LINE-1 Integration Preferences in Human Somatic Cells

by

Diane Alexandra Flasch

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Human Genetics)
in the University of Michigan
2017

Doctoral Committee:

Professor John V. Moran, Co-Chair
Professor Thomas E. Wilson, Co-Chair
Professor Michael Boehnke
Assistant Professor Jeffrey M. Kidd
Assistant Professor Ryan E. Mills
Emeritus Professor Edward D. Rothman

Diane A. Flasch

daflasch@umich.edu

ORCID iD: 0000-0003-4825-2621

I dedicate this thesis to my parents, Joseph and Karen Flasch,
who instilled in me the relevance of critical thinking, logic, and deductive reasoning.

Acknowledgements

First and foremost I would like to thank my mentors, Dr. John V. Moran and Dr. Thomas E. Wilson. Both have dedicated much of their time and efforts towards this project, providing critical feedback that has only strengthened the integrity of this study. I'd like to personally thank Dr. John V. Moran for creating a rigorous, enthusiastic, and welcoming environment in his laboratory. I will forever be astonished by John's ability to recite detailed facts of any experiment performed in his laboratory, or that in a published work. I only hope that one day I too will be such an expert in my future field of study. I have truly appreciated John's support, both financial and moral, throughout this process.

To Dr. Thomas E. Wilson I thank profusely for his patience and ability to explain a complicated concept in such simple and pertinent terms. I'm thankful that Tom took on computer programing as a 'hobby' so that I could benefit from such knowledge he gained. While deeply humble, I'm convinced that Tom knows everything. I have yet to find a topic that he isn't already quite familiar with. Tom has suggested, provided, and helped perform a number of pertinent data sets for analysis in this project.

I would also like to thank my committee members, Dr. Michael Boehnke, Dr. Edward Rothman, Dr. Jeffrey Kidd, and Dr. Ryan Mills. They all provided helpful feedback both individually and collectively that helped shape this final project. To each and every single one of them I thank for their time and feedback!

I would like to thank all the members of the Moran laboratory, both past and present. They have all helped me throughout my graduate career by providing helpful feed back at the bench, or helping to lift my spirits. I have enjoyed working with each and every one of them, and hope to continue to keep in touch as future colleagues.

I would have never known my deep desire to explore genetics if it weren't for Mrs. Eileen Cairo at St. Viator High School who taught me both Honors Biology and AP Biology. Her two-week section devoted to genetics is what initially sparked my curiosity. I thank her for helping me find that curiosity! I also need to thank Dr. Howard Laten at Loyola University of Chicago who first introduced me to the world of repetitive elements in his basic Genetics course. He would later allow me the opportunity to study SIRE-1, a soybean repetitive element, in his laboratory for undergraduate research. I have appreciated Dr. Laten's continued support and words of encouragement from the very beginning. If it weren't for his encouragement, I may have never applied to as competitive a program as at the University of Michigan.

Finally I would like to thank all of my family and friends. To the Michigan friendship circle, I love you all! You all helped me get through graduate school by keeping me sane! To my family, thank you for always believing in me, even when I had doubts. I would have never been able to do this without your support and love.

Table of Contents

List of Figures

List of Tables

List of Appendices

List of Abbreviations

APE	Apurinic/apyrimidinic-like endonuclease
ATM	Ataxia-telangiectasia mutated
ATR	Ataxia-telangiectasia and Rad3 related
ATRIP	ATR interaction protein
Bru-seq	Bromouridine sequencing
CCS	Circular consensus sequence
CDF	Cumulative distribution function
CHO	Chinese hamster ovary
dsDNA	Double-strand DNA
EGFP	Enhanced green fluorescent protein
EN	Endonuclease
ENCODE	Encyclopedia of DNA Elements
ENi	Endonuclease-independent
ERV	Endogenous retrovirus
FA	Fanconi anemia
FANCD2	Fanconi anemia complementation group D2
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
HERV	Human-specific endogenous retrovirus
hESC	human embryonic stem cell
HGWD	Human genome working draft sequence
ICL	Inter-strand DNA crosslink
LINE-1 or L1	Long Interspersed Element-1
LRE3	example of L1Hs specific sequence
LTR	Long terminal repeat
L1Hs	LINE-1 *Homo Sapiens*; the only active L1 subfamily to date
L1.3	example of L1Hs specific sequence
MLV	Moloney murine leukemia virus
MYA	Million years ago
NAHR	Non-allelic homologous recombination
NHEJ	Non-homologous end joining
NPC	Neuronal progenitor cells
OH	Hydroxyl
OK-seq	Okazaki fragment sequencing
ORF	Open reading frame
PacBio	Pacific Biosciences sequencing platform
PCNA	Proliferating cell nuclear antigen
PIP	Proliferating cell nuclear antigen interacting domain
PLE	*Penelope*-like element
RFD	Replication Fork Direction

RPA	Replication protein A
RPKM	Reads Per Kilobase of transcript per Million mapped reads
RRM	RNA recognition motif
RT	Reverse transcriptase
SINE	Short INterspersed Element
SMRT	Single molecule real time
ssDNA	Single-strand DNA
SVA	SINE-R/VNTR/Alu-like
TE	Transposable elements; mobile pieces of DNA
TIR	Terminal inverted repeat sequence
TPRT	Target-site primed reverse transcription
TSD	Target-site duplication
UTR	Untranslated regions
VLP	Virus-like particle
WGA	Whole genome amplification

Abstract

Long INterspersed Element-1 (LINE-1 or L1) is the only autonomously active transposable element in the human genome. The vast majority of L1s are inactive, but a small number (~80-100 per human genome) retain the ability to mobilize by a 'copy and paste' mechanism called retrotransposition. L1 encodes two proteins (ORF1p and ORF2p) required for retrotransposition. ORF2p is a 150kDa protein that has endonuclease (EN) and reverse transcriptase (RT) activities that are responsible for initiating L1 integration by a mechanism termed target-site primed reverse transcription (TPRT). During canonical TPRT, the L1 EN makes a single-strand endonucleolytic nick at a double-stranded genomic DNA target sequence (typically 5'-TTTT/A-3' and variants of that sequence), to expose a 3'-hydroxyl group that is used as a primer by the L1 RT to reverse transcribe L1 messenger RNA.

Different types of transposable elements (TEs) have evolved convergent strategies to target genomic 'safe havens,' where TE insertions are predicted to have relatively minimal effects on host fitness and gene expression. Whether L1 integrates into specific genomic regions requires elucidation. In this thesis, I have examined L1 integration preferences in four human cell lines that are proxies for *in vivo* cell types known to accommodate endogenous *de novo* L1 retrotransposition events. By combining cultured cell, molecular biological, the Pacific Bioscience sequencing platform, and computational approaches, I characterized 65,079 *de novo* engineered human L1 integration sites. I compared our L1 insertion dataset to a weighted random model, which assumes that L1 integration preferences are mediated solely by the presence of a degenerate L1 EN consensus cleavage site in the human genome. The data suggest that gene content, transcriptional activity, strand bias, epigenetic environment, and DNA

replication status have minimal effects on L1 integration. Thus, L1 EN is the principal determinant of L1 integration.

In contrast to canonical EN-dependent L1 retrotransposition, previous studies indicated that L1s could also integrate at sites of DNA damage, including dysfunctional telomeres, by an endonuclease-independent (ENi) mechanism in certain cultured cell lines that contain mutations in genes that render the non-homologous end-joining (NHEJ) pathway of DNA repair and p53 inactive. Here, we explored whether the disruption of other DNA repair pathways influence ENi L1 integration. We observed ENi retrotransposition in certain tissue culture cell lines containing defects in the Fanconi anemia (FA) DNA repair pathway. Since defects in the FA pathway can lead to the accumulation of inter-strand DNA crosslinks that, if left unrepaired, can interfere with DNA replication, we hypothesized that lesions arising at stalled DNA replication forks may provide substrates for enhanced ENi retrotransposition. Indeed, the examination of L1 EN mutant integration sites in FANCD2-deficient cells, suggests that a 3'-hydroxyl group present at Okazaki fragments and/or double-strand DNA breaks generated at collapsed DNA replication forks might be used as a primer to initiate ENi L1 retrotransposition.

In sum, our results suggest that ENi L1 retrotransposition may represent an ancestral mobilization mechanism used by LINE-like retrotransposons prior to the acquisition of an endonuclease domain. Under this scenario, LINE-like elements were reliant upon genomic features (*e.g.*, sites of genomic DNA damage, replication forks, and, less frequently, dysfunctional telomeres) to initiate TPRT in the absence of an endonuclease. Indeed, we posit that the acquisition of an endonuclease domain allowed L1 to autonomously insert throughout the genome and, as originally implied by its name, become an interspersed retrotransposon.

CHAPTER 1

Transposable Elements Mobilize in Genomes

Overview

This thesis aims to determine if LINE-1 displays preferential integration sites in the human genome. Chapter one provides a general overview of transposable elements and their means of mobilization in genomes. I then primarily focus on LINE-1 and previously published methods to characterize endogenous LINE-1 integration events in the human genome. Finally, I explore the integration preferences of several transposable elements and discuss what is known about LINE-1 integration preference. Chapter two provides the bulk of my thesis work and describes the generation, capture, and analysis of thousands of engineered LINE-1 integration events. In Chapter 2, I also describe the creation of a weighted model data set, which we used in our analysis to determine the influence of the LINE-1 endonuclease on integration preference in the human genome with respect to several examined genomic features. In Chapter 3, I explore the influence of DNA repair proteins on LINE-1 integration preference. Chapter 4 provides a summary of my findings, conclusions, and discusses areas for future study.

Abstract

Eukaryotic genomes are littered with the remnants of parasitic mobile pieces of DNA known as transposable elements (TEs). For the most part, transposable elements exist in a delicate balance with their host genome – promoting their own replication while minimizing detrimental mutations that can damage their host. TEs are mutagenic, and their insertion can disrupt normal gene expression, create genomic structural variation, and alter epigenetic marks within the genome, which can lead to intra-species, intra-individual, and inter-individual genetic diversity. From a teleological standpoint, it is

imperative that TEs do not become too 'greedy,' as their relentless mobility may adversely affect genome integrity. As such, it is not surprising that many TEs have evolved strategies to insert into specific genomic regions, thereby minimizing damage to their respective host genomes.

In this chapter, I provide a brief background about general TE biology, and then delve further into the biology of the only autonomously active retrotransposable element in the human genome, Long INterspersed Element-1 (LINE-1 or L1). I then discuss previous studies, which highlight methods used to discover polymorphic germline and somatic human TE insertions within early developmental precursor cells, cancers, and neurons. Finally, I discuss different integration mechanisms and known insertion site preferences of several TEs, and then provide additional information about LINE-1 to address the over-arching question in this dissertation: Do LINE-1 retrotransposition events display integration preferences within the human genome?

A Plethora of Repetitive Elements

The completion of the human genome working draft sequence (HGWD) led to the realization that protein-coding exons only comprise ~1.5% of the human genome (Consortium, 2012; Lander et al., 2001). We now know that sequences outside of exonic and intronic sequences, such as promoters and enhancers, are also critical genomic components that are needed to ensure proper gene expression (Khoury and Gruss, 1983; Smale and Kadonaga, 2003). However, these sequences only compose between 8 and 15% of our genome (Kellis et al., 2014; Ponting and Hardison, 2011; Rands et al., 2014).

In 1968, Britten and Kohne performed DNA hybridization experiments (*i.e.*, C_0t analyses) to determine the relative proportions of repetitive DNA sequence in several mammalian genomes (Britten and Kohne, 1968; Waring and Britten, 1966). Surprisingly, at least 50% of the human genome can be classified as repetitive DNA (Lander et al., 2001). These repetitive sequences can be divided into four categories: (1) interspersed repeats, also known as transposable element (TE) derived repeats; (2) short k-mer nucleotide repeats, such as microsatellite and minisatellite DNAs; (3) segmental

duplications, which generally range in size from 10-300 kb; and (4) tandem repeat sequences, such as the α-satellite DNA present at many centromeres.

Transposable element derived repeats comprise ~45% of the human genome (Lander et al., 2001) and can be subdivided into 2 groups: DNA transposons and retrotransposons (Figure 1.1). DNA transposons comprise 3% of the human genome, while retrotransposons comprise the remaining 42% of our genome (Lander et al., 2001). Elements in both groups can be further subdivided into autonomous and non-autonomous elements. Autonomous elements encode proteins required for their mobility, whereas non-autonomous elements rely upon the proteins of the autonomous elements to mobilize throughout the genome (Craig, 2014; Richardson et al., 2015).

Since the cells in our body must expend energy during DNA replication to create new cells to maintain the life of the organism, the following questions arise: (1) why would our genome retain all this extra sequence if it is not 'required' for survival; and (2) why does an organism 'keep' this extra 'baggage' [*i.e.*, commonly referred to as 'junk' DNA (Ohno, 1972)] if it requires more effort to maintain?

DNA Transposons

In the 1940's, almost two decades prior to the completion of the seminal experiments of Britten and Kohne, Barbara McClinktock discovered that DNA transposable element activity is responsible for the variegated corn kernel color phenotypes observed in certain strains of *Zea mays* (McClintock, 1950). Subsequent work uncovered two DNA transposable elements responsible for this phenomenon: the autonomous *Activator (Ac)* DNA transposon, and its non-autonomous partner, *Dissociation (Ds)* (Fedoroff et al., 1983). *Activator* encodes an enzyme that makes a double-strand break (*i.e.*, transposase). When *Activator* creates double-strand breaks at a specific locus harboring *Dissociation*, it leads to genomic instability that effects the expression of a pigmentation gene. Thus, the phenotypic variation leading to kernel color variegation was due to TE activity during different stages of kernel development. These data provided the first evidence suggesting that genomes are unstable entities, and that mobile DNA sequences can lead to genomic instability resulting in phenotypic

variation. In recognition of this discovery, Barbara McClintock won the Nobel Prize in Physiology or Medicine in 1983 (McClintock, 1950).

DNA transposons typically consist of a pair of terminal inverted repeat sequences (TIRs) that surround a transposon-encoded open reading frame (ORF) that encodes an enzyme named transposase (Figure 1.1). Transposase is a member of the $DD_{35}E$ superfamily of integrase proteins. In eukaryotes, transposase binds the transposon TIRs within the nucleus, 'cuts' the element from its existing genomic location, and 'pastes' the excised DNA into a new genomic location. Transposition leads to the generation of signature length, short 4-6 bp target-site-duplications (TSDs) that flank the TIRs in genomic DNA (Feschotte and Pritham, 2007; Munoz-Lopez and Garcia-Perez, 2010). As a consequence of this non-replicative 'cut and paste' transposition mechanism, the copy number of DNA transposons within a genome remains relatively constant—the DNA transposon sequence is simply 'cut' out of one genomic location and moved (or 'pasted') into a new genomic location (Kleckner, 1990). While DNA transposons thrive in bacteria and simple eukaryotes, they appear to be extinct in most mammalian, including human, genomes (Lander et al., 2001).

The maize *Activator* (*Ac)*, *Drosophila melanogaster hobo*, and *Antirrhinum majus* (a flowering plant) *Tam3* DNA transposons are members of the hobo-*Ac*-Tam3 (hAT) superfamily of eukaryotic DNA transposons (Calvi et al., 1991; Feldmar and Kunze, 1991). The hAT transposons are the most abundant DNA transposons found in humans; however, they have been inactivated by mutational processes and were rendered immobile approximately 40 million years ago (Pace and Feschotte, 2007) (Figure 1.1). Interestingly, while DNA transposons are extinct in most mammalian genomes, recent studies suggest that some hAT DNA transposons remain active in the little brown bat, *Myotis lucifigus* (Ray et al., 2008; Ray et al., 2007), indicating that transposable elements follow different evolutionary trajectories in different mammalian lineages. Additional examples of active DNA transposons are the *Drosophila Melanogaster* P-element (Kidwell, 1992), the bacterial Tn7 transposon (Craig, 1996), and insect *piggyBac* transposons (Cary et al., 1989).

HUH DNA transposons (where H represents a histidine and U represents a hydrophobic residue) are DNA transposons found in all domains of life. HUH transposons can be distinguished from other classes of mobile elements because they mobilize by a replicative process that employs a single-strand DNA (ssDNA) intermediate (Lam and Roth, 1983). Indeed, many HUH elements preferentially integrate into the lagging strand during DNA replication, leading to a copy of the transposon at the initial donor site and newly inserted copy in genomic DNA. HUH DNA transposons elements also carry sub-terminal palindromic structures, instead of TIRs, at their termini and insert 3' to specific AT-rich tetra- or penta- nucleotides without duplicating nucleotides at the genomic DNA target site (Figure 1.2). Examples of HUH DNA transposons include: IS608, originally discovered in the pathogen *Helicobacter pylori* (Kersulyte et al., 2002) and ISDra2 from *Deinococcus radiodurans* (Ton-Hoang et al., 2010). Integration of IS608 always occurs immediately downstream of a tetra-nucleotide sequence (5'-TTAC) (Kersulyte et al., 2002), whereas the target sequence of ISDra2 is a penta- nucleotide sequence (5'-TTGAT) (Islam et al., 2003).

Retrotransposons

The second category of TEs, retrotransposons, mobilize by a replicative 'copy and paste' mechanism in which the progenitor retrotransposon sequence is first transcribed into RNA, 'copied' into complementary DNA (cDNA) by a reverse transcriptase (RT), and then 'pasted' into a new genomic location (Cost et al., 2002; Mager and Stoye, 2015). Thus, the progenitor element is left unaltered and a reverse transcribed copy is inserted into a new genomic location. This replicative mechanism bestows retrotransposons with the potential to undergo exponential amplification in the genome.

Phylogenetic analyses led to the identification of two distinct groups of RT-containing retrotransposons (Malik et al., 1999; Xiong and Eickbush, 1988). The first group consists of retroviruses, certain DNA viruses, and long terminal repeat (LTR) retrotransposons. The second group consists of RT-containing sequences of fungal mitochondrial introns and non-LTR retrotransposons. Indeed, mobile group II introns in

bacterial and organellar genomes are proposed predecessors of non-LTR retrotransposons (Lambowitz and Zimmerly, 2004).

Retrotransposons can be further subdivided into two distinct classes: LTR retrotransposons and non-LTR retrotransposons. LTR retrotransposons are structurally similar to simple retroviruses, but since most contain a non-functional envelope (*env)* gene, they are confined to replicate within the cell (Bannert and Kurth, 2006). Non-LTR elements move by a fundamentally different mechanism termed target-site primed reverse transcription (TPRT), which is discussed in greater detail below.

Long Terminal Repeat Retrotransposons: Intracellular Retroviral-like Mobilization

LTR retrotransposons structurally resemble simple retroviruses and consist of direct sequence repeats (*i.e.*, LTRs) that flank an internal coding region. The LTRs consist of U3 (unique to the 3' end); R (repeated); and U5 (unique to the 5' end) sequences. Transcription of the LTR initiates in the 5' LTR at the U3/R junction and terminates in the 3' LTR at the R/U5 junction (Craig, 2014; Mager and Stoye, 2015). Thus, RNA polymerase II generates a full-length LTR-retrotransposon RNA that contains the following structure: 5'-R-U5-internal coding regions-U3-R-3'.

The internal coding region of LTR retrotransposons contain at least two open reading frames (ORFs), which encode structural and enzymatic proteins. The first ORF encodes gag, a structural protein that plays a key role in the formation of cytoplasmic virus-like particles (VLPs). The second ORF encodes pol, a protein with four, distinct enzymatic activities (protease, reverse transcriptase, integrase, and RNase H) (Craig, 2014). Like DNA transposons, LTR retrotransposon integrase proteins generally are members of the $DD_{35}E$ superfamily of proteins. The third open reading frame, if present, encodes envelope (env), a protein crucial for retroviral packaging and cellular export. Since env is either absent or not functional in most LTR retrotransposons, LTR retrotransposons are generally relegated to intracellular replication (Bannert and Kurth, 2006).

The mechanism of LTR retrotransposition involves packaging the element encoded RNA and additional host factors [*e.g.*, a specific transfer RNA (tRNA)] into

VLPs (Craig, 2014). Within cytoplasmic VLPs, a specific host encoded tRNA binds to a complementary sequence located downstream of the R/U5 junction, creating a tRNA/RNA hybrid. The 3' hydroxyl group at the end of the tRNA is then used as a primer by the reverse transcriptase activity encoded by pol to convert the retrotransposon RNA into a double-stranded cDNA. Integrase then binds to the LTR sequences at the end of the resultant double-stranded cDNA, transports it to the nuclear DNA, and integrates the cDNA at a new site in genomic DNA. Integration occurs by a mechanism similar to that employed by DNA transposons and leads to the generation of short, signature length, TSDs (usually 4-6 bp in size) that flank the LTRs in genomic DNA (Telesnitsky and Goff, 1997).

Unlike DNA transposons, LTR retrotransposons are quite prevalent in eukaryotes. For example, the mouse genome contains multiple active LTR-retrotransposon families. Indeed, it is estimated that approximately 10-12% of sporadic mutations in the mouse are due to the retrotransposition of autonomous and non-autonomous LTR retrotransposons (Maksakova et al., 2006; Kazazian and Moran, 1998). Human-specific endogenous retroviruses (HERVs) comprise ~8% of human genomic DNA, have been rendered inactive by mutations, and are no longer active within our genomes (Lander et al., 2001). However, certain members of the HERV-K family (where the K designates a lysine tRNA that is critical for priming first-strand HERV-K cDNA synthesis) are polymorphic with respect to presence/absence in the human population (Belshaw et al., 2005; Moyes et al., 2007). These findings suggest that endogenous retroviruses (ERVs) have been active since the divergence of humans and chimpanzees, but that HERV activity has since decreased in the human lineage. Other examples of autonomous LTR retrotransposons include: *Saccharomyces cerevisiae Transposon yeast 1* (Ty1), Ty2, Ty3, and Ty5 (Boeke et al., 1985), *Schizosaccharomyces pombe Tf1* (Esnault and Levin, 2015), and autonomously active mouse intracisternal A particles (IAP) and MusD (Mager and Stoye, 2015).

Non-Long Terminal Repeat (non-LTR) Retrotransposons: Target-site Primed Reverse Transcription

Non-LTR retrotransposons lack LTR sequences, but like LTR retrotransposons, require an RT-containing protein to mediate their retrotransposition (Gilbert et al., 2005; Lander et al., 2001; Luan et al., 1993; Moran et al., 1996). Non-LTR retrotransposons also encode an endonuclease (EN) activity that is required for integration. While the reverse transcriptase domain is shared among different non-LTR elements, the endonuclease domain varies between elements. In some cases, this nuclease has a profound effect on non-LTR retrotransposon target-site integration preference (see below).

Non-LTR elements are able to 'copy and paste' themselves into a new genomic location by a process termed target-site primed reverse transcription (TPRT). TPRT requires both endonuclease and reverse transcriptase activities (Luan et al., 1993). In its simplest form, TPRT involves cleavage of one strand of target site genomic DNA, exposing a free 3' hydroxyl (OH) group that can be used as a primer for reverse transcription of the element-encoded RNA. Second-strand DNA cleavage typically occurs downstream of the initial single-strand endonucleolytic nick, exposing a 3'-OH group that is presumably used by the element-encoded RT to mediate second-strand DNA synthesis (Christensen and Eickbush, 2015). The completion of integration generally leads to the generation of variably sized TSDs flanking the newly inserted element, which for human L1s range in size from ~7-20bp (Gilbert et al., 2005; Lander et al., 2001).

Types of non-LTR Retrotransposons

Penelope-like Elements

Penelope-like elements (PLE) are a widespread class of retroelements named after *Penelope*, a transposable element originally isolated from *Drosophila virilis* (Evgen'ev et al., 1997). PLEs contain a single open reading frame (ORF) that encodes a reverse transcriptase (RT) and GIY-YIG endonuclease (EN) domain, which is also found in the proteins encoded by certain bacterial group I introns (Arkhipova et al.,

2003; Belfort and Perlman, 1995; Evgen'ev and Arkhipova, 2005). Interestingly, several PLE members encode spliceosomal introns, which is unusual for an element that is believed to move through an RNA intermediate (Figure 1.2C). The ends of PLEs usually contain several hundred base pair long direct repeats that encode a self-splicing hammerhead ribozyme, which is believed to mediate the excision of the *Penelope* RNA from a larger, precursor pre-mRNA transcript (Cervera and De la Pena, 2014).

Through the process of TPRT, the EN activity of PLE creates a single-strand endonucleolytic cleavage at a double-strand DNA target, resulting in the generation of a free 3'-OH group, which can be used as a primer by the element-encoded RT to initiate reverse transcription of Penelope mRNA (Pyatkov et al., 2004). The EN exhibits some sequence preference, preferring AT-rich targets in genomic DNA, but otherwise contains no other pronounced sequence specificity (Evgen'ev and Arkhipova, 2005; Pyatkov et al., 2004). The examination of *Penelope* integration sites in several geographical strains of *D. virilis* reveals an integration preference for euchromatic chromosome arms; Penelope elements are rarely found in heterochromatic regions of chromatin (Evgen'ev et al., 2000; Evgen'ev and Arkhipova, 2005; Zelentsova et al., 1999). Interestingly, 45% of the *Penelope* integration sites found in multiple *D. virilis* strains, termed 'hot spots', coincide or lie in close vicinity of inversion breakpoints (Evgen'ev and Arkhipova, 2005). PLEs have also been discovered in crustaceans, echinoderms, fish, amphibians, flatworms, roundworms, and rotifers (Dalle Nogare et al., 2002; Lozovskaya et al., 1990; Lyozin et al., 2001; Volff et al., 2001).

Phylogenetic analysis of the PLE reverse transcriptase reveals that it forms a sister clade with telomerase reverse transcriptase (TERT), which is distinct from the non-LTR and LTR clades of retrotransposons (Gladyshev and Arkhipova, 2007). These data have fueled speculation that the RTs of PLEs are the progenitor of telomerase (Curcio and Belfort, 2007; Gladyshev and Arkhipova, 2007). Notably, TERTs are not TEs and lack a recognizable endonuclease domain. Instead, they are specialized ribonucleoprotein (RNP) enzymes that maintain telomeres by reiteratively copying a short segment of an unlinked template RNA (commonly called TERC) onto the 3'-OH group present at the end of a linear chromosome (Greider and Blackburn, 2004).

Mobile Group II Introns: A Proposed Ancestor of Non-Long Terminal Repeat Retrotransposons

Mobile group II introns are catalytic self-splicing RNAs that are found in bacterial, eukaryotic organelle, and certain archaebacteria genomes (Novikova and Belfort, 2017). Mobile group II introns are thought to be the evolutionary precursors of splicesomal introns (Lambowitz and Zimmerly, 2004). They typically encode a single ORF, which encodes a protein that is critical for splicing (*i.e.*, a maturase activity) and/or intron mobility (*i.e.*, reverse transcriptase and DNA endonuclease activities) (Figure 1.2C) (Gorbalenya, 1994; Kennell et al., 1993; Michel and Lang, 1985; Moran et al., 1994). Notably, some group II introns (*e.g.* aI1 and aI2 in the yeast mitochondrial *coxI* gene and LI.LtrB in the bacteria *Lactococcus lactis*) are mobile genetic elements (Moran et al., 1995; Cousineau et al., 1998). Group II intron mobilization can occur by two distinct mechanisms: retrohoming and retrotransposition (Ichiyanagi et al., 2002; Lambowitz and Zimmerly, 2004).

Retrohoming enables group II introns to insert into an intronless allele of their cognate gene. During the first step of retrohoming, a ribonucleoprotein complex, which consists of the intron encoded protein (IEP) and spliced intron lariat RNA, uses both RNA-DNA base pairing and IEP-DNA interactions to reverse splice into the sense strand of a double-strand DNA target present in the intronless allele of its cognate gene (Lambowitz and Zimmerly, 2004). By a process similar to TPRT, the IEP endonuclease then cleaves the opposite strand of DNA, exposing a 3' hydroxyl group that is used by the IEP RT to reverse transcribe the reverse spliced intron RNA (Zimmerly et al., 1995). Bacterial group II introns most likely rely upon cellular RNase H to digest the RNA template and host DNA polymerase to complete second-strand synthesis (Cousineau et al., 1998). Additional host factors (*e.g.*, DNA ligase) likely act to complete retrohoming.

Retrotransposition allows group II introns to insert into new genomic locations, allowing intron dispersal. Retrotransposition does not require EN activity, but instead occurs when group II introns insert into an ectopic site (generally ssDNA or ssRNA) that resembles the normal retrohoming substrate (Ichiyanagi et al., 2002). Like retrohoming, the first step of retrotransposition involves reverse splicing of the group II intron RNA

into an ectopic target site. The integrated intron RNA is then reverse transcribed by the IEP RT, but integration is completed by recombination repair, using the original donor intron as a repair substrate.

Autonomous Non-LTR Retrotransposons

Autonomous non-LTR retrotransposons consist of one or two open reading frames (ORFs) that encode endonuclease (EN) and reverse transcriptase (RT) activities critical for retrotransposition. Some non-LTR retrotransposons (*e.g.*, mammalian LINE-1s) end in a poly(A) tract, whereas others (*e.g.*, the African clawed frog Tx1L element, *Drosophila* I-factor, and zebrafish LINE-2 elements) end in a series of simple repeated DNA sequence.

Autonomous non-LTR retrotransposons replicate by TPRT (Cost et al., 2002; Feng et al., 1996; Luan et al., 1993; Moran et al., 1996). During TPRT, the element-encoded endonuclease makes a single-strand endonucleolytic nick in target site DNA to expose a 3' hydroxyl group that is used by the element encoded RT as a primer to reverse transcribe the transposable element mRNA.

Although they share an evolutionary conserved RT domain, different autonomous non-LTR retrotransposons encode for different endonucleases. For example, some site-specific non-LTR retrotransposons (*e.g.*, insect R2 elements) contain a type-II restriction endonuclease (RE)-type domain, whereas others (*e.g.*, mammalian LINE-1s, insect R1, Waldo, and MinoAg1, silkworm SART1 and TRAS1, frog Tx1L, and green algae DRE retrotransposons) encode apurinic/apyrimidinic-like endonucleases (APEs) (Figure 1.3) (Zingler et al., 2005). The different endonuclease activities likely mediate the different integration preferences for each type of retrotransposon (see later in Chapter).

Non-autonomous Non-LTR Retrotransposons: Alu, SVA, and Some Cellular RNAs

Non-autonomous non-LTR retrotransposons, such as Short INterspersed Elements (SINEs), include Alu and SINE-R/VNTR/Alu-like (SVA) elements comprise ~11% human genomic DNA (Figure 1.1) (Lander et al., 2001). Alu is named after the *AluI* restriction site within the element (Houck et al., 1979). Full-length Alus are ~280bp

in length and consist of two related monomers that are derived from 7SL RNA (Rubin et al., 1980; Ullu and Tschudi, 1984). The left monomer contains an internal RNA polymerase III promoter and is followed by an A-rich linker sequence, which separates it from the right monomer (Ullu and Weiner, 1985). Likes L1s, Alus typically end in a poly(A) tract. However, the poly(A) tract is encoded by genomic DNA and may be extended during TPRT by a template switching mechanism (Roy-Engel et al., 2002); it is not added to Alu RNA during post-transcriptional processing.

SVA elements comprise ~0.2% of human genomic DNA (Richardson et al., 2015). A typical SVA element is ~2kb in length, appears to be transcribed by RNA polymerase II, and has a composite structure that consists of: (1) a hexameric CCCTCT repeat; (2) an inverted Alu-like element repeat; (3) a set of GC-rich variable nucleotide tandem repeats (VNTRs); (4) a SINE-R sequence derived from HERV-K10, an inactive retrotransposon; (5) a canonical RNA polymerase II cleavage polyadenylation specificity factor binding site that is followed by a poly(A) tract (Richardson et al., 2015).

Non-autonomous non-LTR retrotransposon elements do not encode proteins. Instead, they must rely upon the endonuclease and reverse transcriptase activities encoded by autonomous non-LTR retrotransposons, such as L1 for Alu and SVA, to mediate their retrotransposition (Dewannieux et al., 2003; Hancks et al., 2011; Raiz et al., 2012). Since these elements are mobilized via TPRT, they are typically flanked by variable-length target-site duplications (Grimaldi and Singer, 1982; Rubin et al., 1980).

The LINE-1 encoded proteins have been implicated in the mobilization of other cellular RNAs. For example, ORF1p and ORF2p can occasionally act in *trans* to mobilize mature cellular mRNAs to new genomic locations, resulting in processed pseudogene formation (Esnault et al., 2000; Vanin, 1985; Wei et al., 2001; Weiner et al., 1986). Similarly, ORF1p and/or ORF2p can mobilize non-coding RNAs such as U6 uracil-rich small nuclear RNAs (U6 snRNA) and small nucleolar RNAs (snoRNAs) (*i.e.,* U3 snoRNA) to new locations within the genome (Buzdin et al., 2003; Buzdin et al., 2002; Garcia-Perez et al., 2007; Gilbert et al., 2005). Unlike the mechanism of processed pseudogene formation, the structures of chimeric U6/L1 pseudogenes

suggest that U6 snRNA is conjoined to a 5' truncated LINE-1 cDNA during the process of LINE-1 integration (Buzdin et al., 2002; Garcia-Perez et al., 2007). U6/LINE-1 chimeras have been discovered in several primate genomes (Hasnaoui et al., 2009) and in cultured cell retrotransposition assays (Garcia-Perez et al., 2007; Gilbert et al., 2005), suggesting their formation is ongoing in the human population.

LINE-1 in the Human Genome

L1s are responsible for ~17% of human genomic DNA and are the only known autonomously active retrotransposons in the human genome (Lander et al., 2001). The majority of L1s have been rendered inactive by mutational processes [*i.e.*, 5' truncations (Grimaldi and Singer, 1983), internal rearrangements (Ostertag et al., 2001), and/or point mutations that disrupt the activities of the L1-encoded proteins (ORF1p and ORF2p)]. However, experiments in cultured cells indicate that ~80-100 L1s per individual retain the ability to retrotranspose (Brouha et al., 2003; Moran et al., 1996; Sassaman et al., 1997).

A Brief History of L1 Evolution

L1s have been replicating in mammalian genomes since the marsupial/placental divergence, which occurred ~170 MYA (Burton et al., 1986; Smit et al., 1995; Yang et al., 2014). The vast majority of L1s present in the human genome amplified since the divergence of the ancestral mouse and human lineages ~65 to 75 MYA. Sequence comparisons of primate-specific L1s identified sixteen subfamilies (termed PA1 to PA16) (Khan et al., 2006; Smit et al., 1995). These L1 subfamilies have a monophyletic origin and have undergone an amplification process known as subfamily succession, where one L1 subfamily is replaced over evolutionary time by an emergent L1 subfamily. This evolutionary trajectory is a signature of an 'arms race' and suggests that the emergent L1 subfamily replicates more efficiently than its predecessor, perhaps because it evades the effects of host factors that repress retrotransposition (Jacobs et al., 2014). As a result of subfamily succession, only one L1 subfamily preferentially amplifies in the human genome at any given time.

The youngest L1 subfamily (termed PA1 or L1Hs) has amplified since the divergence of humans and chimpanzees, which occurred ~6 MYA (Goodman et al., 1998). The majority of active L1Hs elements belong to a small subset termed transcribed-active L1s (the Ta-subset) (Skowronski et al., 1988), which arose ~4 MYA (Beck et al., 2011; Brouha et al., 2003; Moran et al., 1996; Sassaman et al., 1997). Ta-subset L1s contain a defining 5-'ACA' trinucleotide sequence and G nucleotide in their 3' UTR (at positions 5930-5932 and 6015, respectively based on the sequence L1.2, an active L1: GenBank accession number M80343) (Dombroski et al., 1991). These nucleotides allow one to distinguish Ta-subset L1s from chimpanzee and older L1s in the human genome using BLAT or BLAST searches (Altschul et al., 1997; Kent, 2002) and PCR-based molecular biological approaches. Together, these approaches have been used to identify human-specific, polymorphic, and *de novo* L1 insertions in human genomes (Beck et al., 2010; Beck et al., 2011) (see below).

The Ta-subset can be further sub-divided into two subgroups, the older Ta-0 and newer Ta-1 L1s (Boissinot et al., 2000). Approximately 80% of Ta-0 elements inserted into the human genome over ~1.6 MYA (Boissinot et al., 2000). The newer Ta-1 subset appears to have arisen about 2.5 MYA, and most (~75%) of the Ta-1 L1s having been generated during the last ~1.6 million years (Boissinot et al., 2000). While Ta-0 contains some active elements, Ta-1 contains the greatest number of active LINE-1s (Beck et al., 2010). It is believed that ~69% of a small number of Ta-1-containing loci are polymorphic with respect to presence/absence in the genome (Boissinot et al., 2000). Examination of the human genome reference reveals that approximately 30% to 35% Ta-subset L1s are full-length, while 65% to 70% contain 5' truncations. Notably, approximately 25% of 5' truncated L1s contain internal rearrangements known as inversion/deletions (Boissinot et al., 2000; Lander et al., 2001; Myers et al., 2002; Ostertag and Kazazian, 2001).

LINE-1: Structure of a Full-length Element

A full-length human L1 sequence is approximately 6 kb in length (Dombroski et al., 1991; Scott et al., 1987). Full-length L1s contain a 5' untranslated region (UTR), two ORFs, and a 3' UTR that is typically ends in a poly(A) tract (Figure 1.1). The 5'UTR

contains *cis*-acting DNA binding sites for multiple transcription factors (*e.g.*, YY1, Sp1, RUNX3, and SRY-like) that are important for L1 transcription (Athanikar et al., 2004; Becker et al., 1993; Kuwabara et al., 2009; Minakami et al., 1992; Swergold, 1990; Tchenio et al., 2000; Yang et al., 2003). The human L1 5'UTR also contains both a sense and an antisense RNA polymerase II promoter. Transcription from the sense strand promoter can generate full-length L1 RNAs (Swergold, 1990). Transcription from the L1 antisense promoter (ASP) can influence the expression of genes residing upstream of a full-length L1 (Macia et al., 2011; Speek, 2001). The ASP also encodes an open reading frame (ORF-0) (Denli et al., 2015); however, whether ORF-0 plays any role in L1 retrotransposition requires further elucidation.

Two open reading frames (ORF1 and ORF2) that are separated by a 63bp inter-ORF spacer (Dombroski et al., 1991) follow the L1 5'UTR. Human ORF1 encodes a ~40kDa RNA binding protein (ORF1p) that contains an amino-terminal protein-protein interaction domain, a centrally located RNA recognition motif (RRM) (Khazina et al., 2011; Khazina and Weichenrieder, 2009), and a carboxyl-terminal basic domain (Moran et al., 1996). Biochemical studies indicate that ORF1p forms a trimer (Khazina et al., 2011) that can bind ~50 base pairs of unstructured single-strand RNA and DNA substrates in a sequence independent manner (Callahan et al., 2012; Khazina and Weichenrieder, 2009; Martin and Bushman, 2001; Martin et al., 2005). Interactions between the ORF1p RRM and carboxyl-terminal basic domain mediate nucleic acid binding (Hohjoh and Singer, 1996, 1997; Holmes et al., 1992; Januszyk et al., 2007; Khazina et al., 2011; Khazina and Weichenrieder, 2009; Martin, 1991). ORF1p also contains nucleic acid chaperone activity that may play an important role in the initial steps of TPRT (Khazina and Weichenrieder, 2009; Martin and Bushman, 2001). ORF1p is required for retrotransposition in cultured human cells (Holmes et al.,1992; Leibold, et al., 1990;Moran et al., 1996).

Human ORF2 is translated from a bicistronic L1 RNA by an unconventional termination/reinitiation mechanism, leading to the production of a ~150kDa protein (ORF2p) (Alisch et al., 2006; Doucet et al., 2010; Ergun et al., 2004). ORF2p contains endonuclease (EN) (Feng et al., 1996) and reverse transcriptase (RT) activities (Hattori

et al., 1986; Mathias et al., 1991). ORF2p also contains a carboxyl-terminal cysteine-rich (C) domain of unknown function (Fanning and Singer, 1987; Feng et al., 1996; Moran et al., 1996). Mutational analyses demonstrate that the ORF2p EN, RT, and C-domains are required for L1 retrotransposition in cultured cells (Moran et al., 1996; Wei et al., 2001).

The L1 3'UTR is ~206bp in length and contains a conserved polypurine tract that is predicted to form a G-quadruplex structure (Usdin and Furano, 1989). While this polypurine tract is evolutionary conserved amongst mammalian L1s (Howell and Usdin, 1997; Usdin and Furano, 1989), it is dispensable for L1 retrotransposition in cultured cells (Moran et al., 1996). Thus, how the polypurine tract functions in L1 biology requires elucidation. The L1 3'UTR also contains a functional RNA polymerase II polyadenylation signal. This poly(A) signal is relatively weak and is often bypassed in favor of adjacent downstream poly(A) signals present in 3' flanking genomic DNA sequences (Moran et al., 1999). The use of these genomic polyadenylation sequences can lead to addition of 3' genomic sequences to the L1 RNA; the retrotransposition of the resultant RNAs can lead to L1-mediated 3' sequence transductions [(Holmes et al., 1994; Moran et al., 1996; Moran et al., 1994) (also, see below)].

Mobilization of L1 in the Human Genome

The L1 retrotransposition cycle begins with transcription of a full-length genomic L1 from the internal RNA polymerase II sense strand promoter located within its 5'UTR (Figure 1.4). The L1 mRNA is then exported from the nucleus to the cytoplasm, where its translation results in the production of ORF1p and ORF2p. Within the cytoplasm, ORF1p and ORF2p bind to their encoding mRNA by a process termed *cis*-preference (Esnault et al., 2000; Wei et al., 2001). The ORF1p/ORF2p/mRNA complex forms a ribonucleoprotein particle (RNP), which is necessary for L1 retrotransposition (Esnault et al., 2000; Hohjoh and Singer, 1996; Kulpa and Moran, 2005, 2006; Martin, 1991; Wei et al., 2001).

