resistant colonies. The rescue cassette contains a neomycin resistance gene in the opposite orientation of the L1. The neomycin resistance gene is interrupted by a sense intron similar to that found in the standard L1 retrotransposition cassette. The one addition to this system that allows for recovery of new L1 inserts is the addition of a CoIE1 bacterial origin which allows for extraction of genomic DNA from transfected neomycin resistant
cells and bacterial transformation and recovery of individual retrotransposition events (Gilbert, Lutz-Prigge et al. 2002, Gilbert, Lutz et al. 2005).

Although this method of L1 insert recovery is extremely tedious and time-consuming, it allows for the detection and analysis of the precise L1 insertion event and yields the 5' and 3' junctions of the insert, providing information about target site duplication and polyA-tail length. Additionally, because the new L1 element is assayed by primer walking, information is obtained about the length of the L1 insert and the degree of 5' truncation, 5' inversion, or genomic deletion at the site of insertion.

Using the L1 rescue system has allowed the characterization of typical de novo L1 insertion characteristics from 100 rescued L1 insertion events. Typical L1 insertion events were found to insert at the predicted L1 endonuclease cleavage site (TTTT/A), were flanked by target site duplications typically 15-20 base pairs in length, and ended with polyA tails typically between 60 and 100 adenines in length. Additionally, frequent genomic deletions and L1 inversions were observed, consistent with what is observed in the genome. Approximately 6% of L1 inserts were full length (ie, included the entire L1 and the neomycin selection cassette), while approximately 11% of the inserts incorporated greater than 6kb of sequence indicating they have retrotransposed a full length L1 equivalent of sequence. Using this, the percentage of newly inserted L1 elements expected to be full length range from 6-11% (Gilbert, Lutz-Prigge et al. 2002, Gilbert, Lutz et al. 2005).
Figure 13
Figure 13 Recovery of *de novo* L1 insertions using the L1 rescue cassette.

The L1 rescue cassette is a modified version of the L1 retrotransposition reporter cassette (Figure 7) that contains, in addition to a reporter gene for antibiotic selection, a bacterial origin of replication. Following selection for retrotransposition events in mammalian cells, genomic DNA is harvested, digested, and self-ligated. Circular fragments containing L1 retrotransposition events are able to grow as plasmids when transformed into *E. coli* allowing for isolation and sequencing of the genomic insertion site.
1.10.B Transposon insertion profiling chip (TIP-chip)

An additional low-throughput method for the detection of transposable elements in human genomes is transposon insertion profiling chip (TIP-chip) developed by the Boeke lab at Johns Hopkins initially to identify the yeast retrotransposon Ty1 (Figure 14). TIP-chip is a genomic tiling microarray that begins with restriction digested genomic DNA and subjects it to vectorette PCR using a transposon specific primer, and then fluorescently labels the resulting products and hybridizes them to the TIP-chip microarray. Using TIP-chip, Whelan et al. successfully identified 94% of the known Ty1 elements in a specific lab strain of the yeast Saccharomyces cerevisiae (Wheelan, Scheifele et al. 2006). However, the yeast genome is significantly smaller than the human genome and much less replete with transposable elements (3% of the S. cerevisiae genome compared with at least 45% of the human genome).

In 2010, Huang et al. modified TIP-chip to detect L1 elements in human genomes. In this study, the authors were able to successfully detect a large fraction of all human genomic L1 elements, however their specificity and sensitivity were limited due to their approach (Huang, Schneider et al. 2010). Because TIP-chip relies on an L1 element being within a given distance from the end of a digested DNA fragment to be detected, it requires a combination of restriction enzymes be used if an even distribution across the genome is desired.
Even still, using the three restriction enzymes reported by Huang et al., approximately 10% of the genome will remain outside the range of detection of TIP-chip. Further, analysis of microarray data is tedious and provides little or nothing in terms of sequence information of the mobile element insertion event.
Figure 14

RE site → Genomic → Transposon → Genomic

PCR

Hybridize to transposon microarray
**Figure 14 TIP-ChIP.** Genomic DNA is subject to restriction digest fragmentation followed by digest-specific adapter ligation. Vectorette PCR amplifies ligated fragments and libraries are hybridized to a microarray specific for transposons.
1.10.C Fosmid-based approaches to full length L1 detection

All of the methods discussed prior to this point have only focused predominantly on 5’ truncated L1 fragments. While the majority of L1 insertion events are 5’ truncated, and this strategy is a good one for detecting de novo insertions, few studies have specifically targeted detection of full length L1 elements. Because all retrotransposition in the genome is the result of the activity of full length L1 elements, it is useful to know the extent of full length L1s in a given genome.

The first study to systematically attempt to detect full length L1 elements on the genome scale was performed in 2010 in the Moran Lab at the University of Michigan (Figure 15). In a prior study, the Kazazian lab demonstrated that the vast majority of L1 retrotransposition was the result of a number of particularly active L1 elements termed “hot” L1s. In this study, the authors investigated the 90 L1 elements in the human genome reference sequence with intact open reading frames. Among these, 44% are polymorphic with respect to presence in individuals and that ~50% are active in human culture. However, the vast majority of the retrotranspositional activity observed in culture was the result of six highly active L1 elements present in the human genome reference sequence (Brouha, Schustak et al. 2003).
Because the human genome reference build represents a panel of individuals rather than a single genome, Beck et al. sought to isolate full length L1 elements from single individual genomes using a fosmid-based approach. After obtaining 40 kb fragments of genomic DNA and creating a fosmid library, Beck and colleagues used paired end sequencing to map both ends of each fosmid and looked for discordantly mapping pairs that map to the reference genome 34 kb apart rather than the expected 40 kb. These events represent fosmids that may contain a 6 kb full length L1 in the fosmid but not in the reference genome and therefore represent a potential polymorphic L1 element in that individual (Beck, Collier et al. 2010).

Using this approach, Beck et al. reported 68 full-length L1 elements that are differentially present among six individuals and are not present in the human genome reference build. Of these 68 polymorphic L1 elements, the majority (37/68, 54%) are highly active in a cell culture model (Beck, Collier et al. 2010). This rate of “hotness” is significantly increased from that observed by Brouha et al. for reference full length L1 elements (6/82, 7%) and indicates that those L1 elements differentially present between individuals are more likely to be hot than those in the reference build of the human genome (Brouha, Schustak et al. 2003, Beck, Collier et al. 2010). The study by Beck et al. estimates the presence of approximately 3-9 hot L1s per typical individual, however one individual analyzed was found to have at least 14 potentially hot L1 elements (Beck, Collier et al. 2010).
Due to the screening method employed by Beck et al., it is not clear if all full length L1 elements in an individual were assayed. While the fosmid-based approach does not rely on PCR based methods for the detection of L1, it does require shearing genomic DNA to a tight size fractionation. It would be easy for fosmids containing full length L1 elements to escape detection if the size fractionation of the experiment was slightly off. Additionally, creation of a fosmid library followed by extensive screening is costly and time consuming and lacks the benefits of massive parallelization of high throughput sequencing-based approaches.
Figure 15

~40 kb

Concordant fosmid

~40 kb, LINE1?

Insertion in fosmid

ASP; Southern

Fosmid seq.

Full length L1 detection
Figure 15 Fosmid-based screening for full length L1s. Fosmid libraries designed to detect large insertion or deletion events were screened versus the human genome reference sequence to identify fosmids containing ~6kb inserts relative to the reference genome. Fosmids containing inserts were screened by hybridization to an L1Hs-specific probe and those positive for L1 sequence were confirmed by Southern blotting with an L1 5’ UTR-specific probe. Those positive for both L1 5’ and 3’ ends were sequenced for the presence of full length L1 elements.
1.11 Hypothesis and directions

The work described in this dissertation is the result of a need for a way to accurately and easily detect the panoply of full length L1 elements in any given genome. Because active full length L1 elements are the driver of all transposon-mediated mutagenesis, it is vital to know their abundance in any given individual. Further, since polymorphic full length L1 elements are more likely to be highly active than fixed full length L1 elements, a detailed knowledge of the degree of full length L1 polymorphism that exists in the population is vital to the study of L1 biology (Brouha, Schustak et al. 2003, Beck, Collier et al. 2010). Additionally, we believe that individuals with defects in repair processes that have been shown to limit L1 mobilization will experience a higher burden of L1 insertional mutagenesis. As a result, we would expect individuals defective for one such repair process, NER (XPA mutation) (see Section 1.7) to exhibit a higher overall rate of L1 insertions than a non-affected individual.
CHAPTER 2:

PCR of Anchored L1 Extensions (PALE): A METHOD FOR DETECTION OF POTENTIALLY ACTIVE FULL LENGTH L1 ELEMENTS

2.1. Introduction

Retrotransposon-mediated mutagenesis is responsible for ~0.5% of new genetic diseases. The sole driver of retrotransposon insertional mutagenesis events in humans is the retroelement L1. In order to mobilize itself as well as other retroelements, L1 must be expressed as an intact, full length 6kb element that is capable of coding for the two proteins essential for retrotransposition. There are ~90 such intact elements that are fixed in the population (ie, universally present in all humans). Of these, the majority have been inactivated by the accumulation of mutations in their open reading frames, rendering them unable to retrotranspose. However, ~40 fixed L1 elements remain modestly active (Brouha, Schustak et al. 2003).

The bulk of human retrotransposition, however, is the result of the mobilization of young, polymorphic L1 elements (ie, elements that are differentially present between individuals in a population). Because they have
inserted more recently in evolutionary time, these elements are more likely con
tain intact open reading frames lacking any inactivating mutations. Six such
elements identified in the human genome reference build have been reported to
account for >80% of all the retrotransposition from elements in reference genome
(Brouha, Schustak et al. 2003). These elements, however, while polymorphic,
likely represent fairly prevalent insertions and likely exist at moderately high allele
frequency because they were present in most of the individuals sequenced to
build the human genome reference sequence.

A more likely source of active L1s are the more rare allelic variants that
have inserted more recently in evolutionary time. Only very limited efforts have
been made to identify such variants, and none in a high-throughput manner. One
study identified 68 polymorphic full length L1 elements and found that >60% of
these were active, with >50% representing highly active L1 elements. By-and-
large, these elements represented low allele-frequency elements that were
relatively recently inserted into the human lineage (Beck, Collier et al. 2010). As
a result of this study, it is estimated that there are about 90 active L1 elements in
any given genome, however this number is based on estimations that have not
been validated in a systematic manner. Aside from this limited study, no attempts
have been made to determine the number of full length L1 elements in any given
individual. This information is critical to help understand the extent of variation
between individuals for the capacity to support L1 mobilization.

The recent advances in high-throughput sequencing have enabled a
previously unparalleled ability to analyze retrotransposon insertions. The vast
majority of these advances however, have focused on determination of the rates of de novo insertion events in various somatic tissues, particularly various types of cancer (Ewing and Kazazian 2010, Iskow, McCabe et al. 2010, Witherspoon, Xing et al. 2010, Baillie, Barnett et al. 2011, Stewart, Kural et al. 2011, Lee, Iskow et al. 2012, Solyom, Ewing et al. 2012, Shukla, Upton et al. 2013, Helman, Lawrence et al. 2014). While these studies provide vital information on the rate of insertions and the potential role of L1 in tumorigenesis, they lack information about which L1 elements are capable of mobilization in a given individual or tumor.

Particularly active L1 elements have been described in a number of independent cancer samples. Interestingly, specific L1 elements appear to be especially active with one example of a hot L1 on chromosome 22 resulting in 83 independent somatic insertion events in a sample of colon cancers (Helman, Lawrence et al. 2014, Pitkanen, Cajuso et al. 2014). The cohort of active L1 elements appears to differ between individual tumors with the hot L1s in one tumor, differing from the hot L1s in another tumor. This has led to the growing belief that variation in an individual’s hot L1s is responsible for differential rates of retrotransposition.

In order to determine the extent of variation between individuals for the number of full length L1s present as well as to get an idea of the average number of full length L1 elements, we sought to develop a high-throughput method for the detection of full length L1 elements from human genomes. We describe a method called PALE (PCR of Anchored L1 Extensions) that detects virtually all
full length L1 elements from a given genome with a high degree sensitivity and specificity (Figure 16).
Figure 16
**Figure 16 PCR of Anchored L1 Extensions (PALE).** Genomic DNA is randomly sheared by sonication. Sheared fragments undergo linear amplification by primer extension. Only those fragments containing the 5’ promoter region of full length L1 elements (large red arrow) will allow the primer to sit down and extend. Fragments containing truncated L1s (small red arrow) or no L1 will not primer extend from the L1 promoter-specific primer. During primer extension, untemplated adenines are added to the 3’ end of the primer extended fragment allowing for t-linker mediated ligation. Nested PCR using L1 5’ UTR-specific and linker-specific primers followed by gel-based size fractionation allows for generation of a library compatible with sequencing on an Illumina Genome Analyzer.
2.2. PCR of Anchored L1 Extensions (PALE) allows for high throughput detection of virtually all full length L1s in a given genome

2.2.A Development of PALE

Various anchored PCR strategies have been effectively employed to identify mobile element insertions in a whole-genome setting. They all make use of one primer within the known sequence of the repetitive element, and use different strategies to target a second primer outside of the element to allow for PCR-based amplification of the genomic fragment containing the repetitive element insertion. However, these methods are not without significant drawbacks.

Some methods ligate a linker of known sequence to a restriction site located randomly outside of the mobile element insertion. This method only creates a single amplified fragment size from each unique element in the genome due to the specific location of the restriction fragment end. Thus, it is impossible to tell when using restriction digested fragments whether duplicate sequences are independent ligations or duplicates created during the PCR cycles required for library generation. This becomes a problem when trying to determine the robustness of a technique. Without knowledge about whether a specific L1 element is detected multiple independent times, there is no way of knowing whether an amplicon is the result of a potential spurious ligation, limiting
the ability of such a technique to differentiate potentially rare insertion events from artifacts of library preparation.

A similar argument can be made for the use of an arbitrary primer to prime outside the mobile element. Much like with restriction digested fragments, the fragments generated by random priming may not be able to differentiate artifacts from legitimate element insertions. Furthermore, restriction sites located at different lengths away from the element will amplify with differential efficiencies, as will different arbitrary primers. Thus, these methods require pooling of multiple experiments using different restriction enzymes or primers to come close to saturating the potential insertions.

Random shearing of target DNA may help eliminate some of the problems described above, but even protocols that utilize randomly sheared DNA to apply anchors make use of a major ligation step during library preparation. These ligation steps, such as those used in the preparation of Illumina sequencing libraries lead to low levels of chimeric fragment ligation between genomic fragments that can confound data analysis (Quail, Kozarewa et al. 2008). In fact, a recent study purporting to show high levels of somatic retrotransposition in the human brain was in fact reporting only improper ligation events that occurred during library preparation (Baillie, Barnett et al. 2011).

PCR of Anchored L1 Extensions (PALE) is a unique L1-detection method based on the principles of random shearing combined with t-linker ligation-mediated PCR that takes advantage of the massive parallelization offered by
high throughput next generation sequencing technology (Glenn 2011, Quail, Smith et al. 2012) (Figure 16). PALE makes use of a primer specific to the promoter region of full length L1 elements that gets used in a single round of primer extension. Specific extension off of full length L1 elements allows for priming on all of the ~5000 full length L1 elements in the human genome. Priming off of full length L1s continues through the unique genomic sequence flanking the L1 element and generates a single adenine overhang. The vast majority of the genome does not contain a full length L1 element on which to prime, and as such will remain single-stranded so that little besides the extended mobile element fragments will have the adenine overhang required for duplex linker ligation via a 3’ thymidine overhang. Following ligation, PCR amplification using L1 promoter- and linker-specific primers allows for specific amplification of only those DNA fragments anchored by an L1 extension event. Additional Illumina adapter sequences are added in a subsequent PCR reaction and the PALE library is size fractionated on an agarose gel before final library amplification and library verification by BioAnalyzer (Figure 17).
Figure 17

A

B
Figure 17 PALE library quality control. PALE libraries are checked for size and quality by agarose gel electrophoresis (A) as well as capillary electrophoresis on an Agilent BioAnalyzer (B). Libraries in the size range of ~400-700bp are multiplexed and sent for high-throughput sequencing on an Illumina HiSeq 2000 Genome Analyzer.
After preparation, PALE libraries can be directly loaded onto the next generation sequencing platform for sequence determination and analysis (Figure 18). The benefit of PALE is each sample is generated from a randomly sheared fragmentation, making it easy to differentiate PCR duplicates from authentic detection of the same element multiple times, allowing for improved reliability of sequencing data (Figure 19). In addition, we reason that the ligation step used in PALE would be less likely to create artificial chimeras than a traditional linker ligation because of the nature of the 3’ overhangs generated by primer extension. Rather than random linker ligation to both ends of the DNA fragment, ligation of amplifiable fragments will only occur after successful primer extension has taken place using the L1 promoter-specific primer. Those fragments that do not properly ligate linker sequence to L1 ends will not amplify by future rounds of PCR and will therefore be able to achieve incredibly low rates of false positives.
Figure 18

Hybridize to flowcell

Bridge Amplicification

Dye-terminator sequencing
Figure 18 Illumina sequencing reaction. Denatured Illumina libraries are hybridized to an Illumina flow-cell using adapter sequences on either end of the sequencing library. Bridge amplification of fragments generates double strands which are again denatured. This generates a series of clusters of DNA on the flowcell which then undergo sequencing by a modified chain terminator reaction using four fluorescently labelled dNTPs which allow for detection of DNA sequences by a sequence reader. Massive parallelization allows for the generation of hundreds of Gb of data per flowcell.
Figure 19
Figure 19 Distinguishing PCR duplicates from independent ligation events.

(A) Genomic fragments subject to restriction digestion or those having undergone degenerate priming will result in a population of PCR fragments for which the sequencing read begins at the same location for each fragment. In this case, it is not possible to distinguish duplicates created during PCR from independent linker ligation events, which confounds analysis. (B) PALE benefits from random DNA shearing which allows unique ligation events to be detected from sequence mappings.
2.2.B PALE sensitivity

To determine the ability of PALE to accurately detect full length L1 elements from genomic DNA, we sought to determine the sensitivity of our assay. To achieve this, we sought to determine the efficiency of PALE at pulling out known, reference full length L1 elements that are fixed in the human population. These L1 elements represent evolutionarily established L1 elements that inserted in a primate genome before the split of *Homo sapiens*, and thus represent L1 elements universally present in all human genomes at a diploid level. A method that successfully targets full length L1 elements should be able to identify the vast majority of the fixed full length L1 elements in the genome.

We employed PALE to determine how many known, fixed full length L1 elements on three randomly selected chromosomes we could detect and to what degree we could detect them (Figure 20). We limited our initial analysis to three chromosomes to allow a more exhaustive manual analysis of any variants. Of the 543 fixed full length L1 elements on three randomly selected chromosomes, PALE successfully detected the vast majority with an average read depth of ten independent (having a different linker location) PALE reads per element. Of these, 511/543 (94%) were detected by at least one PALE read, with 502/543 (92%) detected by more than three independent PALE events (reads) (Figure 20). The 32/543 (6%) fixed full length L1 elements not detected by PALE were
located entirely in regions of repetitive DNA. Because they are located in repetitive regions, their unique genomic flank sequence which is used to identify the mapping location of a particular element is actually not unique, and instead lies within a repeat. Because of this, it is impossible to unambiguously determine the exact mapping location of these elements.
Figure 20

Histogram of known, fixed full-length L1s on Chromosomes 4, 12 and 17
Figure 20 PALE detects fixed full length L1s with high sensitivity. Histogram depicting the number of independent linker ligations for all of the known, fixed full length L1 elements on chromosomes 4, 12 and 17. The median number of ligation events per fixed full length L1 is nine. PALE detects 94% of known fixed full length L1 elements and those L1 elements not detected are missed because they lie in highly repetitive genomic regions.
To determine how PALE performed for non-reference full-length L1 elements (ie, those elements either present as a single copy or not present in any given individual) we assessed the detection rate of previously described polymorphic full length L1 elements on the same three chromosomes (Figure 21). These represent lower-frequency L1 alleles that may or may not be present in a given individual, and if present, are likely to be so only as a single allelic copy. We performed a similar verification experiment as described above, using three randomly selected chromosomes for analysis. While the three chromosomes selected were found to have 152 previously described full length polymorphic elements on them, we would only expect a small proportion of these elements to be assayed in our selected individual by PALE given the polymorphic nature of these elements. In fact, PALE determined that only 23 of these known polymorphic full length L1 elements were present in our tested individual. As with fixed L1 elements, PALE performed well for polymorphic elements, with a median of seven independent PALE reads per polymorphic element (Figure 21).

As an additional layer of quality control, the results of our PALE analysis were compared to a low-throughput method which identified 68 novel full length L1 elements in the genomes of six individuals of diverse backgrounds (Beck, Collier et al. 2010). Using PALE, we confirmed the presence of 29/68 of these non-reference full length L1 elements in at least one of our seven individuals, including the identification of elements with estimated allele frequencies as low as 2%, many of which were known to be hot for retrotransposition (Table 1) (Beck, Collier et al. 2010). Together, these results indicate PALE is able to detect
nearly all full length L1 elements in a given individual with an incredibly high degree of sensitivity.
Figure 21