The resultant L1 RNP gains access to the nucleus, by a mechanism that does not strictly require nuclear envelope breakdown (Kubo et al., 2006), where TPRT

occurs. During TPRT, the ORF2p L1 EN activity makes a single-strand endonuclease nick at a double-stranded DNA target sequence in genomic DNA (Cost and Boeke, 1998; Cost et al., 2001; Feng et al., 1996). This endonucleolytic nick is typically made at an L1 EN degenerate consensus cleavage site [(5'-TTTT/A-3', where '/' represents the location of the scissile phosphate)] (Gilbert et al., 2002; Morrish et al., 2002; Symer et al., 2002; Szak et al., 2002). The L1 EN endonucleolytic nick exposes a free 3' hydroxyl group that can be used as a primer by the ORF2p L1 RT activity to reverse transcribe the associated L1 RNA (Cost et al., 2002; Feng et al., 1996; Kulpa and Moran, 2006; Luan et al., 1993). Through a mechanism that is still incompletely understood, second-strand DNA cleavage typically occurs downstream of the initial single-strand endonucleolytic nick, exposing a 3'-OH group that is likely used as a primer by the L1 RT DNA-dependent DNA polymerase activity to synthesize second-strand L1 cDNA (Christensen and Eickbush, 2005). The completion of L1 integration, which likely requires host factors (*e.g.*, DNA ligase), leads to a newly inserted L1 that is generally flanked by variable-length TSDs (~7-20 bp) or, on fewer occasions, small target site deletions (Gilbert et al., 2005; Gilbert et al., 2002; Symer et al., 2002). If the new L1 insertion is full-length, it can continue to amplify, thereby increasing the repertoire of active L1 copies in the genome.

LINE-1 TPRT Associated Rearrangements

LINE-1 mediated retrotransposition events are occasionally accompanied by intra-L1 rearrangements (*e.g.*, 5' truncations and 5' truncations associated with inversions/deletion events) or larger genomic structural rearrangements. In an anthropomorphic sense, L1 would 'desire' to generate full-length copies that can undergo subsequent rounds of retrotransposition. Consistent with this idea, *in vitro* studies suggest that the L1 reverse transcriptase (Piskareva and Schmatchenko, 2006), as well as the reverse transcriptase encoded by the *Bombyx mori* R2 retrotransposon (Bibillo and Eickbush, 2002), are efficient enzymes that are more processive than those encoded by retroviruses. Thus, it would be reasonable to hypothesize that the host would evolve pathways to 'restrict' unabated L1 retrotransposition. An extension of this

logic suggests the host may have evolved factors that recognize DNA structures generated during TPRT to combat retrotransposition.

It is hypothesized that DNA structures associated with TPRT intermediates (*e.g.*, a single-strand nick, a 3' flap structure, and/or a double-strand break) are recognized as substrates by host DNA repair proteins. For example, 5' truncated L1s are proposed to occur via a process termed abortive retrotransposition (Gilbert et al., 2005), where the L1 RT becomes dissociated from the L1 cDNA during TPRT. How this dissociation occurs requires further study; however, it is notable that the over-expression of proteins involved in the nucleotide excision repair (*e.g.*, ERCC1/XPF) inhibits the retrotransposition of engineered L1s in cultured human cells (Gasior et al., 2008; Servant et al., 2017). Similarly, the APOBEC3A protein can deaminate cytidine residues, which converts them to uracil residues, in transiently exposed single-strand L1 cDNAs that arise during TPRT (Richardson et al., 2014); repair processes that remove the uracil residues from single-strand DNA and the completion of L1 integration may lead to 5' truncation. Finally, the ataxia telangiectasia mutated (ATM) protein is hypothesized to recognize lesions (perhaps double-strand DNA breaks) generated during TPRT to inhibit L1 retrotransposition (Coufal et al., 2011). Given this data, it is intriguing to speculate that other DNA damage sensing pathways, such as damage recognized by the ataxia telangiectasia mutated-related (ATR) protein also play a role in the inhibition of L1 integration in the genome. Indeed, the initial steps in L1 integration likely represent a battleground between L1 and its host to limit retrotransposition.

A variation of TPRT, termed "twin-priming," can lead to the formation of L1 inversion/deletion structures (Ostertag and Kazazian, 2001). During twin priming, it is proposed that sequences in the L1 poly(A) tail and an internal segment of L1 RNA anneal to short (~4-6 bp) single-strand sequences in target site DNA, leading to the formation of a double-strand break. The 3'-OH groups present at the 3' termini of the of the resultant RNA/DNA hybrids then serves as primers for the L1 RT RNA-dependent DNA polymerase activity to mediate the convergent synthesis of two L1 cDNAs from different regions of the L1 RNA template (presumably through a template switching mechanism). The resolution of the convergent L1 cDNAs can occur through two distinct

mechanisms. Most often, microhomology-mediated annealing between the cDNAs, followed by the completion of second-strand synthesis by either a L1 RT DNA-dependent DNA polymerase activity or a host-encoded DNA polymerase then leads to the formation of inversion/deletion structures. Less frequently, one of the L1 cDNAs can use non-allelic L1 copies located on the same chromosome as repair templates, by a process termed synthesis dependent strand annealing (SDSA), to generate more complex genomic rearrangements (Beck et al., 2011; Gilbert et al., 2005).

TPRT can also lead to small or larger structural rearrangements in genomic DNA. For example, if second-strand cleavage is directly opposed to the initial endonucleolytic nick, TPRT can lead to a 'blunt' insertion that does not contain target site alterations. By comparison, if second-strand cleavage occurs upstream of the initial endonucleolytic break, the completion of TPRT can lead to the generation of small deletions at the target site (Gilbert et al., 2002).

TPRT can also occasionally lead to large genomic deletions. In one model, large (> 10 kb) deletions can be generated if the L1 cDNA invades a double-strand break upstream of the initial integration site (Gilbert et al., 2005; Gilbert et al., 2002). In a second model, termed single-strand annealing, the newly synthesized first-strand L1 cDNA can undergo non-allelic homologous recombination (NEHJ) with an L1 located upstream of the insertion site. The resolution of TPRT intermediates can result in the generation of a 'chimeric' L1 and the concomitant deletion of DNA (typically on the order of hundreds of base pairs to ~3.1 kb) that resides between the newly integrated L1 cDNA and the genomic L1. Finally, although far less frequent, newly integrated L1 cDNAs may be able to undergo inter-chromosomal recombination events, leading to the formation of translocations (Gilbert et al., 2005). Thus, structural variants generated during TPRT can lead to the formation of large structural variants in the genome. Genomic deletions generated during TPRT can lead to various human diseases. For example, a 46 kb L1 insertion/deletion event into the *PHDX* gene has resulted in a case of pyruvate dehydrogenase deficiency (Mine et al., 2007), and several L1 insertion/deletion deletions into *NF1*, have led to sporadic cases of neurofibromatosis (Kazazian and Moran, 2017; Vogt et al., 2014).

L1-Mediated Transductions

On occasion, L1s are able to mobilize genomic DNA sequences flanking their 3', and 5' ends to new genomic locations by a process termed L1-mediated transduction. L1-mediated 3' transductions occur when RNA polymerase II bypasses the 'weak' L1 polyadenylation site present in the L1 3'UTR, and instead uses a 'stronger' polyadenylation site in flanking genomic DNA to generate a full-length L1 mRNA (Holmes et al., 1994; Moran et al., 1999; Moran et al., 1996). Retrotransposition of the resultant mRNA can lead to the mobilization of 3' flanking genomic sequence, resulting in 3' transductions (Goodier et al., 2000; Pickeral et al., 2000). Approximately 20% of human-specific L1s harbor 3' transduction events, and these sequences have been used to infer progenitor offspring relationships among L1Hs subfamily members (Beck et al., 2010; Goodier et al., 2000; Holmes et al., 1994; Kidd et al., 2010; Macfarlane et al., 2013; Pickeral et al., 2000; Tubio et al., 2014a).

L1-mediated 5' transductions occur when RNA polymerase II initiates transcription from genomic sequences that reside upstream of the L1 5'UTR. Retrotransposition of the resultant RNA can then lead to a full-length L1 copy, which contains additional genomic sequence at its 5' end. Although L1-mediated 5' transductions are quite rare, they have been identified in the HGWD (Lander et al., 2001), cultured cells (Symer et al., 2002; Wei et al., 2001), a mutagenic mouse L1 insertion (Chen et al., 2006), and a somatic L1 insertion in the human brain (Evrony et al., 2012).

Endonuclease Independent LINE-1 Retrotransposition

While endonuclease activity is generally required for L1 retrotransposition, under certain cellular environments L1s lacking endonuclease activity are capable of mobilizing throughout the genome by an alternative integration pathway termed endonuclease-independent (ENi) L1 retrotransposition. Engineered L1s containing missense mutations in the L1 EN domain can undergo ENi retrotransposition in Chinese hamster ovary (CHO) cells that are deficient in components of the non-homologous end-joining (NHEJ) pathway of DNA double-stranded break repair (Morrish et al., 2007;

Morrish et al., 2002). ENi retrotransposition events are distinct from canonical TPRT-mediated L1 insertions in that they are frequently both 5' and 3' truncated, do not occur in typical L1 endonuclease cleavage sites, generally lack TSDs, and often are associated with genomic deletions at the integration site (Morrish et al., 2002). Additionally, some ENi retrotransposition events contain the addition of short cDNA fragments at their L1/genomic DNA junctions, which appear to be derived from the reverse transcription of cellular mRNAs (Morrish et al., 2007; Morrish et al., 2002). Indeed, in DNA protein kinase catalytic subunit-deficient (DNA-PKcs) CHO cells, ENi retrotransposition events can occur at dysfunctional telomeres (Morrish et al., 2007). In this cellular environment, chromosome ends likely provide the free 3' OH required for initiating priming of reverse transcription. Thus, endonuclease-deficient L1s can use genomic lesions to initiate TPRT (Morrish et al., 2002).

Although the vast majority of L1 retrotransposition proceeds via TPRT, putative ENi retrotransposition events have been identified in the human genome (Sen et al., 2007; Srikanta et al., 2009). Moreover, an ENi retrotransposition event into *EYA1*, which is accompanied by a ~17kb genomic DNA deletion, is responsible for a sporadic case of human oto-renal syndrome (Morisada et al., 2010). It remains to be determined whether mutations in double-strand break repair pathways other than NHEJ may also provide permissible means for ENi retrotransposition.

Post-integration Genomic Rearrangements

Non-allelic homologous recombination (NAHR) events between genomic L1s can also lead to genomic structural variation. NAHR between genomic L1s, as well as between genomic Alus has been observed in several sporadic cases of disease and has been shown to play a role in the generation of CNVs in the genome (Burwinkel and Kilimann, 1998; Kazazian and Moran, 2017; Lehrman et al., 1985; Segal et al., 1999; Startek et al., 2015; Temtamy et al., 2008). For example, NAHR between genomic L1s has led to sporadic cases of phosphorylase kinase deficiency, Alport syndrome, and Ellis-van Creveld syndrome (Burwinkel and Kilimann, 1998; Segal et al., 1999; Temtamy et al., 2008).

L1 as a Mutagen

The first notion that L1 is still actively moving within the human genome was highlighted by the discovery of two independent mutagenic L1 insertions found within exon 14 of the *Factor VIII* gene, which caused hemophilia A in two unrelated males (Kazazian et al., 1988). A few years later, an L1 insertion within the *Dystrophin* gene was found to be the cause of a case of muscular dystrophy (Holmes et al., 1994). Since that time, ~130 mutagenic L1-mediated retrotransposition events (which include L1, Alu, SVA, and to a lesser extent processed pseudogenes insertions) have been reported in various diseases (Hancks and Kazazian, 2016; Kazazian and Moran, 2017). Indeed, a comprehensive study characterizing mutations in the *NF1* gene suggested that L1-mediated retrotransposition events are responsible for approximately 1 in 250 disease-producing mutations in humans (Wimmer et al., 2011).

L1-mediated retrotransposition events are estimated to occur, at a minimum, in 1 of 20 meioses for Alu, 1 of 20 to 200 meioses for LINE-1, and 1 of 900 meioses for SVA (Cordaux and Batzer, 2009; Kazazian and Moran, 2017). L1-mediated retrotransposition events can cause mutations by a variety of mechanisms. For example, insertions into exons can disrupt the coding potential of a gene, whereas insertions into introns can result in exon skipping or mis-splicing, leading to the generation of hypomorphic or null expression alleles (Awano et al., 2010; Holmes et al., 1994; Kagawa et al., 2015; Kazazian and Moran, 1998; Kondo-Iida et al., 1999; Musova et al., 2006; Narita et al., 1993; Rodriguez-Martin et al., 2016). The L1 antisense promoter also can generate transcripts that can affect gene expression (Speek, 2001). Similarly, *in* vitro experiments indicate that full-length L1 retrotransposition events into introns can either introduce premature polyadenylation sites or RNA polymerase II transcriptional pause sites into genes, thereby disrupting gene expression (Han et al., 2004; Perepelitsa-Belancio and Deininger, 2003).

Overview: Polymorphic and Somatic L1 Insertions

Comparative genomic analyses between the human genome reference sequence and the draft chimpanzee genome led to the identification of ~11,000

species-specific transposable elements, including 5,530 Alu, 1,174 L1 and 864 SVA elements specific to humans (Mills et al., 2006). Several sequencing studies have focused on capturing L1Hs specific integration events (*e.g.*, Sheen et al., 2000; Badge et al. 2003; Beck et al., 2010). In sum, these studies have identified L1s that are polymorphic with respect to presence/absence in the human population. Indeed, the genomes of any two unrelated individuals contain ~300 polymorphic L1s (Ewing and Kazazian, 2010). Thus, an individual genome only provides a 'snapshot' of L1 diversity in the population, leading to an under-representation of polymorphic, and active, L1Hs sequences in the current human genome draft sequence (Beck et al., 2010) (Table 1.2). Notably, recent studies have also found, as originally predicted by Barbara McClintock's studies, that active L1s retrotranspose during early development and in a subset of somatic cells, leading to the presence of L1 insertions that are present in some somatic lineages, but not others (*i.e.*, creating somatic mosaicism) (Table 1.3). In this section, I focus on methods to discover polymorphic and somatic L1 integration events and then discuss the developmental timing of such L1 retrotransposition events.

Approaches to Identify Polymorphic and Somatic LINE-1 Insertions

The diagnostic 'ACA' tri-nucleotide and downstream 'G' in the 3'UTR L1Hs sequences (Badge et al., 2003; Ewing and Kazazian, 2010; Huang et al., 2010; Iskow et al., 2010; Ovchinnikov et al., 2001; Rodic et al., 2015; Sheen et al., 2000; Solyom et al., 2012) have been exploited using PCR-based approaches to identify polymorphic L1Hs sequences. For example, Sheen *et al.* used targeted PCR approaches to identify six L1Hs elements, which were polymorphic with respect to presence/absence in the human population. Similarly, a derivation of classical transposon-based display methods [amplification typing of L1 active subfamilies (ATLAS)], which involves the generation of libraries containing short sequences of genomic DNA (typically 100-1,000 bp in length) containing artificial DNA linkers at their termini in conjunction with suppression PCR methodology, led to the isolation of eight full-length polymorphic L1Hs sequences. Three of these (3 out of 8) were highly active in cultured cell retrotransposition assays (Badge et al., 2003). Thus, ATLAS provided a means to capture active, polymorphic L1Hs sequences that were absent from the human genome draft sequence.

Variations of the general PCR-based approach used in ATLAS, which use the targeted capture of L1Hs sequences (Baillie et al., 2011; Shukla et al., 2013), microarray based approaches to identify L1Hs sequences (Cardelli et al., 2012; Huang et al., 2010) and the identification of size differences in individual fosmid DNA libraries (Beck et al., 2010; Kidd et al., 2008; Tuzun et al., 2005), have been used in conjunction with high-throughput sequencing and sophisticated bioinformatics analysis pipelines to capture both polymorphic germline and somatic L1 integration events. Indeed, the ever-increasing availability of individual human DNA sequences (*e.g.*, such as those from the 1000 Genomes Project) has also allowed the identification of polymorphic and somatic L1Hs sequences using post-sequencing bioinformatics tools that allow the identification of L1-mediated insertions in short-read whole-genome sequencing data (Figures 1.5 and 1.6) (Genomes Project et al., 2015; Helman et al., 2014; Lee, 2012; Stewart et al., 2011; Tubio et al., 2014a). An extensive list of techniques used to identify polymorphic and somatic L1Hs integration events is presented in Tables 1.2, 1.3, and 1.4.

Advantages and disadvantages exist for targeted capture/PCR-based and whole genome-based approaches. In general, PCR-based capture approaches allow the identification and 'calling' of L1Hs integration sites from relatively small amounts of genomic DNA input. However, chimeric artifacts arising during the PCR process can lead to a high false discovery rate, complicating downstream bioinformatics analysis. Moreover, short-read DNA sequencing technologies can complicate efforts of mapping genomic DNA that flank L1Hs sequences to unique regions of the human genome, especially in areas of the genome replete with repetitive DNA sequences. Similarly, microarray chip-based approaches are limited by probe design/coverage and often have difficulty in identifying L1Hs sequences in repetitive areas of the genome. Finally, fosmid-based approaches, while representing the 'gold standard,' are costly and effort intensive, which limits throughput, and are unable to detect severely 5' truncated L1 integration events. By comparison, whole-genome approaches offer the advantage of leveraging the plethora of DNA sequence data available to the scientific community. However, low sequence coverage, combined with the abovementioned difficulty in assembling short read DNA sequences containing repetitive DNA are notable technical

hurdles that hamper the identification of polymorphic (and somatic) L1Hs sequences in the genome.

The advent of whole genome single-cell DNA amplification (WGA) followed by targeted capture/and high-throughput DNA sequencing offers promise to detect both low frequency and/or 'private' L1Hs germline polymorphism and somatic L1Hs insertions within individual cells. Indeed, these approaches have been used to confirm that L1Hs elements can retrotranspose in somatic cells (Erwin et al., 2016; Evrony et al., 2012; Evrony et al., 2015; Upton et al., 2015). However, WGA is still in its infancy; different WGA approaches lead to different efficiencies in genome amplification, and often result in uneven genomic DNA amplification and allelic drop-out in repetitive regions of the genome. This can further lead to chimeric artifacts that complicate downstream analyses, and may not faithfully amplify and/or map genomic regions rich in repetitive DNA, as they are often filtered out in downstream bioinformatics pipelines. The advantages and disadvantages aside, a combination of the above methodologies, as well as classical genetic and molecular biological approaches, have led to important insights about when and where L1s retrotranspose during development.

Endogenous LINE-1 Retrotransposition

Evidence for L1 Mobilization in Germline and During Early Embryogenesis

The first evidence that retrotransposition can occur during early human embryonic development involved the discovery of a 5' truncated L1 containing a 3' transduction in exon 4 of the X-linked *CYBB* gene of a male chronic granulomatous disease patient (Brouha et al., 2002). The L1 3' transduction allowed the identification of the progenitor source element, *LRE3*, on chromosome 2q24.1. Interestingly, these data suggested that RNA derived from the maternal *LRE3* allele retrotransposed into the *CYBB* gene in a primary oocyte prior to the onset of maternal meiosis II, and that independent assortment during meiosis II resulted in the inheritance of the *de novo* insertion, but not the progenitor *LRE3* allele, in the patient.

Initial studies in transgenic mice further indicated that L1 can retrotranspose during gamete generation. Evidence from a male transgenic mouse containing an

engineered human L1 containing an acrosin signal peptide/EGFP retrotransposition indicator cassette whose expression was driven by a pre-proacrosin promoter indicated that *de novo* L1 retrotransposition events were present in 2 of 135 offspring. Thus, these events likely occurred in the male germline prior to the onset of meiosis II at a low frequency (Athanikar et al., 2002; Ostertag et al., 2002).

An emerging body of evidence suggests that L1s may also retrotranspose in the female germline. For example, genetic and molecular biological experiments revealed that mutagenic *de novo* L1 retrotransposition that led to hemophilia A in a male patient likely occurred in his mother's germline (Kazazian and Moran, 1998; Richardson et al., 2017). Similarly, studies using engineered human L1s suggest that L1 can retrotranspose in female oocytes (Georgiou et al., 2009), whereas L1 expression and or retrotransposition may lead to oocyte attrition in mice (Malki et al., 2014). Finally, whole-genome DNA sequencing of representative mice in two and three generation pedigrees uncovered eleven *de novo* full-length L1 insertions (Richardson et al., 2017). Subsequent PCR-based validation experiments from parental mice and their offspring on genomic DNA derived from reproductive organs, as well as tissues representing the three distinct primary germ layers, revealed that the *de novo* L1 insertions occurred in primordial germ cells, later in germline development, and in the early embryo (Richardson et al., 2017). Thus, it is clear that L1 can retrotranspose in both the male and female germlines.

Post-zygotic Insertions

In addition to the Richardson *et al*. (2017) study mentioned above, it is now apparent that L1 retrotransposition events can occur after fertilization during early zygotic development. For example, the characterization of a full-length L1 insertion into the *CHM* gene, which led to the X-linked recessive disorder choroideremia in a male patient, provided the initial evidence that L1 retrotransposition could occur during early embryogenesis (van den Hurk et al., 2007). This mutagenic insertion contained two 3'-transductions events, which allowed the identification of the progenitor L1 in the patient's mother. Intriguingly, the patient and his sisters shared the same X-chromosome haplotype, but only one sister inherited the mutagenic L1 insertion,

indicating that the mother was a germline mosaic. Subsequent experiments, using the mother's lymphoblast DNA, revealed that she also was a somatic mosaic with respect to this mutagenic L1 insertion. Thus, this mutagenic L1 insertion must have occurred early during embryogenesis in the mother, prior to the segregation of the germline and somatic lineages.

Additional support for L1 retrotransposition in the early embryo comes from cell culture studies in human embryonic stem cells (hESCs), a model of early human embryonic development. Approximately 20% of expressed L1Hs elements in cultured hESCs are retrotransposition competent and both ORF1p and endogenous L1 ribonucleoprotein particles are highly expressed in hESCs (Garcia-Perez et al., 2007; Macia et al., 2011). Moreover, engineered L1 constructs tagged with a retrotransposition indicator cassette can mobilize in cultured hESCs, although at lower levels when compare to L1 mobilization in HeLa cells (Garcia-Perez et al., 2007; Moran et al., 1996; Wei et al., 2000). Thus, these data provide orthogonal pieces of experimental evidence to support the hypothesis that heritable retrotransposition events can occur in the early embryo.

Somatic L1 Retrotransposition Events: Cancers

Early studies revealed that cell lines derived from epithelial tumors exhibited higher levels of L1 RNA and ORF1p expression than primary cell types. Moreover, the increase in L1 expression correlated with the methylation status of the L1 5'UTR—CpG residues within the L1 5'UTR tended to exhibit more hypomethylation in tumor cell lines than primary cells (Alves et al., 1996; Doucet-O'Hare et al., 2016; Hata and Sakaki, 1997; Nur et al., 1988; Thayer et al., 1993; Woodcock et al., 1996; Yoder et al., 1997).

The realization that L1 retrotransposition events may play a role in tumor initiation and/or progression was realized upon the discovery of a mutagenic L1 insertion that disrupted the *APC* tumor suppressor gene in a colon tumor (Miki, 1992). Notably, the mutagenic insertion was not present in the surrounding normal colonic tissue, suggesting that it may have been a 'driver' mutation responsible for the initiation of tumorigenesis. Indeed, a recent study identified a 5' truncated somatic L1 insertion

located 388bp upstream of the aforementioned L1 insertion in the *APC* gene (Scott et al., 2016). This mutagenic L1 insertion disrupted the open reading frame of the *APC* tumor suppressor gene (Miki, 1992; Scott et al., 2016). The other *APC* allele contained a stop codon, which most likely led to its inactivation (Scott et al., 2016). Thus, these mutations likely represent two 'driver' mutations required for APC inactivation and initiation of tumorigenesis.

During the past 17 years, a number of PCR-based targeted capture sequencing and whole-genome sequencing approaches have identified somatic L1 insertions in several epithelial tumor cancer types (Baillie et al., 2011; Ewing and Kazazian, 2010; Helman et al., 2014; Iskow et al., 2010; Lee, 2012; Rodic et al., 2015; Tubio et al., 2014a), and have been reviewed in Scott *et al.* (2017). A number of emerging themes have resulted from these studies, which are summarized below.

First, it is clear that some L1 insertions occur into exons of known tumor suppressor genes (*e.g.*, *PTEN*) (Helman et al., 2014) or into the untranslated regions of candidate genes that have been implicated as putative tumor suppressor genes in cancers (*e.g.*, *ROBO2*, *CDH12*, and *CDH11*) (Lee, 2012; Solyom et al., 2012). Second, a proof-of-principle experiment revealed that a mutagenic L1 insertion disrupted a transcriptional repressor site in the *ST18* gene (Shukla et al., 2013). Since *ST18* is frequently amplified in hepatocellular carcinomas, these data suggest that the L1 insertion may result in ST18 over-expression and oncogenic activation. Thus, it is possible that some L1 insertions may serve as driver mutations in cancers.

Second, the rate of L1 retrotransposition seems to vary greatly between these different tumor types (Doucet-O'Hare et al., 2015; Lee, 2012; Rodic et al., 2015; Tubio et al., 2014a). For example, individuals with Barrett's esophagus contain on average five L1 insertions per person, while individuals with esophageal adenocarcinoma observe a rate almost five times that of 23.5 L1 insertions per person (Doucet-O'Hare et al., 2015). Indeed, one study identified 102 candidate L1 insertions in a colorectal tumor (Lee, 2012). However, no somatic insertions have been discovered to date within brain (Carreira et al., 2016; Helman et al., 2014; Iskow et al., 2010) or blood (Helman et al.,

2014; Lee et al., 2012) based tumors. Thus, it seems that the L1 mutagenic load varies within and between cancers.

Third, and consistent with what has been learned from germline L1 insertions (Beck et al., 2010; Holmes et al., 1994; Kidd et al., 2008; Moran et al., 1999; Moran et al., 1996), approximately 24% of somatic retrotransposition events in tumors contain 3' transductions, and these transduced sequences have been used to identify highly active progenitor L1s in cancers (Macfarlane et al., 2013; Tubio et al., 2014a). For example, a comprehensive study found that two highly active L1Hs elements account for more than a third of all somatic transductions identified in several tumor sample types, and that 95% of all the identified 3' transductions could be attributed to 72 germline L1Hs loci (Tubio et al., 2014b). Since that time, elegant methods have been developed to exploit DNA sequences between L1Hs elements, instead of transductions, to infer progenitor/progeny L1 relationships and to map RNA-seq reads to individual L1 loci (Scott et al., 2016). Clearly, L1 retrotransposition occurs in several cancers, and on occasion, can result in 'driver' mutations that affect cell growth. However, given the evidence at hand, it seems likely that the bulk of the resultant insertions represent 'passenger' mutations that have little contribution to the tumor phenotype (Kazazian and Moran, 2017; Tang et al., 2017).

The Role of Transposable Elements During Neurogenesis

Epithelial tumor cells are not the only cells harboring somatic L1 retrotransposition events. Indeed, recent studies have demonstrated that L1s can undergo somatic retrotransposition in neuronal progenitor cells, and leads to somatic mosaicism in the brain that contributes to intra-individual genetic variation. Since neurons are among the longest-lived cells in the body, it has been hypothesized that somatic L1 retrotransposition events may contribute to individual phenotypic differences and disease susceptibilities. Below, I highlight some of the seminal findings in this area of research.

Discovery of L1 Retrotransposition in Neurons

In 2005, Muotri *et al.* discovered that an engineered human L1 containing an EGFP retrotransposition indicator cassette could retrotranspose in adult rat neural progenitor cells (NPCs) *in vitro* and in the brains of transgenic mice *in vivo*, leading to EGFP-positive cells within the brain. Notably, EGFP-positive cells co-localized with neuronal marker proteins, but not oligodendrocyte or astrocyte markers, indicating that L1 retrotransposition occurred in neuronal progenitor cells (NPCs) rather than in a common precursor cell early during early embryogenesis. The characterized L1 integration events revealed that 5/17 (29%) integrated within neuronally-expressed genes. Intriguingly, one antisense L1 insertion within the 5'UTR of *Psd-93* (also referred to as *DLG2*), led to its overexpression, which promoted the differentiation of precursor cells to neuronal fates. In agreement with previous studies, EGFP-positive cells were also detected in germ cells (ovary and testes) of transgenic mice (Ostertag et al., 2002). Taken together, these data indicate that a engineered human L1 can retrotranspose in the mouse brain and, at least in one case, the insertion could influence the expression of the resultant gene, leading to a measurable cell-based phenotype.

Subsequent studies revealed that engineered L1s can retrotranspose at low levels in human fetal brain stem cells, at high levels (*e.g.*, in up to 20% of cells) in human embryonic stem cell (hESC)-derived NPCs when compared to fetal NPCs (Coufal et al., 2009), and at similarly high levels in a human embryonic carcinoma cell line, PA-1, which shares attributes with NPCs (Garcia-Perez et al., 2010). Thus, human NPCs also can accommodate engineered L1 retrotransposition *in vitro*—sometimes at surprisingly high levels.

Similar to studies in cancer cells, bisulfite conversion analyses on human adult genomic DNA revealed that the L1 5'UTR exhibits significantly less methylation in brain samples when compared to matched skin samples (Coufal et al., 2009). Moreover, sensitive quantitative PCR assays revealed that L1 DNA content is reproducibly elevated in certain regions of the brain (containing ~80 additional copies of ORF2 DNA per cell) when compared to heart and liver DNA from the same individuals. These data suggested that endogenous L1s could possibly contribute to intra-neuronal genetic

diversity; however, it remained possible that other cellular processes, besides retrotransposition, were responsible for the increase in ORF2 copy number (Coufal et al., 2009).

Discovery of Somatic Retrotransposition Events in the Brain

Bulk Tissue Experiments

A targeted-capture PCR-based approach performed in conjunction with next generation DNA sequencing [*i.e.*, retrotransposon capture sequencing (RC-seq)] led to the identification of 7,743 putative somatic L1 insertions, 13,692 somatic Alu insertions, and 1,350 SVA insertions that were present in bulk DNA samples derived from the hippocampus and caudate nucleus of three donors, which were absent from their blood DNA (Baillie et al., 2011). The characterization of 14 L1 integration events revealed typical L1 structural hallmarks, suggesting that at least some represent *bona fide* instances of *de novo* endogenous L1 retrotransposition. The authors further claimed that ~3.4% of insertions were within coding exons, whereas ~43.3% were within the introns of annotated genes. Gene ontology analyses further suggested an enrichment of L1 retrotransposition events within genes involved in neurogenesis in synaptic function. Although a groundbreaking study that established L1-mediated retrotransposition events could occur within the brain, some have suggested that many (or perhaps the majority) of the calls represent false-positives. If so, the conclusions of many downstream analyses (*e.g.*, understanding the proportion of insertion into genes, *etc.*) may require refinements.

Single Cell Experiments

A number of laboratories have utilized whole genome amplification (WGA) approaches to amplify DNA from single neurons. The WGA products are then subject to targeted capture and/or whole genome sequencing to detect somatic L1 retrotransposition events in neurons. For example, Evrony et al., isolated single neurons from postmortem and surgically resected human brains and used WGA to analyze 300 single neurons from the cerebral cortex and caudate nucleus of three neurologically normal individuals (Evrony et al., 2012). L1Hs sequences were then identified in the

WGA samples using L1 insertion profiling (L1-IP), an approach previously used by Ewing and Kazazian to identify polymorphic L1Hs sequences in 15 individual genomes. Intriguingly, the authors identified a full-length somatic insertion within intron 4 of *ICQH*. This insertion was identified in two cortical neurons, but not single caudate neurons, and was present at a low level in an unsorted 50,000 nuclei, as well as in bulk brain DNA. The observed low-level mosaicism of this somatic insertion, and its detection only in cortical neurons suggests the insertion occurred during cortical development.

Subsequent studies, which used droplet digital PCR to analyze DNA samples from 32 different brain regions for two of the somatic L1 retrotransposition events identified above, revealed that one L1 insertion most likely occurred in a neocortical progenitor of the left middle frontal gyrus, giving rise to mostly neurons (Evrony et al., 2015). The second L1 insertion seems to have mobilized considerably earlier during nervous system development in a progenitor for both neurons and glia. Taken together, the above studies confirmed that L1 retrotransposition occurs, albeit at a low level (~0.04-0.6 unique somatic L1 insertions occur per neuron), in the human brain, and indicated that L1 retrotransposition may occur during different times of brain development.

Faulkner and colleagues have also adapted RC-seq to identify somatic L1 insertions in single hippocampal and cortical neurons and glial cells (Upton et al., 2015). In contrast to the above studies, they identified over 2,000 putative somatic L1Hs insertions in at least one hippocampal neuron, resulting in a predicted rate of 13.7 somatic L1 insertions per hippocampal neuron. Four of these insertions were identified in both neurons and glial cells, suggesting that L1 retrotransposition events can arise in proliferating neural stem cells prior to glial or neuronal commitment, but that glia cells generally support less L1 retrotransposition than neurons. Intriguingly, once again, Faulkner and colleagues found that 1.2% of L1 insertions were within exons, whereas ~36.8% were within introns. They also reported an 1.8-fold enrichment for L1 insertions into highly transcribed neuronal genes and active enhancer elements in neuronal stem cells. However, some have criticized this study, claiming that chimeric artifacts arising during WGA have led to large number of false-positive calls (Evrony et al, 2016).

Finally, Gage and colleagues used a targeted single-cell sequencing approach and machine learning-based analysis to identify somatic L1-associated variants (SLAVs) on single nuclei derived from the frontal cortex and hippocampus of three healthy individuals (Erwin et al., 2016). A SLAV is defined as a somatic L1Hs insertion, or a deletion (as large as 792kb) in the genome believed to be due to a post-integration recombination or homology-mediated event. The authors identified 46 putative SLAVs. Interestingly, an endogenous SLAV was discovered in *DLG2*, in hippocampal progenitor cells. This was the same gene in which an engineered human L1 retrotransposition event was discovered in rat NPCs previously (Muotri et al., 2005). As above, SLAVs are generated during a variety of neural development stages, including in an early progenitor cell that contributes to the development of both the hippocampus and frontal cortex. Together, these data led the authors to suggest that SLAVs occur at a rate of ~0.58 to 1 events per neuron and/or glial cell and are predicted to affect ~44% to 63% of cells in the healthy brain.

Discrepancies in Rate of L1 Integration in the Brain

In general, the field now agrees that somatic L1-mediated retrotransposition events within the brain contribute to intra-individual genetic variation. However, the discrepancies among the above studies have sparked a spirited debate regarding the 'true' rate of somatic L1 retrotransposition in the brain. Indeed, it is unlikely that one group is fully correct, while another is completely wrong. Thus, how can one account for these observed differences in values?

Some of the discrepancies could be attributed to the difference in cell types examined in the above studies (hippocampal versus cortical neurons). It remains possible that different types of neurons accommodate L1 retrotransposition at different efficiencies. Moreover, the stringency used in bioinformatics calling pipelines could contribute to variability. For example, while Faulkner and colleagues have been criticized for having a high-false positive rate in their putative call sets, it also is possible that high false-negative rates, which are harder to detect, contribute to the lower estimates in the Evrony and Erwin studies. Finally, technical issues that arise using different WGA methodologies in single cell-based studies likely lead to differences in

genome coverage and allelic dropout (which are generally assessed using single copy sequences), which would confound downstream analyses. Moreover, time will tell whether WGA actually provides an accurate depiction of the repetitive DNA content of the genome—a significant problem that has not satisfactorily been addressed in the literature. Indeed, given the above caveats, the field should critically assess whether WGA single cell-based methods provide a suitable means to make quantitative conclusions about L1, and repetitive DNA content, in the human genome. Finally, given these controversies, the field should be wary about downstream claims regarding the integration preferences of L1s in cells. Indeed, a high false positive rate (as well as exaggerated claims made on 'call sets' containing a small number of events) could confound, and perhaps skew, results of downstream analyses.

Integration Preferences of Transposable Elements

Transposable element (TE) insertions are mutagenic. From a teleological standpoint, the unabated mobility may adversely affect genome integrity, thereby harming the host. As such, it is not surprising that many TEs have evolved strategies to insert into specific genomic regions, thereby minimizing damage to their respective host genomes. Below, I describe convergent evolutionary strategies used by different TEs to 'minimize' damage to their host genomes. Sultana *et al.* (2017) recently published a similar review on integration site selection by eukaryotic retroviruses and transposable elements, while Levin and Moran (2011) also addressed several of the topics presented here. I then briefly describe what is known about L1 integration preferences and address the main question in this dissertation: Do L1 retrotransposition events display integration preferences within the human genome?

Integration into Ribosomal DNA

Work on the *Bombyx mori* R2 non-LTR retrotransposon has been instrumental in generating the TPRT model of retrotransposition. The R2 protein, like L1 ORF2p, has both EN and RT activities, but the EN belongs to the type-II RE family of proteins (Yang et al., 1999). The R2 EN cleaves a specific sequence within the 28S ribosomal RNA (rRNA) gene to mediate its integration (Eickbush, 2002). Interestingly, insect R1

retrotransposons encode an APE-type EN that also cleaves a specific sequence within the 28S rRNA gene, which is located 74 bp downstream of the R2 cleavage site, to mediate their integration (Xiong and Eickbush, 1988). These findings demonstrate non-LTR retrotransposons possessing different types of EN domains can integrate into similar genomic compartments. Indeed, the redundant nature of rDNA genes in an organism may provide a safe haven for retrotransposon insertions (*i.e.*, knocking out a small number of rDNA genes may not have a dramatic effect on rRNA synthesis). Consistent with these thoughts, it is notable that R1 and R2 occupy anywhere from a few percent to over half of the rDNA units of most insects (Jakubczak et al., 1991).

Integration into transfer RNAs

The examination of whole genome sequences and subsequent experimental work indicates that some *Saccharomyces cerevisiae* LTR-retrotransposons (*i.e.*, Ty1 and Ty3) preferentially integrate upstream of RNA polymerase III-transcribed genes, such as tRNAs, which are generally located in gene-poor environments (Bushman, 2003; Devine and Boeke, 1996; Lesage and Todeschini, 2005; Sandmeyer, 2003). For example, Ty3 integration occurs in a position-specific manner inserting in either orientation 16 to 19 nucleotides upstream of tRNA genes (Chalker and Sandmeyer, 1990). Interactions between Ty3 integration complexes and components of the RNA III polymerase complex mediate precise integration at the tRNA transcription start site (Sandmeyer et al., 2015). By comparison, Ty1 integration occurs within a nucleosome bound ~700 bp window upstream of tRNA and other RNA polymerase III genes at a periodicity of ~80 bp (Devine and Boeke, 1996). Indeed, the finding that highly transcribed tRNAs are preferential targets suggest that interactions between the Ty1 integration complex and RNA polymerase III, which appear to differ from those used by Ty3, are important components that influence Ty1 target site specificity.

Transfer RNA genes are also targeted by non-LTR retrotransposons. For example, *Dictyostelium* repeat element (DRE or TRE5-A), an APE-type non-LTR retrotransposon found in *Dictyostelium discoideum* (a soil-living amoeba), occurs ~48 nucleotides upstream and in the opposite orientation of tRNA genes (Marschalek et al., 1992; Spaller et al., 2017). The stable tRNA gene transcription complex, TFIIIB, is

postulated to play a role in the position-specific integration of DRE upstream of various tRNA genes (Chung et al., 2007; Marschalek et al., 1992). By comparison, a related element, Tdd-3, appears to integrate in a position-dependent manner downstream ~100 bp downstream from tRNA genes (Szafranski et al., 1999). Thus, different types of retrotransposons (LTR and non-LTR) have developed convergent mechanisms to target integration near tRNA genes.

Telomeres as Favorable Substrates for Integration

Terminal repeat sequences at chromosomal ends, referred to as telomeres, shorten with each cell division, since lagging strand DNA synthesis cannot fully copy the end of a chromosome (Fujiwara et al., 2005; Greider and Blackburn, 2004). Most eukaryotic telomeres consist of short, five to six base pairs, repeated nucleotides sequences called telomeric repeats. The lengths of telomeric repeats are maintained by a specialized RNP, termed telomerase, which uses a reverse transcriptase activity to reiteratively copy a small sequence in its associated template RNA onto chromosome ends after the completion of DNA replication, thereby maintaining chromosomal stability (Blackburn, 1991; Lundblad and Wright, 1996). When telomeric repeat lengths become shorter than a critical threshold, cells can cease to undergo cellular division.

Intriguingly, a variety of retrotransposons either target telomeres or have been co-opted by the host for telomere maintenance. For example, *Drosophila melanogaster* does encode telomerase and their chromosome ends do not contain canonical telomeric repeats. Instead, *Drosophila* relies upon the following three non-LTR retrotransposons to maintain telomeres: the non-autonomous HeT-A element, and the autonomous TART, and TAHRE (Telomere-associated and *HeT-A*-related element) elements. These elements are primarily found in an array at Drosophila telomeres. TART and TAHRE encode a reverse transcriptase that uses the retrotransposon RNAs as templates to mediate end replication at chromosome ends (Biessmann et al., 1990; Traverse and Pardue, 1988). TART and TAHRE also contain an AP-like EN domain; however, how or if this activity functions in telomere maintenance requires elucidation (Arkhipova, I.R. et al., 2017.

Similar to *Drosophila*, the silkworm *Bombyx mori* has two telomere-specific APE-type non-LTR retrotransposons, telomeric-repeat-associated sequence 1 (TRAS1) and SART1, which is 'TRAS' inverted, as it is transcribed and found inserted in a reverse orientation relative to TRAS1 (Takahashi, et al., 1997; Takahashi and Fujiwara, 1999). However, unlike the situation observed for *Drosophila*, the ends of *Bombyx mori* chromosomes contain a telomeric repeat sequence of $(TTAGG)_n$. Thus, it appears that TRAS1 and SART1 preferentially integrate into a 6-8 kb stretch of the $(TTAGG)_n$ telomeric repeats.