Histogram of known, polymorphic full-length L1s on Chromosomes 4, 12 and 17.
Figure 21 PALE detects polymorphic full length L1s at similar levels as fixed L1s. Histogram depicting the number of independent linker ligations for known polymorphic full length L1 elements on chromosomes 4, 12 and 17. The median number of ligation events per fixed full length L1 is seven, indicating that PALE performs similarly for polymorphic elements as it does for fixed elements.
<table>
<thead>
<tr>
<th>flL1 ID</th>
<th># indiv.</th>
<th>Allele Freq (%)</th>
<th>Activity (% of L1.3)</th>
<th>flL1 ID</th>
<th># indiv.</th>
<th>Allele Freq (%)</th>
<th>Activity (% of L1.3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>11</td>
<td>0</td>
<td>36</td>
<td>2</td>
<td>9</td>
<td>&lt;1</td>
</tr>
<tr>
<td>2</td>
<td>?</td>
<td>21</td>
<td>69</td>
<td>37</td>
<td>2</td>
<td>43</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
<td>44</td>
<td>107</td>
<td>38</td>
<td>1</td>
<td>NI</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>119</td>
<td>39</td>
<td>0</td>
<td>NI</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>NI</td>
<td>118</td>
<td>40</td>
<td>0</td>
<td>0.35@</td>
<td>74</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>18</td>
<td>141</td>
<td>41</td>
<td>7</td>
<td>NI</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1.2/7*</td>
<td>0</td>
<td>42</td>
<td>6</td>
<td>58##</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>NI</td>
<td>0</td>
<td>43</td>
<td>?</td>
<td>NI</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>22</td>
<td>130</td>
<td>44</td>
<td>0</td>
<td>NI</td>
<td>69</td>
</tr>
<tr>
<td>10</td>
<td>5</td>
<td>27</td>
<td>101</td>
<td>45</td>
<td>0</td>
<td>NI</td>
<td>161</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>21</td>
<td>133</td>
<td>46</td>
<td>0</td>
<td>NI</td>
<td>54</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>NI</td>
<td>0</td>
<td>47</td>
<td>2</td>
<td>NI</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>9</td>
<td>46</td>
<td>48</td>
<td>6</td>
<td>NI</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>21</td>
<td>118</td>
<td>49</td>
<td>3</td>
<td>NI</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>50</td>
<td>0</td>
<td>NI</td>
<td>79</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>NI</td>
<td>NI</td>
<td>51</td>
<td>0</td>
<td>NI</td>
<td>38</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>6</td>
<td>86</td>
<td>52</td>
<td>0</td>
<td>NI</td>
<td>11</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>NI</td>
<td>3</td>
<td>53</td>
<td>7</td>
<td>NI</td>
<td>57</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>NI</td>
<td>0</td>
<td>54</td>
<td>0</td>
<td>NI</td>
<td>191</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>NI</td>
<td>63</td>
<td>55</td>
<td>4</td>
<td>NI</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>0</td>
<td>NI</td>
<td>0</td>
<td>56</td>
<td>3</td>
<td>NI</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>0</td>
<td>NI</td>
<td>0</td>
<td>57</td>
<td>0</td>
<td>NI</td>
<td>21</td>
</tr>
<tr>
<td>23</td>
<td>4</td>
<td>26</td>
<td>15</td>
<td>58</td>
<td>3</td>
<td>NI</td>
<td>13</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>0.25</td>
<td>98</td>
<td>59</td>
<td>7</td>
<td>NI</td>
<td>23</td>
</tr>
<tr>
<td>25</td>
<td>7</td>
<td>NI**</td>
<td>&lt;1</td>
<td>60</td>
<td>?</td>
<td>NI</td>
<td>NI</td>
</tr>
<tr>
<td>26</td>
<td>0</td>
<td>0.25</td>
<td>126</td>
<td>61</td>
<td>0</td>
<td>NI</td>
<td>8</td>
</tr>
<tr>
<td>27</td>
<td>7</td>
<td>High***</td>
<td>0</td>
<td>62</td>
<td>2</td>
<td>NI</td>
<td>&lt;1</td>
</tr>
<tr>
<td>28</td>
<td>0</td>
<td>NI</td>
<td>62</td>
<td>63</td>
<td>7</td>
<td>NI</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>0</td>
<td>0****</td>
<td>&lt;1</td>
<td>64</td>
<td>0</td>
<td>NI</td>
<td>157</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>0.25#</td>
<td>170</td>
<td>65</td>
<td>0</td>
<td>NI</td>
<td>36</td>
</tr>
<tr>
<td>31</td>
<td>0</td>
<td>0.17@</td>
<td>120</td>
<td>66</td>
<td>0</td>
<td>NI</td>
<td>&lt;1</td>
</tr>
<tr>
<td>32</td>
<td>0</td>
<td>7.8#</td>
<td>68</td>
<td>67</td>
<td>0</td>
<td>NI</td>
<td>129</td>
</tr>
<tr>
<td>33</td>
<td>4</td>
<td>NI</td>
<td>6</td>
<td>68</td>
<td>0</td>
<td>NI</td>
<td>139</td>
</tr>
<tr>
<td>34</td>
<td>1</td>
<td>2</td>
<td>41</td>
<td>69</td>
<td>0</td>
<td>NI</td>
<td>73</td>
</tr>
<tr>
<td>35</td>
<td>0</td>
<td>9#</td>
<td>0</td>
<td>70</td>
<td>0</td>
<td>NI</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 1 PALE-detection of previously described polymorphic full length L1s. PALE successfully detects many previously described full length L1 elements including some with allele frequencies as low as 2% and those with high levels of retrotranspositional activity. flL1 ID = unique flL1 identification number. # indiv. = how many out of the seven individuals tested by PALE contained that full length L1.
2.2.C PALE specificity

We also sought to determine the rate of false positives (specificity) detected using PALE. To this end, we applied PALE to seven independent individuals to identify novel full length L1 elements that had not previously been described. We randomly chose 15 novel non-reference full length L1 elements detected by PALE to perform PCR- and Sanger sequencing-based validation using L1 junction PCR to amplify across the 5’ L1 junction. For all analysis using PALE, we limited our selections to only elements that were hit by at least two independent PALE ligations. This method allowed us to determine the 5’ junction of the novel non-reference L1 element. We successfully validated 15/15 (100%) PALE-detected L1 polymorphisms by 5’ junction PCR and confirmed the insertions by Sanger sequencing of the 5’ junction (Figure 22, Table 2). Additionally, manual inspection of PALE reads detected no false positives.

Together, these data indicate PALE is a powerful tool to detect virtually all full length L1 elements (both fixed and polymorphic) in a given individual in a single experiment, allowing for the detection of novel non-reference full length L1s that may contribute to mutagenesis by mobilization in these individuals (Figure 23).
Figure 22
Figure 22 PCR validation of 5’-PALE events. (A) PCR primers were designed within the 5’ promoter region of L1 and within the unique genomic flanking sequence upstream of the element. (B) Representative gels showing PCR-based validation of two novel full length L1 elements detected by PALE. NTC, no template control. M, 1kb DNA ladder. Numbers indicate the individual in which presence or absence of the specific L1 element was tested.
<table>
<thead>
<tr>
<th>fL1 ID</th>
<th>Chr</th>
<th>5’ jcn</th>
<th>Ind. 1</th>
<th>Ind. 2</th>
<th>Ind. 3</th>
<th>Ind. 4</th>
<th>Ind. 5</th>
<th>Ind. 6</th>
<th>Ind. 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL022</td>
<td>2</td>
<td>82098043</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>FL043</td>
<td>3</td>
<td>80590165</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
</tr>
<tr>
<td>FL052</td>
<td>3</td>
<td>151148645</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FL054</td>
<td>3</td>
<td>169269964</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>FL055</td>
<td>3</td>
<td>186372121</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FL057</td>
<td>4</td>
<td>10632678</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FL107</td>
<td>6</td>
<td>142450103</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FL139</td>
<td>10</td>
<td>14310827</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FL166</td>
<td>13</td>
<td>19557911</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FL168</td>
<td>13</td>
<td>27339037</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FL178</td>
<td>14</td>
<td>71197795</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FL181</td>
<td>15</td>
<td>47507328</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>FL184</td>
<td>16</td>
<td>5196792</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FL202</td>
<td>18</td>
<td>15095248</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>FL206</td>
<td>18</td>
<td>49216219</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 2 Validation of PALE by 5’ L1 junction PCR. Fifteen randomly selected L1 elements detected by PALE were validated by PCR across the 5’ junction. The table indicates which of the seven individuals tested contain (X) or do not contain (-) the given full length L1 element.
Figure 23
Figure 23 PALE detects known and novel full length L1 elements. PALE reads were aligned to the genome and analyzed using CLC Genomics workbench. PALE detects both known (A) and novel (B) full length L1 elements. Green lines represent PALE library reads mapped to the human genome reference sequence. In (A) a known full length L1 element lies downstream of the reads indicating detection of a known element. In (B) there is no known L1 annotation downstream of the reads, indicating detection of a novel full length L1 element in this individual.
2.3 Discussion

Previous attempts to determine the extent of active L1s in human genomes have either relied on specific detection of particularly active L1 elements via tracking their insertion into new genomic loci, or have used low-throughput techniques to identify full length L1 elements. Development of a robust tool for the detection of potentially active L1 elements will allow for determination of the extent that active L1 elements populate human genomes and will allow prediction of L1-associated genomic risk by identifying individuals or tumors that may be particularly prone to high levels of L1 mobilization because of an increased burden associated with a greater number of active L1s.

We have developed a robust tool, PALE, to identify novel non-reference active L1 elements in individual genomes in a high-throughput manner. PALE makes significant improvements over previous L1 detection methods. Because DNA fragmentation for PALE occurs by random shearing of genomic DNA by sonication, PALE is able to detect L1 elements that may be located too distally from restriction enzyme cut sites used by other high-throughput sequencing methods. Additionally, random shearing allows for discrimination of independent ligation events from PCR duplicates and avoids biases that result from restriction digestion or random priming (Figures 16 and 19).
Because it is incredibly sensitive at detecting full length L1s, virtually all known and unknown active L1 elements will be discernible following application of PALE (Figures 20 and 21). The benefit of having an essentially non-existent false detection rate also allows PALE to be used to make confident determinations about the extent of active L1 elements in any given genome (Figure 22).

PALE also has benefits over other fosmid-based approaches used in the past in that it is extremely high throughput, allowing the detection of essentially all genomic L1 elements in a single PALE reaction. Additionally, the power of high-throughput sequencing allows for the multiplexing of multiple samples into a single lane of sequencing providing a cost-effective way to determine the full complement of active L1s in many individuals in a single experiment.
2.4 Materials and Methods

2.4.A Cell lines and oligonucleotides

Fibroblast cell lines GM01631, GM01632, GM05510, GM05568, GM16094, GM15983, and GM08207 were obtained from the Coriell Institute (Camden, NJ). All cell lines were maintained in Eagle’s Minimal Essential Media (EMEM) supplemented with non-essential amino acids, sodium pyruvate, and 10% fetal bovine serum. DNA oligonucleotides and duplex linkers were obtained from Integrated DNA Technologies (Coralville, IA).

2.4.B 5’ PALE library generation

Genomic DNA (gDNA) from fibroblast cell lines was extracted using the DNEasy Blood and Tissue Kit (Qiagen, Germantown, MD). DNA was sheared using a Diagenode BioRuptor (Denville, NJ) on high, 30 seconds on/30 seconds off for a total cycle time of 12 minutes. 50ng of BioRuptor-sheared gDNA was subject to a primer extension reaction using Taq polymerase (Promega, Madison WI) and L1 promoter-specific primer 1 (5’-GGTGAGGCAATGCCTCGCCCTGCTT-3’). Phosphorylated duplex T-linkers were synthesized by IDT (Coralville, IA) (Top: 5’-/Phos/GATCGGAAGAGCGTCTGAGGAGGAAAGAGTGTAGA-3’; Bottom: 5’-
TCTACACTTTTCCCTACACGACGCTCTTCCGATCT-3') and were ligated to
the primer-extended gDNA using T4 DNA ligase. First round PCR was performed
for 20 cycles using L1 promoter-specific primer 1 and linker primer 1 (5’-
ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’. One-million fold dilution of
PCR 1 was performed and 1uL of this dilution was subjected to a nested PCR
using L1 promoter-specific primer 2 (5’-
CTCGGCATTCCGTCTGAACCGCTCTTCCGATCTCTNNNNNCCACTGTCTGGCA
CTCCCTAGTGAGTGA-3’ and linker primer 1 for 25 cycles. PCR products
were run on an agarose gel and a gel slice at ~500-700bp was extracted using
the Qiagen Gel Extraction kit (Germantown, MD). Following extraction, the final
library was amplified using Phusion polymerase (Thermo, Waltham, MA) for 12
cycles as per the Illumina library generation protocol, and gel purified to yield the
final library.

2.4.C 5’ Library quantitation and Illumina sequencing

Illumina sequencing libraries were quantified by qPCR and checked for
quality by Agilent BioAnalyzer trace at Elim Biopharmaceuticals (Hayward, CA).
Illumina sequencing was performed on an Illumina HiSeq 2000 by Elim
Biopharmaceuticals (Hayward, CA).

2.4.D Reference Genome Masking

Reference repetitive element annotation files were downloaded from
UCSC genome browser as GTF files and used to annotate the GRCh37 (hg19)
reference genome in CLC Genomics Workbench (CLC Bio, Cambridge MA). Two custom GFF annotation files were also generated: 1) representing known polymorphic elements and 2) representing all known full length L1 elements plus 600bp of upstream genomic flanking sequence.

To identify and annotate 5' L1 regions of the genome, the first 300 base pairs of the L1.3 consensus sequence was aligned to the human reference genome (CRCh37/hg19) via NCBI BLAST (blastn algorithm). Initial search parameters were relaxed to increase sensitivity for older, more mutated elements (word size=7, gap cost=3, gap extension cost=3, match score=2, mismatch penalty=-3, minimum E score of 10)(Scott, Schmeckpeper et al. 1987, Altschul, Gish et al. 1990). Blast output was generated in tabular format. Because many of the hits obtained from this search consisted of isolated smaller regions (20-30bp) that were independent of any identifiable L1 element, we further filtered for those hits where alignments were made across >=250bp of the 300bp query. This greatly improved our specificity while maintaining sensitivity to older, more mutated elements. Tabular format blast results were converted to GFF via Perl scripting, and an additional 600bp of flanking sequence was added upstream and downstream to each matching position.

2.4.E 5’ PALE Mapping Strategy

FASTQ sequencing files generated by the Illumina HiSeq 2000 were analyzed for quality using FastQC. PCR duplicate reads were removed using custom in-house Perl scripting. Sequence read mappings were performed using
CLC Genomics Workbench. Single end genomic flank reads were mapped using a two-pronged mapping strategy. First, reads that either mapped to within 600bp of known, fixed full length L1 elements or those reads that did not map unambiguously were discarded, as they represented either known elements or unmappable reads. The remaining reads were then mapped uniquely to an annotated GRCh73/hg19 reference genome to identify novel polymorphic or potentially de novo full length L1 elements.

2.4.F PCR validation of PALE reads

Primers specific to the L1 promoter region and the unique genomic flanking region were designed using Primer3 software. PCR validation was performed using GoTaq (Promega, Madison WI) and Sanger sequencing was performed by Elim Biopharmaceuticals (Hayward CA). Sequence junctions were manually aligned to the L1.3 promoter region consensus sequence and junctions were determined. Unique flanking sequence was mapped to the human genome using BLAT (UCSC) to identify the exact L1-genome junction point (Kent 2002).
CHAPTER 3:

PALE discovery of novel full length L1 elements

3.1. Introduction

Current estimates of the number of active L1 elements that exist per individual come from two different experimental estimations. The first uses the total number of L1 elements fixed in the human genome reference sequence that are active as a starting point to develop an estimate of the total number of active L1s per individual. This estimation was done by starting with the known 40 active fixed reference L1s and adding to them the known polymorphic active elements and correcting for allele frequency to estimate the number of polymorphic elements likely to occur in any given individual. Using this estimation, the authors arrive at a number of approximately 90 active L1 elements per individual (with the number of hot, particularly active L1 elements around 10) (Brouha, Schustak et al. 2003).

The second group to estimate the total number of active L1s in a given individual used an in silico genotyping estimate in a single individual to identify how many of 68 known polymorphic L1 elements were present in a single
individual. After performing this analysis, the authors determined that this single individual contains ~25 polymorphic full length L1s, 14 of which are hot (Beck, Collier et al. 2010).

Both studies, however, have weaknesses that prevent them from being able to provide confident assessments of the numbers of active L1s per individual. The study by Brouha and colleagues only looked at a very small subset of active L1 elements (those present in the human genome reference build) and used mathematical modeling to attempt to make predictions about how many non-reference polymorphic L1s might potentially exist (Brouha, Schustak et al. 2003). Such predictions are for obvious reasons not ideal and rely highly on estimation. The study by Beck and colleagues also only genotypes for the presence of known polymorphic full length L1s and only does so in a single individual. Additionally, this study uses a low-throughput fosmid-screening approach that is likely to miss a good number of full length L1s that may be present in a given individual. To their credit, the authors of both studies acknowledge that given the limited capacity of their analysis methods it is difficult to get a true handle on the degree of full length L1s in the extant population.

PALE represents a high-throughput solution to overcome the problems associated with previous L1-detection techniques. Because of its incredibly high specificity and sensitivity, PALE allows for detection of essentially all full length L1 elements in a single individual with high confidence and allows for estimations of L1 activity levels (Figures 16-23). Application of PALE to a sample of unrelated individuals will allow for determination of the degree of variation seen between
individuals with respect to full length L1 elements as well as detection of the average number of full length L1 copies that exist in a single genome. In order to determine this information, we have applied PALE to seven unrelated individuals (Table 3) to answer two questions: 1) how many full length L1 elements are in any given individual? and 2) how much variation exists between the number of full length L1s between different individuals?. This information will allow us to draw conclusions about the degree to which these L1 elements are likely to be active and will permit estimations of the L1-induced mutagenic burden of any given individual or tumor sample.
### Table 3

<table>
<thead>
<tr>
<th>Individual #</th>
<th>Coriell ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GM01631</td>
</tr>
<tr>
<td>2</td>
<td>GM01632</td>
</tr>
<tr>
<td>3</td>
<td>GM05510</td>
</tr>
<tr>
<td>4</td>
<td>GM05568</td>
</tr>
<tr>
<td>5</td>
<td>GM16094</td>
</tr>
<tr>
<td>6</td>
<td>GM15983</td>
</tr>
<tr>
<td>7</td>
<td>GM08207</td>
</tr>
</tbody>
</table>
Table 3 List of cell lines used in this study. List of Coriell identification numbers for cell lines used in this study.
3.2. Full length L1 elements are more abundant than previously estimated and numbers vary considerably between individuals

To determine the extent of human full length L1 polymorphism between unrelated individuals, we applied PALE to seven individual genomic DNA samples from non-related individuals to assess the typical number of full-length L1 elements present in each of these individuals. For ease of analysis, PALE reads mapping to known full length L1 elements were discarded, and only those reads that did not represent reference elements were taken into account for analysis.

Using PALE, we detected 228 non-fixed full length L1 elements in the genomes of seven unrelated individuals. Of the non-reference full length L1 elements detected by PALE, 160 were elements present in dbRIP or described by previous studies. The remaining 68 non-reference full length L1 elements detected by PALE in these individuals were novel elements that had not previously been described (Table 4). We performed 5’ junction PCR-based sequence validation of 15 randomly selected novel non-reference L1 elements and achieved a 100% validation rate. Additionally, we also validated the precise
3’ junction of seven randomly selected novel non-reference L1 elements and confirmed these junctions with PCR (Figures 22 and 24, Table 5).

The average individual’s genome contained 96 non-fixed (polymorphic) full length L1 elements (Table 6). Because rare L1 alleles represent elements that are more likely to be active, we also wanted to get an idea of the number of rare L1 alleles present in each individual. Of the 68 novel L1 elements detected by this study, 60% represented rare alleles that were only detected in a single individual out of the seven tested in our study. In fact, the vast majority (90%) of novel L1 elements detected by our study were limited to two or fewer individuals, indicating that the majority of novel L1 elements detected are low-allele frequency elements that are most likely to be active (Figure 24). Interestingly however, as many as 10% of novel non-reference L1 elements are shared between at least two individuals in our sample, indicating that L1 elements with higher allele frequencies still remain to be discovered (Figure 24).
### Table 4

<table>
<thead>
<tr>
<th>Chr</th>
<th>Posn</th>
<th>Ind. 1</th>
<th>Ind. 2</th>
<th>Ind. 3</th>
<th>Ind. 4</th>
<th>Ind. 5</th>
<th>Ind. 6</th>
<th>Ind. 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56405300</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>149308700</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>120060000</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>111361000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>365704000</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>620982000</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>113261300</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>132769100</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>141557500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>201638200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>11</td>
<td>610830000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>12</td>
<td>805901000</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>109787900</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>123590900</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>146511000</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>169270200</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>205006000</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>111361000</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>141557500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>20</td>
<td>201638200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chr</th>
<th>Posn</th>
<th>Ind. 1</th>
<th>Ind. 2</th>
<th>Ind. 3</th>
<th>Ind. 4</th>
<th>Ind. 5</th>
<th>Ind. 6</th>
<th>Ind. 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>56405300</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>149308700</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>120060000</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>111361000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>365704000</td>
<td>X</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>6</td>
<td>620982000</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>7</td>
<td>113261300</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>132769100</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>141557500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td>201638200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>11</td>
<td>610830000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>12</td>
<td>805901000</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>109787900</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>123590900</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>146511000</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>169270200</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>205006000</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>111361000</td>
<td>-</td>
<td>-</td>
<td>X</td>
<td>X</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>141557500</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
<tr>
<td>20</td>
<td>201638200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>X</td>
</tr>
</tbody>
</table>
Table 4 Novel full length L1 elements identified in this study. Table showing the 68 novel full length L1 elements identified in this study and indicating which individuals the elements were identified in as well as their genomic location.
<table>
<thead>
<tr>
<th>flID</th>
<th>ID</th>
<th>Chr</th>
<th>5' jcn</th>
<th>3' jcn</th>
<th>TSD Length</th>
<th>TSD</th>
<th>EN Cut Site</th>
<th>pA length</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL043</td>
<td>3</td>
<td>80590165</td>
<td>80590176</td>
<td>10</td>
<td>TTCCATTTTC</td>
<td>TTTC/CA</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td>FL054</td>
<td>3</td>
<td>169269964</td>
<td>169269952</td>
<td>13</td>
<td>AAAAGGGTATTTT</td>
<td>TTTT/AA</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>FL107</td>
<td>6</td>
<td>142450103</td>
<td>142450110</td>
<td>8</td>
<td>AATAGTTT</td>
<td>TTTT/TTT</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>FL168</td>
<td>13</td>
<td>27339037</td>
<td>27339023</td>
<td>9</td>
<td>AAAATGGATTTT</td>
<td>AATTT/CA</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>FL178</td>
<td>14</td>
<td>71197795</td>
<td>71197777</td>
<td>19</td>
<td>AAGAATAATGACATTTG</td>
<td>TCTT/CA</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>FL206</td>
<td>18</td>
<td>49216219</td>
<td>49216232</td>
<td>14</td>
<td>CAGAGATTTGTTT</td>
<td>TTTT/CA</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>FL139</td>
<td>10</td>
<td>14310827</td>
<td>14310841</td>
<td>15</td>
<td>CTAGCTAATTTC</td>
<td>TTTT/CT</td>
<td>26</td>
<td></td>
</tr>
</tbody>
</table>
Table 5 Validation of PALE-detected full length L1 elements by 3' junction PCR. Table showing the seven full length L1 elements detected by PALE that were validated at their 3' junction in addition to the 5' junction. This allows determination of the size of the target site duplication (TSD), the sequence of the TSD, the L1 endonuclease (EN) cut site sequence, and the length of the polyA tail of the L1 element.
Figure 24
**Figure 24 Detection of novel full length L1 elements in seven individuals.** In total, 68 novel full length L1 elements were detected among the seven individuals tested. The vast majority (90%) were elements that were limited to one or two individuals with most (62%) being isolated only from a single individual. Very few novel elements were present in greater than two individuals and none were present in greater than five individuals. This indicates that most of the novel L1 elements described in this study are relatively low allele frequency elements and are most likely to be active.
To determine the variation between individuals in the extent of full length L1s, we looked at how many potentially active full length L1 elements exist in each individual. The range of non-reference full length L1 elements varied from a low of 73 per individual to a high of 134 per individual (Table 6). This surprising variation indicates that different individuals within a population are likely prone to different degrees of L1-associated damage.

To get a better understanding of the frequency distribution of non-reference full length L1 elements detected by PALE, we performed pair-wise comparisons of the PALE-detected non-reference L1 elements between each of the seven individuals (Figures 25-28). Our data reveal any two individuals shared between 29 and 94 non-reference full length L1 elements with the average number of non-reference L1s shared between any two individuals being 52 (Figure 25). Similar pairwise comparison revealed any one individual has between 13 and 91 non-reference full length L1 elements that are not present in another single individual with an average of 44 non-reference full length L1 elements not shared between any two individuals (Figure 26). Analysis of novel non-reference full length L1 elements (ie, those not previously reported) revealed individuals contained between nine and 26 novel non-reference full length L1 elements, with an average of 15 novel non-reference elements per individuals (Figure 27). Additionally, any two individuals share between five and 24 novel full length L1s (Figure 28). Taken together, these data indicate an individual genome
may contain upwards of 100 non-reference full length L1 elements with the potential to be active, and there is considerable variation between individuals in the overall numbers of potentially active L1 elements.

We next analyzed the chromosomal distribution of non-reference full length L1 elements detected in our study. PALE successfully detected non-reference L1 elements on all 22 autosomes and X with no apparent bias for any chromosome to harbor either a particularly high or particularly low number of non-reference full length L1 elements. Novel full length L1 elements were detected on all chromosomes except 8 and 19 (Figure 29). Additionally, our data show that any individual contains on average 10 unique non-reference full length L1 elements with more than half of those (on average, six) having not been previously described (Figure 30 and Table 7). These elements represent full length L1 elements that are specific to a single individual in this study and likely represent low-allele frequency elements most likely to be hot.