Although telomerase activity has not been observed at any developmental stage of *B. mori* somatic (fat bodies, silk glands) or germ-line cells (testes), a putative telomerase reverse transcriptase (TERT) has been identified within its genome, termed *Bmo*TERT (Fujiwara et al., 2005). *Bmo*TERT contains various structural differences when compared to conventional eukaryotic TERT genes (*e.g.* it lacks introns, has lost an N-terminal domain, and contains five putative AUG codons in its 5' UTR), which is thought to result in the repression of *Bmo*TERT expression and enzymatic activity. By comparison, genomic copies of both TRAS1 and SART1 appear to be conserved among telomere regions in a wide variety of Lepidopteran insects (butterflies and moths), are generally full-length, and undergo abundant transcription (Kubo et al., 2001; Mandrioli, 2002; Mita et al., 2004; Okazaki et al., 1995; Takahashi and Fujiwara, 1999; Takahashi et al., 1997). Thus, TRAS1 and SART1 may play an important role in telomere maintenance.

The green algae *Chlorella* genome harbors multiple copies of an 8.9kb full-length non-LTR retrotransposon, *Zepp* (Higashiyama et al., 1997). Although the majority of *Zepp* retrotransposons are distributed at non-telomeric regions of different chromosomes, two chromosomes contain a cluster of *Zepp* sequences near or at the telomere region. These finding suggest that Zepp elements may either target subtelomeric regions for integration and/or possibly play a role in telomere protection (Higashiyama et al., 1997).

Some Penelope-like (PLE) retrotransposons encode a reverse transcriptase domain, but lack an overt endonuclease domain. Interestingly, a subset of PLEs

comprise telomeric ends in bdelloid rotifers, basidiomycete fungi, stramenopiles, and plants (Gladyshev and Arkhipova, 2007). Since most PLEs also carry short stretches of telomeric repeats at or near their 3' ends, they appear to use a 3'OH group at G-rich telomeric termini to reverse transcribe their RNA onto chromosome ends (Gladyshev and Arkhipova, 2011). This mechanism is akin to endonuclease-independent retrotransposition of L1 EN mutants which can use dysfunctional telomeres as integration substrates (Curcio and Belfort, 2007; Morrish et al., 2007).

Telomeric targeting is not unique to non-LTR retrotransposons. For example, approximately 90% of the *Saccharomyces cerevisiae* Ty5 LTR-retrotransposon are preferentially located within silent heterochromatic—68% of Ty5 elements are found in sub-telomere regions, whereas 22% are found in the silent mating type loci (Zou et al., 1996). Ty5 integration specificity has been mapped to a nine amino acid targeting domain that binds to Sir4, a structural component of heterochromatin, to target integration (Levin and Moran, 2011). Intriguingly, under conditions of stress, such as nitrogen deprivation, the failure to phosphorylate an amino acid (Ser_{1095}) within the Ty5 integrase protein, allows Ty5 to integrate into expressed, gene-rich genomic regions (Dai et al., 2007). Thus, the integration preferences of certain transposable elements are not static and may be altered by environmental stress—a result that harkens back to McClintock's 'genome shock' hypothesis (McClintock 1950).

Replication Targeted for Integration

Other TEs appear to take advantage of components involved in DNA replication to target integration. For example, original studies suggested that *Drosophila* P elements preferentially insert near a gene promoter; however more recent studies revealed that they preferentially integrate into origin recognition complex protein-binding sites (ORCs) (Spradling et al., 2011). Indeed, promoters associated with ORC binding sites exhibit 29-fold enriched for P element insertions than those lacking ORC binding sites (16-fold enrichment for P elements). Thus, P elements may be able to effectively increase their copy number in *Drosophila* by inserting into ORC binding sites prior to DNA replication (Spradling et al., 2011).

IS*608* and IS*Dra*2 (types of HUH DNA transposons) exhibit a strong bias for integration into the lagging-strand template (Ton-Hoang et al., 2010). Moreover, other genomic features, such as stalled DNA replication forks containing lagging-strand templates, are a 'hot spot' for IS*608* integration, raising the possibility that access to lagging-strand templates may increase when replication forks stall.

The targeting protein, TnsE, encoded by the *E.coli* DNA transposon *Tn7* recognizes and targets components found on the lagging strand template during replication. The preferred binding substrate of TnsE is a structure that contains a 3' recessed end (Peters and Craig, 2001), a structure that is abundant during DNA replication. Subsequent investigations revealed that the sliding-clamp processivity factor is required for TnsE-mediated transposition (Parks et al., 2009). Sliding-clamp proteins are deployed with each new priming cycle on the lagging strand template, and residual sliding clamps are believed to be important for recruiting proteins involved in maturation of the lagging strand template, including RNA primer removal, ligation of Okazaki fragments, and mismatch repair (Lopez de Saro and O'Donnell, 2001; Pluciennik et al., 2009). Interestingly, biochemical and genetic studies uncovered a putative clamp-interacting motif within TnsE that was important for interactions with the sliding clamp. Mutations in TnsE that disrupted the TnsE-sliding clamp interaction eliminated or severely reduced transposition. Intriguingly, TnsE appears to specifically target Tn7 transposition into lagging strands, where DNA replication forks tend to stall (Peters and Craig, 2000, 2001). Another possibility is that Tn7 targets the double-strand break repair intermediates created by DNA repair proteins working to resolve stalled replication forks (Peters and Craig, 2000; Shi et al., 2008).

Certain retrotransposons also appear to target DNA replication intermediates as integration substrates. For example, blocking the endonuclease activity of the L1.LtrB group II intron still allowed ENi retrohoming into leading strand templates of DNA replication forks, suggesting 3'OH groups present at the nascent leading DNA strand can serve as a primer for reverse transcription of the reverse spliced intron RNA. Likewise, the mobilization of L1.LtrB via the EN-independent mechanism of retrotransposition exhibit an integration preference into DNA replication forks,

specifically into the lagging strand templates (Ichiyanagi et al., 2003; Ichiyanagi et al., 2002). These findings suggest a mechanism in which the intron reverse splices into single-stranded DNA at a replication fork, followed by priming from a nascent lagging DNA strand with a free 3-'OH to mediate cDNA synthesis of the reverse spliced RNA. Finally, the RmInt1 group II intron, which encodes an IEP lacking an EN domain also shows the same integration preference into the lagging strand template (Martinez-Abarca et al., 2004).

Integration Preference into RNA polymerase II Transcribed Genes

Some transposable elements preferentially target RNA polymerase II transcribed genes for integration. For example, the *Schizosaccharomyces pombe Tf1* LTR retrotransposon preferentially integrates into the promoters of RNA polymerase transcribed genes (Chatterjee et al., 2009; Guo and Levin, 2010; Leem et al., 2008; Majumdar et al., 2011). Approximately 95% of Tf1 integration sites generated *in vitro* occur upstream of ORFs and some of the most frequently targeted promoters are associated with stress responsive genes (Guo and Levin, 2010). Furthermore, other studies suggest that *Tf1* targeting requires Switch Activating Protein 1 (Sap1), a protein that is involved in DNA replication (de Lahondes et al., 2003) that is enriched at nucleosome-free regions (Tsankov et al., 2011), and the presence of an active replication fork barrier (Jacobs et al., 2015).

Several retroviruses also display integration preferences for RNA polymerase II transcribed genes. For example, human immunodeficiency virus-1 (HIV-1) and other lentiviruses display strong preferences for transcription units with a distinct bias towards highly expressed and intron-rich genes (Lesbats et al., 2016). Moloney murine leukemia virus (MLV) preferentially integrates near the start of transcriptional units (Wu et al., 2003), as well as at strong enhancers and active promoters (LaFave et al., 2014). This integration preference appears to benefit the retrovirus as it inserts itself into an area of the genome that is likely to be transcribed quite efficiently.

Targeted Integration into Microsatellite Repeats

Certain APE-type non-LTR retrotransposons can integrate within multi-copy microsatellite repeats. For example, Waldo has been shown to integrate near 5'-AC-3' repeats in *Drosophila melanogaster* (Busseau et al., 2001), whereas Waldo elements in *A. gambiae (mosquito)* and *F. scudderi (earwig)* specifically integrate into 5'-ACAY-3' (where Y can be a C or T) repeats (Kojima and Fujiwara, 2003). In addition, the *A. gambiae* APE-type non-LTR retrotransposon, MinoAg1, preferentially integrates into 5'-AC-3'repeats (Kojima and Fujiwara, 2003).

L1 Integration Preference

While it is clear that several DNA transposons and retrotransposons preferentially target certain areas of the genome for integration, it remains unclear whether L1s specifically integrate into certain regions of the genome. Below I briefly summarize what is known about L1 integration preferences.

Known L1 Accumulation Preferences

The release of the human genome reference sequence revealed that L1s are predominantly found in AT-rich genomic regions (Lander et al., 2001). Curiously, while Alus rely upon L1 ORF2p for retrotransposition (Dewannieux et al., 2003), older Alu subfamilies show a strong bias towards GC-rich DNA. In contrast, Alu elements from younger subfamilies (*e.g.*, Ya5 and Yb8) are found in AT-rich regions of the genome (Lander et al., 2001). Thus, it appears that post-integrative selection process played an important role in shaping L1 and Alu distributions in the human genome. As such, simply looking at where L1s and Alus accumulate in the genome over evolutionary time may not accurately reflect their initial insertion preferences.

L1s are also dramatically overrepresented on the X-chromosome (Lander et al., 2001; Smit, 1999). Whether this overrepresentation is due to the preferential targeting of L1s to the X-chromosome or suggests a function for L1s in X-chromosome biology remains a spirited debate. The most significant increase of L1s is on Xq13, which contains the X-inactivation center (Bailey et al., 2000). Interestingly, genomic loci that

escape X-chromosome inactivation are reduced in L1 content when compared with loci that are subject to X-inactivation. This data has led to the hypothesis that L1 sequences may function as *cis*-acting elements that propagate X-inactivation along the X chromosome (Bailey et al., 2000; Gartler and Riggs, 1983; Lyon, 1998; Riggs, 1990). However, others have suggested that the accumulation of L1s on the X-chromosome is primarily due to the fact that unlike autosomes, the X chromosome is incapable undergoing recombination with its homolog in male meiosis (Langley et al., 1988; Wichman et al., 1992).

L1s tend to be more prevalent in intergenic regions of the genome, though many occur within the introns of genes. As expected from purifying selections, L1 insertions are highly underrepresented in protein coding exons and generally do not occur near intron/exon boundaries (Medstrand et al., 2002; Zhang et al., 2011). Finally, the L1s that reside within genes exhibit a strong tendency to be oriented in the opposite transcriptional orientation of the resident gene (Smit, 1999; Zhang et al., 2011).

Regions of the Human Genome that Lack TEs

There are a number of 'transposon-free regions' that do not accumulate TE insertions in the human genome (Simons et al., 2006). In general, these regions are at least 5kb in length, are conserved between the human and mouse genomes, lack recognizable transposable element sequences, and are composed of non-protein coding DNAs that may have functions in gene regulation. Moreover, several 'transposon-free regions' overlap with ultra-conserved regions of the genome, which contain 100% sequence identity between human, mouse and rat genomes over a minimum of 200bp (Bejerano et al., 2004). Some of these 'transposon free regions' are clearly important for development. For example, homeobox gene clusters (*i.e.*, HOX genes) completely lack TE sequences (Lander et al., 2001). Whether 'transposon-free regions' are inaccessible for transposon integration or whether TE integration within these regions is catastrophic to the host and are therefore subject to strong negative selection remains an open question. I will further explore these possibilities in Chapter 2.

Evidence for Nonrandom Integration in the Human Genome

Anecdotal evidence from disease producing L1-mediated retrotransposition events has led to the idea that L1-mediated retrotransposition events might target particular genes. For example, two unrelated hemophilia B patients contained mutagenic Alu insertions at adjacent nucleotides in the *Factor IX* gene (Vidaud et al., 1993; Wulff et al., 2000). Similarly, two unrelated families with X-linked agammaglobulinemia contained an Alu or SVA retrotransposon insertions at the same nucleotide within a coding exon of the *BTK* gene (Conley, 2005). Finally, a comprehensive analysis of 18 L1 mediated *de novo* mutations within the *NF1* gene revealed that 12 occurred at unique positions in the gene (Wimmer et al. 2011). However, three sites within the gene contained two independent insertions. Thus, although ascertainment biases must be considered in the studies, it is possible that there are 'hot spots' for L1-mediated insertions in the genome.

Closing Remarks and Overview of Thesis

Whether L1 integrates into preferential sequences within the human genome requires elucidation. It is not unreasonable to assume that L1 displays integration preference since several non-LTR retrotransposons, which also contain an APE-type EN like L1, display sequence-specific integration preference. Thus, the following questions come to mind: (1) Does L1 integration preferentially occur into expressed genes or those open and accessible regions of the genome? (2) Does L1 target multi-copy sequences in the genome so as to minimize damage to the host? (3) Are there cellular mechanisms that L1 may target to aid in the process of retrotransposition? These are just a few of the questions I will address in the following chapters of my thesis.

In this thesis, I focus on examining several thousands of *de novo* engineered human L1 integration events, under several different conditions, to determine if L1 preferentially integrates within the human genome. The data presented in Chapter 2 suggests that the L1 EN is the primary determinant of L1 integration site preference in the human genome, allowing L1 to disperse throughout the genome. In Chapter 3, I

examine how disruption of two different DNA repair processes in the cell may influence L1 integration preference, and discuss how and when these DNA repair proteins may act during L1 retrotransposition. In Chapter 4, I discuss the results presented in Chapters 2 and 3, discuss remaining questions, and provide suggestions for future experimental directions. Appendices present slight modifications made to our capture technique to isolate *de novo* engineered Zebrafish L2 integration events (Appendix A), and polymorphic or somatic L1Hs integration events in the human genome (Appendix B).

Figure 1.1: Types of mobile elements in the human genome.

Types of transposable elements (TEs) in the human genome and an example structure of each is shown. DNA transposons mobilize by a DNA intermediate in a 'cut and paste' mechanism utilizing their coded enzymatic transposase activity. Transposons are flanked by inverted terminal repeats. Autonomous retrotransposons, which encode their own proteins, mobilize in a 'copy and paste' fashion utilizing an RNA intermediate. Long Terminal Repeat (LTR) retrotransposons, such as the human endogenous retrovirus (ERV), contain a protein with group-specific antigen (GAG), protease (Pro), and polymerase (Pol) activities and an envelope protein (Env). The most abundant, and currently active autonomous non-LTR retrotransposon in our genome is Long INterspersed Element-1 (LINE-1). LINE-1 contains an internal promoter within it's 5' untranslated region (UTR) and encodes and ORF1p with RNA binding properties and ORF2p with both endonuclease (EN) and reverse transcriptase (RT) activities. The element ends with a 3'UTR, followed by a poly (A) tail and is typically flanked by variable sized target sized duplications (TSDs). Non-autonomous retrotransposons must rely upon the L1 encoded proteins for mobilization. An example of a Short INterspersed Element (SINE), still active within our genome to date, is Alu. Alu is roughly 280bp containing an RNA polymerase III promoter with A and B components, followed by two 7SL monomers separated by an adenosine-rich (AR) segment. Alus and processed pseudogenes are also followed by a poly(A) tail and flanked by TSDs as they are mobilized by L1, which moves via target-site primed reverse transcription (TPRT).

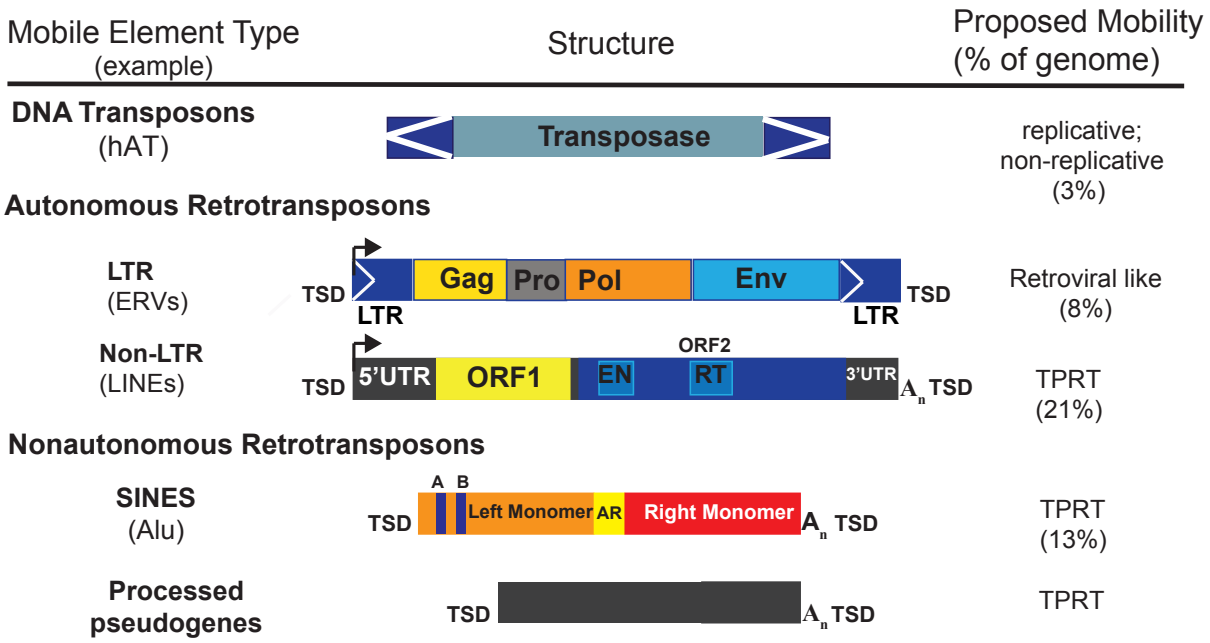


Figure 1.1: Types of mobile elements in the human genome.

Figure 1.2: Specific types of transposable elements.

A) *HUH Transposons:* HUH Transposons contain two ORFs, tpnA and tpnB. TpnA is required for mobilization as it contains the Histidine-hydrophobic-hystidine (HUH) transposase. The element is flanked by two subterminal palindromes; one on the left end (LE) and one on the right end (RE).

B) *Retroviral-like Retrotransposons:* The main difference between Ty1 versus Ty3 and Tf1 is the order of the integrase (IN), reverse transcriptase (RT), and RNaseH (RH) domains in the polymerase (pol) domain. The Ty5 element's group-specific antigen (gag) and pol domains are fused. There is an additional RNA-binding domain (RB) as well.

C) *Additional TPRT Elements:* Penelope-like elements (PLE) utilize TPRT for mobilization and contain flanking LTR-like sequences. Occasionally some PLEs contain an intron within the 5' LTR. PLEs encode a reverse transcriptase (RT) and a GIY-YIG endonuclease (EN). The mobile group II intron, L1.LtrB, is flanked by a 5' and 3' exon (E1 and E2, respectively) disrupted by intron RNA sequence. L1.LtrB encodes an intron-encoded protein (IEP) that contains RNA maturase that facilitates splicing, as well as reverse transcriptase and DNA endonuclease activities.

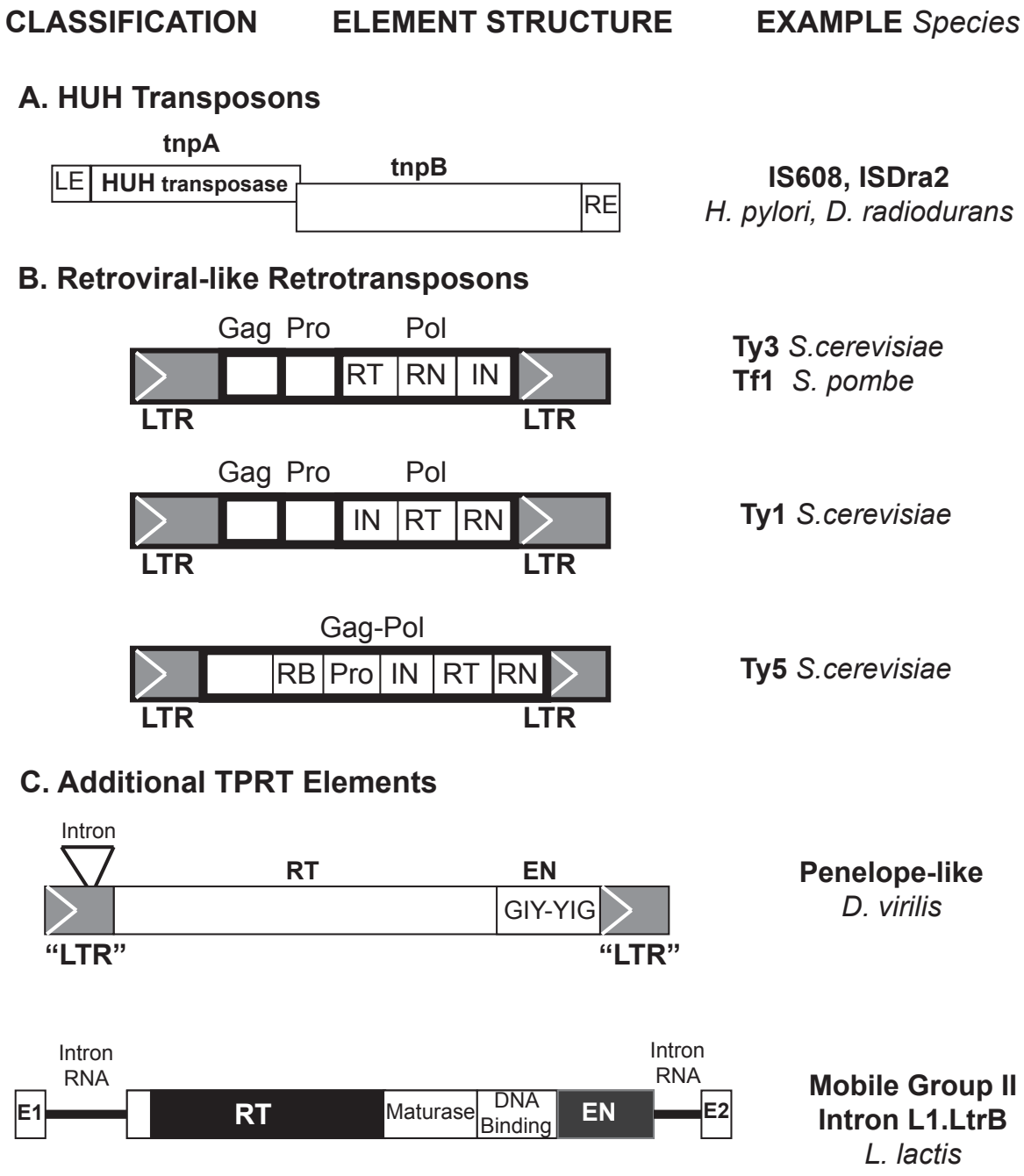


Figure 1.2: Specific types of transposable elements.

Figure 1.3: Specific types of non-LTR retrotransposons.

In total there are 11 clades of non-LTR retrotransposons phylogenetically divided by RT domain (CRE, R2, R4, L1, RTE, Tad1, R1, LOA, Jockey, CR1, and I). For brevity, examples of 4 of the 11 clades are shown, many of which have known target-site or position-specific integration preference. The R2 clade contains elements which encode a single ORF containing a site-specific restriction endonuclease-like (REL) domain that is located after the reverse transcriptase domain. All elements of the L1 clade contain two ORFs, where ORF2 encodes a apurinic/apyrimidinic (AP)-type endonuclease (APE) domain before the RT domain. The human specific L1 sequence is followed by a poly(A) tail. The TX1L element is flanked by terminal inverted repeat sequences and two types of tandem internal repeats (PTR-1 and PTR-2). DRE is flanked by non-symmetric LTRs containing a combination of three specific repeat sequences (A,B, and C) and is followed by a poly(A) tail. Members of the R1 clade differ by the presence or absence of a poly(A) tail. TRAS1 has acquired an additional RNaseH domain. The Jockey clade contains elements that have LTR-like domains similar to the gag and pol domains. HeT-A is a non-autonomous retrotransposon and lacks reverse transcriptase activity and thus must rely upon TART or TAHRE for mobilization in the genome.

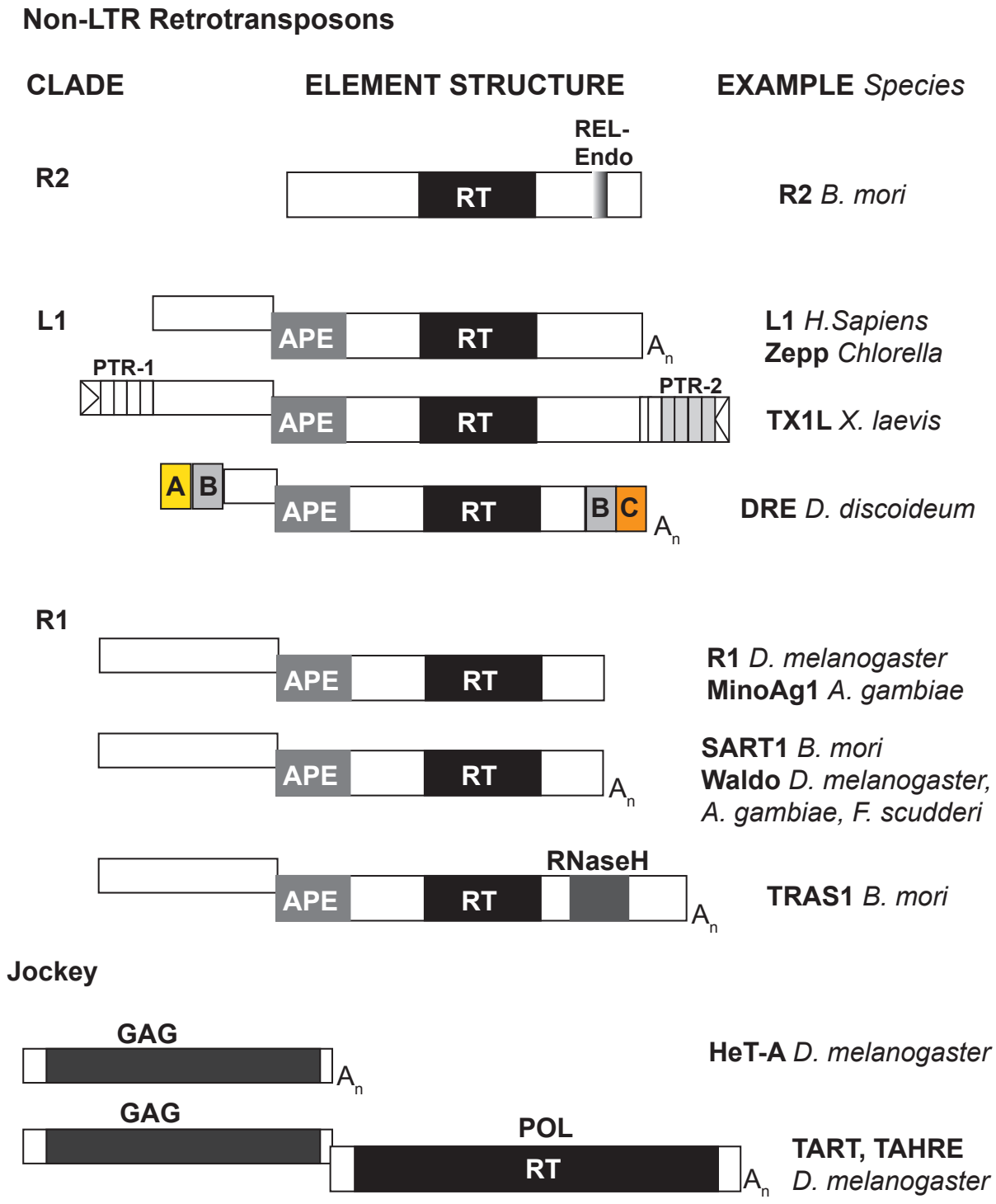


Figure 1.3: Specific types of non-LTR retrotransposons.

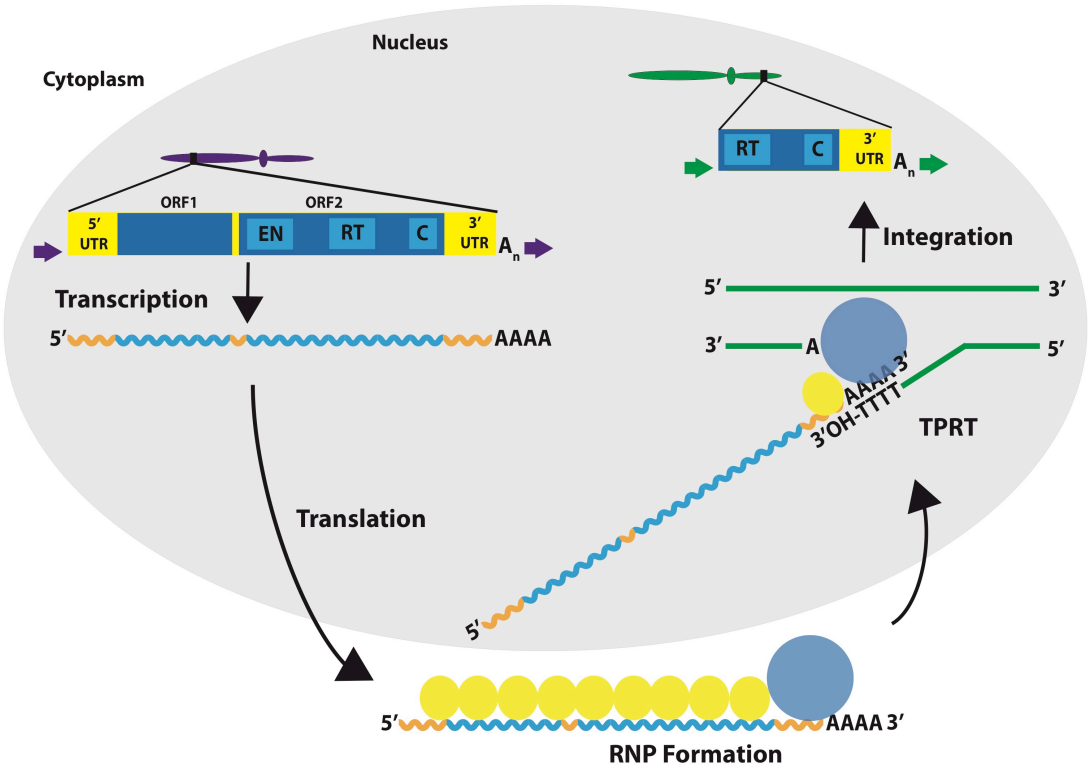


Figure 1.4: LINE-1 retrotransposition cycle.

Within the nucleus of the cell a full length LINE-1 sequence is transcribed. This mRNA is then exported to the cytoplasm where ORF1p and ORF2p are translated and bind back to their encoded transcript in '*cis*-preference', forming a ribonucleoprotein particle (RNP). The RNP then returns to the nucleus where the EN activity of ORF2p creates a single-strand endonucleolytic nick at a double-stranded site in the human genome (5'-TTTT/A). This cut creates a free 3'-OH which can then be used as a primer for reverse transcription of the L1 mRNA in a process termed target-site primed reverse transcription (TPRT). After cDNA synthesis, second-strand cleavage and second-strand synthesis, now at a new chromosomal location (green ellipses) there is a new integration site in the genome. This new integration site contains TPRT characteristics; 5' truncated, contains a poly(A) tail, and flanked by target site duplications (green arrows).

Figure 1.5: A general outline of L1Hs capture techniques.

L1Hs capture techniques involve (1) collection of gDNA from at least a blood source as a control, and then depending upon the study, from a tumor sample, brain sample, heart and or liver sample. (2) Genomic DNA is then randomly sheared by mechanical means or digested with restriction enzymes to smaller fragment sizes. (3) Known adapter sequences of varying designs are then ligated on to processed DNA ends. (4) L1Hs specific sequences are amplified with at least one L1Hs specific sequence containing the 'ACA' trinucleotide in the 3'UTR sequence, and a primer specific to the ligated adapter sequence in order to successfully amplify the 3' end of L1Hs sequences and their 3' flanking gDNA. A second round of PCR utilizing primers specific to the 'G' in the L1Hs 3'UTR may also be performed. (5) Amplified 3' L1Hs/3' flanking gDNA products are then identified in the genome by Southern blotting, microarray, or DNA sequencing methods.

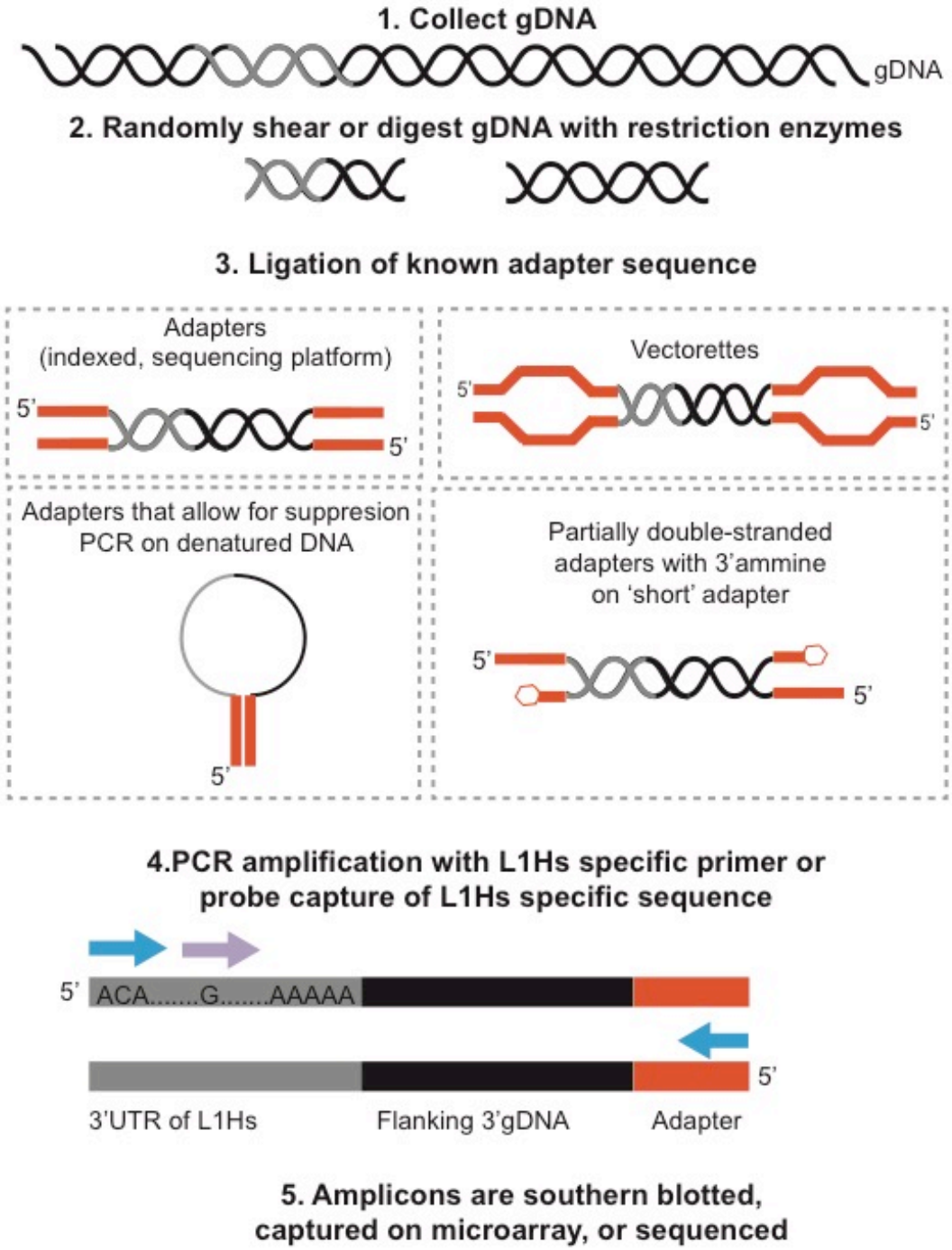


Figure 1.5: A general outline of L1Hs capture techniques.

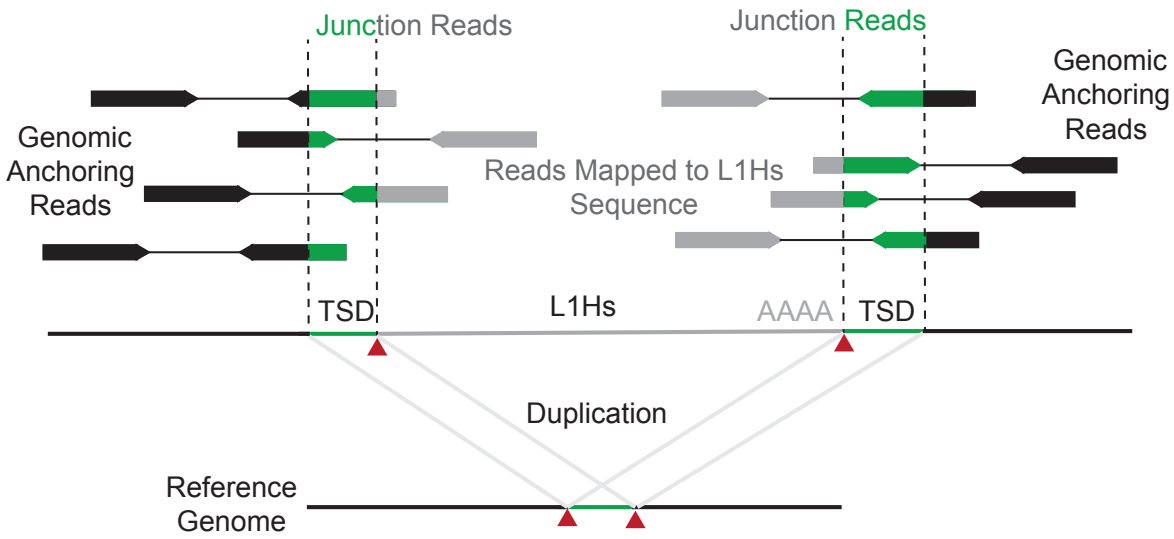


Figure 1.6: Paired-end sequencing reads analysis to identify L1Hs insertions.

This image is adapted and modified from Lee *et al.* 2015. This figure displays the basic rationale of whole genome sequencing algorithms identifying *de novo* L1Hs polymorphic or somatic insertions not present in the Reference Genome. Paired-end sequencing reads (long arrows pointing towards each other) in which one read aligns to the genome (Genomic Anchoring Reads in black) and the other read aligns to L1Hs sequences (Reads Mapped to L1Hs Sequence in grey) help identify the location of the *de novo* L1Hs sequence in the genome. Junction reads (reads containing grey and green and/or black) are those reads that contain part of the L1Hs sequence and the exact 'junction' sequence where the L1 sequence ends and the genomic flanking sequence begins. Junction reads can be used to determine L1Hs integration sites to single nucleotide base-pair resolution. Reads in green represent the target site duplication sequence.

Table 1.1: Retrotransposable elements and their preferential integration sites.

TE Type	Name	Species	Target Site	Target Sequence	Reference
Sister Clade of TERT	PLE	*Drosophila virilis*	45% of 'hot spots' within or near breakpoints of inversion	A + T rich	Evgen'ev, MB et al. 2005 Pyatkov, K et al. 2004
non-LTR, R2 clade	R2Bm	*Bombyx mori*	28S rDNA	5′–TCTCTTAAGGTAGCCAAA 3′–AGAGAATTCCATCGGTTT	Xiong, Y et al. 1998
non-LTR, L1 clade	DRE	*Dictoystelium discoideum*	47- 53 bp upstream of tRNA genes	X	Marschalek, R et al. 1992
	Tx1L	*Xenopus laevis*	Internal tandem repeat of Tx1D elements	5′–GTAACTTCAGCTAATGAAAAATCAACACATTGACTGCATTT 3′–CATTGAAGTCGATTACTTTTTAGTTGTGTAACTGACGTAAA	Garrett, J et al. 1986 Christensen, S et al. 2000
non-LTR, R1 clade	R1	*Drosophila Melanogaster, Bombyx mori*	28S rDNA	5′–CTGTCCCTATCTACTA 3′–GACAGGGATAGATGAT	Xiong, Y. et al. 1998
	MinoAg1	*Anopheles gambiae*	$(AC)_n$	5′–ACACACACACACACAC 3′–TGTGTGTGTGTGTGTG	Kojima, KK and Fujiwara H 2003
	SART1	*Bombyx mori*	telomere	5′–TTAGGTTAGGTTAGG 3′–AATCCAATCCAATCC	Takahashi, H et al. 1997
	Waldo	*Drosophila Melanogaster, Anopheles gambiae, Forficula scudderi*	$(ACAY)_n$	5′–ACAYACAYACAYACAY 3′–TGTRTGTRTGTRTGTR	Kojima, KK and Fujiwara H 2003
	TRAS1	*Bombyx mori*	telomere	5′–CCTAACCTAACCTAA 3′–GGATTGGATTGGATT	Okazaki, S et al. 1993 Okazaki, S et al. 1995
non-LTR, jockey clade	TART, HeT-A	*Drosophila*	telomere	X	Lewis, RW et al. 1993 Biessmann, H 1992
	Tahre	*Drosophila Melanogaster*	telomere	X	Abad, JP et al. 2004

Column 1: Transposable element clade. Column 2: Specific example and name of transposable element in designated clade in column 1. Column 3: The species in which the transposable element is found in the genome. Column 4: Description of the preferential target site of the given transposable element. Column 5: If applicable, the known specific targeted sequence in the genome. If not applicable an 'X' is shown. Column 6: Reference.

Table 1.2. L1-mediated capture techniques identifying polymorphic insertions.

Column 1: The published name of the technique. Column 2: Experimental details of the technique. Column 3: The number of polymorphic insertions identified by the given technique. Column 4: References (Note: Number of polymorphic insertions identified in this table reflect *de* novo reported polymorphic sites that are not found in dbRIP or by other documented studies at the time).

Table 1.2. L1-mediated capture techniques identifying polymorphic insertions.