Additionally, we conducted a population study to determine the allele frequency of a subset of the full length L1 elements detected by PALE. To this end, 40 full length L1 elements were tested for presence or absence in 20 independent individuals from each of four geographically diverse populations (African American, Asian, German Caucasian, and South American) for a total of 80 independent individuals. This analysis revealed the vast majority (81%) of
PALE-detected full length L1 elements are present at an allele frequency less than 50%, with the majority existing at allele frequencies less than 25% in the population. In fact, 31% of polymorphic full length L1s tested existed at an allele frequency of less than 10%, indicating that the majority of the full length L1 loci detected by PALE are likely to be active (Figure 31). Taken together, these data imply that the extent of non-reference full length L1 polymorphism in the population is greater than previously appreciated.
Table 6

<table>
<thead>
<tr>
<th>Indiv.</th>
<th>Total polymorphics</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73</td>
</tr>
<tr>
<td>2</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td>134</td>
</tr>
<tr>
<td>4</td>
<td>120</td>
</tr>
<tr>
<td>5</td>
<td>96</td>
</tr>
<tr>
<td>6</td>
<td>91</td>
</tr>
<tr>
<td>7</td>
<td>80</td>
</tr>
<tr>
<td>Avg.</td>
<td>96</td>
</tr>
</tbody>
</table>
Table 6 Number of polymorphic full length L1 elements detected in each of seven individuals by PALE. PALE analysis of seven independent individuals revealed the typical individual contains 96 polymorphic full length L1 elements on average with a range in the population from a low of 73 full length L1 elements per individual to a high of 134 full length L1 elements per individual.
Figure 25

<table>
<thead>
<tr>
<th>Indiv.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>57</td>
<td>46</td>
<td>49</td>
<td>33</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>43</td>
<td>47</td>
<td>34</td>
<td>34</td>
<td>29</td>
<td>75</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>94</td>
<td>64</td>
<td>61</td>
<td>52</td>
<td>134</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>66</td>
<td>60</td>
<td>51</td>
<td>120</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>73</td>
<td>63</td>
<td>96</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>67</td>
<td>91</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>80</td>
</tr>
</tbody>
</table>

| 73 | 75 | 134 | 120 | 96 | 91 | 80 |
Figure 25 Pairwise comparison showing all polymorphic full length L1 elements shared between any two individuals. This pairwise comparison shows all of the polymorphic full length L1 elements detected in seven independent individuals and indicates how many are shared between any two individuals. Total numbers of full length L1 elements detected per individual are identified as column and row totals at the bottom of each column or right side of each row. Numbers in each of the interior boxes represent the total number of polymorphic full length L1 elements shared between the two individuals indicated by the intersection of the row and column specific to those individuals.
<table>
<thead>
<tr>
<th>Indiv.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>16</td>
<td>27</td>
<td>24</td>
<td>40</td>
<td>42</td>
<td>42</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>-</td>
<td>32</td>
<td>28</td>
<td>41</td>
<td>41</td>
<td>46</td>
</tr>
<tr>
<td>3</td>
<td>88</td>
<td>91</td>
<td>-</td>
<td>40</td>
<td>70</td>
<td>73</td>
<td>82</td>
</tr>
<tr>
<td>4</td>
<td>71</td>
<td>73</td>
<td>26</td>
<td>-</td>
<td>54</td>
<td>60</td>
<td>69</td>
</tr>
<tr>
<td>5</td>
<td>63</td>
<td>62</td>
<td>32</td>
<td>30</td>
<td>-</td>
<td>23</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>57</td>
<td>30</td>
<td>31</td>
<td>18</td>
<td>-</td>
<td>24</td>
</tr>
<tr>
<td>7</td>
<td>49</td>
<td>51</td>
<td>28</td>
<td>29</td>
<td>18</td>
<td>13</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>73</td>
<td>75</td>
<td>134</td>
<td>120</td>
<td>96</td>
<td>91</td>
<td>80</td>
</tr>
</tbody>
</table>
Figure 26 Pairwise comparison showing all polymorphic full length L1 elements unique between any two individuals. This pairwise comparison shows all of the polymorphic full length L1 elements detected in seven independent individuals and indicates how many are unique between any two individuals. Total numbers of full length L1 elements detected per individual are identified as column and row totals at the bottom of each column or right side of each row. Numbers in each of the interior boxes represent the total number of polymorphic full length L1 elements unique between the two individuals indicated by the intersection of the row and column specific to those individuals.
Figure 27

<table>
<thead>
<tr>
<th>Indiv.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

|     | 15 | 18 | 25 | 26 | 9  | 10 | 13 |

18

25

26

9

10

13
Figure 27 Pairwise comparison showing all novel full length L1 elements shared between any two individuals. This pairwise comparison shows all of the novel polymorphic full length L1 elements detected in seven independent individuals and indicates how many are shared between any two individuals. Total numbers of novel full length L1 elements detected per individual are identified as column and row totals at the bottom of each column or right side of each row. Numbers in each of the interior boxes represent the total number of novel polymorphic full length L1 elements shared between the two individuals indicated by the intersection of the row and column specific to those individuals.
Figure 28

<table>
<thead>
<tr>
<th>Indiv.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>7</td>
<td>12</td>
<td>12</td>
<td>15</td>
<td>14</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>-</td>
<td>15</td>
<td>15</td>
<td>18</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>22</td>
<td>-</td>
<td>17</td>
<td>24</td>
<td>23</td>
<td>21</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>13</td>
<td>8</td>
<td>-</td>
<td>14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>-</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>9</td>
<td>9</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>12</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>-</td>
</tr>
</tbody>
</table>

15 18 25 26 9 10 13
Figure 28 Pairwise comparison showing all novel full length L1 elements unique between any two individuals. This pairwise comparison shows all of the novel polymorphic full length L1 elements detected in seven independent individuals and indicates how many are unique between any two individuals. Total numbers of novel full length L1 elements detected per individual are identified as column and row totals at the bottom of each column or right side of each row. Numbers in each of the interior boxes represent the total number of novel polymorphic full length L1 elements unique between the two individuals indicated by the intersection of the row and column specific to those individuals.
Figure 29
Figure 29 Chromosomal distribution of polymorphic full length L1 elements across seven individuals. Polymorphic full length L1 elements were detected by PALE across all autosomes and X without any apparent bias for one genomic region over another. Novel full length L1 elements (grey) were also detected across nearly all autosomes and X without apparent bias.
Figure 30
Figure 30 Novel full length L1 elements are detected in all seven individuals tested. Histogram showing previously undescribed full length L1 elements detected in each of the seven independent individuals tested in this study. The typical individual contains, on average, 10 novel full length L1 elements, the majority of which are unique to that individual. Significant variation in novel full length L1 elements exists between individuals with individual seven containing three novel full length L1s and individual three containing 28 novel L1s.
Figure 31
**Figure 31 Histogram depicting population study of 40 polymorphic full length L1 elements detected by PALE.** Histogram showing polymorphic full length L1 elements tested in a population based study of 80 independent individuals from diverse genetic backgrounds. The vast majority (81%) of polymorphic elements detected by PALE exist at allele frequencies <50% with the majority existing at allele frequencies <25%. Polymorphic elements detected by PALE largely represent young, low allele frequency elements that are most likely to be active.
<table>
<thead>
<tr>
<th>Indiv.</th>
<th>Total Unique (New)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7 (6)</td>
</tr>
<tr>
<td>2</td>
<td>8 (8)</td>
</tr>
<tr>
<td>3</td>
<td>28 (14)</td>
</tr>
<tr>
<td>4</td>
<td>13 (6)</td>
</tr>
<tr>
<td>5</td>
<td>6 (3)</td>
</tr>
<tr>
<td>6</td>
<td>4 (2)</td>
</tr>
<tr>
<td>7</td>
<td>3 (2)</td>
</tr>
<tr>
<td>Avg.</td>
<td>10 (6)</td>
</tr>
</tbody>
</table>
Table 7 Novel full length L1s detected in each of the seven individuals. The typical individual contains, on average, ten unique full length L1 elements detected only in that one individual out of the seven tested. The numbers of unique full length L1s between the seven individuals tested range from a low of three per individual to a high of 28 unique full length L1 elements per individual tested. The majority of these unique L1s are novel full length L1 elements described for the first time by this study, indicating they are most likely to be active.
3.3. Discussion

We have used PALE to successfully identify virtually all of the L1 elements capable of mobilization in seven unrelated individuals. The results obtained from PALE analysis of full length L1 elements in this population revealed a few surprising observations. First, the prevalence of full length L1 elements in each individual was greater than estimated by previous estimates. Prior to this study, the belief was that there were approximately 50 polymorphic L1 elements per individual, with only a small proportion of those (about 14 elements per individual) representing particularly active hot L1s. Combined with the 40 known active full length L1 elements in the human genome, this resulted in the idea that there are approximately 90 active L1 elements per individual, with the majority of these elements being only minimally active and the bulk of retrotransposition resulting from the expression of a small number of particularly hot L1 loci.

In this study, we show that the typical individual has approximately 100 polymorphic L1 elements, about twice what has previously been estimated. Because these elements represent relatively young L1 insertion events, a large majority of these elements (>60%) are likely to be active, with most (>50%) likely to be particularly active hot L1s. This finding supports a model of L1 expression by which each individual harbors unique active L1 loci, resulting in the bulk of
retrotransposition in one individual originating from different source L1s than insertion events in another individual.

Additionally, our data show that there is significant variation in the number and extent of potentially active L1 elements between different individuals. Based on our findings, some individuals appear to contain nearly twice as many potentially active full length L1 elements as others. This implies that certain individuals in the population suffer a greater potential for L1-induced mutagenesis than others. A scenario could be envisioned in which an individual with a particularly large number of active L1 elements may be more prone to develop certain diseases, particularly cancers that appear to have a genetic basis but have no known gene basis.

Further, differing degrees of active L1s within the cells of a tumor may promote tumorigenesis. If a cell in a given tumor acquires the insertion of a full length L1 that has the ability to retrotranspose at a high rate, that hot L1 can then rapidly mobilize within that tumor, resulting in a more rapid progression of tumorigenesis as a result of L1 insertion events.
3.4. Materials and Methods

3.4.A Cell lines and oligonucleotides

Cell lines and oligonucleotides are as described in Chapter 2.

3.4.B 5’ PALE library generation

PALE library generation was as described in Chapter 2.

3.4.C 5’ Library quantitation and Illumina sequencing

Illumina sequencing libraries were quantified by qPCR and checked for quality by Agilent BioAnalyzer trace at Elim Biopharmaceuticals (Hayward, CA). Illumina sequencing was performed on an Illumina HiSeq 2000 by Elim Biopharmaceuticals (Hayward, CA).

3.4.D Reference Genome Masking

Genome masking occurred as described in Chapter 2.

3.4.E 5’ PALE Mapping Strategy

5’ PALE reads were mapped to GRCh37/hg19 as in Chapter 2.
3.4.F PCR validation of PALE reads

PCR validation of 5’ junctions was done as in Chapter 2. Additionally, 3’ junctions were validated in a similar manner except using an L1 primer specific to the 3’ end of the L1.3 consensus sequence instead of one to the promoter region.

3.4.G Allele frequency determination

Allele frequency for 40 randomly selected polymorphic L1 elements detected by PALE was determined by screening a panel of 80 individuals of diverse genetic background using a PCR-based assay to detect the presence or absence of the polymorphic element in each individual.
CHAPTER 4:

PALE discovery of \textit{de novo} L1 insertion events

4.1. Introduction

Insertional mutagenesis by L1 and L1-driven retroelements such as Alu and SVA has been responsible for nearly 100 instances of \textit{de novo} human disease. L1-driven retrotransposition insertions are estimated to be responsible for approximately 0.2\% of all new human germline diseases and have resulted in diseases ranging from hemophilia and muscular dystrophy to hypercholesterolemia and cancer (Belancio, Hedges et al. 2008, Belancio, Deininger et al. 2009, Belancio, Roy-Engel et al. 2010). The vast majority of studies looking at retroelement insertion events have focused on retrotransposon insertions that occur in the germline resulting in \textit{de novo} instances of new disease in the next generation of progeny. Recently, however, advances in high-throughput sequencing techniques have allowed for analyses of somatic retrotransposition events.

The bulk of studies investigating somatic retrotransposition analyze various tumor samples and compare them to matched normal tissue to identify
retroelements present in the tumor that are not present in the normal tissue. This type of analysis allows for discrimination of retrotransposition insertions that occurred in the tumor itself. Such studies have been done in a number of tumor types, with some tumor types supporting high levels of retrotransposition (colon, lung, breast) and others supporting little to no retrotransposition (leukemias, brain tumors) (Ewing and Kazazian 2010, Iskow, McCabe et al. 2010, Witherspoon, Xing et al. 2010, Baillie, Barnett et al. 2011, Ewing and Kazazian 2011, Lee, Iskow et al. 2012, Solyom, Ewing et al. 2012, Shukla, Upton et al. 2013, Helman, Lawrence et al. 2014).

Studies of somatic retrotransposition in tumors have revealed a number of interesting findings. While L1 retrotransposition appears to be well-supported in a number of tumor types, levels of Alu retrotransposition appear to be significantly lower than expected. Additionally, somatic L1 retrotransposition events seem to be more highly truncated than what is observed from germline events. Perhaps most interestingly, L1 somatic retrotransposition has resulted in insertional mutagenesis directly into known tumor suppressor genes, implicating L1 retrotransposition in tumorigenesis and progression. Additionally, even within tumor types, different tumors appear to support L1 retrotransposition to varying extents (Ewing and Kazazian 2010, Iskow, McCabe et al. 2010, Witherspoon, Xing et al. 2010, Baillie, Barnett et al. 2011, Ewing and Kazazian 2011, Lee, Iskow et al. 2012, Solyom, Ewing et al. 2012, Shukla, Upton et al. 2013, Helman, Lawrence et al. 2014, Pitkanen, Cajuso et al. 2014). For example, one specific colorectal tumor was reported to harbor over 100 unique somatic L1
retrotransposition events; nearly an order of magnitude more than the colorectal tumor with the second highest levels of retrotransposition (Pitkanen, Cajuso et al. 2014).

Rates of retrotransposition in tumors appear to exceed those observed in germline. Conservative estimates of retrotransposition rates predict a de novo L1 insertion event in one out of every 200 live births. Alu retrotransposition is estimated to occur in one out of every 20 live births (Brouha, Schustak et al. 2003). What is not known, however, is where during development retrotransposition occurs. Several options exist to explain germline retrotransposition events: retroelement mobilization can occur in gametes or retrotransposition can occur very early in embryogenesis. In either scenario, the result is the presence of retroelement insertion events present in virtually all cells within an individual.

Because of their detrimental effect on genome architecture, retroelements are subject to a number of cellular processes that control and limit the rates of their mobilization. A long period of co-evolution has allowed a diverse array of cellular processes to develop to limit insertional mutagenesis by retroelements and includes processes such as APOBECs, small interfering RNAs, epigenetic regulation, and various DNA repair processes.

One of the strongest inhibitors of L1 retrotransposition is the nucleotide excision repair (NER) pathway. The NER pathway is responsible for recognizing structural DNA lesions resulting primarily from UV-induced genomic damage.
Patients with defects in various NER proteins suffer from a number of severe disorders classified as either xeroderma pigmentosum or Cockayne syndrome (Reardon and Sancar 2005, Hoeijmakers 2007, Sugasawa, Akagi et al. 2009). Patients with NER defects suffer from a range of disease states including extreme sun sensitivity, mental disabilities, and predisposition to various cancers (Sugasawa 2008, Cleaver, Lam et al. 2009, Nouspikel 2009, Nouspikel 2009).

The NER repair pathway consists of two distinct branches: the global genome repair (GGR) subpathway detects DNA lesions throughout the entire genome while the transcription coupled repair (TCR) subpathway recognizes lesions in actively transcribed areas of the genome (Volker, Mone et al. 2001, Fousteri and Mullenders 2008, Sugasawa 2008). The initial DNA-damage recognition steps of the two NER subpathways make use of different initial sensor proteins: GGR uses XPC and Rad53 to identify bulky DNA lesions, while TCR recruits CSA and CSB to the site of stalled RNA polymerase III forks. Following lesion recognition, the two subpathways converge to a single downstream repair pathway involving the protein complexes XPB, XPD, XPA, XPF/ERCC1, and XPG {reviewed in (Kamileri, Karakasilioti et al. 2012)} (Figure 32).

L1 insertion intermediates result in the formation of bulky DNA distortions that mimic the natural lesion recognized by the NER pathway (Figure 32). We have shown that the NER pathway is capable of recognizing L1 insertion intermediates and is able to significantly limit L1 mobilization events in a tissue culture system (Gasior, Roy-Engel et al. 2008). The absence of a functional NER system in various NER mutants results in high levels of L1 retrotransposition that
are decreased with complementation of the NER deficiency. This phenomenon was not observed for Alu retrotransposition (Servant et al. 2014, manuscript in preparation). As a result, we expect high levels of L1 retrotransposition to occur in individuals defected for NER repair.

In order to identify de novo L1 retrotransposition events in NER-deficient patient cells, we sought to modify PALE (Chapter 2) to detect newly integrated L1 elements. In this chapter, we describe the development of a modified version of PALE called 3’-PALE that can specifically detect de novo L1 retrotransposition events in individual genomes, and the application of 3’-PALE to two unique NER-deficient patient samples.
Figure 32

Transcription-coupled repair (TCR)

Global genome repair (GGR)

Lesion recognition

Helix opening

Dual incision

DNA repair synthesis

RNAPII

CSA

TFIIH (XPB-XPD-TTDA)

XPC

XPG

RPA

ERCC1-XPF
Figure 32 The nucleotide excision repair (NER) pathway. Bulky DNA lesions are repaired by the NER pathway. Two subpathways exist: transcription coupled repair (TCR) repairs DNA lesions in actively transcribed regions of the genome, which global genome repair (GGR) repairs lesions in the remainder of the genome. Each subpathway makes use of unique proteins before converging to make use of a common set of proteins for repair. The DNA structures generated during repair by NER resemble DNA structures generated during L1 insertion by target primed reverse transcription. The NER pathway has been shown to inhibit insertion of L1.
4.2. 3’-PALE can be used for *de novo* L1 retrotransposition detection

4.2.A Development of 3’-PALE

In order to be able to detect *de novo* L1 insertion events in an NER-deficient background, we first had to modify PALE to be able to detect all new L1 insertion events. Because most of the time TPRT results in the creation of severely truncated L1 insertions, we had to modify our L1 element recovery procedure to detect the 3’ most end of L1 elements (Figure 33). This modification to the protocol is not without significant complications. Unlike full length L1 elements which only exist at about 5,000 genomic copies, L1 3’ ends litter the genome with over 500,000 interspersed copies. This 100-fold increase in genomic target copy number can present significant challenges with background detection levels.

In order to limit background associated with the plethora of genomic L1 fragments, we took advantage of slight changes to the L1 consensus sequence between the different L1 subfamilies. Because nearly all human retrotransposition comes from only the youngest L1 subfamilies, we modified our PALE primer extension to take advantage of nucleic acid changes specific to only the youngest, most active human-specific L1 elements (L1Hs) (Figure 34). By
incorporating these very specific sequence changes into our primers, we were able to effectively reduce the genomic background to the approximately 5000 L1Hs elements in the human genome (a level of background nearly equivalent to that in our original PALE experiments) (Lander, Linton et al. 2001).

Aside from changes to primer sequences and slight modifications to PCR cycling conditions, the details of 3’-PALE remain largely identical to those used in PALE library preparation. Following initial primer extension off of L1Hs elements in sheared genomic DNA, duplex t-linkers are ligated to fragments only if they have undergone L1Hs primer extension. Downstream PCR and gel purification as in 5’-PALE, allow for selection of highly specific, targeted libraries representing human specific L1 elements and any potential de novo L1 insertion events.
Figure 33
Figure 33 Modification of PALE to detect human-specific L1 elements at the 3’ end (3’-PALE). Slight modifications were made to PALE to adapt it for use on L1 3’ ends. The primer locations within the L1 element were moved to the 3’ end of the L1, which allows 3’-PALE to detect novel L1 insertion events, which are often 5’ truncated upon insertion. The remaining steps of 3’-PALE are the same as for 5’-PALE.
Figure 34

- ACA x CTC
- Old L1 (~500,000 copies) → No priming
- L1Hs (~5000 copies) → Priming and extension
**Figure 34** 3’-PALE selectively detects only the youngest, most active L1 subfamilies. 3’-PALE makes use of primers specific to the youngest, most active subfamilies of L1 (L1Hs). The primers take advantage of a three nucleotide ACA insertion in L1Hs (left) allowing for priming on the approximately 5000 L1Hs elements in the human genome but not on the remaining ~500,000 copies of older, inactive L1s (right). This enables 3’-PALE to avoid the incredibly high levels of background that would come with amplification of half a million inactive genomic L1 elements.
4.2.B 3’-PALE Sensitivity

As with our initial iteration of PALE, we wanted to determine the sensitivity of 3’-PALE by assessing its ability to detect known L1Hs elements on three randomly selected chromosomes. As before, we determined the ability of 3’-PALE to detect known L1Hs elements on three randomly selected chromosomes via manual inspection of 3’-PALE mapped sequence reads. We find that 3’-PALE accurately and robustly detects L1Hs elements from individual human genomes (Figure 35). Like with 5’-PALE, we determined the sensitivity of 3’-PALE and found that it is equally robust. We were able to accurately detect 93% of all known L1Hs elements on three randomly selected chromosomes (Figure 36). Like with 5’-PALE, those 7% of elements that were not detected reside nearly entirely within highly repetitive genomic regions, rendering them refractory to accurate, unambiguous genome mapping. Additionally, the vast majority of L1Hs elements (90%) were detected by at least two independent 3’-PALE events, indicating 3’-PALE is a robust way to identify de novo L1 insertions (Figure 36).
**Figure 35 3’-PALE detects L1Hs elements from human genomes.** Alignment and analysis of 3’-PALE reads was done by CLC Genomics Workbench. Red lines represent 3’-PALE reads in the reverse orientation obtained after 3’-PALE library sequencing. Reads read into a known L1Hs element (arrow), confirming success of the 3’-PALE strategy.
Figure 36
**Figure 36 3’-PALE detects L1Hs elements with high sensitivity.** Histogram depicting the number of independent linker ligations for all of the known, L1Hs elements on chromosomes 4, 12 and 17. The median number of ligation events per fixed full length L1 is seven, with the vast majority of L1Hs elements detected by at least two independent 3’-PALE reads. 3’-PALE detects 93% of known L1Hs elements, and those L1Hs elements not detected are missed because they lie in highly repetitive genomic regions.
4.2.C 3’-PALE Specificity

Also similarly to 5’-PALE, we sought to determine the false positive rate encountered when using 3’-PALE. To get an idea of the specificity of 3’-PALE we assayed four independent individuals using 3’-PALE to identify novel L1Hs elements within the genomes of these individuals. We randomly selected 10 novel non-reference L1Hs elements that were detected by 3’-PALE and subjected then to PCR- and Sanger sequencing-based validation using L1 junction PCR to amplify across the L1Hs 3’ junction (Figure 37). This method allows us to determine the precise 3’ junction of the novel, non-reference L1Hs. Performing this analysis, we successfully validated 7/10 non-reference L1Hs elements that were detected by 3’-PALE and confirmed these L1Hs elements by Sanger sequencing the PCR product (Figure 37). The remaining 3/10 elements were refractory to PCR amplification. However, at least one of the remaining three L1Hs elements could be confirmed by an empty/filled site PCR comparison, indicating that the inability to successfully amplify L1 3’ junctions by PCR is not necessarily indicative of a false positive event.