Technique	Details	# Insertions	Reference
L1-Display	2 PCRs: 1. L1Hs specific 'ACA' and 10bp arbitrary primer 2. L1Hs nested 3'UTR primer 'G' and same 10bp primer; Products are Southern blotted	6 L1Hs in 6 males	Sheen, FM *et al.* 2000
		53 L1Hs in 91 individuals	Ovchinnikov, I *et al.* 2001
Amplification Typing of L1 Active Subfamilies (ATLAS)	Restriction digested gDNA; Suppression PCR with L1Hs specific primer and linker primer; Linear PCR with radiolabeled L1Hs specific sequence; Products are Southern blotted	18 L1Hs; 24 older L1s	Badge, RM *et al.* 2003
Fosmid-based, paired-end DNA sequencing	Fosmid end sequencing; PCR validation;Southern blotting; Sequencing to identify full-length L1Hs	68 full-length L1Hs	Beck, CR *et al.* 2010
L1-seq	Assymetric PCR with L1Hs specific 'ACA'; Hemi-specific PCR with L1Hs specific 'ACA' and 5-mer degenerate primer; L1Hs nested 3'UTR primer 'G'; paired-end Illumina sequencing	299 L1Hs in 25 individuals	Ewing, AD *et al.* 2010
Mobile Element -Scan (ME-Scan)	gDNA randomly sheared; indexed adapters for pooling; PCR Capture; paired-end Illumina sequencing	487 AluYb8/AluYb9 in 4 individuals	Witherspoon, DJ *et al.* 2010
		2524 AluYb8/AluYb9 in 169 individuals	Witherspoon, DJ *et al.* 2013
Transposon-seq	Restriction digested gDNA; Ligation of partially double-stranded linkers with 3' ammine; L1Hs or Alu specific amplification followed by nested PCR; ABI capillary sequencing or pyrosequencing	1145 Alu and L1Hs in 76 samples	Iskow, RC et al.2010
Transposon Insertion Profiling by microarray (TIP-chip)	Restriction digested gDNA ligated to vectorettes; PCR amplification by L1 or Alu specific primers; Amplicons are hybridized to microarray	125 L1Hs on X chromosome in 75 males	Huang, CR *et al.* 2010
		4 Alus in 10 individuals	Cardelli, M *et al.*2012
Transposable Element Analyzer (Tea)	Bioniformatics tool to analyze paired-end whole genome sequencing data	3521 polymorphic (L1s, Alus, SVAs, ERVKs, ERVL-MaLRs) in 44 individuals	Lee, E *et al.* 2012
Retrotransposon Capture Sequencing (RC-seq)	Hybridization to custom tiling array probes or liquid-phase sequence capture probes targeting 5' and 3' of active retrotransposons; paired-end Illumina sequencing	7644 L1Hs, Alu, SVA, and LTR-flanked in 19 individuals	Baillie, JK *et al.* 2011
			Shukla, R *et al.* 2013
TranspoSeq	Paired-end whole genome and exome sequencing bioinformatics tool	2704 L1Hs, Alu, SVA from 200 matched samples in lung squamous, head and neck, colorectal, and endometrial carcinomas	Helman, E *et al.* 2014

Table 1.3: L1-mediated capture techniques identifying somatic insertions in epithelial cancers.

Column 1: The name of the technique. Column 2: Experimental details of the technique. Column 3: The number of somatic insertions and in which tumor sample type identified by the technique. Column 4: References.

Table 1.3: L1-mediated capture techniques identifying somatic insertions in epithelial cancers.

Technique	Details	# Insertions	Reference
Transposon-seq	Restriction digested gDNA; Ligation of partially double-stranded linkers with 3' ammine; L1Hs or Alu specific amplification followed by nested PCR; ABI capillary sequencing or pyrosequencing	9 L1Hs in 6 lung tumors	Iskow, RC *et al.*2010
L1-seq	Assymetric PCR with L1Hs specific 'ACA'; Hemi-specific PCR with L1Hs specific 'ACA' and 5-mer degenerate primer; L1Hs nested 3'UTR primer 'G'; Paired-end Illumina Seq	69 L1Hs in 16 colorectal tumor	Solyom, S *et al*. 2012
		118 L1Hs in 10 individuals with Barrett's esophagus or esophageal adenocarcinoma	Doucet-O'Hare, TT *et al.* 2015
		104 L1Hs in 18 gastrointenstinal cancer patients	Ewing, AD *et al*. 2015
Transposable Element Analyzer (Tea)	Bioinformatics tool to analyze paired-end whole genome sequencing data	194 somatic insertions (183 L1Hs, 10 Alu, 1 ERV) in epithelial cancers (colorectal, prostate, ovarian) in 43 samples	Lee, E *et al.* 2012
TranspoSeq	Paired-end whole genome and exome sequencing bioinformatics tool	802 L1Hs, 7 Alu,1 SVA from 200 matched samples in lung squamous, head and neck, colorectal, and endometrial carcinomas	Helman, E *et al.* 2014
Retrotransposon Capture Sequencing (RC-seq)	Hybridization to custom tiling array probes or liquid-phase sequence capture probes targeting 5' and 3' of active retrotransposons; paired-end Illumina sequencing	12 L1Hs, 1SVA, and Alu events in 5 individuals with hepatocellular carcinoma	Baillie, JK *et al.* 2011 Shukla, R *et al.* 2013
Tranposome Finder in Cancer (TraFic)	Whole genome sequence pipeline identifying L1 events and transductions	2756 L1Hs and 93 Alu (24% of which are 3' transductions) from 290 samples containing 11 tumor types (Lung, head and neck, colon, prostate, breast, bone, bladder, glioma, melanoma, pancreas, renal)	Tubio, JM *et al.* 2014
Transposon insertion Profiling by Sequencing (TIP-seq)	Restriction digested gDNA; ligation to vectorette oligonucleotides; One sided PCR with L1Hs 'ACA' specific primer amplification; PCR amplicons are sheared and paired-end sequenced	465 L1Hs insertions in 20 Pancreatic ductal adenocarcinoma (PDAC)	Rodic, N *et al.* 2015
		36 L1 insertions in 7 type II ovarian carcinomas	Tang, Z *et al.* 2017

Table 1.4: L1-mediated capture techniques identifying somatic insertions in neurons.

Column 1: The name of the technique. Column 2: Experimental details of the technique. Column 3: The number of somatic events identified in the specified neuronal cells. Column 4: References.

Table 1.4: L1-mediated capture techniques identifying somatic insertions in neurons.

Technique	Details	# Insertions	Reference
Retrotransposon Cature Sequencing (RC-seq)	Hybridization to custom tiling array probes or liquid-phase sequence capture probes targeting 5' and 3' of active retrotransposons; paired-end Illumina sequencing	7,743 L1, 13,692 Alu, 1,350 SVA from bulk hippocampus and caudate nucleus of 3 individuals	Baillie, JK *et al.* 2011
L1Hs Insertion Profiling (L1-IP)	Whole genome amplify genomes of single neurons (MDA); PCR amplification with L1Hs specific-oligos ('ACA' and 'G'); Amplify 3' of L1Hs and 3' flanking gDNA	5 L1Hs identified in 50 single neurons from cerebral cortex and 50 from caudate nucleus from 3 neurologically normal individuals	Evrony, GD *et al* 2012
Single-cell Transposable Element Analyzer (*ScTea*)	Analysis of whole genome amplified single neurons (MDA) with modified version of Tea with high sensitivy and specificity to identify AluY, L1Hs, and SVA elements	2 L1Hs from 16 single-neuron genomes of neurologically normal individual (1 L1Hs limited to left middle frontal gyrus, 1 L1Hs distributed over entire left hemisphere)	Evrony, GD *et al* 2015
Single-cell Retrotransposon Capture Sequencing (RC-seq)	Whole genome amplify genomes of single neurons (MALBAC); Illumina library preparation; Hybridization capture of 5' and 3' ends of L1Hs with locked nucleic acid probes; paired-end sequencing	2,782 somatic L1Hs insertions from glia, and hippocampal and cortical neurons from four individuals	Upton,KR *et al.* 2015
Somatic L1-associated Variants (SLAV) -seq	Targeted MDA amplified single-cell ligation-mediated PCR with biotinylated L1Hs oligo; Illumina paired-end split-read identification of the 3' end of L1Hs and flanking gDNA;data-driven, machine-learning-based prediction	47 SLAVs; 89 single nuclei and bulk samples from frontal cortex and hippocampus of 3 healthy individuals	Erwin, JA *et al*. 2016

References

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

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

Arkhipova, I.R., Pyatkov, K.I., Meselson, M., and Evgen'ev, M.B. (2003). Retroelements containing introns in diverse invertebrate taxa. Nat Genet *33*, 123-124.

Arkhipova, I.R., Yushenova, I.A., Rodriguez, F. (2017). Giant reverse transcriptase-encoding transposable elements at telomeres. Mol Biol Evol *34*, 2245-2257.

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

Athanikar, J.N., Morrish, T.A., and Moran, J.V. (2002). Of man in mice. Nat Genet *32*, 562-563.
Awano, H., Malueka, R.G., Yagi, M., Okizuka, Y., Takeshima, Y., and Matsuo, M. (2010). Contemporary retrotransposition of a novel non-coding gene induces exon-skipping in dystrophin mRNA. J Hum Genet *55*, 785-790.

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

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

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M.*, et al.* (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Bannert, N., and Kurth, R. (2006). The evolutionary dynamics of human endogenous retroviral families. Annu Rev Genomics Hum Genet *7*, 149-173.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159-1170.

Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet *12*, 187-215.

Becker, K.G., Swergold, G.D., Ozato, K., and Thayer, R.E. (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. Hum Mol Genet *2*, 1697-1702.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

Belfort, M., and Perlman, P.S. (1995). Mechanisms of intron mobility. J Biol Chem *270*, 30237-30240.

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

Bibillo, A., and Eickbush, T.H. (2002). High processivity of the reverse transcriptase from a non-long terminal repeat retrotransposon. J Biol Chem *277*, 34836-34845.

Biessmann, H., Mason, J.M., Ferry, K., d'Hulst, M., Valgeirsdottir, K., Traverse, K.L., and Pardue, M.L. (1990). Addition of telomere-associated HeT DNA sequences "heals" broken chromosome ends in Drosophila. Cell *61*, 663-673.

Blackburn, E.H. (1991). Structure and function of telomeres. Nature *350*, 569-573.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491-500.

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

Bonitz, S.G., Coruzzi, G., Thalenfeld, B.E., Tzagoloff, A., and Macino, G. (1980). Assembly of the mitochondrial membrane system. Structure and nucleotide sequence of the gene coding for subunit 1 of yeast cytochrme oxidase. J Biol Chem *255*, 11927-11941.

Britten, R.J., and Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science *161*, 529-540.

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

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A *100*, 5280-5285.

Burton, F.H., Loeb, D.D., Voliva, C.F., Martin, S.L., Edgell, M.H., and Hutchison, C.A., 3rd (1986). Conservation throughout mammalia and extensive protein-encoding capacity of the highly repeated DNA long interspersed sequence one. J Mol Biol *187*, 291-304.

Burwinkel, B., and Kilimann, M.W. (1998). Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. J Mol Biol *277*, 513-517.

Bushman, F.D. (2003). Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. Cell *115*, 135-138.

Busseau, I., Berezikov, E., and Bucheton, A. (2001). Identification of Waldo-A and Waldo-B, two closely related non-LTR retrotransposons in Drosophila. Mol Biol Evol *18*, 196-205.

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. (2003). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Res *31*, 4385-4390.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y., and Sverdlov, E. (2002). A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. Genomics *80*, 402-406.

Callahan, K.E., Hickman, A.B., Jones, C.E., Ghirlando, R., and Furano, A.V. (2012). Polymerization and nucleic acid-binding properties of human L1 ORF1 protein. Nucleic Acids Res *40*, 813-827.

Calvi, B.R., Hong, T.J., Findley, S.D., and Gelbart, W.M. (1991). Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator, and Tam3. Cell *66*, 465-471.

Cardelli, M., Marchegiani, F., and Provinciali, M. (2012). Alu insertion profiling: array-based methods to detect Alu insertions in the human genome. Genomics *99*, 340-346.

Carreira, P.E., Ewing, A.D., Li, G., Schauer, S.N., Upton, K.R., Fagg, A.C., Morell, S., Kindlova, M., Gerdes, P., Richardson, S.R*., et al.* (2016). Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. Mob DNA *7*, 21.

Cary, L.C., Goebel, M., Corsaro, B.G., Wang, H.G., Rosen, E., and Fraser, M.J. (1989). Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. Virology *172*, 156-169.

Cervera, A., and De la Pena, M. (2014). Eukaryotic penelope-like retroelements encode hammerhead ribozyme motifs. Mol Biol Evol *31*, 2941-2947.

Chalker, D.L., and Sandmeyer, S.B. (1990). Transfer RNA genes are genomic targets for de Novo transposition of the yeast retrotransposon Ty3. Genetics *126*, 837-850.

Chatterjee, A.G., Leem, Y.E., Kelly, F.D., and Levin, H.L. (2009). The chromodomain of Tf1 integrase promotes binding to cDNA and mediates target site selection. J Virol *83*, 2675-2685.

Chen, J., Rattner, A., and Nathans, J. (2006). Effects of L1 retrotransposon insertion on transcript processing, localization and accumulation: lessons from the retinal degeneration 7 mouse and implications for the genomic ecology of L1 elements. Hum Mol Genet *15*, 2146-2156.

Christensen, S.M., and Eickbush, T.M. (2005). R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. Mol Cell Biol *25*, 6617-28.

Chung, T., Siol, O., Dingermann, T., and Winckler, T. (2007). Protein interactions involved in tRNA gene-specific integration of Dictyostelium discoideum non-long terminal repeat retrotransposon TRE5-A. Mol Cell Biol *27*, 8492-8501.

Conley, M.E. (2005). Two Independent Retrotransposon Insertions at the Same Site Within the Coding Region of BTK. Hum Mutat.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Cordaux, R., and Batzer, M.A. (2009). The impact of retrotransposons on human genome evolution. Nat Rev Genet *10*, 691-703.

Cost, G.J., and Boeke, J.D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry *37*, 18081-18093.

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

Cost, G.J., Golding, A., Schlissel, M.S., and Boeke, J.D. (2001). Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res *29*, 573-577.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Marchetto, M.C., Muotri, A.R., Mu, Y., Carson, C.T., Macia, A., Moran, J.V., and Gage, F.H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc Natl Acad Sci U S A *108*, 20382-20387.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127-1131.

Cousineau, B., Smith, D., Lawrence-Cavanagh, S., Mueller, J.E., Yang, J., Mills, D., Manias, D., Dunny, G., Lambowitz, A.M., and Belfort, M. (1998). Retrohoming of a bacterial group II intron: mobility via complete reverse splicing, independent of homologous DNA recombination. Cell *94*, 451-462.

Craig, N.L. (1996). Transposon Tn7. Curr Top Microbiol Immunol *204*, 27-48.
Craig, N.L., Chandler, M., Gellert, M., Lambowitz, A. Rice, P.A., and Sadmeyer, S. (2014). Mobile DNA III.

Curcio, M.J., and Belfort, M. (2007). The beginning of the end: links between ancient retroelements and modern telomerases. Proc Natl Acad Sci U S A *104*, 9107-9108.

Dai, J., Xie, W., Brady, T.L., Gao, J., and Voytas, D.F. (2007). Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin. Mol Cell *27*, 289-299.

Dalle Nogare, D.E., Clark, M.S., Elgar, G., Frame, I.G., and Poulter, R.T. (2002). Xena, a full-length basal retroelement from tetraodontid fish. Mol Biol Evol *19*, 247-255.

de Lahondes, R., Ribes, V., and Arcangioli, B. (2003). Fission yeast Sap1 protein is essential for chromosome stability. Eukaryot Cell *2*, 910-921.

Denli, A.M., Narvaiza, I., Kerman, B.E., Pena, M., Benner, C., Marchetto, M.C., Diedrich, J.K., Aslanian, A., Ma, J., Moresco, J.J.*, et al.* (2015). Primate-specific ORF0 contributes to retrotransposon-mediated diversity. Cell *163*, 583-593.

Devine, S.E., and Boeke, J.D. (1996). Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. Genes Dev *10*, 620-633.

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

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

Doucet, A.J., Hulme, A.E., Sahinovic, E., Kulpa, D.A., Moldovan, J.B., Kopera, H.C., Athanikar, J.N., Hasnaoui, M., Bucheton, A., Moran, J.V.*, et al.* (2010). Characterization of LINE-1 ribonucleoprotein particles. PLoS Genet *6*.

Doucet-O'Hare, T.T., Rodic, N., Sharma, R., Darbari, I., Abril, G., Choi, J.A., Young Ahn, J., Cheng, Y., Anders, R.A., Burns, K.H.*, et al.* (2015). LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc Natl Acad Sci U S A *112*, E4894-4900.

Doucet-O'Hare, T.T., Sharma, R., Rodic, N., Anders, R.A., Burns, K.H., and Kazazian, H.H., Jr. (2016). Somatically Acquired LINE-1 Insertions in Normal Esophagus Undergo Clonal Expansion in Esophageal Squamous Cell Carcinoma. Hum Mutat *37*, 942-954.

Eickbush, T.H. (2002). Mobile DNA II (Washington DC: ASM Press).

Ergun, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Stratling, W.H., and Schumann, G.G. (2004). Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. J Biol Chem *279*, 27753-27763.

Erwin, J.A., Paquola, A.C., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I., Butcher, C.R., Herdy, J.R.*, et al.* (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat Neurosci *19*, 1583-1591.

Esnault, C., and Levin, H.L. (2015). The Long Terminal Repeat Retrotransposons Tf1 and Tf2 of Schizosaccharomyces pombe. Microbiol Spectr *3*.

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

Evgen'ev, M., Zelentsova, H., Mnjoian, L., Poluectova, H., and Kidwell, M.G. (2000). Invasion of Drosophila virilis by the Penelope transposable element. Chromosoma *109*, 350-357.

Evgen'ev, M.B., and Arkhipova, I.R. (2005). Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet Genome Res *110*, 510-521.

Evgen'ev, M.B., Zelentsova, H., Shostak, N., Kozitsina, M., Barskyi, V., Lankenau, D.H., and Corces, V.G. (1997). Penelope, a new family of transposable elements and its possible role in hybrid dysgenesis in Drosophila virilis. Proc Natl Acad Sci U S A *94*, 196-201.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A.*, et al.* (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483-496.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J.*, et al.* (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49-59.

Evrony, G.D., Lee, E., Park, P.J., Walsh, C.A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. Elife *5*

Ewing, A.D., Gacita, A., Wood, L.D., Ma, F., Xing, D., Kim, M.S., Manda, S.S., Abril, G., Pereira, G., Makohon-Moore, A.*, et al.* (2015). Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res *25*, 1536-1545.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res *20*, 1262-1270.

Ewing, A.D., and Kazazian, H.H., Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res *21*, 985-990.

Fanning, T., and Singer, M. (1987). The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. Nucleic Acids Res *15*, 2251-2260.

Fedoroff, N., Wessler, S., and Shure, M. (1983). Isolation of the transposable maize controlling elements Ac and Ds. Cell *35*, 235-242.

Feldmar, S., and Kunze, R. (1991). The ORFa protein, the putative transposase of maize transposable element Ac, has a basic DNA binding domain. EMBO J *10*, 4003-4010.

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

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet *41*, 331-368.

Fujiwara, H., Osanai, M., Matsumoto, T., and Kojima, K.K. (2005). Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori. Chromosome Res *13*, 455-467.

Fukuyama, R., Niculaita, R., Ng, K.P., Obusez, E., Sanchez, J., Kalady, M., Aung, P.P., Casey, G., and Sizemore, N. (2008). Mutated in colorectal cancer, a putative tumor suppressor for serrated colorectal cancer, selectively represses beta-catenin-dependent transcription. Oncogene *27*, 6044-6055.

Garcia-Perez, J.L., Marchetto, M.C.N., Muotri, A.R., Coufal, N.G., Gage, F.H., O'Shea, K.S., and Moran, J.V. (2007). LINE-1 retrotransposition in human embryonic stem cells. Hum Mol Genet *16*, 1569-1577.

Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P.*, et al.* (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature *466*, 769-773.

Gartler, S.M., and Riggs, A.D. (1983). Mammalian X-chromosome inactivation. Annu Rev Genet *17*, 155-190.

Gasior, S.L., Preston, G., Hedges, D.J., Gilbert, N., Moran, J.V., and Deininger, P.L. (2007). Characterization of pre-insertion loci of de novo L1 insertions. Gene *390*, 190-198.

Gasior, S.L., Roy-Engel, A.M., and Deininger, P.L. (2008). ERCC1/XPF limits L1 retrotransposition. DNA Repair (Amst) *7*, 983-989.

Genomes Project, C., Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A.*, et al.* (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Georgiou, I., Noutsopoulos, D., Dimitriadou, E., Markopoulos, G., Apergi, A., Lazaros, L., Vaxevanoglou, T., Pantos, K., Syrrou, M., and Tzavaras, T. (2009). Retrotransposon RNA expression and evidence for retrotransposition events in human oocytes. Hum Mol Genet *18*, 1221-1228.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol *25*, 7780-7795.

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

Gladyshev, E.A., and Arkhipova, I.R. (2007). Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci U S A *104*, 9352-9357.

Gladyshev, E.A., and Arkhipova, I.R. (2011). A widespread class of reverse transcriptase-related cellular genes. Proc Natl Acad Sci U S A *108*, 20311-20316.

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

Goodman, M., Porter, C.A., Czelusniak, J., Page, S.L., Schneider, H., Shoshani, J., Gunnell, G., and Groves, C.P. (1998). Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. Mol Phylogenet Evol *9*, 585-598.

Gorbalenya, A.E. (1994). Self-splicing group I and group II introns encode homologous (putative) DNA endonucleases of a new family. Protein Sci *3*, 1117-1120.

Greider, C.W., and Blackburn, E.H. (2004). Tracking telomerase. Cell *116*, S83-86, 81 p following S86.

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

Grimaldi, G., and Singer, M.F. (1983). Members of the KpnI family of long interspersed repeated sequences join and interrupt alpha-satellite in the monkey genome. Nucl Acids Res *11*, 321-338.

Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I.B., Calderaro, J., Bioulac-Sage, P., Letexier, M., Degos, F*., et al.* (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet *44*, 694-698.

Guo, Y., and Levin, H.L. (2010). High-throughput sequencing of retrotransposon integration provides a saturated profile of target activity in Schizosaccharomyces pombe. Genome Res *20*, 239-248.

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

Hancks, D.C., Goodier, J.L., Mandal, P.K., Cheung, L.E., and Kazazian, H.H., Jr. (2011). Retrotransposistion of marked SVA elements by human L1s in cultured cells. Hum Mol Genet *20*, 3386-3400.

Hancks, D.C., and Kazazian, H.H., Jr. (2016). Roles for retrotransposon insertions in human disease. Mob DNA *7*, 9.

Hasnaoui, M., Doucet, A.J., Meziane, O., and Gilbert, N. (2009). Ancient repeat sequence derived from U6 snRNA in primate genomes. Gene *448*, 139-144.

Hata, K., and Sakaki, Y. (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene *189*, 227-234.

Hattori, M., Kuhara, S., Takenaka, O., and Sakaki, Y. (1986). L1 family of repetitive DNA sequences in primates may be derived from a sequence encoding a reverse transcriptase-related protein. Nature *321*, 625-628.

Helman, E., Lawrence, M.S., Stewart, C., Sougnez, C., Getz, G., and Meyerson, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Res *24*, 1053-1063.

Higashiyama, T., Noutoshi, Y., Fujie, M., and Yamada, T. (1997). Zepp, a LINE-like retrotransposon accumulated in the Chlorella telomeric region. EMBO J *16*, 3715-3723.

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

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

Holmes, S.E., Dombroski, B.A., Krebs, C.M., Boehm, C.D., and Kazazian, H.H., Jr. (1994). A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. Nat Genet *7*, 143-148.

Holmes, S.E., Singer, M.F., and Swergold, G.D. (1992). Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. J Biol Chem *267*, 19765-19768.

Houck, C.M., Rinehart, F.P., and Schmid, C.W. (1979). A ubiquitous family of repeated DNA sequences in the human genome. J Mol Biol *132*, 289-306.

Howell, R., and Usdin, K. (1997). The ability to form intrastrand tetraplexes is an evolutionarily conserved feature of the 3' end of L1 retrotransposons. Mol Biol Evol *14*, 144-155.

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T.*, et al.* (2010). Mobile interspersed repeats are major structural variants in the human genome. Cell *141*, 1171-1182.

Ichiyanagi, K., Beauregard, A., and Belfort, M. (2003). A bacterial group II intron favors retrotransposition into plasmid targets. Proc Natl Acad Sci U S A *100*, 15742-15747.

Ichiyanagi, K., Beauregard, A., Lawrence, S., Smith, D., Cousineau, B., and Belfort, M. (2002). Retrotransposition of the LI.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. Mol Microbiol *46*, 1259-1272.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M., and Devine, S.E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell *141*, 1253-1261.

Islam, S.M., Hua, Y., Ohba, H., Satoh, K., Kikuchi, M., Yanagisawa, T., and Narumi, I. (2003). Characterization and distribution of IS8301 in the radioresistant bacterium Deinococcus radiodurans. Genes Genet Syst *78*, 319-327.

Jacobs, F.M., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R., and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature *516*, 242-245.

Jacobs, J.Z., Rosado-Lugo, J.D., Cranz-Mileva, S., Ciccaglione, K.M., Tournier, V., and Zaratiegui, M. (2015). Arrested replication forks guide retrotransposon integration. Science *349*, 1549-1553.

Jakubczak, J.L., Burke, W.D., and Eickbush, T.H. (1991). Retrotransposable elements R1 and R2 interrupt the rRNA genes of most insects. Proc Natl Acad Sci U S A *88*, 3295-3299.

Januszyk, K., Li, P.W., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J.A., Martin, S.L., and Clubb, R.T. (2007). Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. J Biol Chem *282*, 24893-24904.

Kagawa, T., Oka, A., Kobayashi, Y., Hiasa, Y., Kitamura, T., Sakugawa, H., Adachi, Y., Anzai, K., Tsuruya, K., Arase, Y.*, et al.* (2015). Recessive inheritance of population-specific intronic LINE-1 insertion causes a rotor syndrome phenotype. Hum Mutat *36*, 327-332.

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

Kazazian, H.H., Jr., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N Engl J Med *377*, 361-370.

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

Kellis, M., Wold, B., Snyder, M.P., Bernstein, B.E., Kundaje, A., Marinov, G.K., Ward, L.D., Birney, E., Crawford, G.E., Dekker, J.*, et al.* (2014). Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A *111*, 6131-6138.

Kennell ,J.C., Moran, J.V., Perlman, P.S., Butow, R.A., and Lambowitz, A.M. (1993). Reverse transcriptase activity associated with maturase-encoding group II introns in yeast mitochondria. Cell *73*,133-46.

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

Kersulyte, D., Velapatino, B., Dailide, G., Mukhopadhyay, A.K., Ito, Y., Cahuayme, L., Parkinson, A.J., Gilman, R.H., and Berg, D.E. (2002). Transposable element ISHp608 of Helicobacter pylori: nonrandom geographic distribution, functional organization, and insertion specificity. J Bacteriol *184*, 992-1002.

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res *16*, 78-87.

Khazina, E., Truffault, V., Buttner, R., Schmidt, S., Coles, M., and Weichenrieder, O. (2011). Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. Nat Struct Mol Biol *18*, 1006-1014.

Khazina, E., and Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proc Natl Acad Sci U S A *106*, 731-736.

Khoury, G., and Gruss, P. (1983). Enhancer elements. Cell *33*, 313-314.

Kidd, J.M., Cooper, G.M., Donahue, W.F., Hayden, H.S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F.*, et al.* (2008). Mapping and sequencing of structural variation from eight human genomes. Nature *453*, 56-64.

Kidd, J.M., Graves, T., Newman, T.L., Fulton, R., Hayden, H.S., Malig, M., Kallicki, J., Kaul, R., Wilson, R.K., and Eichler, E.E. (2010). A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell *143*, 837-847.

Kidwell, M.G. (1992). Horizontal transfer of P elements and other short inverted repeat transposons. Genetica *86*, 275-286.

Kleckner, N. (1990). Regulation of transposition in bacteria. Annu Rev Cell Biol *6*, 297-327.

Kojima, K.K., and Fujiwara, H. (2003). Evolution of target specificity in R1 clade non-LTR retrotransposons. Mol Biol Evol *20*, 351-361.

Kojima, K.K., and Fujiwara, H. (2004). Cross-genome screening of novel sequence-specific non-LTR retrotransposons: various multicopy RNA genes and microsatellites are selected as targets. Mol Biol Evol *21*, 207-217.

Kondo-Iida, E., Kobayashi, K., Watanabe, M., Sasaki, J., Kumagai, T., Koide, H., Saito, K., Osawa, M., Nakamura, Y., and Toda, T. (1999). Novel mutations and genotype-phenotype relationships in 107 families with Fukuyama-type congenital muscular dystrophy (FCMD). Hum Mol Genet *8*, 2303-2309.

Kubo, S., Seleme, M.C., Soifer, H.S., Perez, J.L., Moran, J.V., Kazazian, H.H., Jr., and Kasahara, N. (2006). L1 retrotransposition in nondividing and primary human somatic cells. Proc Natl Acad Sci U S A *103*, 8036-8041.

Kubo, Y., Okazaki, S., Anzai, T., and Fujiwara, H. (2001). Structural and phylogenetic analysis of TRAS, telomeric repeat-specific non-LTR retrotransposon families in Lepidopteran insects. Mol Biol Evol *18*, 848-857.

Kulpa, D.A., and Moran, J.V. (2005). Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Hum Mol Genet *14*, 3237-3248.

Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol *13*, 655-660.

Kuwabara, T., Hsieh, J., Muotri, A., Yeo, G., Warashina, M., Lie, D.C., Moore, L., Nakashima, K., Asashima, M., and Gage, F.H. (2009). Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nat Neurosci *12*, 1097-1105.

LaFave, M.C., Varshney, G.K., Gildea, D.E., Wolfsberg, T.G., Baxevanis, A.D., and Burgess, S.M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res *42*, 4257-4269.

Lam, S., and Roth, J.R. (1983). IS200: a Salmonella-specific insertion sequence. Cell *34*, 951-960.

Lambowitz, A.M., and Zimmerly, S. (2004). Mobile group II introns. Annu Rev Genet *38*, 1-35.

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

Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N., and Charlesworth, B. (1988). On the role of unequal exchange in the containment of transposable element copy number. Genet Res *52*, 223-235.

Lee, E. (2012). Landscape of Somatic Retrotransposition in Human Cancers. Science *337*, 967-970..

Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K.*, et al.* (2012). Landscape of somatic retrotransposition in human cancers. Science (New York, N Y) *337*, 967-971.

Leem, Y.E., Ripmaster, T.L., Kelly, F.D., Ebina, H., Heincelman, M.E., Zhang, K., Grewal, S.I., Hoffman, C.S., and Levin, H.L. (2008). Retrotransposon Tf1 is targeted to Pol II promoters by transcription activators. Mol Cell *30*, 98-107.

Lehrman, M.A., Schneider, W.J., Sudhof, T.C., Brown, M.S., Goldstein, J.L., and Russell, D.W. (1985). Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. Science *227*, 140-146.

Lesage, P., and Todeschini, A.L. (2005). Happy together: the life and times of Ty retrotransposons and their hosts. Cytogenet Genome Res *110*, 70-90.

Lesbats, P., Engelman, A.N., and Cherepanov, P. (2016). Retroviral DNA Integration. Chem Rev *116*, 12730-12757.

Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nat Rev Genet *12*, 615-627.

Lopez de Saro, F.J., and O'Donnell, M. (2001). Interaction of the beta sliding clamp with MutS, ligase, and DNA polymerase I. Proc Natl Acad Sci U S A *98*, 8376-8380.

Lozovskaya, E.R., Scheinker, V.S., and Evgen'ev, M.B. (1990). A hybrid dysgenesis syndrome in Drosophila virilis. Genetics *126*, 619-623.

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

Lundblad, V., and Wright, W.E. (1996). Telomeres and telomerase: a simple picture becomes complex. Cell *87*, 369-375.

Lyon, M.F. (1998). X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet *80*, 133-137.

Lyozin, G.T., Makarova, K.S., Velikodvorskaja, V.V., Zelentsova, H.S., Khechumian, R.R., Kidwell, M.G., Koonin, E.V., and Evgen'ev, M.B. (2001). The structure and evolution of Penelope in the virilis species group of Drosophila: an ancient lineage of retroelements. J Mol Evol *52*, 445-456.

Macfarlane, C.M., Collier, P., Rahbari, R., Beck, C.R., Wagstaff, J.F., Igoe, S., Moran, J.V., and Badge, R.M. (2013). Transduction-specific ATLAS reveals a cohort of highly active L1 retrotransposons in human populations. Hum Mutat *34*, 974-985.

Macia, A., Munoz-Lopez, M., Cortes, J.L., Hastings, R.K., Morell, S., Lucena-Aguilar, G., Marchal, J.A., Badge, R.M., and Garcia-Perez, J.L. (2011). Epigenetic control of retrotransposon expression in human embryonic stem cells. Molecular and cellular biology *31*, 300-316.

Mager, D.L., and Stoye, J.P. (2015). Mammalian Endogenous Retroviruses. Microbiol Spectr *3*, MDNA3-0009-2014.

Majumdar, A., Chatterjee, A.G., Ripmaster, T.L., and Levin, H.L. (2011). Determinants that specify the integration pattern of retrotransposon Tf1 in the fbp1 promoter of Schizosaccharomyces pombe. J Virol *85*, 519-529.

Maksakova, I.A., Romanish, M.T., Gagnier, L., Dunn, C.A., van de Lagemaat, L.N., and Mager, D.L. (2006). Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet *2*, e2.

Malik, H.S., Burke, W.D., and Eickbush, T.H. (1999). The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol *16*, 793-805.

Malki, S., van der Heijden, G.W., O'Donnell, K.A., Martin, S.L., and Bortvin, A. (2014). A Role for Retrotransposon LINE-1 in Fetal Oocyte Attrition in Mice. Dev Cell *29*, 521-533.

Mandrioli, M. (2002). Cytogenetic characterization of telomeres in the holocentric chromosomes of the lepidopteran Mamestra brassicae. Chromosome Res *10*, 279-286.

Marschalek, R., Hofmann, J., Schumann, G., Gosseringer, R., and Dingermann, T. (1992). Structure of DRE, a retrotransposable element which integrates with position specificity upstream of Dictyostelium discoideum tRNA genes. Mol Cell Biol *12*, 229-239.

Martin, S.L. (1991). Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol Cell Biol *11*, 4804-4807.

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

Martin, S.L., Li, W.L., Furano, A.V., and Boissinot, S. (2005). The structures of mouse and human L1 elements reflect their insertion mechanism. Cytogenet Genome Res *110*, 223-228.

Martinez-Abarca, F., Barrientos-Duran, A., Fernandez-Lopez, M., and Toro, N. (2004). The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. Nucleic Acids Res *32*, 2880-2888.

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

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

Medstrand, P., van de Lagemaat, L.N., and Mager, D.L. (2002). Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res *12*, 1483-1495.

Melander, A., Olsson, J., Lindberg, G., Salzman, A., Howard, T., Stang, P., Lydick, E., Emslie-Smith, A., Boyle, D.I., Evans, J.M.*, et al.* (1999). 35th Annual Meeting of the European Association for the Study of Diabetes : Brussels, Belgium, 28 September-2 October 1999. Diabetologia *42*, A1-A330.

Michel, F., and Lang, B.F. (1985). Mitochondrial class II introns encode proteins related to the reverse transcriptases of retroviruses. Nature *316*, 641-643.

Miki, Y. (1992). Disruption of the APC Gene by a Retrotransposal Insertion of L1 Sequence in a Colon Cancer. Cancer Research *52*, 643-645.

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

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

Mine, M., Chen, J.M., Brivet, M., Desguerre, I., Marchant, D., de Lonlay, P., Bernard, A., Ferec, C., Abitbol, M., Ricquier, D.*, et al.* (2007). A large genomic deletion in the PDHX gene caused by the retrotranspositional insertion of a full-length LINE-1 element. Hum Mutat *28*, 137-142.

Mita, K., Kasahara, M., Sasaki, S., Nagayasu, Y., Yamada, T., Kanamori, H., Namiki, N., Kitagawa, M., Yamashita, H., Yasukochi, Y.*, et al.* (2004). The genome sequence of silkworm, Bombyx mori. DNA Res *11*, 27-35.

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

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

Moran, J.V., Mecklenburg, K.L., Sass, P., Belcher, S.M., Mahnke, D., Lewin, A., and Perlman, P. (1994). Splicing defective mutants of the COXI gene of yeast mitochondrial DNA: initial definition of the maturase domain of the group II intron aI2. Nucleic Acids Res *22*, 2057-2064.

Morisada, N., Rendtorff, N.D., Nozu, K., Morishita, T., Miyakawa, T., Matsumoto, T., Hisano, S., Iijima, K., Tranebjaerg, L., Shirahata, A.*, et al.* (2010). Branchio-oto-renal syndrome caused by partial EYA1 deletion due to LINE-1 insertion. Pediatr Nephrol *25*, 1343-1348.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. (2007). Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature *446*, 208-212.

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

Moyes, D., Griffiths, D.J., and Venables, P.J. (2007). Insertional polymorphisms: a new lease of life for endogenous retroviruses in human disease. Trends Genet *23*, 326-333.

Munoz-Lopez, M., and Garcia-Perez, J.L. (2010). DNA transposons: nature and applications in genomics. Curr Genomics *11*, 115-128.

Muotri, A.R., Chu, V.T., Marchetto, M.C., Deng, W., Moran, J.V., and Gage, F.H. (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature *435*, 903-910.

Musova, Z., Hedvicakova, P., Mohrmann, M., Tesarova, M., Krepelova, A., Zeman, J., and Sedlacek, Z. (2006). A novel insertion of a rearranged L1 element in exon 44 of the dystrophin gene: further evidence for possible bias in retroposon integration. Biochem Biophys Res Commun *347*, 145-149.

Myers, J.S., Vincent, B.J., Udall, H., Watkins, W.S., Morrish, T.A., Kilroy, G.E., Swergold, G.D., Henke, J., Henke, L., Moran, J.V*., et al.* (2002). A comprehensive analysis of recently integrated human Ta L1 elements. Am J Hum Genet *71*, 312-326.

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

Novikova, O., and Belfort, M. (2017). Mobile Group II Introns as Ancestral Eukaryotic Elements. Trends Genet.

Nur, I., Pascale, E., and Furano, A.V. (1988). The left end of rat L1 (L1Rn, long interspersed repeated) DNA which is a CpG island can function as a promoter. Nucleic Acids Res *16*, 9233-9251.

Ohno, S. (1972). So much "junk" DNA in our genome. Brookhaven Symp Biol *23*, 366-370.
Okazaki, S., Ishikawa, H., and Fujiwara, H. (1995). Structural analysis of TRAS1, a novel family of telomeric repeat-associated retrotransposons in the silkworm, Bombyx mori. Mol Cell Biol *15*, 4545-4552.

Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L., and Kazazian, H.H., Jr. (2002). A mouse model of human L1 retrotransposition. Nat Genet *32*, 655-660.

Ostertag, E.M., and Kazazian, H.H., Jr. (2001). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res *11*, 2059-2065.

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

Pace, J.K., 2nd, and Feschotte, C. (2007). The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. Genome Res *17*, 422-432.

Pardue, M.L., and DeBaryshe, P.G. (2003). Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu Rev Genet *37*, 485-511.

Parks, A.R., Li, Z., Shi, Q., Owens, R.M., Jin, M.M., and Peters, J.E. (2009). Transposition into replicating DNA occurs through interaction with the processivity factor. Cell *138*, 685-695.

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

Peters, J.E., and Craig, N.L. (2000). Tn7 transposes proximal to DNA double-strand breaks and into regions where chromosomal DNA replication terminates. Mol Cell *6*, 573-582.

Peters, J.E., and Craig, N.L. (2001). Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. Genes Dev *15*, 737-747.

Pickeral, O.K., Makalowski, W., Boguski, M.S., and Boeke, J.D. (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res *10*, 411-415.

Piskareva, O., and Schmatchenko, V. (2006). DNA polymerization by the reverse transcriptase of the human L1 retrotransposon on its own template in vitro. FEBS Lett *580*, 661-668.

Pluciennik, A., Burdett, V., Lukianova, O., O'Donnell, M., and Modrich, P. (2009). Involvement of the beta clamp in methyl-directed mismatch repair in vitro. J Biol Chem *284*, 32782-32791.

Pon, J.R., and Marra, M.A. (2015). Driver and passenger mutations in cancer. Annu Rev Pathol *10*, 25-50.
Ponting, C.P., and Hardison, R.C. (2011). What fraction of the human genome is functional? Genome Res *21*, 1769-1776.

Pyatkov, K.I., Arkhipova, I.R., Malkova, N.V., Finnegan, D.J., and Evgen'ev, M.B. (2004). Reverse transcriptase and endonuclease activities encoded by Penelope-like retroelements. Proc Natl Acad Sci U S A *101*, 14719-14724.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Lower, J., Stratling, W.H., Lower, R., and Schumann, G.G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res *40*, 1666-1683.

Rands, C.M., Meader, S., Ponting, C.P., and Lunter, G. (2014). 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genet *10*, e1004525.

Ray, D.A., Feschotte, C., Pagan, H.J., Smith, J.D., Pritham, E.J., Arensburger, P., Atkinson, P.W., and Craig, N.L. (2008). Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. Genome Res *18*, 717-728.

Ray, D.A., Pagan, H.J., Thompson, M.L., and Stevens, R.D. (2007). Bats with hATs: evidence for recent DNA transposon activity in genus Myotis. Mol Biol Evol *24*, 632-639.

Richardson, S.R., Doucet, A.J., Kopera, H.C., Moldovan, J.B., Garcia-Perez, J.L., and Moran, J.V. (2015). The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. Microbiol Spectr *3*, MDNA3-0061-2014.

Richardson, S.R., Gerdes, P., Gerhardt, D.J., Sanchez-Luque, F.J., Bodea, G.O., Munoz-Lopez, M., Jesuadian, J.S., Kempen, M.H.C., Carreira, P.E., Jeddeloh, J.A.*, et al.* (2017). Heritable L1 retrotransposition in the mouse primordial germline and early embryo. Genome Res.

Richardson, S.R., Narvaiza, I., Planegger, R.A., Weitzman, M.D., and Moran, J.V. (2014). APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. Elife *3*, e02008.

Riggs, A.D. (1990). Marsupials and Mechanisms of X-Chromosome Inactivation. Aust J Zool *37*, 419-441.

Rodic, N., Steranka, J.P., Makohon-Moore, A., Moyer, A., Shen, P., Sharma, R., Kohutek, Z.A., Huang, C.R., Ahn, D., Mita, P.*, et al.* (2015). Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat Med *21*, 1060-1064.

Rodriguez-Martin, C., Cidre, F., Fernandez-Teijeiro, A., Gomez-Mariano, G., de la Vega, L., Ramos, P., Zaballos, A., Monzon, S., and Alonso, J. (2016). Familial retinoblastoma due to intronic LINE-1 insertion causes aberrant and noncanonical mRNA splicing of the RB1 gene. J Hum Genet *61*, 463-466.

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

Rubin, C.M., Houck, C.M., Deininger, P.L., Friedmann, T., and Schmid, C.W. (1980). Partial nucleotide sequence of the 300-nucleotide interspersed repeated human DNA sequences. Nature *284*, 372-374.

Sandmeyer, S. (2003). Integration by design. Proc Natl Acad Sci U S A *100*, 5586-5588.

Sandmeyer, S., Patterson, K., and Bilanchone, V. (2015). Ty3, a Position-specific

Retrotransposon in Budding Yeast. Microbiol Spectr *3*, MDNA3-0057-2014.
Sassaman, D.M., Dombroski, B.A., Moran, J.V., Kimberland, M.L., Naas, T.P., DeBerardinis, R.J., Gabriel, A., Swergold, G.D., and Kazazian, H.H., Jr. (1997). Many human L1 elements are capable of retrotransposition. Nat Genet *16*, 37-43.

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., and Margolet, L. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics *1*, 113-125.

Scott, E.C., and Devine, S.E. (2017). The Role of Somatic L1 Retrotransposition in Human Cancers. Viruses 9.

Scott, E.C., Gardner, E.J., Masood, A., Chuang, N.T., Vertino, P.M., and Devine, S.E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res *26*, 745-755.

Segal, Y., Peissel, B., Renieri, A., de Marchi, M., Ballabio, A., Pei, Y., and Zhou, J. (1999). LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. Am J Hum Genet *64*, 62-69.

Sen, S.K., Huang, C.T., Han, K., and Batzer, M.A. (2007). Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. Nucleic Acids Res *35*, 3741-3751.

Servant, G., Streva, V.A., Derbes, R.S., Wijetunge, M.I., Neeland, M., White, T.B., Belancio, V.P., Roy-Engel, A.M., Deininger, P.L. (2017). The nucleotide excision repair pathway limits L1 retrotransposition. Genetics *1*, 139-153.

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

Shi, Q., Parks, A.R., Potter, B.D., Safir, I.J., Luo, Y., Forster, B.M., and Peters, J.E. (2008). DNA damage differentially activates regional chromosomal loci for Tn7 transposition in Escherichia coli. Genetics *179*, 1237-1250.

Shukla, R., Upton, K.R., Munoz-Lopez, M., Gerhardt, D.J., Fisher, M.E., Nguyen, T., Brennan, P.M., Baillie, J.K., Collino, A., Ghisletti, S.*, et al.* (2013). Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell *153*, 101-111.

Simons, C., Pheasant, M., Makunin, I.V., and Mattick, J.S. (2006). Transposon-free regions in mammalian genomes. Genome Res *16*, 164-172.

Skowronski, J., Fanning, T.G., and Singer, M.F. (1988). Unit-length LINE-1 transcripts in human teratocarcinoma cells. Mol and Cell Biol *8*, 1385-1397.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annu Rev Biochem *72*, 449-479.

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

Smit, A.F., Toth, G., Riggs, A.D., and Jurka, J. (1995). Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol *246*, 401-417.

Solyom, S., Ewing, A.D., Rahrmann, E.P., Doucet, T., Nelson, H.H., Burns, M.B., Harris, R.S., Sigmon, D.F., Casella, A., Erlanger, B.*, et al.* (2012). Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res *22*, 2328-2338.

Spaller, T., Groth, M., Glockner, G., Winckler, T. (2017). TRE5-A retrotransposition profiling reveals putative RNA polymerase III transcription complex binding sites on the Dictyostelium extrachromosomal rDNA element. PLoS One. *12(4)*.

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol *21*, 1973-1985.

Spradling, A.C., Bellen, H.J., and Hoskins, R.A. (2011). Drosophila P elements preferentially transpose to replication origins. Proc Natl Acad Sci U S A *108*, 15948-15953.

Srikanta, D., Sen, S.K., Huang, C.T., Conlin, E.M., Rhodes, R.M., and Batzer, M.A. (2009). An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. Genomics *93*, 205-212.

Startek, M., Szafranski, P., Gambin, T., Campbell, I.M., Hixson, P., Shaw, C.A., Stankiewicz, P., and Gambin, A. (2015). Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. Nucleic Acids Res *43*, 2188-2198.

Stewart, C., Kural, D., Stromberg, M.P., Walker, J.A., Konkel, M.K., Stutz, A.M., Urban, A.E., Grubert, F., Lam, H.Y., Lee, W.P*., et al.* (2011). A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet *7*, e1002236.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719-724.

Swergold, G.D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol *10*, 6718-6729.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327-338.

Szafranski, K., Glockner, G., Dingermann, T., Dannat, K., Noegel, A.A., Eichinger, L., Rosenthal, A., and Winckler, T. (1999). Non-LTR retrotransposons with unique integration preferences downstream of Dictyostelium discoideum tRNA genes. Mol Gen Genet *262*, 772-780.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol *3*, research0052.

Takahashi, H., and Fujiwara, H. (1999). Transcription analysis of the telomeric repeat-specific retrotransposons TRAS1 and SART1 of the silkworm Bombyx mori. Nucleic Acids Res *27*, 2015-2021.

Takahashi, H., Okazaki, S., and Fujiwara, H. (1997). A new family of site-specific retrotransposons, SART1, is inserted into telomeric repeats of the silkworm, Bombyx mori. Nucleic Acids Res *25*, 1578-1584.

Tang, Z., Steranka, J.P., Ma, S., Grivainis, M., Rodic, N., Huang, C.R., Shih, I.M., Wang, T.L., Boeke, J.D., Fenyo, D*., et al.* (2017). Human transposon insertion profiling: Analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. Proc Natl Acad Sci U S A *114*, E733-E740.

Tchenio, T., Casella, J.F., and Heidmann, T. (2000). Members of the SRY family regulate the human LINE retrotransposons. Nucleic Acids Res *28*, 411-415.

Telesnitsky, A., and Goff, S.P. (1997). Reverse Transcriptase and the Generation of Retroviral DNA. In Retroviruses, J.M. Coffin, S.H. Hughes, and H.E. Varmus, eds. (Cold Spring Harbor (NY)).

Temtamy, S.A., Aglan, M.S., Valencia, M., Cocchi, G., Pacheco, M., Ashour, A.M., Amr, K.S., Helmy, S.M., El-Gammal, M.A., Wright, M.*, et al.* (2008). Long interspersed nuclear element-1 (LINE1)-mediated deletion of EVC, EVC2, C4orf6, and STK32B in Ellis-van Creveld syndrome with borderline intelligence. Hum Mutat *29*, 931-938.

Thayer, R.E., Singer, M.F., and Fanning, T.G. (1993). Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. Gene *133*, 273-277.

Ton-Hoang, B., Pasternak, C., Siguier, P., Guynet, C., Hickman, A.B., Dyda, F., Sommer, S., and Chandler, M. (2010). Single-stranded DNA transposition is coupled to host replication. Cell *142*, 398-408.

Totoki, Y., Tatsuno, K., Yamamoto, S., Arai, Y., Hosoda, F., Ishikawa, S., Tsutsumi, S., Sonoda, K., Totsuka, H., Shirakihara, T.*, et al.* (2011). High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet *43*, 464-469.

Traverse, K.L., and Pardue, M.L. (1988). A spontaneously opened ring chromosome of Drosophila melanogaster has acquired He-T DNA sequences at both new telomeres. Proc Natl Acad Sci U S A *85*, 8116-8120.

Tsankov, A., Yanagisawa, Y., Rhind, N., Regev, A., and Rando, O.J. (2011). Evolutionary divergence of intrinsic and trans-regulated nucleosome positioning sequences reveals plastic rules for chromatin organization. Genome Res *21*, 1851-1862.

Tubio, J.M., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K.*, et al.* (2014a). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

Tubio, J.M., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K.*, et al.* (2014b). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

Tuzun, E., Sharp, A.J., Bailey, J.A., Kaul, R., Morrison, V.A., Pertz, L.M., Haugen, E., Hayden, H., Albertson, D., Pinkel, D.*, et al.* (2005). Fine-scale structural variation of the human genome. Nat Genet *37*, 727-732.

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

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

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sanchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M.*, et al.* (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228-239.

Usdin, K., and Furano, A.V. (1989). The structure of the guanine-rich polypurine:polypyrimidine sequence at the right end of the rat L1 (LINE) element. J Biol Chem *264*, 15681-15687.

van den Hurk, J.A.J.M., Meij, I.C., Seleme, M.d.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S.*, et al.* (2007). L1 retrotransposition can occur early in human embryonic development. Hum Mol Genet *16*, 1587-1592.

Vanin, E.F. (1985). Processed pseudogenes: characteristics and evolution. Annu Rev Genet *19*, 253-272.

Veitenhansl, M., Stegner, K., Hierl, F.X., Dieterle, C., Feldmeier, H., Gutt, B., Landgraf, R., Garrow, A.P., Vileikyte, L., Findlow, A.*, et al.* (2004). 40th EASD Annual Meeting of the European Association for the Study of Diabetes : Munich, Germany, 5-9 September 2004. Diabetologia *47*, A1-A464.

Vidaud, D., Vidaud, M., Bahnak, B.R., Siguret, V., Gispert Sanchez, S., Laurian, Y., Meyer, D., Goossens, M., and Lavergne, J.M. (1993). Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. Eur J Hum Genet *1*, 30-36.

Vogt, J., Bengesser, K., Claes, K.B., Wimmer, K., Mautner, V.F., van Minkelen, R., Legius, E., Brems, H., Upadhyaya, M., Hogel, J.*, et al.* (2014). SVA retrotransposon insertion-associated deletion represents a novel mutational mechanism underlying large genomic copy number changes with non-recurrent breakpoints. Genome Biol *15*, R80.

Volff, J.N., Hornung, U., and Schartl, M. (2001). Fish retroposons related to the Penelope element of Drosophila virilis define a new group of retrotransposable elements. Mol Genet Genomics *265*, 711-720.

Waring, M., and Britten, R.J. (1966). Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. Science *154*, 791-794.

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

Wei, W., Morrish, T.A., Alisch, R.S., and Moran, J.V. (2000). A Transient Assay Reveals That Cultured Human Cells Can Accommodate Multiple LINE-1 Retrotransposition Events. Analytical Biochemistry *284*, 435-438.

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

Wichman, H.A., Van den Bussche, R.A., Hamilton, M.J., and Baker, R.J. (1992). Transposable elements and the evolution of genome organization in mammals. Genetica *86*, 287-293.

Wimmer, K., Callens, T., Wernstedt, A., and Messiaen, L. (2011). The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. PLoS Genet *7*, e1002371.

Woodcock, D.M., Williamson, M.R., and Doherty, J.P. (1996). A sensitive RNase protection assay to detect transcripts from potentially functional human endogenous L1 retrotransposons. Biochem Biophys Res Commun *222*, 460-465.

Wu, X., Li, Y., Crise, B., and Burgess, S.M. (2003). Transcription start regions in the human genome are favored targets for MLV integration. Science *300*, 1749-1751.

Wulff, K., Gazda, H., Schroder, W., Robicka-Milewska, R., and Herrmann, F.H. (2000). Identification of a novel large F9 gene mutation-an insertion of an Alu repeated DNA element in exon e of the factor 9 gene. Hum Mutat *15*, 299.

Xing, J., Zhang, Y., Han, K., Salem, A.H., Sen, S.K., Huff, C.D., Zhou, Q., Kirkness, E.F., Levy, S., Batzer, M.A.*, et al.* (2009). Mobile elements create structural variation: analysis of a complete human genome. Genome Res *19*, 1516-1526.

Xiong, Y., and Eickbush, T.H. (1988). Similarity of reverse transcriptase-like sequences of viruses, transposable elements, and mitochondrial introns. Mol Biol Evol *5*, 675-690.

Yang, J., Malik, H.S., and Eickbush, T.H. (1999). Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. Proc Natl Acad Sci U S A *96*, 7847-7852.

Yang, L., Brunsfeld, J., Scott, L., and Wichman, H. (2014). Reviving the dead: history and reactivation of an extinct I1. PLoS Genet *10*, e1004395.

Yang, N., Zhang, L., Zhang, Y., and Kazazian, H.H., Jr. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. Nucleic Acids Res *31*, 4929-4940.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet *13*, 335-340.

Zelentsova, H., Poluectova, H., Mnjoian, L., Lyozin, G., Veleikodvorskaja, V., Zhivotovsky, L., Kidwell, M.G., and Evgen'ev, M.B. (1999). Distribution and evolution of mobile elements in the virilis species group of Drosophila. Chromosoma *108*, 443-456.

Zhang, Y., Romanish, M.T., and Mager, D.L. (2011). Distributions of transposable elements reveal hazardous zones in mammalian introns. PLoS Comput Biol *7*, e1002046.

Zimmerly, S., Guo, H., Perlman, P.S., and Lambowitz, A.M. (1995). Group II intron mobility occurs by target DNA-primed reverse transcription. Cell *82*, 545-554.

Zingler, N., Weichenrieder, O., and Schumann, G.G. (2005). APE-type non-LTR retrotransposons: determinants involved in target site recognition. Cytogenet Genome Res *110*, 250-268.

Zou, S., Ke, N., Kim, J.M., and Voytas, D.F. (1996). The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. Genes Dev *10*, 634-645.

Chapter 2

The LINE-1 Endonuclease is the Principal Determinant of LINE-1 Integration Preference in the Human Genome

This chapter represents a working draft of a manuscript in preparation. Dr. Ángela Macia and Ms. Laura Sanchez, while in the laboratory of Dr. José L. García-Pérez, performed retrotransposition assays in hESCs and NPCs. Bru-seq was performed on PA-1 cells by Michelle Paulsen and the resultant data were compiled in Dr. Mats Ljungman's laboratory. PacBio sequencing was performed at the University of Michigan Sequencing Core. I performed all retrotransposition assays in HeLa and PA-1 cells, the targeted LINE-1 sequencing, and final data analyses.

Abstract

Long INterspersed Element-1 (LINE-1 or L1) encodes two proteins (ORF1p and ORF2p) required for its mobilization (*i.e.,* retrotransposition) in the human genome. Human ORF2p, a 150kDa protein, has endonuclease (EN) and reverse transcriptase (RT) activities. L1 EN is classified as an apurinic/apyrimidinic-like endonuclease (APE) and is responsible for initiating target-site primed reverse transcription (TPRT) during the L1 retrotransposition cycle. During canonical TPRT, the L1 EN activity makes a single-strand endonucleolytic nick at a double-strand DNA target sequence in genomic DNA. The endonucleolytic nick is typically made at an L1 EN degenerate consensus cleavage sequence (5'-TTTT/A-3'), exposing a free 3' hydroxyl group that can be used as a primer by the ORF2p L1 RT activity to reverse transcribe the associated L1 RNA. Here, we characterized greater than 65,000 *de novo* engineered L1 retrotransposition events in four different cultured human cell lines to determine if L1 exhibits overt integration preferences in the human genome. To determine if L1 EN is the prominent determinant of L1 integration preference, we compared our insertion data

sets to a weighted random model that accounts for the ability of L1 EN to direct integration to degenerate L1 EN consensus sequences in genomic DNA. Our data suggest that L1 EN drives L1 integration into ~100 base pair AT-rich regions of the genome and is the principal determinant in directing L1 integration. Furthermore, we found that L1 integration is not targeted to transcribed regions of the genome and exhibits a slight integration bias toward lagging strand DNA templates. Importantly, engineered L1s readily integrate into protein coding exons and 'transposon free' regions of the genome; thus, no region of the genome appears to be 'off limits' for L1 integration. Thus, we conclude that L1 has earned the title of "interspersed element," as engineered L1 insertions are generally dispersed throughout the genome.

Introduction

Long INterspersed Element-1 (LINE-1 or L1) is the only known human autonomous non-Long Terminal Repeat (non-LTR) retrotransposon, and L1-derived sequences account for ~17% of human genomic DNA (Lander et al., 2001). In addition to moving itself throughout the human genome by a 'copy and paste' mechanism termed retrotransposition (Boeke et al., 1985), the L1-encoded proteins can mobilize Short INterspersed Element (SINE) RNAs (*e.g.* Alu and SINE-R/VNTR/Alu-like (SVA) elements) (Dewannieux et al., 2003; Hancks et al., 2011; Raiz et al., 2012), non-coding RNAs (*e.g.*, U6 snRNA) (Buzdin et al., 2003; Buzdin et al., 2002; Garcia-Perez et al., 2007; Gilbert et al., 2005), and messenger RNAs to new genomic locations (Esnault et al., 2000; Wei et al., 2001); the latter process results in the formation of processed pseudogenes. Collectively, the movement of other cellular RNAs by the L1-encoded proteins is termed L1-mediated retrotransposition. L1-mediated retrotransposition events are responsible for greater than 130 documented cases of human disease (Hancks and Kazazian, 2016; Kazazian and Moran, 2017).

A full-length L1 element is 6kb in length and contains a 5' untranslated region (UTR) with a sense and antisense promoter, two open reading frames (ORF1 and ORF2), and a 3'UTR that is typically followed by a poly(A) tract (Dombroski et al., 1991; Scott et al., 1987; Speek, 2001; Swergold, 1990). Human L1 ORF1 encodes a ~40kDa RNA binding protein (ORF1p) that contains nucleic acid binding and nucleic acid

chaperone activities (Basame et al., 2006; Hohjoh and Singer, 1996; Januszyk et al., 2007; Khazina and Weichenrieder, 2009; Kolosha and Martin, 1997; Martin and Bushman, 2001; Moran et al., 1996). Human L1 ORF2 encodes a 150kDa protein (ORF2p) (Doucet et al., 2010; Ergun et al., 2004) that has endonuclease (EN) (Feng et al., 1996) and reverse transcriptase (RT) activities (Mathias et al., 1991). Activities associated with both ORF1p and ORF2p are required for L1 retrotransposition (Feng et al., 1996; Moran et al., 1996).

The amino terminus of ORF2p contains an apurinic/apyrimidinic (AP) endonuclease (APE)-like domain (Feng et al, 1996; Martin et al, 1995). L1 EN initiates the process of L1 integration by creating a single-strand endonucleolytic cleavage at a double-strand L1 EN degenerate target sequence in genomic DNA (5'-TTTT/A-3' and variants of that sequence, with the '/' indicating the scissile phosphate bond where cleavage occurs) (Cost and Boeke, 1998; Cost et al., 2001; Feng et al., 1996; Gilbert et al., 2002; Jurka, 1997; Morrish et al., 2002; Symer et al., 2002; Szak et al., 2002). This endonucleolytic nick liberates a free 3' hydroxyl group that acts as a primer for the ORF2p RT activity to copy its associated L1 mRNA template (Cost et al., 2002; Feng et al., 1996) by a process termed target-site primed reverse transcription (TPRT) (Cost et al., 2002; Feng et al., 1996; Kulpa and Moran, 2006; Luan et al., 1993). How L1 integration is completed requires elucidation; however, it is proposed that L1 EN generally cleaves the 'top' strand of genomic DNA downstream of the single-strand endonucleolytic nick, and that the resultant 3'-OH group is used as a primer by an ORF2p DNA-dependent polymerase activity to copy L1 (-) strand cDNA. Host proteins (*e.g.*, DNA ligase) are then thought to ligate the L1 cDNA to genomic DNA. TPRT generally leads to the generation of variable-length target site duplications (TSDs) in genomic DNA that flank the newly inserted L1 (Christense and Eickbush, 2005; Gilbert et al., 2005; Gilbert et al., 2002; Maita et al., 2004; Symer et al., 2002).

L1s can retrotranspose in the germline (Brouha et al., 2002; Ostertag et al., 2002; Richardson et al., 2017), during early development (Garcia-Perez et al., 2007; Richardson et al., 2017; van den Hurk et al., 2007), and in select somatic cells, such as epithelial tumors (Doucet-O'Hare et al., 2015; Ewing et al., 2015; Helman et al., 2014;

Iskow et al., 2010; Lee, 2012; Miki, 1992; Rodic et al., 2015; Scott et al., 2016; Tubio et al., 2014) and neuronal progenitor cells (Baillie et al., 2011; Coufal et al., 2009; Evrony et al., 2012; Evrony et al., 2015; Upton et al., 2015).

Different types of transposable elements (TEs) have evolved elegant convergent strategies to target genomic 'safe havens', where TE insertions are predicted to have relatively minimal effects on host fitness and gene expression (Levin and Moran, 2011; Sultana et al., 2017) (see Chapter 1). However, whether L1 integrates into specific genomic regions requires elucidation.

Previous studies have revealed that L1 EN targets integration into local, ~100 bp A-T rich regions of the genome (Gasior et al., 2007). However, it is clear that selective forces act over evolutionary time to skew the initial distribution of L1-mediated retrotransposition events to different genomic regions (Lander et al., 2001). For example, while L1s accumulate in A-T rich genomic regions, 'older' Alu elements, which require ORF2p to retrotranspose (Bennett et al., 2008; Dewannieux et al., 2003), accumulate in G-C rich genomic regions (Lander et al., 2001). As such, simply analyzing the resident L1s and Alus present in genomic DNA, many of which are tens of millions of years old, may not give an accurate picture of the initial insertion preferences of these retrotransposons.

The analysis of disease producing L1-mediated retrotransposition events has led to the suggestion that certain regions within genes could be integration 'hot-spots.' For example, recent data demonstrates that mutagenic L1-mediated retrotransposition events can occur either at the same nucleotide or within a small size window within the *Factor IX* (Li et al., 2001; Vidaud et al., 1993; Wulff et al., 2000), *BTK* (Conley, 2005), *NF1* (Wimmer et al., 2011), and *APC* tumor suppressor genes (Miki, 1992; Scott et al., 2016). However, the number of disease cases are limited and whether these findings represent a general rule for L1 insertions or whether ascertainment biases account for these data (*i.e.*, the insertion sites within genes likely are L1 EN substrates and must result in gene activation to produce disease) requires further study.

Here, we asked whether L1 preferentially integrates into specific regions of the human genome by generating a relatively unbiased dataset of engineered L1 integration events in human cultured cells that serve as a proxy for cellular environments known to support endogenous L1 retrotransposition (Moran et al., 1996). We then used Pacific Bioscience (PacBio) sequencing-based strategies to capture the 3' ends of >65,000 engineered L1 insertions and their respective 3' flanking genomic DNA sequences and mapped the genomic DNA to the human genome reference sequences. The integration preferences of the engineered L1s were then compared to a weighted random model, which assumes that L1 integration preferences are mediated solely by the presence of a degenerate L1 EN consensus cleavage site in the human genome. Remarkably, the genomic features analyzed (*e.g.*, gene content, transcriptional activity, strand bias, epigenetic environment, and DNA replication status) have minimal effects on L1 integration. Thus, we conclude that L1 has earned the title of 'interspersed' and that L1 EN is the principal determinant predicting L1 integration.

Results

Generation of Libraries of Engineered Human L1 Retrotransposition Events

We utilized a cultured cell L1 retrotransposition assay (Moran et al., 1996) to generate over 65,000 *de* novo engineered L1 retrotransposition events in the following four biologically relevant female-derived cell lines: HeLa, an embryonic carcinoma cell line PA-1 (Garcia-Perez et al., 2010), human embryonic stem cells (hESCs), and hESC-derived neuronal progenitor cells (NPCs). We choose these four cell lines because L1 can retrotranspose *in vivo* during early development (hESCs) (Brouha et al., 2002; Garcia-Perez et al., 2007; Richardson et al., 2017; van den Hurk et al., 2007), in cancer cells (HeLa) (Iskow et al., 2010; and reviewed in Scott and Devine, 2017), and in neuronal progenitor cells (PA-1 and NPCs) (Baillie et al., 2011; Coufal et al., 2009; Evrony et al., 2012; Evrony et al., 2015; Garcia-Perez et al., 2010; Muotri et al., 2010; Upton et al., 2015) (Figure 2.1B**)**. We reasoned that the use of engineered L1 insertions would allow us to generate a relatively unbiased dataset of L1 insertions that would be minimally affected by selective pressures that influence L1 accumulation preferences in the genome over evolutionary time.

Briefly, a subset of HeLa cells were transfected with a retrotransposition-competent engineered human L1 sequence (pJM101/L1.3) (Sassaman et al., 1997) containing a retrotransposition indicator cassette (*mneoI*) in its 3'UTR (Freeman et al., 1994; Moran et al., 1996). The use of the retrotransposition indicator cassette ensures that we are characterizing cells (in this case, G418-resistant cells) that contain *de novo* L1 retrotransposition events. Similar assays were conducted in hESCs [using a modified human engineered L1 (pKUB102/L1.3-sv+) (Wissing et al., 2012) containing an *mneoI* retrotransposition indicator cassette], in PA-1 and a subset of HeLa cells [using a retrotransposition-competent L1 containing an *EGFP* retrotransposition indicator cassette (Ostertag et al., 2000) in its 3'UTR (pCEP4/LRE3-*mEGFPI*), which leads to EGFP-positive cells upon a successful L1 retrotransposition event (Brouha et al., 2002; Garcia-Perez et al., 2010)], and in H9 hESC-derived NPCs [using a retrotransposition-competent L1 containing a modified *EGFP* retrotransposition indicator cassette in its 3'UTR (pCEP99/UB-LRE3-*mEGFPI*) (Coufal et al., 2009)] (Figure 2.1A and 2.8).

We next developed an approach to specifically capture the 3' ends of L1 integration events and their associated 3' flanking genomic DNA sequence. Briefly, genomic DNA was isolated from cells containing engineered L1 retrotransposition events. The resultant DNAs were then randomly sheared to ~3kb, the ends were repaired, a dA nucleotide was added to the DNA, and adapters were ligated onto the repaired ends (Figure 2.1C). The adapter sequences are similar to those previously used by Iskow *et al*. 2010. They contain a partially double-stranded DNA sequence containing a 5' overhang; the 3'-OH group is blocked by an amine group to prevent it from being used as a primer in subsequent PCR reactions.

Aliquots of the resultant DNA library were used as substrates to capture the engineered L1 retrotransposition events. First, a linear extension reaction was conducted with a dual biotinylated primer complementary to a sequence specific to the engineered L1 construct [*i.e.*, termed the LEAP sequence (Kulpa and Moran, 2006) (Figure 2.1C)]. The resultant products were then captured on magnetic streptavidin beads, washed to remove un-biotinylated products, and used as input DNA templates for subsequent PCR reactions using a primer specific to the SV40-polyA start signal of

the engineered L1 and with a primer specific to the ligated adapter sequence (Figure 2.1D). The resultant PCR products were then processed to add PacBio 'dumbbell' linker sequences to their ends and were sequenced on a PacBio instrument at the University of Michigan Sequencing Core Facility using the single molecule real-time (SMRT) circular consensus sequence (CCS) sequencing mode.

We chose PacBio CCS Sequencing because: (1) the generation of CCS reads should minimize DNA sequencing errors that arise during PacBio sequencing; and (2) the longer reads of PacBio sequencing should improve our ability to map L1 sequence reads to repetitive regions of the human genome. On average, we obtained ~600 bp PacBio CCS reads from our various samples. (Figure 2.7D). Moreover, as an additional control, all sample library preparations were performed in parallel with a clonal cell line, PC39, which contains three engineered L1 insertions (Garcia-Perez et al., 2010) (Figure 2.10).

Sequence Identification of L1 Integration Events

We developed a computational pipeline to identify PacBio CCS containing *de novo* engineered L1 integration events. Briefly, the PacBio (CCS) reads were first consistently oriented to ensure that the 5' ends of a read began with the SV40-polyA-start primer sequence. To "call" a PacBio CCS as an L1 integration sites required: (1) The presence of the SV40-polyA-start primer sequence at its 5' end, followed by 3' flanking genomic DNA and the 3' adapter primer sequence. (2) The presence of a >14 bp poly(A) tract following the SV40-polyA-start primer. The SV40-polyA-start primer ends in seven adenosine residues; thus, an additional eight adenosines, which were added during the post-transcriptional processing of L1 RNA *in vivo*, were required to ensure we were analyzing *bona fide de novo* engineered L1 retrotransposition events. The hard-clipped reads with removed flanking primer sequences aligned to 'unique' regions in both the GRCh37/hg19 and GRCh38/hg38 (http://genome.ucsc.edu; Kent et al. 2002) versions of the human genome reference sequence (Lander et al., 2001) using Bowtie2 (Langmead and Salzberg, 2012) (Figures 2.7A, 2.7B, and 2.7C).

The poly(A) tract in sequence reads can disrupt alignment to the genome to base-pair resolution; thus, we devised an approach to address this issue. When defining an L1 insertion site to base-pair resolution, we used an 'A-sliding' approach, where we always assigned a poly(A) tract to the genome (Figure 2.7A). Reads were further refined to base-pair resolution using the Smith-Waterman algorithm. The genomic sequence of the mapped position predicted by Bowtie2 was then realigned to the corresponding CCS read under the Smith-Waterman algorithm. The alignment resulting in the highest score was chosen as the base-pair location of the genomic insertion. We then re-verified the presence of a poly(A) tract of at least 15bp in size that was not attributable to the corresponding mapped 3' flanking genomic sequence. The proportion of filtered PacBio CCS reads that led to identification of an L1 insertion site call set is shown in Figure 2.7B.

We tested the accuracy of mapping sequences containing a poly(A) tract to the genome by randomly selecting 100,000 positions in the human genome and capturing genomic sequence lengths that mimicked the distribution of reads lengths in our empirical insertion dataset (Figure 2.7D). We also added poly(A) tracts of similar lengths observed in our insertion dataset to these randomly picked genomic sequence reads (Figures 2.1G and 2.7F). These randomly selected sequences were mapped to the genome reference (GRCh37/hg19) with the same Bowtie2 settings used to map our empirical insertion dataset, and we applied the same algorithms and criteria mentioned above to 'call' the insertions. Only 2.13% of these 100,000 reads could not be uniquely mapped to one genomic location; 0.12% represent incorrect calls that mapped to another genomic location. Thus, 97.75% of the simulated reads were called correctly, giving us confidence that our calling algorithms are accurate and efficient (Figure 2.20).

Engineered L1 Insertions Display Known L1 Integration Characteristics

Two or more independent CCS sequencing reads were identified for many of our engineered L1 insertions [4.7% of HeLa insertions; 20.1% in PA-1 insertions; 8.9% of NPC; 60.6% of hESC (Figure 2.1F and Figure 2.7E)], indicating that we sequenced at least two independent DNA molecules containing the same L1 insertion. The 5' end of the sequences are required to map to the L1 poly(A) tract/3' genomic sequence junction

site in the genome; thus, sequences are considered as independent CCS reads for the same insertion site if the 3' adapter ligated to different regions of 3' flanking genomic DNA. As expected, the numbers of insertions represented by two or more sequencing reads reflects the numbers of independent engineered L1 retrotransposition events in our initial population of cells (*i.e.*, hESCs had the fewest number of insertions but displays the largest proportion of independent CCS reads). Fewer total engineered insertions were identified in hESCs because as observed previously (Garcia-Perez et al., 2007), L1 retrotransposition efficiently in hESCs is at least an order of magnitude lower than in transformed cell lines.

We next sought to verify that our method results in identification of authentic L1 retrotransposition events. The *de novo* insertions exhibit poly(A) tail lengths longer than those typically observed in endogenous L1 insertions, due to the strong SV40 polyadenylation signal present at the 3' end of the engineered L1 constructs (Figure 2.1G and Figure 2.7F). Furthermore, examination of 25bp upstream and downstream of integration events revealed that the L1s integrated into a previously defined L1 EN degenerate consensus cleavage site (5'-TTTT/A-3') (Feng et al., 1996; Gilbert et al., 2002; Morrish et al., 2002; Symer et al., 2002; Szak et al., 2002) (Figure 2.12A). Indeed, logo plots suggest that L1 actually recognizes a 7mer, (5'-TTTTT/AA-3'), as opposed to the previously published 5mer (Figure 2.1I, Figure 2.7G, Figure 2.11B, and Figure 2.12A). Since the L1 insertions are, in some cases, represented by independent CCS reads, contain a 3' poly(A) tail, and integrate into a L1 EN consensus sequence cleavage site, we are quite confident that our dataset contains a population of authentic *de novo* L1 integration events.

The above efforts resulted in the identification of 65,079 independent L1 insertion sites (Figure 2.1E and Tables 2.1-2.4). Examination of the sequence upstream and downstream of these L1 integration events demonstrates that the sequence directly surrounding L1 integration sites (up to a 100bp window span) is adenosine, and particularly thymidine-rich (Figures 2.1H,, 2.11, and 2.11B). Resident endogenous genomic L1Hs sequences (the human-specific L1 subfamily that contains 'active' L1s) also display this strong AT-rich preference (Figure 2.15A). By comparison, this AT-rich

bias appears to be less pronounced for evolutionary 'older' L1 subfamilies (Figure 2.15A). Thus, the engineered L1 insertions mimic the features of the human-specific, but not 'older' L1s. These data suggest that mutational processes have eroded the original AT-rich integration environment of sequences flanking 'older' L1s (Lander et al., 2001). These findings further serve as an orthogonal piece of evidence that that we have captured *bona fide de novo* engineered L1 integration events from cultured cells.

A Weighted Random Model of L1 Insertions Built on the EN Cleavage Site

To determine whether the L1 endonuclease cleavage site is the sole factor influencing L1 integration, we created a weighted random model based upon the frequencies of 7mer L1 EN cleavage sites observed in our empirical L1 insertion dataset. Since the L1s in each of the four cell types examined in our study contain nearly identical distributions of observed L1 EN integration site variants (Figures 2.2A and 2.12C), we created a simulated L1 insertion model based upon the total insertion dataset [(as opposed to creating an independent model for each specific cell type) (Figure 2.2A and 2.12C)].

To generate our model, we first asked whether each base pair in the 7mer EN consensus cleavage sequence are an independent variable or whether certain positions in the 7mer were co-dependent. For the 7bp EN consensus cleavage site (5'-TTTTT/AA), we refer to the base-pairs from 5' to 3' as positions 1 through 7, with position 1 referring to the first 5' most T, and position 7 referring to the 3' most A. Our large dataset allowed us to subset (*e.g.* EN sites with a C present in position 5) and observe specific L1 EN consensus cleavage sites and resulting logo plots.

We compared the initial variability of the EN consensus cleavage site (Figure 2.1I) to the variability observed in 'conditional' EN sites that lack a T or contain a C at position 5 (Figure 2.2B). We observed a slight shift in nucleotide probabilities at position 1 on the conditional logo plots (Figure 2.2B). Under these same conditions, we observe no change in nucleotide probabilities in positions 6 and 7 on the logo plot. However, the nucleotide probabilities at positions 2 through 5 changes, favoring a higher preference for T, when we require a C residue at either position 3 or position 5. Moreover, when we

required a T at position 3, we observe a higher probability of observing a C in the remaining 3 bps of positions 2 through 5 (Figure 2.2B). Over three-quarters of the L1 insertions display a 7mer EN consensus cleavage site in which positions 2 through 5 contains all Ts or three Ts and one C (Figure 2.2C). Thus, as previously described, L1 insertions are found within a degenerate L1 EN consensus cleavage site.

Based on the above analyses, we concluded that positions 2 through 5 of the L1 cleavage site are co-dependent, whereas the nucleotide residing at position 1 is independent of the sequence at positions 2 through 5 (Figure 2.2B). Likewise, positions 6 and 7 are independent of the sequence within the first five positions of the L1 cleavage site. Thus, instead of creating a model that treats every base pair in the EN consensus as an independent variable, we created a 1-4-2 model (where position 1 represents a 1bp unit at position 1, positions 2 through 5 represent a 4bp unit, and positions 6 and 7 represent a 2bp unit).

We determined the frequency of each 12,288 possible 7mer L1 EN consensus cleavage site variants observed in our empirical insertion dataset. Notably, since our mapping methods always favors a genomic poly(A) sequence to map L1s, position 6 cannot be a T residue, making the 7mer sequence 5'-NNNNN/VN-3' (where N is any nucleotide and V represents an A, C, or G, nucleotide). To calculate the predicted L1 EN cleavage site frequencies, we counted the frequency of observing an A, C, G, or T base present at position 1 in our L1 insertion dataset. For positions 2-5, we counted the frequency of the 256 possible 4mers. For positions 6-7, we counted the frequency of the 12 possible 2mers. The modeled L1 EN insertion frequencies are then calculated as the products of the position probability matrix (Figure 2.2E).

We next sought to identify an appropriate insertion count threshold for when we should use predicted L1 integration frequencies *vs*. the empirical data L1 integration frequencies. The frequencies of each L1 integration sequence was then normalized to the most commonly observed L1 EN consensus cleavage site, 5'-TTTTT/AA-3' (Table 2.5; uncorrected model). After comparing weights calculated from the predicted model frequencies to our observed frequencies (Figure 2.2D), we determined the modeled and empirical data begin to converge when the same L1 EN site is observed in three or

more independent insertions (Figure 2.2D; 97% of our modeled data utilizes observed L1 insertion frequencies; the gray line represents 1:1 ratio between calculated and observed weights). For L1 EN sites observed in fewer than three independent insertions (~3% of cases), the modeled data tends to provide lower weight values than the weight values observed in the actual data. On the other hand, L1 EN site variants in which we observed no engineered L1 integration sites will undoubtedly be modeled higher.

We next used the weighted random model to create a simulated dataset of L1 insertions in the human genome. For every base pair in the genome (excluding gapped regions, and certain highly repetitive sequence regions), we determined the L1 EN cleavage 7mer if an insertion occurred at that nucleotide position, and assigned the corresponding 7mer weight. If the calculated position resulted in a T at the 6^{th} position of the L1 EN site, we 'slid' the cleavage site to the next base pair position in which a T is absent from the 6^{th} position. This procedure was done in the same manner used to call our mapped empirical L1 insertions to base pair nucleotide resolution (Figure 2.7A).

We simulated L1 integration sites by randomly selecting sites in the human genome based upon the weighted probabilities calculated above. Our simulated datasets contained the same numbers of L1 integration events (65,079) as those observed in our empirical dataset. We then repeated the process of generating weighted random simulation datasets for 10,000 iterations. As expected, L1 EN consensus sequences from our simulated integration sites generated using the weighted random model were very similar to those observed in our empirical L1 insertion dataset (Figure 2.12D and Figure 2.2F). For experiments where we limited our analyses to individual cell types, our weighted random simulated datasets contained the same numbers of insertions obtained from an individual cell line (*i.e.*, we compared 10,000 iterations of 27,777 simulated *vs.* 27,777 actual L1 EN sites in PA-1 experiments).

We also created a 'corrected' weighted random model that calculated 7mer site frequencies as described above, but these frequencies are divided by the frequency of the same 7mer found in the human genome reference (GRCh37/hg19) sequence. This 'corrected' model thus adjusts 7mer frequencies with respect to how often they occur in

the genome. Since the 5'-TTTTT/AA-3' 7mer occurs most frequently in the genome, when we 'correct' the observed frequencies and apply the weighting scheme, this 7mer becomes the 21st highest weighted 7mer (Table 2.5). This 'corrected' weighted random model was applied in transcription and replication data analyses and appears as 'PWM Corrected' in plots. Otherwise the uncorrected model was used in data comparisons (box plots) and is also listed as 'Simulated Insertions' in plots.

Insertion Preference Observed on X Chromosome

We next examined where our engineered L1 insertions integrate within the genome. We plotted the chromosomal location of each empirical L1 and compared those data to the same numbers of simulated insertions generated from 10,000 iterations of our weighted random model. L1 insertions are located throughout the genome and do not appear to integrate at distinct 'hot spots' (Figures 2.3B and 2.14). L1 insertion counts are positively correlated with chromosome size, with larger chromosomes containing more insertions than smaller chromosomes [(Spearman's rho ranges from 0.927 to 0.948) (Figure 2.3A and 2.13A)]. HeLa cells display a distinct increase of insertions on chromosomes 1 and 5, which can be explained by previous SKY-FISH experiments, which revealed more than 2 copies of each of these chromosomes (data not shown). Intriguingly, the X-chromosome contains more L1 insertions than predicted from our weighted random model in PA-1s, hESCs, and NPCs, but not HeLa cells. Since PA-1 cells have been shown to have attributes similar to neurons (Garcia-Perez et al., 2010), and NPCs are derived from the hESCs, we hypothesize that these somewhat related cell types may have similar cell type-specific mechanisms that allow for an increased number of L1 insertions on the X-chromosome, which are absent in HeLa cells.

LINE-1 EN Consensus Sequence Influences Antisense Integration in Genes

Previous somatic studies in the brain have suggested that L1 preferentially integrates into expressed genes (Baillie et al., 2011; Coufal et al., 2009; Upton et al., 2015). Thus, we explored whether genes represent preferential L1 integration sites. In agreement with previous studies (Beck et al., 2010; Gilbert et al., 2005; Gilbert et al.,

2002; Moran et al., 1999; Symer et al., 2002), we observed that engineered L1s can readily integrate into the introns and exons of genes; however, genes are not preferentially targets for L1 integration (Figure 2.3C and 2.3D). Indeed, in PA-1 cells we observe significantly fewer genic L1 insertions than predicted by our random model; instead, we observe an increase of L1 integration events into intragenic regions of the genome (χ^2 test p-value for exons: 5.782×10^{-8}; χ^2 test p-value for introns: $<2.2 \times 10^{-16}$). Interestingly, in all the cell types examined, except hESCs, we observe fewer insertions into introns than is expected by our weighted random model (χ^2 test p-value for HeLa: 1.48×10^{-9}; χ^2 test p-value for NPC: $<2.2 \times 10^{-16}$). Furthermore, we did not observe a significant enrichment of NPC L1 insertions within neuronal genes via DAVID analysis (Table 2.15).