Taken together, these data indicate 3’-PALE represents a robust tool that is capable of accurately identifying nearly all L1Hs elements in an individual genome, and will be useful for determination of de novo L1Hs insertion events in individuals or tumor samples.
Figure 37 PCR validation of 3’-PALE events. (A) PCR primers were designed within the 3’ UTR of L1Hs and within the unique genomic flanking sequence downstream of the element. (B) Representative gel showing PCR-based validation of a polymorphic L1Hs element detected by 3’-PALE. NTC, no template control. M, 1kb DNA ladder. Numbers indicate the individual in which presence or absence of the specific L1Hs element was tested.
4.2.D Detection of L1Hs events in two NER-deficient family groups

In order to determine whether NER deficiency is responsible for higher levels of L1 retrotransposition in vivo, we applied 3’-PALE to two independent family groups with offspring suffering from XPA deficiency (Figure 38). As a control, one family group contained a child that did not suffer from XPA deficiency, which can be used as a control for the overall rate of expected de novo L1 retrotransposition. By comparing non-reference L1Hs elements in the children to those in both parents, we can determine the rate of de novo L1 retrotransposition in the XPA-deficient background.

Genomic DNA from two XPA-deficient family groups was subjected to 3’-PALE to identify all L1Hs elements present in these individuals. After 3’-PALE library generation and Illumina sequencing, data sets were analyzed for the presence of novel non-reference L1Hs elements in each individual that could represent a de novo L1 insertion event. In order to classify a non-reference L1 element as a de novo L1 insertion, it must be present in the child but not in either of the parental DNA samples.

3’-PALE analysis of the two XPA-deficient individuals identified over 100 previously unidentified polymorphic (non-reference L1 elements) in each of the
two XPA-deficient individuals. However, comparison of these L1Hs elements with the L1Hs elements identified in both parents confirmed that 100% of these L1Hs elements were inherited germline polymorphics and not proper de novo L1Hs insertion events (Table 8). Despite high sequencing depth, we failed to identify any de novo L1 insertion events in either of our two XPA-deficient individuals tested.
Figure 38

A

GM01632

GM01631

GM01630

B

GM05510

GM05568

GM05566

GM05509
Figure 38 XPA-deficient family groups used in this study. Cell lines from two family groups each with a child suffering from XPA deficiency (GM01630 and GM05509) were obtained from Coriell Cell Repositories. Genomic DNA was isolated from each member of each family group and 3'-PALE was performed on each of the seven individuals to determine if de novo L1 insertion events could be detected in an NER deficient background. Individual GM05566 (B) from the second family is an unaffected sibling from family two that can be used as a control to determine if de novo retrotransposition occurs more frequently in NER-deficient individuals than in normal individuals.
Table 8

<table>
<thead>
<tr>
<th></th>
<th># novel L1Hs detected</th>
<th># de novo L1Hs</th>
</tr>
</thead>
<tbody>
<tr>
<td>XPA-deficient 1</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td>XPA-deficient 2</td>
<td>106</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 8 L1Hs elements detected in two XPA-deficient individuals by 3’-PALE. 3’-PALE was used to detect non-reference L1Hs elements in two XPA-deficient individuals to attempt identification of de novo L1 integration events. 3’-PALE detected an average of 102 novel L1Hs elements per individual, however comparison of 3’-PALE data from both sets of parents confirmed these events were inherited polymorphisms and not de novo L1 retrotransposition events.
4.3 Discussion

*De novo* L1 insertion events are responsible for approximately 0.2% of new human genetic diseases. Additionally, somatic L1 retrotransposition in human tumors has shown fairly high levels of L1 retrotransposition in some tumor types compared to others. Further, even within two tumors of the same type, there is variation in an individual tumor's ability to support L1 retrotransposition. Some evidence exists to support the idea that such inter-tumor differences are the result of varying expression levels of cellular pathways that are known to repress L1 retrotransposition. For example, a single colorectal tumor displayed nearly an order of magnitude greater levels of L1 retrotransposition than the next highest colorectal tumor sample. Upon gene expression analysis of this high retrotransposing sample, it was discovered that the mismatch repair gene MLH1 was downregulated in this sample, resulting in low levels of mismatch repair (Lee, Iskow et al. 2012). The mismatch repair pathway is one of the DNA repair pathways that has been shown to limit levels of L1 retrotransposition (Faber *et al.*, unpublished data). This finding implies that tumor subtypes may exist in which different levels of L1 retrotransposition may be the result of differential expression of DNA repair machinery in the tumors.
Based on our preliminary *in cellulo* findings we predicted rates of L1 retrotransposition in an XPA-deficient background would be significantly increased. Using this assumption, we attempted to identify *de novo* L1 retrotransposition insertion events in two XPA-deficient patient samples. Despite the ability to detect nearly all L1Hs elements in an individual using 3’-PALE, we were unsuccessful at detecting *de novo* L1 retrotransposition events in either of the two individuals tested.

Initially, these results seemed unexpected given what seemed like a large increase in rates of L1 retrotransposition in an XPA-deficient background. However closer analysis of the results yield a potential explanation for the lack of discernible L1 retrotransposition in our XPA-deficient patient samples. Based on tissue culture retrotransposition experiments, we expect an approximately 30-fold increase in L1 retrotransposition in the absence of functional XPA protein. When the established estimated rate of *de novo* L1 retrotransposition is taken into account (one retrotransposition event in every 200 live births), we would only expect a *de novo* L1 insertion event to occur in approximately one in every six XPA-deficient individuals (ie, 1/200 x 30).

While we were unable to demonstrate *de novo* somatic L1 retrotransposition events in an NER-deficient background, we have developed a powerful tool that can be used in the future to test additional samples from
additional XPA-deficient family groups to see if L1 retrotransposition is being upregulated in the absence of a function NER system. Additionally, 3'-PALE can be used in various tumors, particularly those shown to be deficient in DNA repair machinery, to assay the degree to which L1 mobilizes within different tumor samples.
4.4 Materials and Methods

4.4.A Cell lines and oligonucleotides

Cell lines derived from XPA-deficient family groups were as follows: GM01630, GM01631, GM01632, GM05509, GM05510, GM05568, and GM05566 and were obtained from the Coriell Institute (Camden, NJ). All cell lines were maintained in Eagle’s Minimal Essential Media (EMEM) supplemented with non-essential amino acids, sodium pyruvate, and 10-15% fetal bovine serum. DNA oligonucleotides and duplex linkers were obtained from Integrated DNA Technologies (Coralville, IA).

4.4.B 3’-PALE library generation

3’-PALE library generation was performed in a manner similar to 5’-PALE, but with the following modifications. L1Hs-specific primer 1 was used for Primer extension and PCR 1 and was as follows (5’- ATAGCATTGGGAGATATACCTA-3’). First round PCR was for 22 cycles. L1Hs-specific primer 2 was used for PCR 2 and was as follows (5’-CTCGGCATTCTGCTGAACCGCTTTCCGATCTNNNNNATACCTAATGCTAGATGACACA-3’). The remaining portions of the library preparation were identical to that for 5’-PALE (Chapter 2).
4.4.C 5’ Library quantitation and Illumina sequencing

Illumina sequencing libraries were quantified by qPCR and checked for quality by Agilent BioAnalyzer trace at Elim Biopharmaceuticals (Hayward, CA). Illumina sequencing was performed on an Illumina HiSeq 2000 by Elim Biopharmaceuticals (Hayward, CA).

4.4.D Reference Genome Masking

Reference genome masking was performed similarly to 5’-PALE but with some modifications. Reference repetitive element annotation files were downloaded from UCSC genome browser as GTF files and used to annotate the GRCh37 (hg19) reference genome in CLC Genomics Workbench (CLC Bio, Cambridge MA). Two custom GFF annotation files were also generated: 1) representing known polymorphic L1Hs elements and 2) representing all known L1Hs elements plus 600bp of downstream genomic flanking sequence.

To identify and annotate L1Hs 3’ regions of the genome, the last 300 base pairs of the L1Hs consensus sequence (Lee, Cordaux et al. 2007) was aligned to the human reference genome (CRCh37/hg19) via NCBI BLAST (blastn algorithm). Initial search parameters were relaxed to increase sensitivity for older, more mutated elements (word size=7, gap cost=3, gap extension cost=3, match score=2, mismatch penalty=-3, minimum E score of 10). Diagnostic ACA trinucleotides indicative of L1Hs elements were required during alignment. Blast output was generated in tabular format. Because many of the hits obtained from
this search consisted of isolated smaller regions (20-30bp) that were independent of any identifiable L1 element, we further filtered for those hits where alignments were made across >=250bp of the 300bp query. This greatly improved our specificity while maintaining sensitivity to older, more mutated elements. Tabular format blast results were converted to GFF via Perl scripting, and an additional 600bp of flanking sequence was added upstream and downstream to each matching position.

4.4.E 3’-PALE Mapping Strategy

3’ PALE reads were mapped to GRCh37/hg19 as for 5’-PALE in Chapter 2, but with modifications. FASTQ sequencing files generated by the Illumina HiSeq 2000 were analyzed for quality using FastQC. PCR duplicate reads were removed using custom in-house Perl scripting. Sequence read mappings were performed using CLC Genomics Workbench. Single end genomic flank reads were mapped using a two-pronged mapping strategy. First, reads that either mapped to within 600bp of known L1Hs elements or those reads that did not map unambiguously were discarded, as they represented either known elements or unmappable reads. The remaining reads were then mapped uniquely to an annotated GRCh73/hg19 reference genome to identify novel polymorphic or potentially de novo L1Hs elements. Finally, novel L1Hs loci detected by 3’-PALE in an XPA-patient sample were intersected with those recovered from both parents to determine the presence of any de novo L1Hs elements that were present in the XPA-deficient patient samples.
CHAPTER 5:

Discussion

5.1 Genomic impact of L1 retrotransposition

The retrotransposon L1 is a major endogenous source of human genetic variation. L1 is ubiquitously present in human genomes and its 500,000 genomic copies constitute 17% of the total mass of the human genome (Lander, Linton et al. 2001). Because it mobilizes by a copy and paste retrotransposition mechanism, L1 is expanding the genome through its insertion (Luan, Korman et al. 1993, Luan and Eickbush 1995, Cost, Feng et al. 2002). Successful L1 mobilization requires expression of a full length L1 element, which results in the expression of two L1-encoded proteins that are required for L1 mobilization (Moran, Holmes et al. 1996). In addition to self-mobilization, L1 is responsible for driving the retrotransposition of other cellular RNAs, particularly of the two non-autonomous retrotransposons Alu and SVA (Esnault, Maestre et al. 2000, Dewannieux, Esnault et al. 2003, Ostertag, Goodier et al. 2003, Garcia-Perez, Marchetto et al. 2007). Despite reported preference for L1 to mobilize itself over other RNAs, trans mobilization of SVA and Alu has been particularly
successful with over one million copies of Alu elements littering the genome (Lander, Linton et al. 2001, Wei, Gilbert et al. 2001).

Through its activity, L1 has resulted in over 80 instances of identified human diseases as the result of insertion of itself or another non-autonomous retrotransposons, usually into exonic regions of genes. Occasionally, however, L1-driven insertions are also associated with deletions, aberrant splicing, and reduced RNA levels (Babushok and Kazazian 2007, Belancio, Hedges et al. 2008, Belancio, Deininger et al. 2009, Belancio, Roy-Engel et al. 2010). In addition to insertional mutagenesis, L1 expression is also associated with other types of genome damage. The expression of L1 ORF2 protein in cells decreases cellular proliferation and results in cellular toxicity (Wallace, Belancio et al. 2008). Additionally, the endonuclease domain of L1 ORF2 protein results in the generation of DNA double strand breaks, a particularly damaging type of DNA lesion (Gasior, Wakeman et al. 2006).
5.2 Regulation of L1 expression

While the majority of L1 expression is deleterious to cells, retrotransposition has resulted in the creation of novel pseudogenes, which provides the genome with a back-up copy of a gene, allowing for the potential for genetic drift. However, despite this, the vast majority of L1 expression results in outcomes that are deleterious for the cell. As a result, co-evolution between organism and retroelement has allowed for a plethora of cellular mechanisms that limit the mobilization of transposable elements.