Previous studies demonstrated that polymorphic L1Hs insertions display a preference for accumulating in the antisense transcriptional orientation of the genes in which they reside (Beck et al., 2010; Ewing and Kazazian, 2010; Huang et al., 2010; Smit, 1999; Zhang et al., 2011). Thus, we explored whether we observe this same antisense preference in our datasets of L1 insertions. We first identified L1 insertions within genes and then calculated the antisense to sense ratio by calculating the numbers of insertions that integrated into the antisense transcriptional orientation of a gene divided by the numbers of insertions that integrated into the sense transcriptional orientation of a gene (Figures 2.3E and 2.3F). Unexpectedly, the middle line of the boxplots, which represents the median observation from 10,000 iterations of the weighted random model, is greater than 1.0, suggesting that L1 EN consensus cleavage sites are slightly enriched on the coding strand of a gene, which would lead to a greater number of antisense L1 insertions. L1 insertions in hESCs exhibit a pronounced preference for antisense integration (χ^2 test p-value: 0.00272).

Expressed Genes are not L1 Preferential Integration Sites

We next performed RNA-seq on two biological replicates of each of our cell types to test if L1 preferentially targets expressed genes. We found that L1 integration is generally depleted in expressed regions of the genome (Figure 2.13B). Insertions from HeLa, PA-1s, and NPCs are significantly overrepresented in unexpressed regions of the

genome, and the rate of expression also has no influence on integration (χ^2 test p-values: 1.776×10^{-12}, $< 2.2 \times 10^{-16}$, $< 2.2 \times 10^{16}$ respectively) (Figure 2.13C). PA-1 and NPC insertions contain significantly more L1 insertions in low-level expressed regions of the genome (Figure 2.13C). Furthermore, PA-1s have a significant depletion of L1 insertions in highly expressed regions of the genome. These data suggest that L1 integration is actually depleted in highly expressed regions of the genome. Human embryonic stem cells are the only cell type that does not contain a significant reduction in L1 insertions in expressed regions of the genome when compared to the weighted random dataset. Thus, our data suggest that L1 insertions: (1) do not preferentially target genes for integration; (2) display an antisense insertion preference into genes due to a bias of the L1 EN cleavage site on the coding strand; (3) do not preferentially target expressed genes for integration.

L1 Does Not Preferentially Insert within Transcribed Regions of the Genome

While the UCSC genome browser, RefSeq database, and many other databases have predicted notations and boundaries for genes, we know that transcription of a gene can actually initiate before, terminate beyond, or occur completely outside of these annotated genomic regions (Paulsen et al., 2013). Thus, we asked whether transcription influences L1 integration preferences in the human genome. Since transcribed regions of the genome often contain open chromatin, we hypothesized that LINE-1 may take advantage of and target open chromatin that arises during RNA transcription. More specifically, we hypothesized that the ORF2p L1 EN activity may specifically target the accessible coding strand during the process of transcription, as RNA polymerase is engaged with the noncoding strand (Figure 2.4A). If so, we would expect to observe an antisense bias of insertions into transcribed genes.

While RNA-seq is a sensitive approach to measure gene expression, it only provides a steady state snapshot of RNA abundance in a cell. RNAs that either undergo rapid degradation or are transcribed in very low quantities may be missed in RNA-seq experiments. Thus, we performed Bru-seq, a method that assesses nascent RNA synthesis, on two biological replicates of our PA-1 cell line, and analyzed two previously generated Bru-seq datasets from two biological replicates of HeLa cell lines (Paulsen et

al., 2014; Paulsen et al., 2013). The two independent HeLa Bru-seq samples are highly correlated with each other (Spearman's rho of 0.8536), as are the two PA-1 samples (Spearman's rho of 0.935). Thus, we performed subsequent comparisons with both HeLa Bru-seq sample datasets; since the observed trends were the same using both datasets, only one was used in comparisons to our empirical L1 insertion database. Similarly, since the PA-1 datasets were highly correlated, only one dataset was used in comparisons to our empirical L1 insertion database.

As observed in Figures 2.4B and 2.4C, transcribed regions of the genome contain fewer L1 insertions (χ^2 Test: HeLa p-value: 6.795×10^{-6}; PA-1 p-value: $< 2.2 \times 10^{-16}$) than expected when compared to our weighted random dataset (32.6% of HeLa insertions, and 19.4% of PA-1 insertions reside within transcribed regions of the genome). Thus, contrary to our original hypothesis, L1s do not preferentially integrate into transcribed regions of the genome. Notably, this effect is not due to a lack of L1 EN degenerate consensus sites in transcribed regions (Figures 2.4B and 2.4C). We next analyzed insertions that reside within transcribed regions of the genome to determine whether the rate of transcription influences L1 integration (Figures 2.4C and 2.16A). Consistent with the RNA-seq data, highly transcribed genes are not preferential L1 integration sites; indeed there are fewer L1 insertions in highly expressed genes than predicted from our weighted random dataset (Kolmogorov-Smirnov bootstrap test p-value $< 1 \times 10^{-6}$ for PA-1 and HeLa).

Finally, because our insertion data is stranded, we asked whether the L1 EN preferentially cleaves the predominant coding or noncoding strand in the genome during transcription. We calculated transcription bias as the absolute value of the difference in RPKM expression values between the top and bottom strand expression of a transcribed region and divided that number by total RPKM expression value (by adding the combined top and bottom strand expression values). A transcription bias of 0 indicates that both strands are transcribed at the same level, while a value of 1.0 means only one strand (either the top or the bottom) is transcribed (Figure 2.16B). We subdivided the transcribed genome into eleven bins of transcription bias values from 0 to 1.0 based upon the transcription bias of the region in the genome, and plotted the

observed fraction of insertions in which the L1 EN cleaved the predominant coding strand in the genome for integration in each transcription bias ratio bin and compared these values to data obtained from 10,000 iterations of our weighted random dataset (Figure 2.4D). While HeLa insertions display a slight L1 EN cleavage preference on the coding strand regardless of the magnitude of transcription bias, we did not observe a strong L1 EN cleavage preference on the predominant coding strand in the genome (Figures 2.16C and 2.16D). This slight coding strand cleavage preference is even harder to distinguish in PA-1 insertions (Figures 2.4D and 2.16D). Together, the available data lead us to conclude that L1 does not specifically target transcribed regions of the genome for integration. That being stated, L1 insertions found within transcribed regions do exhibit slight L1 EN cleavage preference on the coding strand. This preference for cleavage on the coding strand is small and only exerts a minimal effect on L1 integration preferences (Figure 2.16D Kolmogorov Smirnov bootstrap test HeLa Top and Bottom: pvalue < 0.05; PA-1 Top: p-value < 0.01).

Transposon Free Regions Accommodate L1 Integration

Several genic and intergenic regions of the genome overlap with ultraconserved regions of the genome. Ultraconserved regions are at least 200bp in size contain 100% sequence identity between the human genome and other mammalian genomes (*i.e.*, mouse, rat and dog) (Bejerano et al., 2004; McCole et al., 2014). Several ultraconserved regions are associated with essential genes that play critical roles during early development, such as the *SOX*, *HOX*, and *FOX* gene families. Additionally, several of these ultraconserved regions overlap with regions of the genome that lack transposon sequences, which are termed 'transposon free regions' (Simons et al., 2006). Since several transposon free regions coincide with ultra conserved regions of the human genome, we sought to discriminate whether these regions are transposon free because they are inaccessible for transposable element integration or whether transposable elements insertions into these regions result in catastrophic mutations, that are subject to strong negative selection (*e.g.*, they cause early embryonic lethality).

To address the above possibilities, we examined our insertion dataset to determine if any L1 insertions reside within ultraconserved regions, ultraconserved non-

coding regions, or 'transposon free' genomic regions (Bejerano et al., 2004; Dimitrieva and Bucher, 2013; McCole et al., 2014; Simons et al., 2006). We discovered anywhere from one to four insertions per cell type within ultra-conserved regions of the genome (Table 2.6), which is in stark comparison to the complete absence of RepBase L1Hs sequences within these same genomic regions. Of all the LINE-1 sequences in the human genome, only one severely 5' truncated element overlaps with an ultra-conserved noncoding region. Since our L1 insertion dataset contains *de novo* engineered insertions within these ultraconserved regions, the data suggest that there is not an active mechanism that prevents L1 integration into these regions.

Likewise, as expected, young L1Hs reference sequences do not reside within transposon free genomic regions (Table 2.6). However, we observed 1,282 of our total engineered L1 integration events into transposon free regions. Thus, transposon free regions can accommodate L1 integration events at a cellular level, but as above, these events are likely subject to strong negative selection, which effectively removes the L1 containing alleles from the population.

L1 Integration is Interspersed Throughout the Genome

Our RNA-seq and Bru-seq analyses suggest that L1 EN is not preferentially targeting open chromatin regions in the genome. Thus, we sought to determine if closed regions of the genome are preferential integration sites. We compared our L1 insertion datasets to hidden Markov modeled 15 and 18 chromatin state data generated as part of the Roadmap Epigenomics Project (Roadmap Epigenomics et al., 2015). In general, L1 insertions were not enriched in any of the examined chromatin states (Figure 2.5 and Figure 2.18). HeLa and hESC insertions show a small (less than 2-fold) enrichment in some enhancer states. This enrichment is minimal in comparison to the enrichment MLV insertions observed in transcriptional start sites, strong enhancers, and active promoters of the genome (Figure 2.5) (LaFave et al., 2014). We also discovered minimal (less than 2.5-fold) enrichment of HeLa and hESC insertions in super enhancers and typical enhancers (Figure 2.23A). We compared the chromatic states with their respective GC content to ensure that any observed enrichments are not due to the presence of confounding AT-rich sequences; these regions are not overly AT-rich

(Figures 2.22 and 2.23B). As with our Bru-seq analyses, we once again observed a depletion of insertions in transcribed regions of the genome, as defined by this chromatin state analysis (Figures 2.4B and 2.5).

To verify that the enrichment we observed in the MLV dataset is not simply due to presence of over 300,000 events (almost 10x as many events as we observe in PA-1s), we next downsized the MLV insertion dataset, so that it contained the same numbers of insertions observed in our PA-1 dataset, and reanalyzed the data. As above, we still observe enrichment of MLV integration events in transcriptional start sites, promoters, and enhancer regions (Figure 2.17A).

Finally, we examined our HeLa and hESC L1 insertion datasets to look for overlaps with ENCODE DNaseI hypersensitive sites, which serve as a proxy for accessible chromatin structures in the genome. While we observed an enrichment of L1 insertions into DNaseI hypersensitive regions of the genome when compared to the weighted random dataset, L1s are not preferentially targeting DNaseI hypersensitive sites for integration (minimally 1-4% of total insertions are found in DNaseI hypersensitive sites; Table 2.7). Thus, we conclude that L1 integration is generally interspersed throughout the genome and that L1 does not preferentially target open or closed states of the genome, or other repetitive element sequences for integration (Table 2.14).

Lagging Strand Template Preferentially Cleaved by L1 Endonuclease for Integration

To determine whether DNA replication influences L1 integration preferences, we compared our L1 integration sites to the previously published HeLa cell and GM06990 lymphoblastoid cell line Okazaki fragment sequencing data (Petryk et al., 2016). Briefly, OK-seq allows the capture, sequencing, and mapping of Okazaki fragments that arise during DNA replication. Analyses of the resultant data then allow a means to determine replication fork initiation, replication fork termination, and replication fork directionality throughout the genome (Petryk et al., 2016).

We compared our L1 insertion datasets to both the HeLa and lymphoblastoid GM06990 OK-seq datasets and observed similar trends in all sample set comparisons.

As defined by Petryk *et al.* (2016), Replication Fork Direction (RFD) is a value that ranges from -1 to 1 and is determined by the difference of Okazaki fragments that map to the Crick strand from Okazaki fragments that map to the Watson strand, divided by the total of Okazaki fragments for a given genomic region. We partitioned the genome into eleven bins with RFD values from zero to 1, based on the absolute value of the RFD value of a region in the genome. Thus, an RFD of 0 means that there is no replication direction bias, as replication is occurring equally from both directions in the genome. An RFD of 1 means that replication is only occurring in one direction. Since Petryk *et al.* (2016) performed OK-seq on a population of cells, most areas of the genome undergo bidirectional replication. As described by Petryk *et al.* (2016), only a small portion of the genome (2.4% in HeLa and 5.6% in GM06990) is replicated in only one distinct direction.

The published data by Petryk *et al.* (2016) allowed us to determine which strand in a replication fork is the leading strand template and which is the lagging strand template in areas of the genome in which there is a replication forks directional bias (Figure 2.6A). We plotted our L1 insertion data based upon the absolute value of the RFD of a given genomic segment harboring the L1 insertion. For each of the eleven RFD bins, we then counted the proportion of the L1 insertions where L1 EN cleaved the lagging strand template. Each of the cell types examined displayed a small (Kolmogorov-Smirnov boot strap test P-values: HeLa bottom strand cleavage: $< 1x10^{-6}$; HeLa top strand cleavage: < 0.05; PA-1 bottom and top strand cleavage: $< 1x10^{-6}$; NPC bottom and top strand cleavage: < 0.001; hESC bottom strand cleavage: < 0.05), but similar trend, suggesting that L1 EN has slight preference for cleaving lagging strand template (Figure 2.6B). This preference could be due to accessibility of the lagging strand template, as the leading strand template is actively replicated by the passing polymerase. We also observe no enrichment of insertions in sites of replication fork initiation or termination (Figure 2.19D).

Discussion

We have developed a large-scale approach to characterize greater than 65,000 *bona fide de novo* engineered human L1 integration sites and their associated 3’

flanking genomic DNA sequence from several different human cell lines. This technique has led to the capture, sequencing, and identification of >100-fold more L1 integration sites than previous approaches (Gilbert et al., 2005; Gilbert et al., 2002). We utilized the longer read sequences provide by the PacBio single molecule real time (SMRT) circular consensus sequencing mode to improve our ability to map insertions to repetitive areas of the genome (Table 2.14). Since we observed L1 insertions with long poly(A) tails, it indicates that the PacBio sequencing platform can successfully sequence long polynucleotide stretches of DNA, which frequently prove to be problematic using Illumina-based sequencing approaches (Figures 2.1G and 2.7F). We are quite confident that our insertion dataset represents authentic L1 insertions, as the L1 insertions have long poly(A) tails (Figures 2.1G and 2.7F) and integrate into an L1 EN consensus cleavage site (Figures 2.1I and 2.7G). Moreover, in many cases, more than one independent CCS sequencing read was used to call the L1 integration event (Figures 2.1F and 2.7E).

We designed a weighted random model using the L1 EN consensus cleavage site as the principal determinant of L1 integration. To generate our weighted random model, we dissected the L1 EN consensus cleavage site. We discovered that several of the nucleotides within the 7bp L1 EN degenerate consensus cleavage site (5'-TTTTT/AA-3') are co-dependent variables (Figure 2.2B). Interestingly, we found that nucleotide positions 2 through 5 of the 7bp L1 EN cleavage site are dependent upon each other, and quite frequently contain a string of Ts with a single C nucleotide interspersed (Figure 2.2C). The last two nucleotides of the L1 EN site are also dependent upon each other. We created a model that considered these nucleotide dependencies.

We found that our weighted random model mimics our observed engineered L1 integration sites. The weighted random model involves picking random L1 insertion locations throughout the genome, taking into account the likelihood of whether the 7mer provides a favorable L1 EN degenerate consensus cleavage site to accommodate L1 integration. Our weighted random model mimics the same degenerate L1 EN consensus cleavage site observed in our empirical L1 insertion dataset (Figure 2.2F)

and recapitulates the finding that L1 EN drives insertions towards genomic regions that contain a ~100 bp AT-rich content (Figure 2.12D). In sum, our weighted random model assumes that L1 EN is the principal determinant that dictates L1 integration preferences in the human genome.

We next compared our empirical L1 insertion dataset against the weighted random model simulated L1 integrations to determine if other genomic features influence L1 integration preferences. In general, we find that L1 integration sites are interspersed throughout the genome (Figures 2.3B and 2.14), and the human genome lacks preferential hotspots for integration. As observed previously, we report that L1 insertion counts correlate with chromosome size; however, somewhat unexpectedly, we observed an over-representation of L1 insertions on the X-chromosome in PA-1, NPCs, and hESCs, but not HeLa cells (Figure 2.3A and Figure 2.13A) (Lander et al., 2001; Smit, 1999). The genomic and/or cellular feature(s) responsible for the increase of L1 insertions on the X-chromosome, and whether this is a female-specific phenomenon requires elucidation. However, it is notable, that previous studies have suggested a role for increased L1 content in propagating X-inactivation (Bailey et al., 2000; Gartler and Riggs, 1983; Lyon, 1998; Riggs, 1990).

Unlike previous studies that reported an enrichment of somatic neuronal L1 insertions within neuronal stem cell enhancers and neuronal genes (Baillie et al., 2011; Upton et al., 2015), we did not observe an enrichment of engineered L1 insertions within neuronal genes (Table 2.15). This discrepancy could be due to differences between neurons isolated from the brain and hESC-derived NPCs that are cultured *in vitro*. However, it also is possible that many of the previously characterized L1 insertions are actually false positives (Baillie et al., 2011), and/or result from the amplification of chimeric artifacts generated during WGA in single cell-based experiments (Evrony et al., 2016; Upton et al., 2015). Indeed, a high false positive call rate would complicate downstream analysis pipelines, as these analyses assumed that the L1 somatic insertions represented *bona fide de novo* somatic L1 retrotransposition events. The large size of our empirical L1 database, coupled with the high stringency we used to call

engineered L1 retrotransposition events gives us confidence that we have unprecedented statistical power to analyze L1 insertion preferences in cells.

Engineered L1s can integrate into exons and introns of a number of different genes (Figures 2.3C an 2.3D). However, genes are not preferential targets for L1 integration. Indeed, we observed fewer insertions into exons (in PA-1 cells) and introns (in HeLa, PA-1, and NPCs), given the number of favorable EN consensus cleavage sites in these areas of the genome predicted by our weighted random model. Furthermore, we found that L1 integration sites are less likely to occur into expressed regions of the genome (Figure 2.13B). This result, while counterintuitive at face value, may be indicative of host factors at work, carefully protecting the vulnerable, open, accessible genome from L1 integration during the process of transcription.

L1 insertions within genes (exons or introns) display an antisense integration preference (Figure 2.3F). The L1 EN consensus cleavage site drives this antisense preference, as our weighted random model indicates that the L1 EN consensus cleavage site is not evenly distributed across the genome, but instead, is slightly enriched on the sense strand of genes. Utilization of a sense strand L1 EN cleavage site to initiate TPRT would result in an antisense L1 insertion. The L1 EN consensus cleavage site bias found on the sense strand of genes accounts for the antisense genic integration preference observed.

The significant antisense L1 integration preference observed in hESCs may reflect a higher degree of negative selection acting within these cells. For example, L1 insertions in the same transcriptional orientation of a gene that disrupt the expression of genes required for pluripotency or, less frequently, cellular viability, could lead to differentiation or cell death. Under this scenario, the loss of detrimental L1 insertions in the sense transcriptional orientation could explain the strong antisense bias in hESCs. Notably, the antisense orientation preference observed in hESCs is still well below the 1.8 antisense to sense ratio value of resident human genome LINE-1 events (Smit, 1999). Thus, older L1s also appear to be under strong negative selective forces in the genome, driving this antisense preference even higher. Indeed, experimental studies have suggested that L1 insertions in the same transcriptional orientation of a gene are

more detrimental than those in the opposite transcriptional orientation of a gene (Han et al., 2004).

We also asked whether the L1 EN is more accessible for cleavage on the coding strand of an actively transcribed gene. Although we found a depletion of L1 integration events within transcribed regions of the genome (Figures 2.4B and 2.4C), the insertions into transcribed regions show a significant deviation from the weighted random model suggesting a slight preference for L1 EN cleavage on the coding strand (Figures 2.4D and 2.16D). However, once again, these findings could simply reflect the fact that L1 EN target sequences are slightly enriched on the coding strand of genes.

We utilized previously published Okazaki sequencing data to determine if L1 EN cleavage is favored on the leading or lagging strand template during DNA replication (Petryk et al., 2016). We observe a slight preference for L1 EN cleavage on the lagging strand template, the strand that is most likely to be accessible for cleavage as it is not near the active replicating polymerase (Figure 2.6). We did not find any preference for L1 EN cleavage in DNA replication fork initiation sites or termination sites (Figure 2.19B).

Since we utilized several human cell lines in our studies, we were able to observe differences in L1 integration preferences among different cell lines (*e.g.*, enrichment of X-chromosome insertions in PA-1s, NPCs, and hESCs *vs.* HeLa.). However, some cell types (*e.g.*, NPCs, and PA-1s) have similar neuronal-type characteristics and exhibit similar L1 integration preferences. Notwithstanding these differences, we did not observe a high degree of variability in L1 integration preferences among cell lines, suggesting that L1 integration generally is interspersed throughout the genome.

Our study is not without caveats and potential limitations. For example, some may criticize the use of engineered L1 vectors used to generate our L1 insertion dataset because, in most cases, the identification of *bona fide de novo* L1 retrotransposition events required expression of an indicator gene and the isolation of G418-resistant or EGFP-positive cells. Thus, we could miss L1 integration events that occur into non-transcribed or heterochromatic regions of the genome. Such a scenario would be

predicted to lead to an enrichment of engineered L1s in expressed parts of the genome. However, counter to these predictions, we found that engineered L1s do not preferentially integrate into actively transcribed genes, and actually show a slight preference for non-expressed regions of the genome. Moreover, we did not observe a strong preference for engineered L1 insertions into chromatin states associated with expressed genes or open chromatin. As such, our findings would actually over-estimate the frequency of L1 integration events into expressed genes. Simply stated, our data differs from those in previous studies that claimed L1 preferentially targets actively expressed genes and promoters for integration.

The use of PacBio sequencing to characterize our L1 insertion dataset is another area that could be criticized in our studies. For example, many of our L1 insertion calls are only supported by one DNA sequence read. However, in hESCs, which support lower levels of engineered L1 retrotransposition (Garcia-Perez et al., 2007), ~60% of our L1 insertion calls are supported by two or more reads. Moreover, our stringent filtering algorithm, coupled with the fact that our L1 insertion calls have long poly(A) tails and integrate into a degenerate L1 EN consensus cleavage site are in agreement with smaller scale functional studies of L1 biology. Indeed, we generated a dataset of *bona fide de novo* engineered L1 retrotransposition events and our validation efforts indicate that this large dataset has afforded us an unprecedented opportunity to characterize L1 integration preferences in a statistically robust manner.

Conclusion

In sum, we conclude that L1 EN is the principal determinant in driving L1 integrations sites into degenerate, AT-rich, L1 EN consensus cleavage sites in the human genome. While it remains possible that other genomic features affect L1 integration preferences, our random weighted model indicates that any such features would only have minor effects on L1 integration. In an anthropomorphic sense, these findings suggest that L1 is not fickle and opportunistically uses any accessible L1 EN consensus cleavage site to integrate into the genome. Indeed, we posit that the acquisition of the L1 EN domain allowed L1 to live up to its name as an interspersed element.

Materials and Methods

Plasmids

All plasmids were grown in DH5α (F-ϕ80*lac*ZΔM15Δ(*lac*ZYA-*arg*F) U169 *rec*A1 *end*A1 *hsd*R17 (rk-, mk+) *pho*A *sup*E44 λ- *thi*-1 *gyr*A96 *rel*A1) competent *E.coli* (Invitrogen; Carlsbad, CA). Prepared in house as described in (Inoue et al., 1990). Plasmids were prepared using the Qiagen Plasmid Midi Kit (Qiagen; Hilden, Germany, #12125) according to the manufacturer's protocol.

pCEP4/GFP: has been described previously (Alisch et al., 2006). It consists of a pCEP4 backbone (Invitrogen/Life Technologies; Carlsbad, CA #V04450) that contains the coding sequence of the humanized Renilla green fluorescent protein (hrGFP) from phrGFP-C (Stratagene) driven by a CMV promoter.

LINE-1 Expression Constructs

pJM101/L1.3: has been described previously (Dombroski et al., 1993; Sassaman et al., 1997) and consists of the pCEP4 backbone (Invitrogen/Life Technologies; Carlsbad, CA #V04450) containing a full-length copy of the L1.3 element (Sassaman et al., 1997) with the *mneoI* retrotransposition indicator cassette in the 3'UTR (Freeman et al., 1994; Moran et al., 1996).

pJM105/L1.3: has been described previously (Wei et al., 2001). It consists of the pCEP4 backbone (Invitrogen/Life Technologies; Carlsbad, CA #V04450) containing a full-length copy of the L1.3 element with the *mneoI* indicator cassette in the 3'UTR. The L1.3 contains a missense mutation (D702A) in the RT domain of the ORF2 protein resulting in its inefficiency to retrotranspose.

pJJ101/L1.3: This plasmid was described previously (Kopera et al., 2011). It contains a full-length retrotransposition-competent L1 element, L1.3 (Sassaman et al., 1997), tagged with a *mblastI* retrotransposition indicator cassette in the 3'UTR and was subcloned in pCEP4 (Invitrogen).

pJJD205A/L1.3: This plasmid was described previously (Kopera et al., 2011). This construct is similar to pJJ101/L1.3 but contains a D205A mutation in the ORF2p EN domain.

pJCC9/JM105/L1.3: has been described previously (Beck et al., 2010). This construct is similarly to pJM105/L1.3 except that it is a pBluescript (Stratagene) vector.

pCEP4/LRE3-*mEGFPI*: has been described previously (Garcia-Perez et al., 2010). It consists of the pCEP4 backbone (Invitrogen/Life Technologies; Carlsbad, CA #V04450) containing a full-length RC-L1 (LRE3)(Brouha et al., 2002) and the *mEGFPI* indicator cassette (Ostertag et al., 2000) driven by a CMV promoter subcloned into the 3'UTR of LRE3. LRE3 is driven by its native 5'UTR. The pCEP4 hygromycin selectable marker is replaced by a puromycin selectable marker.

pCEP4/JM111/LRE3-*mEGFPI*: is identical to pCEP4/LRE3-*mEGFPI* except that it contains two missense mutations in ORF1 (RR261-262AA), resulting in a retrotransposition incompetent LRE3 (Moran et al., 1996). Dr. William Giblin (University of Michigan Medical School) constructed the plasmid (Zhang et al., 2014).

pKUB102/L1.3-sv+: has been described previously (Wissing et al., 2012). It consists of a modified backbone version of pBSKS-II (Stratagene) that contains a human ubiquitin C promoter (nucleotides 125398319-125399530 of human chromosome 12) upstream of an L1.3 element (Sassaman et al., 1997). The L1.3 element contains the *mneoI* indicator cassette (Freeman et al., 1994) and the SV40 late polyadenylation signal 3' of the engineered L1.

pCEP99/UB-LRE3-*mEGFPI*: has been described previously (Coufal et al., 2009)**.** This construct comprises a pCEP4 backbone (Invitrogen/Life Technologies; Carlsbad, CA #V04450) containing a full-length RC-L1, LRE3 (Brouha et al., 2002), driven by an ubiquitin C promoter (a 1.2-kb fragment of the human UBC gene nucleotides 125398319-125399530 of human chromosome 12) and the *mEGFPI* indicator cassette

(Ostertag et al., 2000) driven by CMV followed by a SV40 late polyadenylation signal 3' of the engineered L1. The hygromycin marker in pCEP4 is replaced by a puromycin marker.

pCEP99/JM111/UB-LRE3-*mEGFPI*: is a derivative of pCEP99/UB-LRE3-*mEFFPI* that has been described previously (Coufal et al., 2009). It contains a full-length retrotransposition defective L1 (Kimberland et al., 1999) that contains two engineered missense mutations in L1-ORF1p (RR261-262AA) that abolishes retrotransposition activity (Moran et al., 1996; Ostertag et al., 2000). The LRE3 is tagged with the *mEGFPI* retrotransposition indicator cassette (Ostertag et al., 2000), and is cloned in the same modified pCEP4 vector as pCEP99/UB-LRE3-*mEGFPI*.

Cell Culture

All cell lines were grown at 37˚C in the presence of 7% CO_2 and 100% humidity.

HeLa-JVM Cells

Cells were grown in Dulbecco's Modified Eagle Medium (DMEM) (Invitrogen) supplemented with 10% fetal bovine calf serum (FBS) (Invitrogen) and 1X penicillin/streptomycin/glutamine (Invitrogen) to create DMEM-complete medium as described previously (Moran et al., 1996).

PA-1 Cells

Cells were cultured in Minimum Essential Media (MEM) (Invitrogen) supplemented with 10% heat-inactivated Fetal Bovine Serum (FBS) (Invitrogen), 1X penicillin/streptomycin/glutamine (Invitrogen) and 0.1mM non-essential amino acids (Invitrogen).

H9 hESC Cells

All reagents were purchased from Invitrogen's GIBCO-Life Technologies unless otherwise indicated. Human Foreskin Fibroblasts (HFFs, passage 3-10, from ATCC) were used to prepare HFF-conditioned media (HFF-CM) as described (Macia et al.,

2011). HFFs were grown following the provider's instructions in Iscove's Modified Dulbecco's Medium (IMDM) supplemented with 25 mM HEPES, 2mM L-glutamine and 10% heat-inactivated fetal bovine serum (FBS).

Human embryonic stem cells (hESCs) were grown as previously described using HFF-CM (Garcia-Perez et al., 2007; Macia et al., 2011). WA09/H9-hESCs were obtained from Wicell and maintained in HFF-conditioned media (HFF-CM) using matrigel coated plates. To prepare HFF-CM, HFFs were mitotically inactivated by γ-irradiation with 3000-3200 rads, seeded at $3x10^6$ cells in T-225 flask and cultured with hESC media (DMEM KnockOut supplemented with 4 ng/ml β-FGF, 20% KnockOut serum replacement, 1mM L-Glutamine, 0.1 mM b-mercaptoethanol and 0.1mM non-essential amino acids) for 24 h. HFF-CM was collected 24h later and frozen at -80°C until used. We collected HFF-CMs during four consecutive days. *Mycoplasma spp*. was confirmed at least once a month using a PCR-based assay (Minerva). STR-genotyping (at least once a year) was used to control the identity of cell lines used in this study (LorGen, Granada, Spain).

H9 hESC-derived NPC Cells

The differentiation of neuronal progenitor cells (NPCs) from hESCs was carried out using a previously established methodology (Coufal et al., 2009; Muotri et al., 2010). Briefly, H9 hESCs that were grown on Matrigel for at least 5 passages were cultured during 2 days in N2 media (DMEM/F12 + N2 supplement) containing 1μm of Dorsomorphin (Calbiochem) and 10μm of SB-431542 (Yirmiya et al.). Undifferentiated hESCs were detached using a cell-scraper and transferred to low-attachment plates to allow for Embryo Body (EB) formation using the same culture media. Once EBs formed (4-6 days), these were then plated in a 60mm matrigel-coated plate, and cultured for 5-7 days using NB medium (0.5x N2, 0.5x B-27, 20ng/ml of FGF-2 (Miltenyi Biotec) and 1% P/S), changing the media every other day. Rosettes were collected, dissociated and plated on poly-L-ornithine (Yirmiya et al.)/Laminin (Invitrogen) plates using NPC medium (KnockOut DMEM/F-12 with Stem Pro Neural Supplement, 1mM L-Glutamine and Penicillin-Streptomycin (10,000 U/mL)). NPCs were expanded, when confluent, using

StemPro Accutase Cell Dissociation Reagent (Invitrogen) and were not used for more than 15 passages. NPC culture medium is: KnockOut™ DMEM/F-12 media supplemented with 1x StemPro Neural Supplement, 10 ng/mL EGF (R&D), 1x Glutamax and 20 ng/mL FGF-2 (R&D). To induce neural differentiation from confluent NPCs, 1mM *all-trans* Retinoic Acid (Yirmiya et al.) was added to the NPC culture media. *Mycoplasma spp*. was confirmed at least once a month using a PCR-based assay (Minerva). STR-genotyping (at least once a year) was used to control the identity of cell lines used in this study (LorGen, Granada, Spain).

PC39

As previously described (Garcia-Perez et al., 2010) PC39 is a PA-1 clonal cell line that contains 2 previously characterized L1 insertions (pc-39-A and pc-39-B). This cell line was grown in the same conditions as those described for PA-1. Genomic DNA isolated from these cells was used as a positive control in LINE-1 retrotransposition capture PCR reactions.

LINE-1 Retrotransposition Assays

HeLa

Retrotransposition assays in HeLa-JVM cells were carried out as previously described (Moran et al., 1996; Wei et al., 2000). Cells were plated 85-90% confluency in T-175 flasks (BD Biosciences) or 150mm x 25mm dishes (Corning; Corning, New York) to obtain quantifiable colonies for the retroelement expression construct used (either pJM101/L1.3 or pCEP4/LRE3-*mEGFPI*). Additionally, simultaneously a 6-well dish was plated with HeLa cells to be transfected 18 hours after plating: 2 wells were co-transfected with pJM101/L1.3 and pCEP4/GFP, 2 wells were transfected with pCEP4/GFP only, and the last 2 wells remained untransfected. An additional 2 wells in a 6-well dish were transfected with pJM1015/L1.3 as a retrotransposition control.

Eighteen hours after plating, transfections were carried out using the FuGene 6 transfection reagent (Roche; Penzberg, Germany) and Opti-MEM (Invitrogen), according to manufacturer's protocol (3 μl FuGENE 6 and 97 μl Opti-MEM per μg of

DNA transfected in 6-well and 19μg of DNA with 58μl FuGENE 6 in T-175 flask). Cell culture media was replaced the following day. Cells were subjected to selection with 400 μg/ml G418 (Invitrogen) 72 hours post-transfection. Additionally, 72 hours post transfection cells in the 6-well dish were collected and subject to flow cytometry to determine transfection efficiency. The transfection efficiency of each sample was determined by the percent of green fluorescent protein (EGFP) expressing cells. On average, transfection efficiency was ~75% for HeLa cells. For the remaining T-175 flasks or 150mm x 25mm dishes, selection was carried out until 14 days after transfection, replacing the selection media every other day. To visualize the number of cells with successful retrotransposition events, a flask was fixed and stained. Media was aspirated from the flask and cells were washed with 10mL of 1x PBS. The PBS was aspirated and then 5.0mLs of Fix Solution (1X PBS, 2% Formaldehyde and 0.2% Glutaraldehyde) coated the flask and was left to sit at 4˚C for 30 minutes. The fixed cells were than flushed with water, and then stained with 0.1% crystal violet at room temperature for 5 minutes. Cells were then rinsed again with water and left to dry.

PA-1

The retrotransposition assay in PA-1 cells were carried out as previously described (Garcia-Perez et al., 2010). Cells were plated at 90-95% confluency in T-175 flasks (BD Biosciences), 150mm x 25mm dishes (Corning; Corning, New York), or T-75 flasks (BD Biosciences) to obtain quantifiable colonies for the pCEP4/LRE3-*mEGFPI* construct used. In addition, a 6-well dish was plated with PA-1 cells for control transfection experiments.

Eighteen hours after plating the 6-well dish (Corning; Corning, New York), 2-wells were co-transfected with equal amounts of a reporter plasmid human enhanced green fluorescent protein pCEP4/GFP and pCEP4/LRE3-*mEGFPI*. Another 2-wells in the 6-well dish contained PA-1 cells transfected only with pCEP4/GFP, and the last 2-wells contained untransfected PA-1 cells. Additionally, as a retrotransposition control, 2 wells of a 6-well dish were transfected with pCEP4/JM111/LRE3-*mEGFI.* For transfections, FuGENE HD transfection reagent (Roche Biochemical) at 8ul per 2.0 ug of plasmid

DNA per 6 well was used. T-175 flasks (BD Biosciences) or 150mm x 25mm dishes (Corning; Corning, New York) were transfected with 32.0μg of plasmid DNA and 128ul FuGENE HD transfection reagent (Roche Biochemical). Forty-eight hours post transfection cells were selected for transfection with media containing 2μg/mL of puromycin and selection continued until five days post transfection. Seventy-two hours post-transfection, the cells in the 6-well tissue culture dish were washed with 1X PBS, trypsinized, collected, and subjected to flow cytometry. The transfection efficiency of each sample was determined by the percent of green fluorescent protein (EGFP) expressing cells. On average, transfection efficiency was ~20% for PA-1 cells. The remaining flasks were fed daily with the puromycin selection media until five days post transfection. At five days post transfection cells were fed with media containing no drugs. Seven days post-transfection, cells were chemically treated for 14-16 hours with 0.5μM Trichostatin A (TSA, Sigma) or 18-24 hours with 2uM anisomycin. Flow cytometry measured percentage of EGFP-positive cells in both the untreated and drug-treated samples. Eight days post-transfection, cells from a set of T-175 flasks that were not drug-treated were collected for isolation of gDNA. A second set of T-175 flasks were drug-treated and subsequently flow-sorted selecting for GFP positive cells (~$1x10^6$ total cells). Those collected GFP positive cells were then plated into a small T-25 flask and once confluent, moved to a T-175 flask. Once confluency was reached in the T-175 flask, cells were collected for gDNA isolation.

H9 hESC

We used a previously described protocol, with minor modifications (Garcia-Perez et al., 2007). Specifically, H9 hESCs were transfected with the indicated plasmid DNA using the Amaxa Human Stem Cell Nucleofector Kit 2 (Lonza VPH-5022) and using the program A-23. Plasmid DNAs were purified using a Midi-kit (Qiagen) and filtered through a 0.22μm filter (Milipore). As described (Watanabe et al., 2007), and to prevent cell death during nucleofection, cells were cultured with HFF-CM containing 10μM iRock (Y-27632, Sigma) for 1 hour prior to harvesting hESCs. Next, cultured H9-hESCs were detached from matrigel-coated plates using TryPle-Select (Thermofisher) following manufacturer´s instructions. Collected H9-hESCs were washed twice with pre-warmed

HFF-CM containing 4ng/ml β-FGF and 10μM iRock. Finally, H9-hESCs were filtered through a strainer (70μ m Nylon, Corning). An aliquot of harvested H9-hESCs was used to calculate the number of cells/ml using 0.05% Trypsin. We routinely used 2-4x10^6 H9-hESCs and 4μg of each plasmid DNA per transfection. As a control, we always transfected an aliquot of hESCs with pCEP4/GFP and determined the percentage of transfected cells using FACS-sorting 48h after nucleofection. After nucleofection, transfected hESCs were slowly recovered from the nucleofection cuvette and seeded on a 10cm matrigel-coated plate. Media was replaced 6-8 hours post-transfection. In non-selection experiments, transfected hESCs were simply grown during 7 days using HFF-CM supplemented with fresh β-FGF (20 ng/ml) and 1μM iRock and the media was changed daily. In experiments where L1 retrotransposition events were selected with G418, transfected hESCs were first cultured during four days using HFF-CM supplemented with fresh β-FGF (20 ng/ml) and 1μM iRock and culture media was changed daily. After four days, H9-hESCs were selected with 50μg/ml G418 during 7 days and with 100μg/ml of G418 for an additional 7 days using HFF-CM supplemented with fresh β-FGF (20 ng/ml) and 1μM iRock.

H9 hESC-derived NPC Cells

We used a previously described protocol (Coufal et al., 2009). NPCs were transfected with the Rat Neuronal Stem Cell Nucleofection kit (Lonza VPG-1004) using the program A-33. Plasmid DNAs were purified using a Midi-kit (Qiagen) and filtered through a 0.22μm filter (Milipore). Confluent cultures of NPCs (with passage number 3-15) were used in nucleofection experiments. Briefly, cells were detached using StemPro Accutase Cell Dissociation Reagent (Thermofisher). Next, NPCs were washed twice using pre-warm NPC media (KnockOut™ DMEM/F-12 media supplemented with 1x StemPro Neural Supplement, 10 ng/mL EGF (R&D), 1x Glutamax and 20 ng/mL FGF-2 (R&D)) and NPCs were filtered through a cell strainer (70μ m Nylon, Corning). An aliquot of NPCs was used to calculate the number of cells/ml using 0.05% Trypsin. We routinely used 1x10^6 H9-hESC-deirved NPCs and 8μg of each plasmid DNA per transfection. After nucleofection, transfected NPCs were slowly recovered from the nucleofection cuvette and seeded on 3 wells of a poly-L-ornithine/Laminine coated 6-

well plate. Media was replaced 6-8 hours post-transfection. Next, transfected NPCs were cultured during 7 days changing the media every day. When indicated, we added 1mg/ml puromycin (Yirmiya et al.) 48h post-transfection to the NPC media to select transfected cells. In experiments conducted with differentiating NPCs, we transfected NPCs using the same method but added 1mM *all-trans* Retinoic Acid (Yirmiya et al.) to the NPC media (starting with the first change of media 6-8h post-transfection).

As above, we transfected an aliquot of NPCs with a plasmid expressing GFP (pCEP4/GFP) to determine the efficiency of transfection using FACS-sorting 48h after nucleofection. Also, we used NPCs transfected with plasmid pCEP99/UB-JM111/LRE3-*mEGFI* to determine the background level of autofluorescence in FACS-sorting analyses.

The retrotransposition efficiency was determined using FACS (BD FACS Aria device); however, to avoid silencing of engineered insertions, as described (Coufal et al., 2009; Garcia-Perez et al., 2010), 500nM Trichostatin A (TSA) was added to transfected NPCs 7 days post-transfection and cells cultured for 18h prior to FACS analyses.

PD20F

PD20F and PD20FD2 (PD20F cells complemented with a retroviral vector containing the human FANCD2 CDNA via the previously described protocol Puslipher *et al.* 1998) cells were transfected using Fugene6 (Roche/Promega) using manufacturer instructions. Briefly, $8x10^4$ cells were plated per 100mm culture plates and transfected 16h later using 10μl of Fugene6 and 4μg of each plasmid DNA using OptiMEM (Invitrogen) following the manufacturer instructions. 24h later, fresh media was added and cells were cultured for the next 4 days, changing the media every other day. Five days after transfection cells were selected with 2μg/ml Blasticidin-S (Invitrogen) for the following 7 days, changing the media every other day. After the selection process, blast-resistant foci were harvested by trypsinization and genomic DNA extracted. Transfection efficiency controls were included, co-transfecting cells in parallel with a GFP expression

vector (pCEP4/GFP) and determining the percentage of GFP-expressing cells 48h post-transfection by FACS.

Genomic DNA Isolation

hESC, NPC, PD20F Retrotransposition Assay gDNA Isolation

Once retrotransposition assays were completed, a cell-scrapper was used to harvest hESC and NPC cells. Cells were then washed twice with 1xPBS and gDNA was extracted and purified using phenol-chloroform extraction or using a DNeasy Blood & Tissue Mini Kit (Qiagen) following the manufacturer's instructions. Genomic DNA concentration was measured with a Nanodrop (Thermofisher).

HeLa and PA-1 Retrotransposition Assay gDNA Isolation

Once retrotransposition assays were completed, HeLa and PA-1 cells were trypsinized and harvested from flasks. Cells were then washed twice with 1xPBS and gDNA was extracted and purified using the Blood and Cell Culture DNA Midi Kit (Qiagen # 13343). Concentrations were determined using a nanospectrometer. Genomic DNA, 1μg, was run on a 0.75% agarose gel to check quality of isolations.

Retrotransposition Cassette PCR

A preliminary test to check if genomic DNA contained retrotransposition events involved PCR with primers flanking the intron of the reporter cassette. Amplification of genomic DNA resulted in two distinct bands, a larger band (*Neo*: 1,396bp band, *EGFP*: 1,245bp) indicative of unspliced DNA with the intron intact, and a smaller band (*Neo*: 493bp band, *EGFP*: 343bp) indicating removal of the intron from the reporter cassette and thus the presence of a successful retrotransposition event (Figure 2.9) (Moran et al., 1996; Ostertag et al., 2000).

LINE-1 Retrotransposition Capture

Primer sequences: All oligonucleotides used in this study were synthesized by Integrated DNA Technologies (IDT; Coralville, Iowa).

PCR Library Preparation

Top strand adapter with T overhang; purified by high-performance liquid chromatography (HPLC):
5'-GGAAGCTTGACATTCTGGATCGATCGCTGCAGGGTATAGGCGAGGACAACT-3'

Bottom strand adapter with 5' phosphorylation and 3' amino modifier; purified by high-performance liquid chromatography (HPLC): 5'-/5Phos/GTTGTCCT/3AmMO/-3'

10μM final concentrations of annealed adapters were made by incubation of both the top and bottom strand adapters at 95˚C for 5 minutes in H_2O and 1x NEB T4 DNA ligase Buffer (NEW England BioLabs), followed by allowing the tube to naturally come to room temperature.

Genomic DNA (15μg) was randomly sheared to 3kb fragments following protocols for the Covaris S220/E220 operating systems. Sheared gDNA was purified following QIA-quick PCR Purification Kit (Qiagen). Purified sheared gDNA was end repaired following NEBNext End Repair Module (New England BioLabs). End repaired gDNA was purified following QIA-quick PCR Purification Kit (Qiagen). A non-templated dAMP was then incorporated on the 3' end of the purified end repaired gDNA as outlined by NEBNext dA-Tailing Module (New England BioLabs). Following this reaction dA-tailed gDNA was subsequently purified with the MinElute PCR Purification Kit (Qiagen). Annealed adapters were ligated onto the final purified DNA molecules in the following conditions: 1μg DNA to 90μM of annealed adapter (Final Adapter concentration of 4.5μM) in a 20μl total volume reaction with 1μl (200U) of T4 DNA ligase (NEW England BioLabs). Ligation reactions were incubated overnight at 16˚C and heat inactivated at 65˚C for 20 minutes. Samples were purified of excess linkers with QIAquick PCR purification kits (Qiagen) and eluted in 50μl EB Buffer.

Uni-linear Biotinylated Amplification

Biotinylated LEAP; purified by high-performance liquid chromatography (HPLC); 5' Dual Biotin; 18bp internal spacer; 5'-/52-Bio//iSp18/GTTCGAAATCGATAAGCTTGGATCC-3'

Linear extension reactions were performed with Roche Expand Long Range dNTP Pack PCR system. Reactions contained 500ng of template gDNA, the Manufacturer's Expand Long Range Buffer including 12.5mM $MgCl_2$, 0.25μM biotinylated LEAP primer, 500μM PCR Nucleotide mix (dATP, dCTP, dGTP, dTTP at 10μM each), 3% DMSO and 3.5U of Expand Long Range Enzyme. Cycling conditions used are as follows: 94˚C for 5 minutes, followed by 30 cycles of 94˚C, 15s; 65˚C, 30s; 68˚C for 3 min., and a seven-minute extension at 68˚C.

The Uni-linear extension products were subsequently column purified with QIA-quick PCR Purification Kit (Qiagen). Purified products were biotin captured following Dynabeads kilobaseBINDER Kit (Invitrogen) for 3 hours at room temperature while rotating the tube. After capture, beads were placed on a magnet and washed twice with the Wash Buffer, and washed a final time with ddH_2O. Final biotin captured products were eluted to 30μl with ddH_2O.

Nested PCR

Adapter primer: 5'-ATCGATCGCTGCAGGGTATAGG-3'

SV40-polyA-start site: 5'-GCAATAAACAAGTTAACAACAAAAAAAAA-3'

Each 30μl biotinylated captured product was divided amongst 3 separate PCR reactions, where 10μl is used as the starting template per PCR reaction. PCR reactions were performed with Roche Expand Long Range dNTP Pack PCR system. PCR reactions contained the Manufacturer's Expand Long Range Buffer including 12.5mM $MgCl_2$, 0.25μM Adapter primer and 0.25μM SV40-polyA-start site primer, 500μM PCR Nucleotide mix (dATP, dCTP, dGTP, dTTP at 10μM each), 3% DMSO and 3.5U of Expand Long Rang Enzyme. Cycling conditions are as follows: 94˚C for 3 minutes, followed by 35 cycles of 94˚C, 10s; 57˚C, 30s; 68˚C for 2 min., and a seven-minute extension at 68˚C.

Final PCR products were column purified with QIA-quick PCR Purification kit (Qiagen) and eluted to a final volume of 50μl with elution buffer. Samples were checked for gDNA

concentration with Invitrogen's Qubit Fluorometer and then sent to the University of Michigan's Sequencing Core for PacBio Single Molecule Real Time (SMRT) circular consensus sequencing.

PCR Product Characterization

PCR products were column purified using the QIAquick PCR Purification Kit (Qiagen). Products were cloned into TA Cloning Kit Dual Promoter (pCR II) cloning vector (Invitrogen), transformed, and plasmid DNA recovered by mini-prep (Promega SV Mini-Prep kit; Promega, Fitchburg, Wisconsin). Individual clones were then sequenced with M13 Forward and M13 Reverse primers. Resulting Sanger sequences were then blatted to the UCSC GRCh37/hg19 Human Genome Browser (http://genome.ucsc.edu/)(Kent, 2002).

PC39 a positive control

As previously described, (Garcia-Perez et al., 2010) PC39 is a clonal cell-line of PA-1, a human embryonic derived carcinoma cell-line, that was transfected with pCEP4/LRE3-*mEGFI*. PC39 contains two previously characterized EGFP silenced insertions, identified as pc-39-A and pc-39-B. Pc-39-A is a 5' truncated insertion within the EGFP cassette, with a 101bp poly(A) tail located on chromosome 5 at base pair position 16173755 (hg19 reference) within the intron of the gene *MARCH11* (Figure 2.10A). Pc-39-B is an inverted/deleted insertion with a 105 bp poly(A) tail inserted on chromosome 1 at base pair position 158239404 (GRCh37/hg19 reference coordinates) (Figure 2.10A). The PC39 clonal cell-line served as a source of gDNA that was used as a positive control in our LINE-1 capture method. Since the insertion locations were characterized previously, instead of randomly shearing gDNA from PC39, we digested the genomic DNA with restriction enzyme *PacI* and *NdeI*, both of which are downstream of the 3' genomic sequence surrounding insertion pc-39-A and pc-39-B respectively. After performing the biotinylated L1 linear amplification, capturing the biotinylated products, and performing nested PCR, this amplified digested PC39 gDNA resulted in two distinct bands ~580bp (pc-39-A) and ~330bp (pc-39-B) on a 0.75% agarose gel (Figure 2.1D).

Additionally, a third distinct band was apparent at ~1.2kb. After further analysis, this third band resulted in the discovery of a third previously unidentified insertion in this PC39 clonal cell-line. Sanger sequencing this distinct band identified a poly(A) tail of 33bp and 1,111bp of flanking gDNA on chromosome 19. This genomic DNA is flanked by an *NdeI* restriction site (Figure 2.10). After performing a number of PCRs, walking 3' to 5' with primer sequences along the engineered LINE-1 sequence, we were finally able to fully characterize this third insertion, following the previous nomenclature we labeled this insertion as pc-39-C. This insertion is 5' truncated, containing the last 100bp of the ORF2p sequence. It is flanked by an 18bp target site duplication, a poly(A) tail of 33bp and has a cleavage site of 5'-TTCTT/GG on chromosome 19 at base pair position 13627881 (GRCh37/hg19 reference coordinates) on the minus strand (Figure 2.10).

Independent Validation PCRs of Identified LINE-1 Insertions

For 4 insertions that were identified by independent CCS reads, validation PCRs were performed by the following technique: Primers were designed flanking the site of the insertion (Table 2.13). The primer 5' of the insertion site was named the "empty site" primer designed to amplify gDNA absent of the insertion. Two primers were designed 3' of the insertion site. The primer furthest from the insertion site was called the "outer" primer and the primer closest to the 3' end of the insertion was named "inner." Nested PCR was performed on the initially purified pooled gDNA. A combination of PCR reactions were performed, the first PCR set included the "outer" primer along with one of the following e primers: a biotinylated primer specific to the cassette CMV promoter, the LEAP sequence (LEAPv1.1), or the "empty" primer. The second PCR, then included the same CMV, LEAP, or "empty" primer with the "inner" primer sequence. Amplified bands were verified by being cloned into Invitrogen's Dual TA promoter cloning vector and Sanger sequenced. PCR reactions were performed with Invitrogen's Platinum Taq. PCR reactions contained the Manufacturer's PCR Buffer including 12.5mM $MgCl_2$, 0.25μM of each primer, 500μM PCR Nucleotide mix (dATP, dCTP, dGTP, dTTP at 10μM each), and 3.5U of Platinum Taq. Cycling conditions were as follows: The first PCR : 96˚C for 12 minutes, followed by 35 cycles of 96˚C, 30s; 60˚C, 45s; 72˚C, 2 min.,

and a four-minute extension at 72˚C. Second PCR: 96˚C for 3 minutes, followed by 25 cycles of 96˚C, 30s; 60˚C, 30s; 72˚C, 1.5 min., and a three-minute extension at 72˚C.

Mapping CCS reads to the human genome

PacBio single molecule real time circular consensus sequencing reads were first aligned to the adapter primer and SV40pA primer sequences with Bowtie2 (Langmead and Salzberg, 2012). Reads that aligned to each primer sequences were then hard-clipped of the primer sequences and orientated so that the most 5'-end of the sequence read contained a poly(A) tract if present. An in-house utility called homopolymer was then used to determine the presence of a poly(A) tail within the 5'-end of the sequence. Briefly, homopolymer performs a Hidden Markov model with 5 states (no homopolymer, or homopolymer of each of the 4 possible nucleotides) to find homopolymer segments in a sequence. We set the ZERO_PROB setting of homopolymer to 10% or 0.1. Reads that contained at least a 15bp poly(A) tract were aligned to the human genome reference GRCh37/hg19 with Bowtie2 allowing for up to 100 possible mapped locations in the output for each single CCS read. We also aligned reads to the GRCh38/hg38 reference with Bowtie2. The best mapped read location was then determined for each read as the alignment that starts within the first 1% of the length of the read near the poly(A) tract and aligns up to the last 2.5% of the read. If multiple alignments fit this criteria then a best mapped read location was chosen if the alignment score was at least greater than 20 as compared to the alignment score of the next best mapped read location (Figure 2.7A). Because all four cell types (HeLa, PA-1, NPC, hESC) were female, we did not keep any insertions in which the best mapped read was to the Y chromosome. If the next best mapped read location and the highest aligning mapped read had an alignment score difference of greater than 20, then the read and insertion was termed 'unmappable' to the human genome.

We found that occasionally alignment of the CCS read sequence to the genome would be disrupted due to the presence of the poly(A) stretch 5' of the sequence. Disruption of alignment resulted in gapped alignments of the sequence to the genome as Bowtie2 attempted to align the entire sequence read as opposed to the largest stretch of

consecutive nucleotides of the sequence read to the genome. In order to account for this mapping disparity we utilized the Smith-Waterman algorithm to further refine alignment to the genome to base-pair resolution. When running the Smith-Waterman algorithm we aligned the sequence read to the best mapped location given by the Bowtie2 results with an additional 50 to 100bp upstream and downstream of the given genomic location. All poly(A)s present in the sequence read that were also present in the human genome sequence at the point of integration were assigned to the genomic point of insertion as opposed to the poly(A) tail length count (Figure 2.7A). Additionally, when mapping insertions to base-pair resolution, if the genomic sequence contained a poly(A) stretch we called the point of insertion as the most 5' A found in the genome sequence. In so doing, this means that the base-pair 5' of the integration site will never be an A, and since the EN consensus cleavage site is the reverse complement of this sequence that means a T will never be in the 6th position of the 7bp endonuclease cleavage sequence.

Once the final base-pair position of alignment to the genome was determined, the poly(A) tail size was then re-calculated to verify that at least a 15bp poly(A) tail was present in the read that could not be attributed to the genomic location of integration. Insertion call sets that came from the same biological replicate were then examined for the presence of insertions that may be within 10bp of each other and in which one site only had one CCS read associated with the call. The insertion site with just one read was then assigned to the other insertion site within 10bp which contained more CCS reads. Situations of this nature most likely represent the same insertion site but due to either PacBio polymerase amplification slipping or polymerase amplification slipping during capture techniques, part of the flanking sequence was lost. Insertion sites that may be at the same integration site in the genome, but come from different biological samples were called as two independent insertions. Finally, certain highly repetitive sequences in the genome such as centromeric or telomeric regions were found to contain large clusters of insertions. These regions of the genome contain large repetitive tandem repeats as identified by the Tandem repeats finder (Benson, 1999), making it difficult to determine if one integration site actually occurred and reads for the

same insertion were mapped to the incorrect reference genome or several insertions actually occurred in these locations of the genome. And we thus, filtered out any insertions called within these regions of the genome as well.

Bowtie2 command line settings to align to adapter and primer sequences:
bowtie2 -N1 -L3 --ma 3 -a -q --local -x ADAPTERS -U fastq.files –S sample_Adapters.SAM

Bowtie2 command line settings to align to human genome: bowtie2 --no-hd --local -k 100 --un $UNALIGNED_FILE -x $BT2_IDX -U -)

We also aligned reads to the human genome with the following setting since we found that Bowtie2 was not aligning longer reads through the local method: bowtie2 --no-hd -k 100 –un $UNALIGNED_FILE –x $BT2_IDX –U -

Smith Waterman scoring used: no trimming of the reads, matched point of 1, mismatch penalty of -1.5, gap open penalty of -2.5, gap extension penalty of -1

Logo Plots

Logo plots of the L1 EN consensus cleavage site were created with Bioconductor's SeqLogo R package. Corrected logo plots were calculated by determining the proportion of each nucleotide at each observed base pair position of the 7bp EN consensus cleavage site and correcting (dividing by) the observed proportion of nucleotides observed at that location in the human genome. (Bembom O (2017). *seqLogo: Sequence logos for DNA sequence alignments*. R package version 1.42.0.).

Modeling L1 integration

We created a weighted random dataset based upon the L1 EN degenerate consensus cleavage site. The consensus cleavage site was determined for the total 65,079 insertions from all four different cell samples. The frequency of occurrence of each of

the possible variations of the 7bp consensus cleavage site (5'-TTTTT/AA) observed within our four samples was determined. Due to our sliding of mapped insertions, favoring any poly-T stretch to the genomic location, we will never observe EN sites with a thymidine in the 6th position, which means we can only observe 12,288 EN sites (*e.g.* 5'-NNNNN/VN; N represents A, C, G, or T and V represents A, C, or G). Next we removed insertions from our dataset that mapped within clusters found in highly repetitive sequences in the genome such as centromeric or telomeric regions. These regions were deemed 'unmappable' due to their large repetitive tandem repeats as identified by the Tandem repeats finder (Benson, 1999). Likewise, removing these regions from the genome as well as gaps and unambiguous 'N' sequences, this dropped our genomic size to 5,669,914,180bp to pick from in the genome.

We determined the EN site for each of these positions in the genome. We applied the same 'T-sliding' approach in our genomic analysis as our mapping of insertion reads. Thus, some positions in the genome will never be picked, and instead their neighboring position will be picked twice or more (Figure 2.7A). We found that regardless of cell type, all insertions displayed a similar distribution of observed EN sites (Figure 2.2A). This similarity across cell types allowed us to create one model for all insertions. For every EN site position observed we calculated the insertion frequency as the number of insertions observed for an EN site of interest over the total number of observed insertions. Of the possible 12,2888 EN site variants, 11,545 EN sites had less than 3 observed integration sites at that sequence. The majority (97.14%) of insertions contain EN sequences in which there are 3 or more insertions which display the given EN site (Figure 2.2D). Thus, when picking sites in the human genome to model our insertion dataset we are only using our predicted model calculations for 3% of the total insertion dataset.

We created a 1-4-2 model in which we treated motif position 1 as independent from motif positions 2-5 which were grouped as one 4bp unit, which is also independent from motif positions 6-7 which were also grouped as one 2bp unit. We then calculated the position frequency for every A, C, G, or T base observed in motif position 1. For motif

positions 2-5 we calculated the frequency of every 256 possible 4-mer observed. And for positions 6-7 we calculated the frequency of the 12 possible 2-mers (T cannot ever be assigned to position 6). The modeled insertion frequency for an EN site in which we observed less than 3 insertions becomes the products of the position probability matrix (Figure 2.2E).

We weighted all EN sites by the most observed EN site, 5'-TTTTT/AA-3' (weights were the insertion frequency of EN site of interest divided by the 5'-TTTTT/AA-3' insertion frequency) (Table 2.5). These weights were then applied when randomly picking sites from the genome so as to mimic the same number of occurrences of EN sites observed in our actual insertion data.

These weighted probabilities were applied to every base pair in the human genome reference (GRCh37/hg19) based upon the nucleotide's corresponding degenerate consensus cleavage site if an L1 insertion were to occur at that precise nucleotide. We then executed a weighted randomization with R, including all sites in the human genome in which a consensus cleavage site could be determined (gap areas and consensus cleavage sites including an ambiguous "N" nucleotide were not included). We then repeated picking sites in the genome based upon this model for 10,000 iterations.

A 'corrected' weighted random model was also created in which observed insertion frequencies were 'corrected' or divided by the observed frequency of the 7mer in the genome. This model accounts for the frequency of the EN site found in the genome, thus 'correcting' for EN cleavage sites that are found more often than others in the genome. This model was also used in data analysis comparisons.

Total RNA isolation

Total RNA was extracted from confluent H9 hESC and H9 hESC derived NPCs using Trizol (Invitrogen) and following the manufacturer's instructions. Total RNA was

extracted from confluent HeLa and PA-1s using an RNeasy Mini kit (Qiagen) following the manufacturer's instructions.

RNA-seq Analysis

Two biological replicates of total RNA was isolated from H9 hESC, NPCs, HeLa and PA-1s each. RNA-sequencing libraries were created at the University of Michigan Sequencing Core. Total RNA samples were first filtered with an Illumina Ribo-Zero rRNA removal kit (#MRZH116). Immediately samples underwent the Illumina TruSeq Stranded mRNA Library Prep Kit (#RS-122-2101) following the Low Sample Protocol and beginning prep at the elute-prime-fragment step. A 1-minute fragmentation was performed to generate a 190bp target insert size and 12 cycles of PCR were performed instead of 15. One biological replicate for each of the four cell types was performed with a 100bp paired-end HiSeq sequencing run at the University of Michigan Sequencing Core. A second biological replicate was performed with 125bp paired-end sequencing, as the 100bp paired-end sequencing kit was no longer available for the second biological replicate run. For each sequencing run, all four samples were multiplexed and run on a single Illumina HiSeq Sequencing cell. Details of sequencing results and read counts from runs can be found in Table 2.8.

RNA-sequencing reads were aligned to the GRCh37/hg19 (GENCODE genome version 19) with Tophat version 2.1.1 and ENSEMBL GRCh37/hg19 transcripts (Aken et al., 2017; Trapnell et al., 2009). The Cufflinks Suite version 2.2.1 was utilized to run Cufflinks with ENSEMBL GRCh37/hg19 transcripts to get assembled transcripts and isoforms (Roberts et al., 2011; Trapnell et al., 2010). Cuffmerge was then performed to get the final transcriptome (ENSEMBL) assembly. Cuffquant was used to quantify gene and transcript expression, and then cuffnorm was used to merge biological replicates and normalize all samples to the same scale for comparison.

Tophat settings: ./tophat --library-type fr- firststrand –GTF
chr_hg19_ENSEMBL_genes.gtf --num-threads 4

Cufflinks settings: ./cufflinks -b GRCh37.p13.genome.fa -u --library-type fr-firststrand --max-bundle-frags 10000 --GTF-guide chr_hg19_ENSEMBL_genes.gtf --label PA1 --num-threads 4 accepted_hits.bam

Cuffmerge settings: ./cuffmerge -g chr_hg19_ENSEMBL_genes.gtf -s hg19.fa -p 4 assembly_GTF_list.txt

Cuffquant settings: ./Cuffquant -p 4 -b hg19.fa -u merged.gtf accepted_hits.bam

Cuffnorm settings: ./cuffnorm -L PA1 -p 4 --library-type fr-firststrand --library-norm-method geometric --output-format cuffdiff merged.gtf abundances.cxb

With the cuff norm data output we first divided the genome into those regions of the genome expressed (FPKM > 0.3) vs. not expressed. Significance was determined using the χ^2 test and comparing sample counts versus the median of the weighted random dataset.

We also divided expressed regions of the genome into 30 (roughly same number of base pairs of the genome) bins and compared the counts of insertions in each bin (Table 2.10) as compared to the weighted random dataset simulations containing the same number of insertions as found in expressed regions of the genome (Sample: number of insertions in expressed regions; HeLa: 6,614; PA-1: 6,125; NPC: 3,379; hESC:1,660). Significance was determined using the χ^2 test. For each simulation we compared the observed amount of insertions in our sample within a given expressed bin, versus the simulated counts of insertions and performed the χ^2 test. Of the 10,000 iterations we then determined the proportion of iterations where the χ^2 test p-value was below a given threshold.

Transcriptional Analysis

Bru-seq experiments and initial transcription FPKM expressions were performed as previously published (Paulsen et al., 2014). The Bru-Seq generated bed files of 1kb

segments were analyzed to determine transcribed regions of the genome in both PA-1 and HeLa cells. We utilized the web base interface, MIBrowser to help determine RPKM threshold of transcription. The RPKM threshold was determined by identifying regions of the genome in which there are clear defined differences in expression between intragenic and intergenic regions of the genome. For PA-1, regions of the genome in which the combined RPKM expression of both the top and bottom strand was above 0.024 were classified as actively transcribed. For HeLa analysis a combined top and bottom strand RPKM expression threshold above 0.022 was used to define transcribed regions of the genome. About 35.3% of the PA-1 genome is transcribed, while 34.1% of the HeLa genome is transcribed. We compared the count of insertions within transcribed regions vs. non-transcribed regions of the genome of the PA-1 and HeLa samples to 10,000 iterations of weighted random datasets that contained the same amount of randomly picked sites as in their respectively compared sample insertion dataset (Figure 2.4B). Significance was determined using the χ^2 test and comparing sample counts versus the median of the weighted random dataset.

We also divided the transcribed regions of the genome into 30 roughly equally sized bins, where each bin represents a different range of RPKM transcription expression (Table 2.9). In PA-1 cells we identified 5,391 insertions in transcribed regions of the genome. We then performed 10,000 iterations of selecting 5,391 sites within transcribed regions of the genome under the weighted random dataset probabilities. These sites, and the corresponding counts of sites within the 30 different transcriptionally expressed bins were compared to the observed PA-1 insertion counts within these bins and significance was determined by χ^2 test (Figure 2.16A). Sample insertion bin counts were determined to be significant if at least 95% of the 10,000 iterations resulted in a χ^2 test p-value of less than 0.05, 99% of the 10,000 iterations for a χ^2 test p-value less than 0.01 *etc.* Likewise, the above analysis was performed for the HeLa transcriptional dataset, except 6,617 sites were randomly selected based on the weighted random dataset, representing the same amount of HeLa insertions found within transcribed regions of the genome.

We also divided the transcribed regions of the genome based upon transcriptional bias. The transcriptional bias was calculated as the absolute value of the ratio of the difference in RPKM expression between the top and bottom strand of a site in the genome over the total RPKM expression of both strands of that site in the genome (Figure 2.16B). Transcriptional bias ratio values were sorted into respective perl integer expression functions resulting in 11 bins from 0.0 to 1.0. For every transcriptional bias bin we counted the total number of insertions that fell within the respective bin, and then the fraction of insertions that inserted into the template strand. We plotted counts of the sample insertions versus the weighted random dataset (Figure 2.4D). To determine significance we tested each simulation count versus the observed insertion count and performed χ^2 test. If 99% of the 10,000 iterations showed a χ^2 test p-value less than 0.01 then the χ^2 test p-value was determined to be < 0.01. We performed this comparison all the way to a p-value <0.00001.

Okazaki Sequencing Analysis

Previously published Edu labeled HeLa and GM06990 Okazaki sequencing (OK-seq) data was downloaded from the NCBI Sequence Read Archive under accession number: SRP065949 (Petryk et al., 2016). Two independent Okazaki-sequencing (OK-seq) datasets were published for HeLa (Spearman's rho of 0.962 between the two sets) and two independent datasets for GM06990 lymphoblastoid cell line (Spearman's rho of 0.954 between the two sets). The two independent cell types, HeLa versus GM06990 lymphoblastoid cell line, when compared to each other show a Spearman's rho of 0.61.
We applied additional smoothing parameters on the data to determine replication fork direction (RFD) and RFD slope values. All insertion datasets were compared to both replicates for the HeLa and GM06990 OK-seq data, and the trends observed were the same in all comparisons analysis made. Figures are shown for one of each of the replicates.

Like transcription data, we plotted both cumulative distribution function (CDF) plots with the data. We also plotted boxplots involving the simulated uncorrected weighted model and our observed insertion dataset with the genome divided into 11 bins based upon

|RFD| values from zero to one. Significance in boxplots was determine using the χ^2 test and testing each of the 10,000 iterations of the weighted random model versus the actual observed insertion data value. We then determined the proportion of iterations in which a significant χ^2 test p-value was observed, to determine the significance of the reported p-value.

Kolmogorov-Smirnov Boostrap test

The Kolmogorov-Smirnov Bootstrap test [ks.boot R function; 10,001 boot iterations] was performed to determine if the observed insertion dataset differed significantly from the corrected model as compared to the simulated insertion dataset difference from the corrected model. This test was performed for all datasets involving a cumulative distribution function (CDF) plot.

Chromatin States Enrichment Analysis

Mnemonics chromatin state bed files (15-state and 18-state) published by the Roadmap Epigenomics Consortium were downloaded (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final) (Roadmap Epigenomics et al., 2015). Specifically, for the 15-state model we downloaded the mnemonics bed file for E008 – H9 hESC cells, E009 – H9 hESC derived neuronal progenitor cultured cells, E010 – H9 hESC derived neuron cultured cells, E065 - Aorta, E066 - Liver, E117 – HeLa-S3 cervical carcinoma, E118 – HepG2 hepatocellular carcinoma, and E123 K562 leukemia.

For the 18-state model, we downloaded the mnemonics bed file (http://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/core_K27ac/jointModel/final/) for E003 – H1 hESC cells, E007 – H1 hESC derived neuronal progenitor cultured cells, E008 – H9 hESC cells, E065 - Aorta, E066 - Liver, E117 – HeLa-S3 cervical carcinoma, E118 – HepG2 hepatocellular carcinoma, E123 K562 leukemia. The E065 and E066 files were downloaded as negative controls. Since there is no 18-state model for female H9 hESC-derived NPCs we used male H1 hESC derived NPCs as a supplement.

As a positive control of strong enrichment and depletion we downloaded the MLV integration events in HepG2 and K562 from LaFave et al. 2014 (LaFave et al., 2014).

The Genome Structure Correction tool was utilized to determine enrichment or depletion of insertions in the respective chromatin states (https://github.com/ParkerLab/encodegsc) (Bickel et al., 2010). The following settings were used after identifying the r and s value that resulted in the least over dispersion: /block_bootstrap.py -1 Sample_insertion.bed -2 Chromatin_State_mnemonic.bed –d Ungapped_reference.bed –r 0.20 –s 0.15 –n 10000 –t rm –B –v. Enrichment was then calculated for insertions in each individual state and a heatmap was created using the ggplot R function. States that covered a very small proportion of the genome and result in observance of less than 30 expected insertions, were masked as grey boxes in the heat plot as these states contain too small of a sample set to determine a true enrichment or depletion.

Chromosomal Ideogram

Chromosomal ideograms (Figure 2.3B and Figure 2.14) were created using PhenoGram from the Ritchie Lab at PennState.
(http://visualization.ritchielab.psu.edu/phenograms/plot)

DAVID

We used DAVID Bioinformatics Resource 6.8 (Huang da et al., 2009a, b) to determine if NPC insertions were enriched in genes associated with neurons. We uploaded the list of genes using the human genome as the default background with the highest stringency settings.

Acknowledgements

Much thanks to Dr. Ángela Macia and Ms. Laura Sanchez for performing retrotransposition assays in NPCs and hESCs as well as isolating gDNA from performed experiments and isolating RNA from these cell lines for RNA-seq. Many thanks to Dr. José L. García-Pérez for his resources and support for this project. A very special thanks to Mrs. Shanda Joe for discussion and suggested improvements to library construction and helping to coordinate use of Covaris for shearing of gDNA. Thanks to Michelle Paulsen for performing Bru-seq on PA-1 cells and to all in Dr. Mats Ljungman lab who compiled the Bru-seq data to the final form ready for analysis. Special thanks to Dr. Brian Manguson in the Ljungman lab for answering technical questions on the Bru-seq data. Thanks to Dr. John Moldovan, Dr. Peter Larson, and Mr. Owen Funk for isolating RNA from HeLa and PA-1 cells for RNA-seq experiments. Special thanks to Mr. Owen Funk for arranging the 1st round of RNA-seq experiments with the University of Michigan Sequencing Core. Thanks to Dr. Robert Lyons and Dr. Christina McHenry at the University of Michigan Sequencing Core for PacBio sequencing sample library preparation and running the PacBio sequence runs. Thanks to Dr. Saurabh Agarwal, in the laboratory of Dr. Shigeki Iwase, for helpful discussions and input on RNA-seq data analysis. To Dr. Stephen Parker and Ms. Arushi Varshney much gratitude for discussions and guidance on analysis of chromatin state data. Lastly, to Dr. Mitsu Nakamura and Dr. Amanda Pendleton, I greatly appreciate your input on figures and text for this chapter.

Figure 2.1: Generation and identification of *de novo* engineered L1 retrotransposition events.

A) General schematic of engineered L1 plasmid constructs to generate L1 retrotransposition events: All four cell types were transfected with episomal plasmids containing an engineered L1 with retrotransposition indicator cassette located within the 3'UTR (represented by green rectangle labeled 'REP', short for 'Reporter'). Reporter cassettes are driven by an upstream promoter (small black arrow). The coding sequence of the reporter cassette is in the opposite orientation of the respective L1 sequence, disrupted by an intron (indicated by black lines in 'v' shape), and followed by a polyadenylation signal (black lollipop). The intron is in the same orientation of the L1 sequence (SD: splice donor site; SA: splice acceptor site). The 'LEAP' (orange) sequence is downstream of the reporter cassette, which is present in these engineered L1 constructs, but is not present in the sequence of endogenous L1 sequences. If the engineered L1 construct is transcribed and the intron is spliced from the reporter cassette, expression of the respective reporter cassette occurs only after successful reverse transcription and integration of the engineered L1 into a new genomic location.

B) Selection or screening of engineered L1 retrotransposition events: (Top) T-175 flask of HeLa L1 retrotransposition events. Proportion of EGFP expressing cells, indicative of *de novo* engineered L1 retrotransposition events, from representative retrotransposition assay in PA-1 cells (Untransfected: Untransfected PA-1 cells, ORF1p Mutant: LRE3 ORF1p mutant pCEP4/JM111/LRE3-*mEGFPI* transfected PA-1 cells, Wild-type L1: wildtype LRE3 pCEP4/LRE3-*mEGFPI* transfected PA-1 cells).

C) Capture of engineered L1 retrotransposition events: Partially double-stranded adapters (red rectangles) are ligated onto end-repaired, dA-tailed sheared genomic DNA collected from cells following completion of the retrotransposition assay. The adapters contain a 3' ammine (indicated by the red asterisk) on the shorter strand to prevent the amplification of spurious genomic DNA by the presence of excess single-stranded adapters. Adapter ligated genomic DNA is subjected to a linear amplification utilizing a dual-biotinylated LEAP primer, a sequence specific to our engineered L1 construct (orange arrow with circle on 5' end indicating biotin). Biotinylated products are streptavidin bead captured (grey circle) and subjected to nested PCR utilizing a primer specific to the SV40pA primer sequence (black arrow) and a primer specific to the adapter sequence (red arrow). Following successful amplification, PacBio CCS SMRTbell adapters (navy dumbbells) are ligated onto products and subjected to PacBio CCS sequencing.

D) Amplified 3' L1 and flanking gDNA: PCR amplified products were run on a 1.0% agarose gel. + Control: PC39, a clonal PA-1 cell line with three known engineered L1 integration events, genomic DNA is used as a positive control. The three distinct bands represents 3 known engineered L1 integration sites in PC39 (pc-39-C ~1150bp; pc-39-B ~580bp; pc-39-A ~330bp). 1 H_2O Control: Uni-linear PCR H_2O template control. 2-1 H_2O Control: PCR2 using uni-linear amplified H_2O control as template. 2 H_2O Control: Nested PCR with H_2O as template. PA-1: amplified library of gDNA from PA-1 cells transfected with pCEP4/LRE3-*mEGFPI* and subjected to capture techniques. The

shmear represents hundreds of successfully amplified retrotransposition events and their flanking 3' gDNA. Ladder: 1kb Plus DNA Ladder (ThermoFisher Scientific).

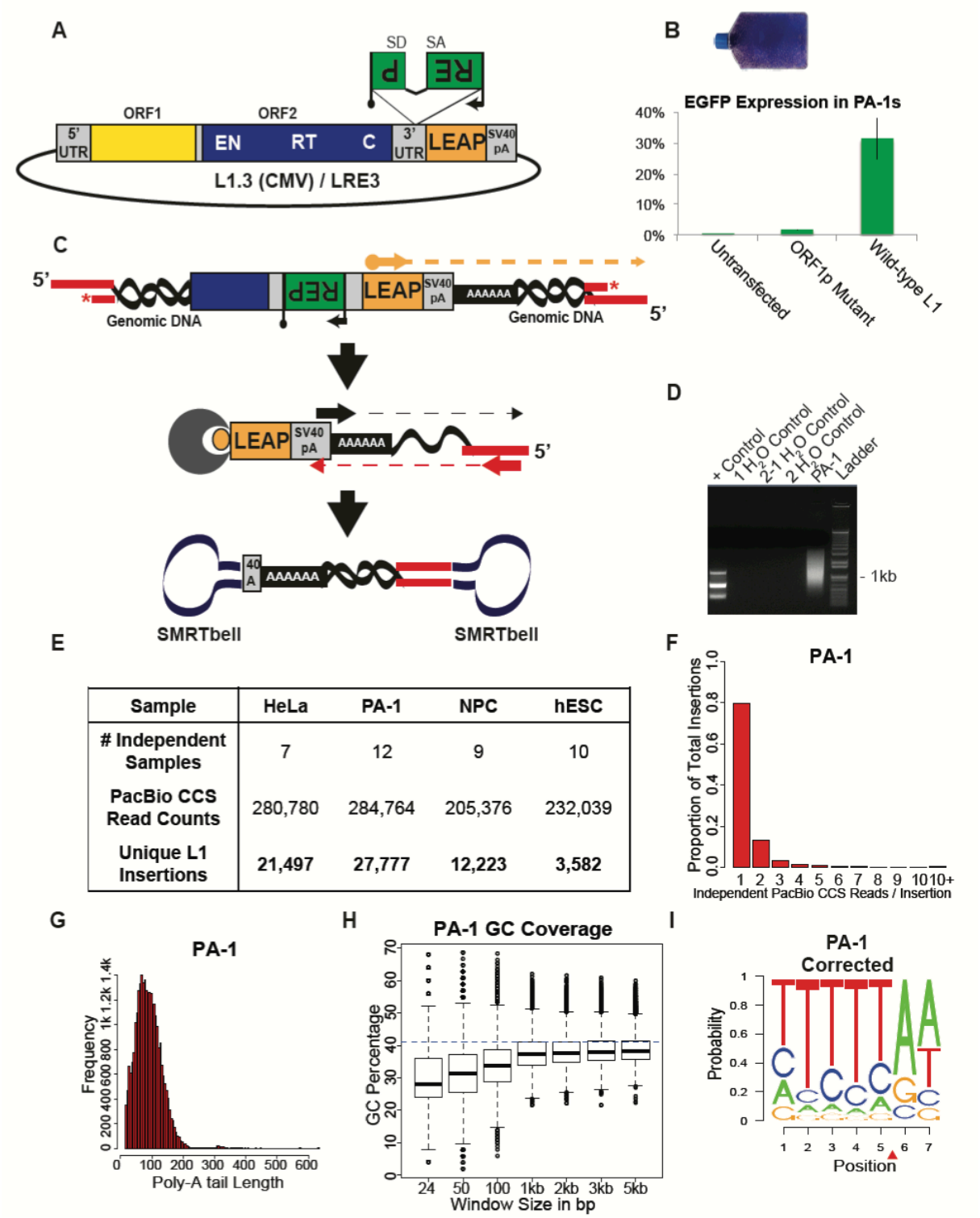
E) Identification of de novo L1 insertions: Independent biological replicates (2nd row) of retrotransposition assays were performed and subjected to the above capture, amplification, and PacBio CCS sequencing. The total number of CCS read counts are provided for each sample (3rd row), and the number of unique L1 insertions identified in each sample (4th row) are shown after filtration.

F) Independent PacBio CCS reads support PA-1 Insertions: The proportion of total PA-1 insertions (y-axis) that are represented by a number of independent CCS reads (x-axis).

G) Engineered insertions display long poly(A) tails: The frequency of PA-1 insertions (y-axis) with the corresponding poly(A) tail length shown in base pairs (x-axis).

H) Engineered insertions are found within AT-rich regions of the genome: Different lengths of genomic sequence upstream and downstream of L1 insertion sites were examined for GC content. A window size of 24bp indicates that 12bp upstream and downstream of the L1 integration site was examined. The blue dotted line represents the average human GC content of 40.94%. Engineered insertions from all four cell types show this same trend as shown here for the PA-1 engineered insertions (Figure 2.7F).

I) Engineered insertions display the degenerate endonuclease consensus cleavage site: A logo plot depicting the 7bp degenerate EN consensus cleavage site of PA-1 insertions.



Sample	HeLa	PA-1	NPC	hESC
# Independent Samples	7	12	9	10
PacBio CCS Read Counts	280,780	284,764	205,376	232,039
Unique L1 Insertions	**21,497**	**27,777**	**12,223**	**3,582**

Figure 2.1: Generation and identification of *de novo* engineered L1 retrotransposition events.

Figure 2.2: A weighted random model based upon L1 EN degenerate consensus cleavage site.

A) All four cell-types show similar degenerate EN consensus sequence distributions: Observed frequency (y-axis) of the 12,288 possible 7bp degenerate EN consensus cleavage sites (x-axis) among the four different cell types. The distribution of observed 7mer variants for the top 80% of the insertions in each sample set is plotted for each cell type individually (respective colors) and the total set of insertions combined (black).

B) Some base pair positions of the L1 EN consensus cleavage site are dependent upon each other: After observing all EN cleavage sequences of insertions in which the 5^{th} nucleotide did not contain a T (top left logo plot), we observed a stronger preference for T in nucleotide positions 2 through 4, but no change in position 1 or positions 6 and 7. Likewise when we select all EN sequences that contain a C in the 5^{th} nucleotide position (top right logo plot) or when we condition for the presence of a T or C in the 3^{rd} position (bottom two logo plots) we observe the same trend. Thus, conditions made upon nucleotides 2-5 do not change the preference of position 1 or of positions 6 and 7. These logo plots indicated to us the need for a 3 group independent model of 1-4-2 where position 1 is a 1bp unit independent from positions 2-5, a 4bp unit, which is independent from positions 6-7, a 2bp unit.

C) Positions 2-5 of L1 EN consensus cleavage site are dependent upon each other: The total percentage of insertions from all four cell-types (column 2) that contained a T or C nucleotide within positions 2 through 5 of the L1 EN consensus cleavage site. N represents any nucleotide of A, T, C, or G. V represents any nucleotide except T. Since we "slide" insertions to determine base-pair resolution, we never observe a T in the 6^{th} position.