One of the most well-studied mechanisms of L1 cellular regulation is hypermethylation of the L1 promoter region (Thayer, Singer et al. 1993). The L1 5’ UTR is a CG-rich region, and methylation of four CpGs in the L1 5’ UTR is capable of repressing L1 transcript levels (Swergold 1990, Hata and Sakaki 1997). It remains possible that mammalian methylation co-evolved with retrotransposons as a way to regulate their detrimental effects (Bird 1997, Yoder, Walsh et al. 1997, Walsh, Chaillet et al. 1998, Bestor 2003).

An additional means of cellular control of retroelements occurs at the post-transcriptional level. Like most cellular mRNAs, the L1 transcript terminates in a stretch of adenine residues typically called a polyA tail. While L1 contains a mammalian polyadenylation signal at its 3’ end, studies have been done that
have shown that L1 also contains numerous internal polyadenylation sites. These internal polyadenylation signals were predicted to be stronger than the one at the 3’ end of the L1 RNA. Northern blot analysis of L1 transcripts confirmed the frequent usage of several internal L1 polyadenylation signals, generating truncated L1 RNAs not capable of retrotransposition (Perepelitsa-Belancio and Deininger 2003). It is interesting to note, however that such truncated RNAs may still retain the ability to generate ORF2 protein (or fragments of ORF2 protein). As a result, trans-mobilization of Alu and SVA would still be expected to occur as would deleterious functions of the L1 endonuclease (Kines et al., NAR 2014, accepted).

An additional regulation of L1 expression also occurs at the RNA level. Analysis of the L1 coding sequence revealed the presence of multiple predicted alternative splice sites, some of which were shown to be functional in mammalian cells (Belancio, Roy-Engel et al. 2008). One particular splice variant results in the generation of functional L1 ORF2 protein, which as above, would allow for the trans-mobilization of Alu and SVA as well as the creation of DNA double strand breaks.

One additional means of cellular control over retrotransposition is through the action of various DNA repair processes that likely act on L1 insertion intermediates to terminate integration of new retrocopies. Numerous DNA repair pathways have been described to inhibit L1 retrotransposition including the mismatch repair (MMR) and nucleotide excision repair (NER) pathways (Gasior, Roy-Engel et al. 2008). It appears that most of these pathways act by
recognizing DNA structures generated during the L1 insertion process and working to limit insertion of new L1 copies.

Together, along with various other cellular mechanisms, these cellular factors work to limit the damage that results from retroelement mobilization. Despite the extensive cellular processes working against them, retrotransposons have succeeded in becoming successful intracellular pathogens responsible for numerous instances of human disease. The source of their success, likely stems from highly active genomic copies of L1 elements (hot L1s) that are responsible for the bulk of human retrotransposition events.
5.3 Active L1s in human genomes

Retrotransposons represent one of the oldest evolutionary forces shaping human genomes. Successful colonization of mammalian genomes by retroelements has resulted in their expansion to over one third of total human genetic content (Lander, Linton et al. 2001). Because of the mutational burden associated with retroelement expression, cells have developed a number of cellular mechanisms to mitigate the effects of retrotransposition.

All modern day retrotransposition is the result of expression of the long interspersed element 1 (L1). L1 uses its protein machinery to mobilize itself, as well as to mobilize all the other currently active human retroelements. Although there are over half a million genomic copies of L1, the vast majority (~99%) are truncated due to early termination of ORF2 retrotransposition upon insertion (Lander, Linton et al. 2001). However, there are a number of full length L1 genomic copies that are capable of retrotransposition, and the emerging consensus of the field is that the bulk of L1 retrotransposition originates from the expression and mobilization of a limited number of particularly active (hot) L1 elements that represent evolutionary recent insertions (Brouha, Schustak et al. 2003, Beck, Collier et al. 2010).
Our data are entirely consistent with this model and add a new layer of complexity. We have shown that the typical human genome contains approximately 100 polymorphic full length L1s, of which we anticipate 63 are active and 54 are highly active hot L1s. These estimations greatly increase the expected number of hot L1s predicted to be present in a single genome by approximately four-fold over previous estimations (Brouha, Schustak et al. 2003, Beck, Collier et al. 2010). Additionally, we have shown that there is a relatively large degree of variation between individuals in terms of the numbers of polymorphic full length L1 elements present in their genomes, with some individuals containing nearly twice the potentially active L1 elements as other individuals.

Taken together, these two findings allow for a model of L1 expression among individuals in a population (Figure 39). Given the inter-individual differences we observed, it stands that some individuals will have a lower than median burden of L1 retrotransposition-induced genomic damage because they contain fewer active L1 loci in their genomes. These individuals represent members of the population with a low chance of experiencing the effects of L1-associated damage, and we can hypothesize that these individuals will be more likely to reach old age without any incidence of cancer formation. On the other edge of the sword, however, are individuals with a significantly higher than median number of potentially active L1 loci. These individuals will suffer from a larger degree of L1-associated genomic damage, and as a result, we hypothesize that they will suffer from a greater incidence of tumorigenesis.
This model is consistent with several published and unpublished experimental observations. Next-generation sequencing based studies by two independent groups of L1 retrotransposition events in colorectal tumors identified specific hot L1 loci in a number of tumor samples that were responsible for the majority of all retrotransposition observed in those tumors (Helman, Lawrence et al. 2014, Pitkanen, Cajuso et al. 2014). An additional study identified the active L1 source elements for a number of somatic L1 retrotransposition events in a variety of tumors including lung, colorectal, and breast, and found that in general, very few L1 loci account for the majority of all somatic L1 retrotransposition, with a third of all somatic L1 transductions resulting from the expression of only two hot L1 loci (Tubio et al., Science 2014, in revision). Additionally, the same study concluded that only a small percentage of active L1 elements identified overlapped with elements previously described in the literature, supporting the hypothesis that the bulk of L1 retrotransposition that occurs in genomes is likely the result of undiscovered intact L1 elements that are polymorphic in the population.
Figure 39

Large number of hot L1s

More extensive L1-associated damage

Cancer? Age-associated diseases?

Familial cancers?

Modest number of hot L1s

Minimal L1-associated damage

Low number of hot L1s

Very limited L1-associated damage

Increased life span?

Decreased incidence of cancer?

L1-associated risk
Figure 39 Model for differing degrees of L1-associated damage in different individuals in the population. Because of natural variation in the number of active L1 loci across individuals in a population, the degree of L1-associated genomic damage will vary between individuals. Individuals with modest numbers of hot L1s (center) will suffer from fairly minimal amounts of L1-generated genomic damage. Individuals fortunate enough to contain lower than average numbers of hot L1s (right) are expected to have decreased rates of L1-associated damage, possibly resulting in increased life span and decreased incidence for cancer. Other individuals will have large numbers of hot L1s which would be predicted to result in more extensive L1-induced damage leading to higher incidences of cancers and other age-associated diseases. Families with individuals containing high levels of hot L1s could explain incidences of familial cancers with no known genic association. This model predicts a spectrum of L1-associated risk ranging from low in individuals with few hot L1 loci to high in individuals with many hot L1s.
These findings can be extended to an additional model for tumorigenesis (Figure 40). In this model, somatic retrotransposition originating at a hot L1 locus can result in the somatic insertion of a full length L1 copy some percentage of the time. Because it is a newly inserted element, this newly inserted full length L1 will be identical (or nearly identical) in sequence to the hot L1 from which it originated. This can result in the establishment of a new active L1 locus, itself capable of further retrotransposition. This model has experimental support from a recent study that identified 29 instances of somatic retrotransposition originating from 17 L1 loci that were themselves somatically acquired during tumorigenesis. Interestingly, in some tumors, the newly retrotransposed L1 loci were responsible for the majority of the de novo somatic L1 insertions in the tumor, and in one instance, a chain of three consecutive somatic retrotranspositions was observed (Tubio et al, Science 2014, in revision).
Figure 40

New somatic hot L1

\[\text{Hot L1 retrotransposition into important growth-related genes}\]

\[\text{Tumorigenesis}\]
Figure 40 Model for hot L1s in tumorigenesis. Somatic retrotransposition of a hot L1 locus in the early stages of a tumor’s progression can result in the integration of a new full length L1 that is hot for retrotransposition. This newly generated hot L1 could go on to retrotranspose into regions of the genome coding for important growth-related genes resulting in tumorigenesis.
A final piece of evidence for a model of L1 retrotransposition like the one described above comes from analysis of high-throughput RNASeq data to investigate L1 RNA expression patterns. Analysis of repetitive element expression is not a trivial undertaking. Ten percent of all cellular RNAs contain a retrotransposon either in their introns or UTRs, so there is an extremely high level of background from non-repeat cellular RNAs that is encountered when analyzing RNA expression data obtained using the short reads generated by high-throughput sequencing. Some degree of L1 expression can be gleaned from analyzing RNASeq data using extremely stringent alignment conditions followed by tedious manual inspection of all reads. This mapping strategy has resulted in L1 expression data that support a model of L1 expression as described above. Different cell lines and tissues express highly from a very small number of L1 loci, and these loci differ from cell line to cell line and tissue to tissue (P. Deininger, unpublished data).

Our data which show large numbers of polymorphic full length L1s that exist at low allele frequency and are most likely to be active strengthen these observations and allow for a model in which a small number of highly active L1 loci are expressed in any one individual or tumor, with the bulk of L1 retrotransposition in that individual or tumor being the result of mobilization of one of the few active L1s present. The L1s most likely to be active are those that are most recently inserted, and as such, are polymorphic within the population (or newly inserted within a tumor). As such, we have described a large number of potentially hot L1 loci that would benefit from further study.
List of References


vast majority of large VHL germline deletions: Molecular characterization and genotype-phenotype correlations in VHL patients." Hum Mutat 30(5): 776-786.


novo retrotransposition of unbiased sequences in a human breast cancer

Sci U S A 103(47): 17632-17637.

23(18): 3208-3216.

Witherspoon, D. J., W. S. Watkins, Y. Zhang, J. Xing, W. L. Tolpinrud, D. J. Hedges,
M. A. Batzer and L. B. Jorde (2009). "Alu repeats increase local
recombination rates." BMC Genomics 10: 530.

Witherspoon, D. J., J. Xing, Y. Zhang, W. S. Watkins, M. A. Batzer and L. B. Jorde
(2010). "Mobile element scanning (ME-Scan) by targeted high-throughput

Witherspoon, D. J., Y. Zhang, J. Xing, W. S. Watkins, H. Ha, M. A. Batzer and L. B.
Jorde (2013). "Mobile element scanning (ME-Scan) identifies thousands of
novel Alu insertions in diverse human populations." Genome Res 23(7):
1170-1181.


create structural variation: analysis of a complete human genome." Genome Res 19(9):
1516-1526.

Yang, N. and H. H. Kazazian, Jr. (2006). "L1 retrotransposition is suppressed by

RUNX3 in human L1 transcription and retrotransposition." Nucleic Acids
Res 31(16): 4929-4940.

BIOGRAPHY

Vincent Anthony Streva was born in New York, New York on July 29, 1985 to Paul and Vita Streva. He graduated as Valedictorian from North Rockland High School in Thiells, NY in 2003, and then received his Bachelors of Science in Biological Sciences at Cornell University in Ithaca, NY in 2007. Vincent entered the Biomedical Sciences Ph.D. program in August 2008 and joined the laboratory of Prescott Deininger in pursuit of a Ph.D.