D) Predicted weights from model are less than weights from observed frequencies for EN sites with insertion counts less than 3: Each point represents one of the 12,288 possible EN consensus cleavage sites. Plot of $\log_{10}$ weights for an EN site when calculated from the position probability matrix **(**x-axis) vs. $\log_{10}$ weights for an EN site calculated when using the actual observed insertion frequencies (y-axis). Light grey line represents 1 to 1 weights in which EN sites that are given the exact same weight by both approaches. Points to the left of the grey line indicate EN sites that are weighted less by the model than by using actual observed frequencies. Points to the right of the grey line indicate EN sites that are weighted more by the model than by using the actual observed frequencies. Red points represent EN sites supported by less than 3 insertions (red) and blue points EN sites observed by 3 or more insertions. Thus, a threshold of 3 or more insertions at an EN site was created. With this 3 insertion threshold we use the model for 11,454 EN 7mer sites, 93.2% of total 7mers, and model 1,860 insertions, only 2.86% of the total insertions. [Note, in order to distinguish points we used the R jitter() function with a factor value of 1 for y-axis values; for weights which had a value of 0 we assigned a value of $1x10^{-10}$].

E) Weighted random model schematic: We created a weighted random model based upon the observed 7pb L1 EN consensus cleavage site. All 7bp iterations of the EN sequence site were determined, with the exception that we will never observe a T in the

6^{th} position because we always favor a T to the human genome reference. Thus, this leaves 12,288 possible 7mer combinations. For each EN site observed we calculated the insertion frequency of that site ($IF_{7mer,obs}$) as the number of insertions at that site ($N_{ins,7mer}$) over the total number of insertions (N_{ins}). Insertion counts from all four cell-types determined the insertion site calculations. For EN sites in which we observed less than 3 insertions we determined the insertion site frequency as the probability of the expected frequency from the position probability matrix determined from observed EN sites. We then weighted the EN sites based on the EN site that had the most insertion counts. These weights were then applied to each respective position in the genome based on the corresponding 7mer and sites in the human genome were then randomly picked based on these weights. We repeated picks for 10,000 iterations.

F) Weighted random model sites display L1 EN degenerate consensus cleavage site: As proof of principle we collected sequence surrounding sites chosen by our weighted random model algorithm and created a logo plot. The logo plot displays that our model is selecting degenerate L1 EN consensus cleavage sites in the human genome, mimicking our insertion dataset.

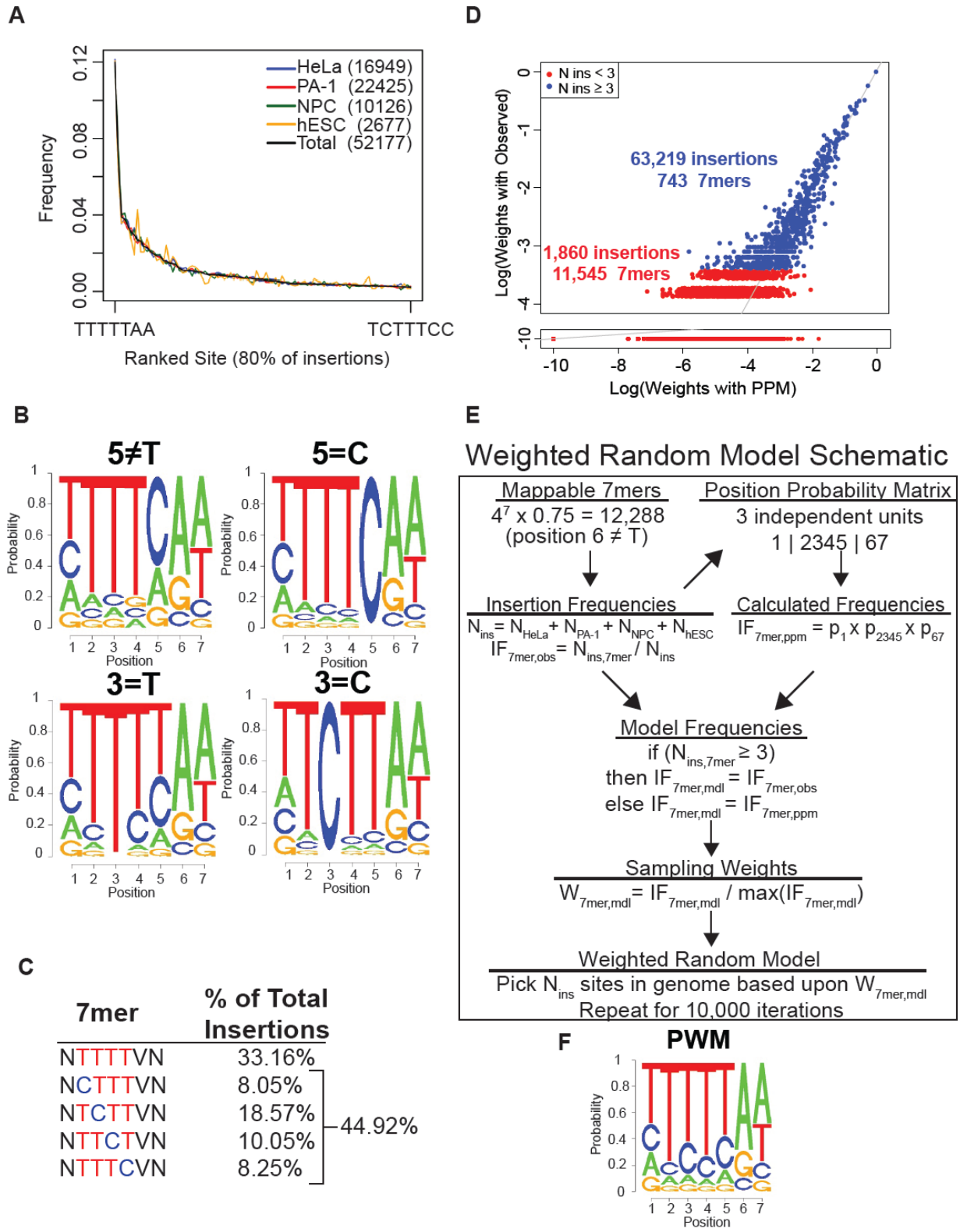


Figure 2.2: A weighted random model based upon L1 EN degenerate consensus cleavage site.

Figure 2.3: LINE-1 is dispersed throughout the human genome.

A) Chromosome insertion counts are correlated with chromosome size: Plots display the shortest chromosome to the largest chromosome from left to right along the x-axis, and the corresponding counts of PA-1 insertions (red circles) per chromosome on the y-axis. Boxplots represent the range of insertion counts observed for the 10,000 iterations of the weighted random model. In PA-1s (red circles), the X chromosome contains more insertions than expected (Outlier test, Bonferroni correction p-value: 0.0046).

B) Engineered insertions are dispersed throughout the human genome: Chromosomal ideogram (from PhenoGram from the Ritchie Lab at PennState) depicting all PA-1 insertions with respect to their mapped location on the corresponding chromosome. Each red horizontal line is a single engineered PA-1 insertion site.

C) L1 integrates into exons: Boxplot of the range of observations from the 10,000 iterations of the weighted random dataset for each corresponding sample within exons. Exons (including 5'UTR, coding exon, and 3'UTR) are defined by the UCSC genome browser's definition. Most samples contain the expected amount of insertions in exons, with PA-1s showing a depletion of insertions within exons (χ^2 test p-value: 5.782×10^{-8}).

D) L1 integration within introns of genes: Boxplot of the range of observations from 10,000 simulations of the weighted random iterations found within introns for each sample. Intron boundaries are defined by the UCSC genome browser's definition. HeLa, PA-1, and NPC insertions all show a depletion of insertions in introns (χ^2 test p-value: 1.48×10^{-9}, $<2.2 \times 10^{-16}$, $<2.2 \times 10^{-16}$ respectively).

E) A schematic of L1 integration into genes: A diagram showing an L1 insertion antisense and sense with respect to the expressed gene (green arrow). The 'top' strand is the coding sequence of the gene. An antisense insertion preference shows a bias of the EN consensus site on the coding strand in the genome.

F) The L1 degenerate EN consensus cleavage site drives insertions towards antisense insertion bias within genes: Utilizing the UCSC Genome Browser annotation of exon (including 5'UTR, coding exon, and 3'UTR) and intron, we considered insertions within an exon or an intron to be within a gene. We calculated the proportion of antisense to sense genic insertions as the number of genic insertions that are in the opposite (antisense) orientation with respect to the expressed gene, over the number of insertions that are in the same (sense) orientation with respect to the expression of the gene. The boxplots show the range of expectations from the 10,000 iterations of the weighted random dataset, showing a median preference of antisense: sense ratio of greater than one. All insertions, except hESC insertions, which show a stronger antisense preference (χ^2 test p-value: 0.00272), are within the expected range of antisense: sense ratio.

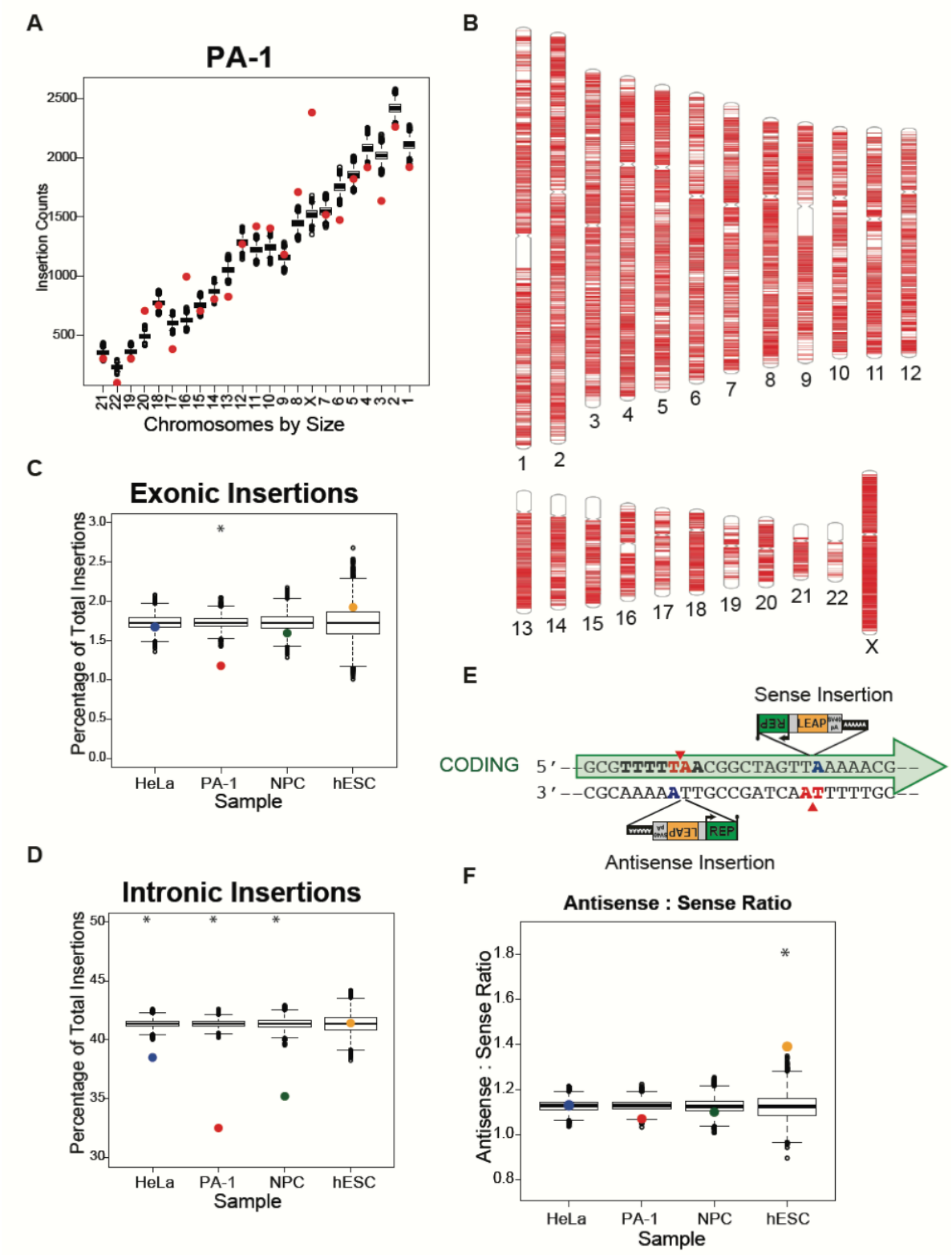


Figure 2.3: LINE-1 is dispersed throughout the human genome.

Figure 2.4: LINE-1 does not target transcribed regions of the genome.

A) Schematic of L1 preferentially targeting the coding strand during transcription: (Top) Schematic of the hypothesize preferential cleavage of L1 EN ORF2p (blue circle) activity at an EN consensus cleavage site of the accessible coding strand during transcription, while RNA polymerase (depicted by green oval) occupies the noncoding strand. (Melander et al.) Bru-seq involves labeling of nascent transcripts in cells with bromouridine (BrU), capture of bromouridine transcripts, and sequencing of transcripts by Illumina sequencing. A visual of PA-1 Bru-seq data in the MIBrowser is shown below for an identified PA-1 insertion antisense of the transcribed *RAVER2* gene [green rectangle depicting 5' to 3' expression on top strand with positive RPKM expression values (blue line)]. *JAK1* (red rectangle), is a gene transcribed and expressed from the bottom strand and thus has negative RPKM expression levels (blue line).

B) Transcription deters L1 integration: Boxplot for the range of observations from the 10,000 iterations of the weighted random dataset observed in transcribed vs. non-transcribed regions of the genome. The blue squares in the top plot represent the observed HeLa insertion counts in transcribed vs. non-transcribed regions of the genome, while the red squares in the bottom plot depict the observed PA-1 insertions. In both samples, we observe significantly more insertions in non-transcribed regions of the genome and significantly less insertions in transcribed regions (χ^2 Test: HeLa p-value: 6.795×10^{-6}; PA-1 p-value: $< 2.2 \times 10^{-16}$).

C) L1 EN preferentially cleaves regions of low transcription levels in the genome: Cumulative distribution functions are plotted for position weighted modeled corrected (red), 10,000 simulated uncorrected weighted model iterations (gray) and actual insertion (blue) datasets, with increasing transcription levels on the x-axis. Both HeLa and PA-1 insertions differ significantly from the corrected model as compared to the simulated insertions when tested against the corrected model (Kolmogorov-Smirnov bootstrap test p-value $< 1 \times 10^{-6}$ for both PA-1 and HeLa plots). HeLa and PA-1 insertions display a faster accumulation of insertions in less transcribed regions of the genome.

D) L1 EN can cleave coding strand during transcription: Transcription bias was calculated as the rate of transcription for the top strand of a region in the genome minus the bottom strand of the same region in the genome over the combined (total expression) of the top and bottom strands (Figure 2.16B). The absolute value of this expression was plotted, and rates were divided into 11 bins from 0 to 1 in 0.1 increments. A transcriptional bias of 0 indicates equal transcription expression from both strands in the genome, while a bias of 1 means only one strand in the genome is actively transcribed. Since Bru-seq measures a population of cells, most regions in the genome have a transcription bias ranging from 0 to 0.8. Box plots represent range of observations observed from each transcriptional bias condition from weighted random iterations. Blue squares represent HeLa insertion observed values and red squares are PA-1 insertion values. In HeLa cells at transcription bias of 0.9 there is significantly more insertions than expected cleaved on the coding strand by L1 EN (χ^2 test p-value: < 0.05).

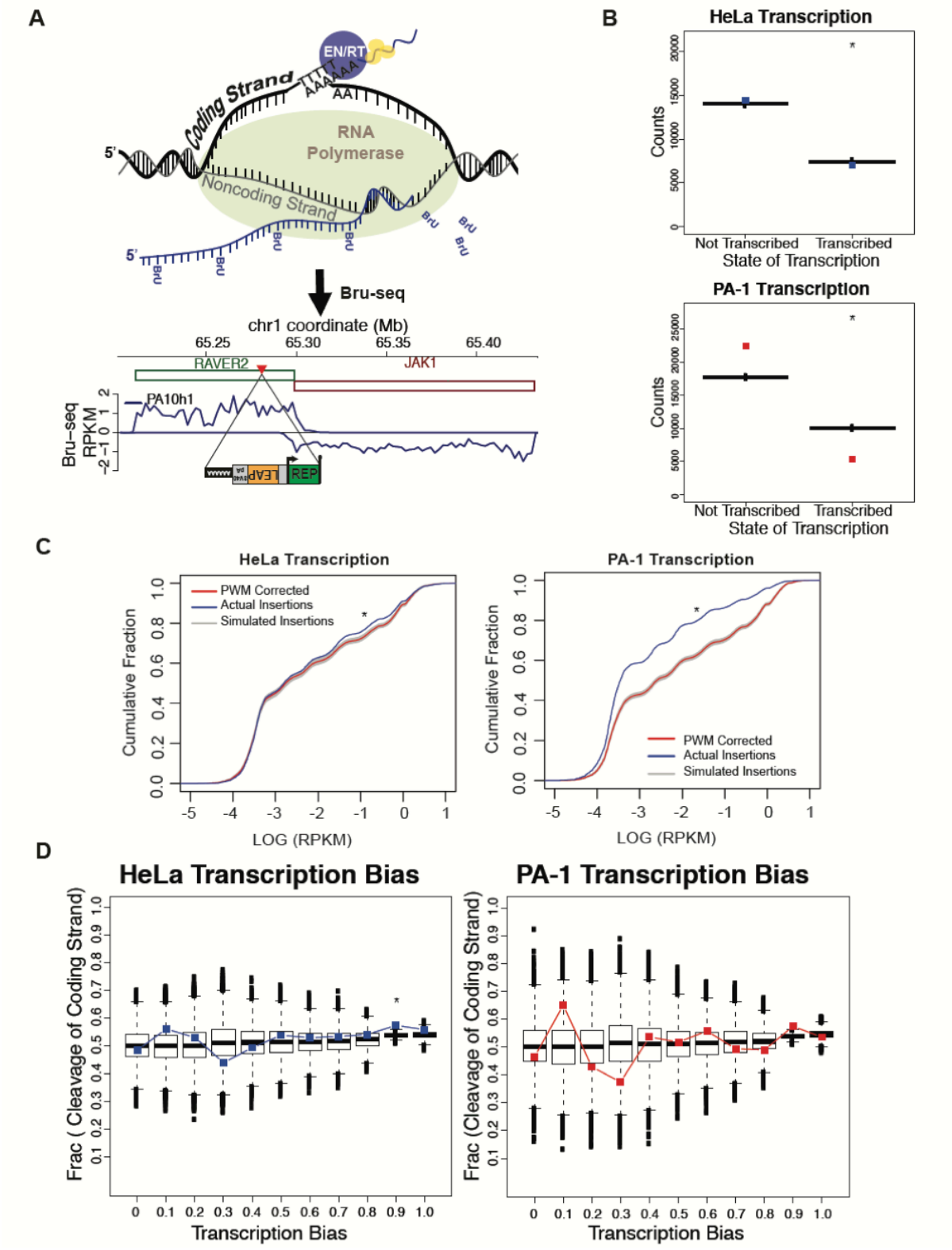


Figure 2.4: LINE-1 does not target transcribed regions of the genome.

Figure 2.5: L1 does not target a specific chromatin state in the human genome.

A) Schematic of L1 integrating throughout the genome: The Roadmap Epigenomics Consortium Hidden Markov Modeled 15 Chromatin State Set separates the genome into 15 different chromatin states, which can roughly be divided into four states: enhancer, promoter, transcription-related, or heterochromatic regions of the genome. We want to test if L1 integration is enriched in any one of these states.

B) L1 integrates throughout the genome, independent of chromatin state: Insertion sample sets were compared to the Roadmap Epigenomics Consortium Hidden Markov Modeled 15 Chromatin State Set. Each box represents fold enrichment of the insertion set across the other cell types. The most relevant cell line to compare to the given insertion dataset is on the leftmost end. Insertions were compared to every cell type tested, as well as the control datasets, E065_Aorta and E066_Liver. Gray boxes are those states in which there are too few insertions (less < 30 insertions) to be able to test for statistically significant enrichment or depletion. As a comparison of extreme enrichment and depletion we included the MLV integration dataset capture from the leukemia K562 cell line (LaFave et al., 2014).

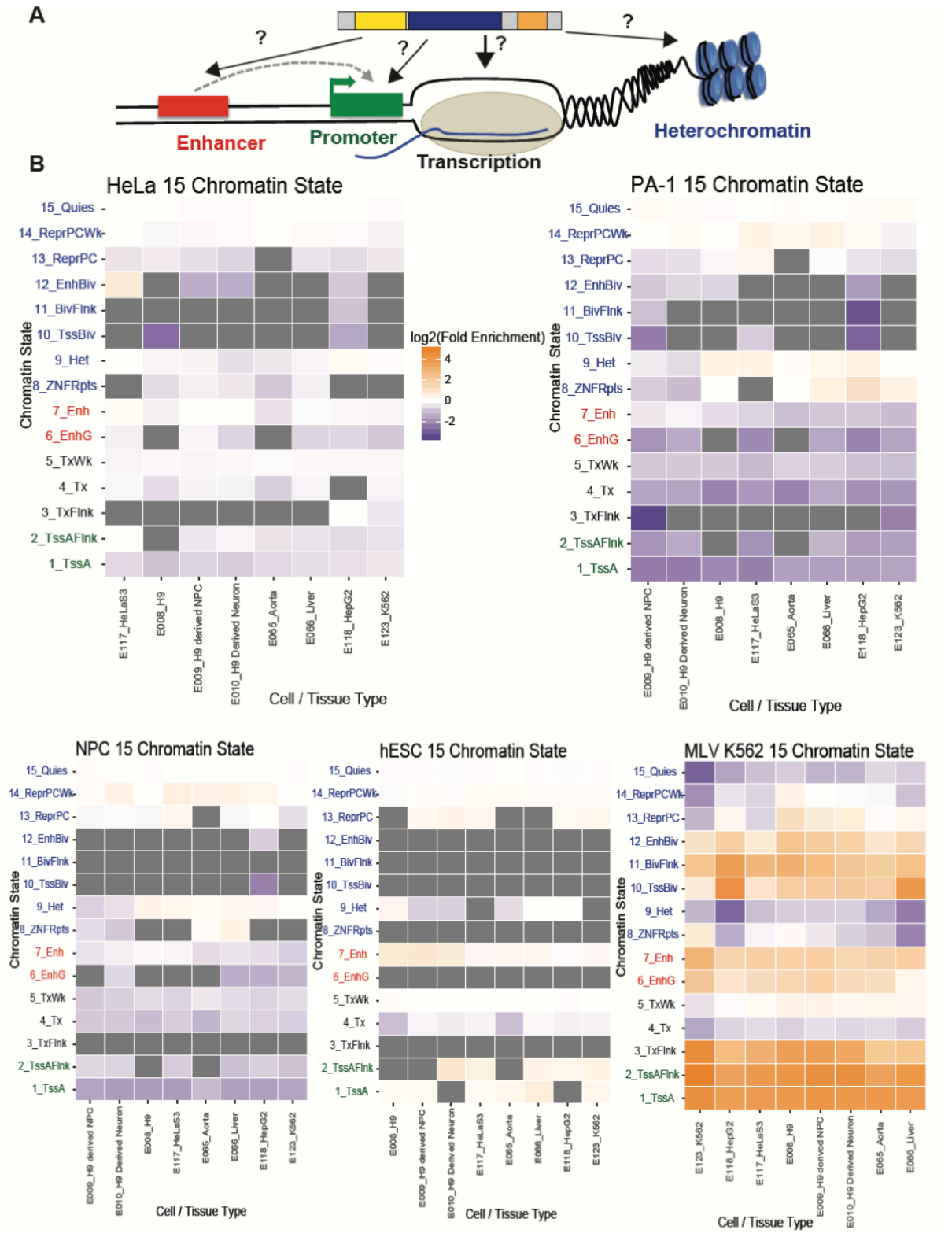


Figure 2.5: L1 does not target a specific chromatin state in the human genome.

Figure 2.6: Replication influences L1 integration in the human genome.

A) Examining the influence of replication on L1 EN cleavage and integration in genome: Depicted is a representation of a replication fork in the genome, where L1 EN has cleaved the lagging strand template (grey). We hypothesize that the lagging strand template is more accessible during replication, thus providing an opportunity for L1 EN cleavage and subsequent L1 integration. Shown below is a MIBrowser screen shot of representative HeLa OK-seq data from chr5 in which a sense L1 insertion was discovered in a region of the genome in which the replication fork was moving towards the left. This is a representative example in which the L1 EN cleaved the lagging strand template for integration to occur in the genome.

B) L1 endonuclease cleavage shows slight bias towards cleavage of lagging strand template during replication: Plots display the absolute value of the Replication Fork Direction (RFD) on the x-axis. Regions of the genome were binned into 11 bins with respect to the replication fork direction measurement of the region of the genome based upon the data published by Petryk *et al.* (2016). An RFD of 0 indicates replication occurring from the right and left, while an RFD of 1 indicates regions of the genome that are only replicated from one direction. Insertions that land within each region of the genome were plotted based upon the fraction of the insertions in each bin in which the L1 EN had to cleave the lagging strand template of replication for integration to have occurred. All datasets were compared to the lymphoblastoid replication fork direction data from Petryk *et al.* (2016), except HeLa, which was compared to the HeLa replication fork direction data. All comparisons were made with all datasets, and all showed the same trends observed here favoring cleavage on the lagging template strand. Box plots represent the distribution of the 10,000 weighted random model for the sample. There is a significant enrichment of HeLa insertions cleaved on the lagging strand template in bin 0.7 (χ^2 p-value: < 0.05). There exists a significant excess of PA-1 insertions in replication fork direction bins 0.3 through 0.8 (χ^2 p-value for bins 0.3 to 0.5: < 0.05; χ^2 p-value for bin 0.6: < $1x10^{-4}$; χ^2 p-value for bins 0.7 and 0.8: < $1x10^{-6}$).

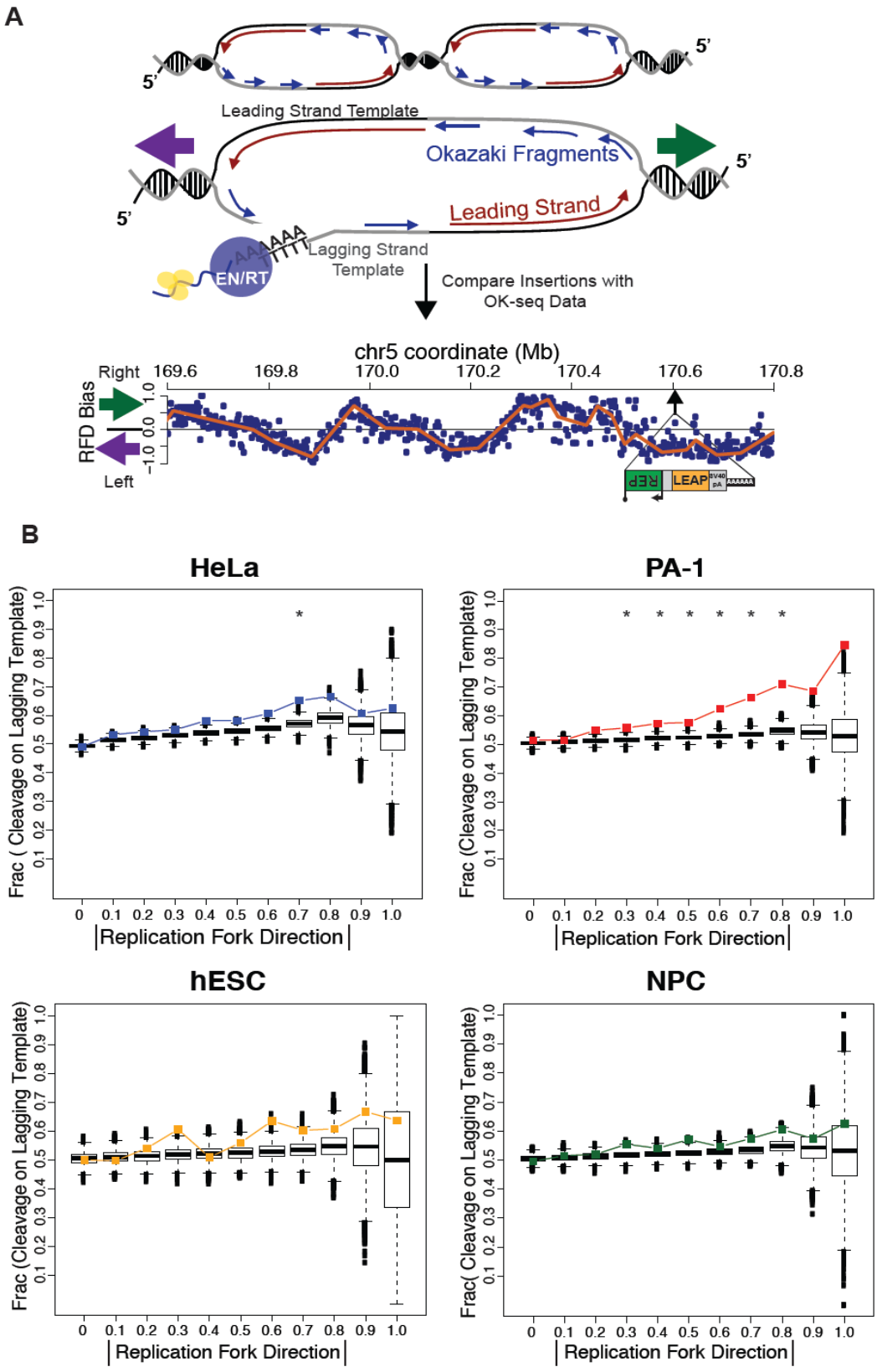


Figure 2.6: Replication influences L1 integration in the human genome.

Figure 2.7: Generation and identification of *de novo* engineered L1 retrotransposition events (Supporting Figure 2.1).

A) Filtering of de novo L1 insertions: Amplified products are sequenced by Pacific Bioscience's (PacBio) single-molecule real time (SMRT) circular consensus sequencing and reads are filtered for *de novo* L1 integration events. Gray box represents SV40pA primer sequence, black box with As represents a poly(A) tail, black wavy line represent 3' flanking gDNA, and red rectangle the adapter primer sequence. Reads that contain both expected primer sequences, as well as presence of a poly(A) tail of at least 15bp (top blue oval) are first clipped of their primer sequences and then mapped to the human genome (GRCh37/hg19 and GRCh38/hg38 via Bowtie2). Reads that are missing a poly(A) tail (orange circle) or lacking both expected primer sequences (yellow circle) are not mapped to the genome. The best mapped read location is chosen (blue oval) as the alignment that covers 96.5% of the total read, and the alignment with the highest alignment score difference (ASDIF) that is at least 20 points higher than the next best alignment. Bottom image: When mapping circular consensus sequences (CCS) to the genome we always favor any poly(A) stretch that may be present in the genome to the insertion call site as opposed to attributing the poly(A) stretch to the poly(A) tail. We classify the most 5' A in the poly(A) stretch the called/annotated insertion site (as shown by the 'A' marked in blue with the asterisk above). The cleaved EN consensus sequence is the sequence on the opposite strand, and EN cleavage occurs between the T and A marked in red, also indicated by the black triangle.

B) Proportion of CCS reads filtered: The initial number of total collected CCS reads are shown in parenthesis. These CCS reads are filtered for the presence of both primer sequences and presence of a poly(A) tail (blue). The proportion of CCS reads that contain both primer sequences, but lack a poly(A) tail, are shown in orange. Reads that do not contain both primers are shown in yellow.

C) Filtered CCS reads mapped to hg19 and hg38 reference: Pie charts detailing mapping states per Bowtie2 for CCS reads. The proportion of reads mapped uniquely to one genomic location are in blue, while multi-mapped reads with one unique location (defined from selecting mapped location with the highest alignment score difference (ASDIF > 20)) are in light blue. In yellow is the proportion of mapped CCS reads that cannot be called to one unique location, and the fraction of reads that did not map at all to the genome is shown in orange.

D) PacBio CCS read lengths: Frequency of insertions (y-axis) at observed CCS read lengths (x-axis). On average, we obtained ~600bp CCS reads.

E) Number of Insertions supported by Independent CCS Reads: Proportion of total insertions (y-axis) and the number of independent CCS reads that support insertions (x-axis). In order for two sequences to be considered independent CCS reads for the same insertion site, the 3' end of the sequence reads ligated to the adapter sequence must have different end sequences, but both reads must uniquely map to the same poly(A) tract/3' genomic sequence junction site in the genome.

F) Engineered L1 insertions display long poly(A) tails: Frequency of insertions (y-axis) with a given poly(A) tail length in base pairs (x-axis). We observe insertions with longer poly(A) tail lengths due to the strong SV40 polyadenylation signal in our engineered L1 constructs.

G) Degenerate EN consensus logo plots: Logo plots generated from the sequence directly surrounding identified L1 insertions. Insertions from all the examined cell types display the L1 EN consensus cleavage site (red triangles).

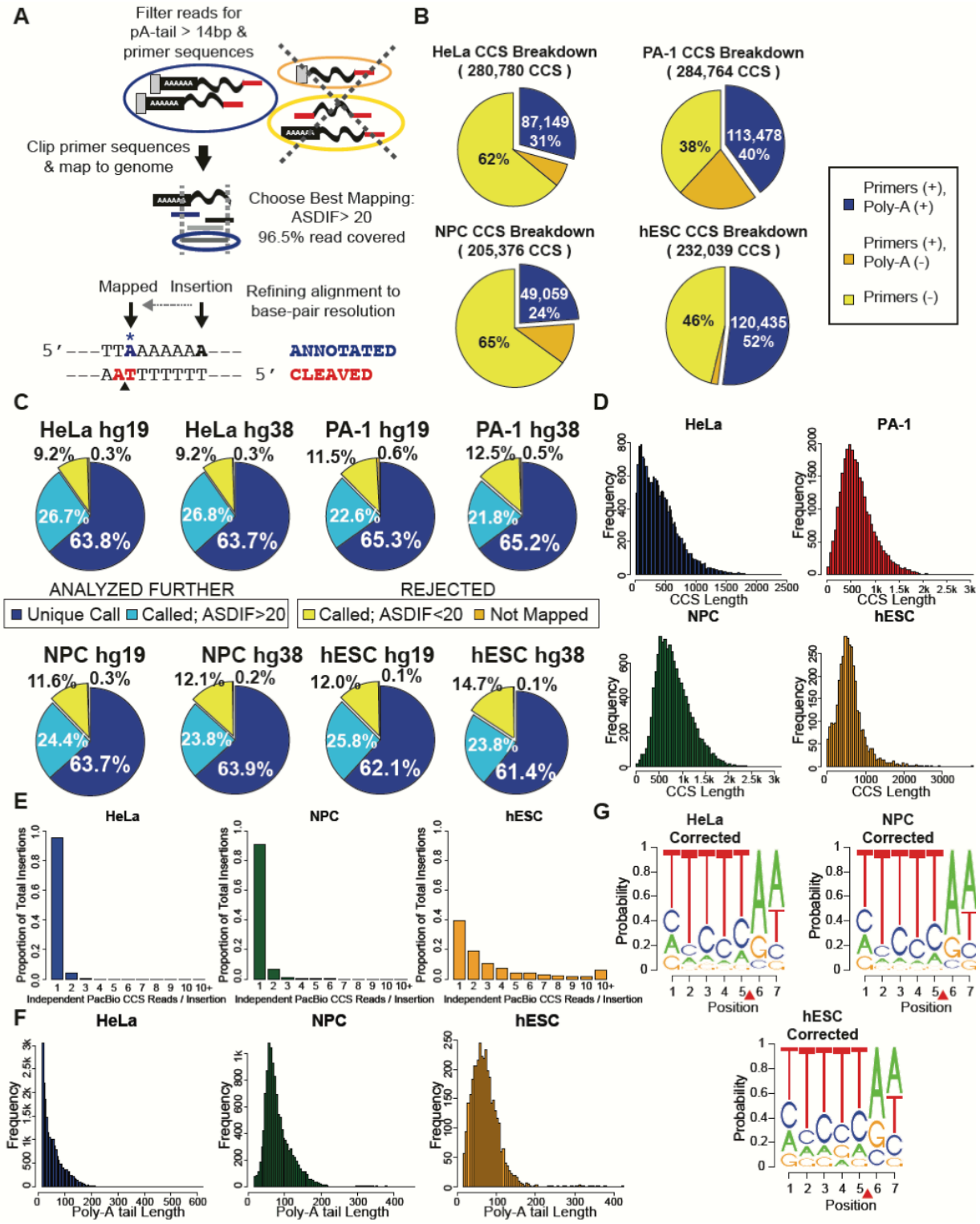


Figure 2.7: Generation and identification of *de novo* engineered L1 retrotransposition events. (Supporting Figure 2.1).

Figure 2.8: Engineered L1 plasmid constructs (Supporting Figure 2.1A).

Schematics of the labeled constructs used in engineered L1 retrotransposition assays. All constructs have a pCEP4 backbone, except pKUB102/L1.3-sv+, which consists of a pBSKS-11 backbone. In pJM101/L1.3 the L1.3 sequence is driven by a CMV promoter and contains a *mneoI* retrotransposition cassette driven by an SV40 promoter (black arrow labeled P'). In pCEP4/LRE3-*mEGFPI*, LRE3 is driven by it's native promoter in the 5'UTR and has an *mEGFPI* retrotransposition cassette driven by a CMV promoter. All constructs with LRE3 contain puromycin resistance. The pCEP99/UB-LRE3-*mEGFPI* construct contains LRE3 driven by an ubiquitin C promoter (pUBC) and the same *mEGFPI* construct described in pCEP4/LRE3-*mEGFPI*. In pKUB102/L1.3-sv+ the 5'UTR of L1.3 has been replaced with pUBC, and contains the same *mneoI* cassette described in pJM101/L1.3.

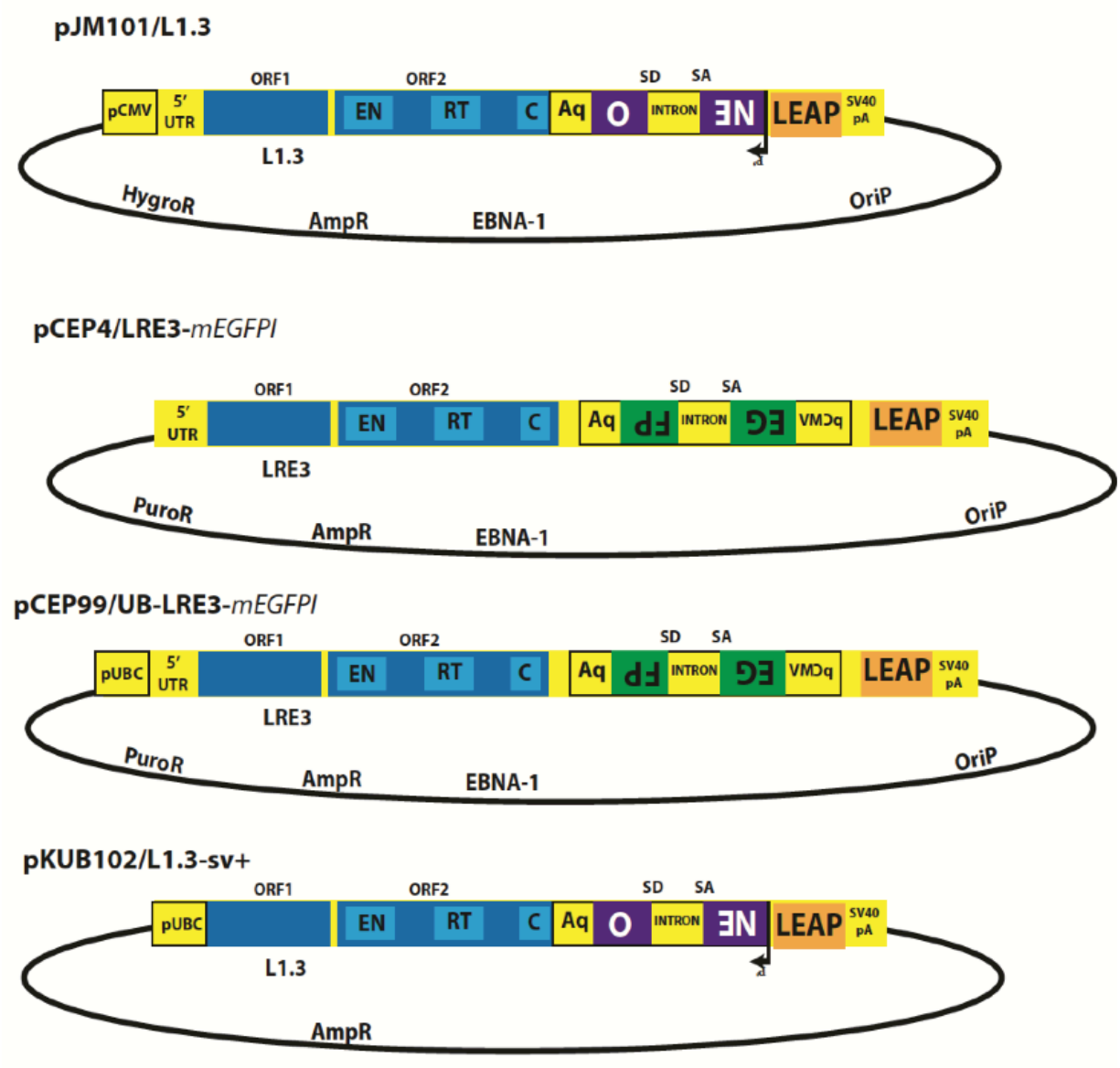


Figure 2.8: Engineered L1 plasmid constructs (Supporting Figure 2.1A).

Figure 2.9: Reporter cassette PCR verifying L1 retrotransposition (Supporting Figure 2.1B).

A) Schematic of the retrotransposition reporter cassette PCR: This PCR setup has been described previously in Moran *et al. (1996)* and Ostertag *et al. (*2000). Primers (magenta arrows) flanking the respective reporter cassette amplify both unspliced and spliced reporter cassette products.

B) *Genomic DNA collected after retrotransposition assays show presence of spliced reporter cassettes:* 493bp product found in HeLa cells transfected with pJM101/L1.3 and selected with G418 for retrotransposition events is observed (lanes labeled HeLa). PA-1 cells transfected with pCEP4/LRE3-*mEGFPI* were either selected for plasmid (PuroR) or screened for GFP expressing cells (GFP). Spliced GFP shows a band at 343bp. (H_2O: Water template for PCR conditions; - Control: respective plasmid template).

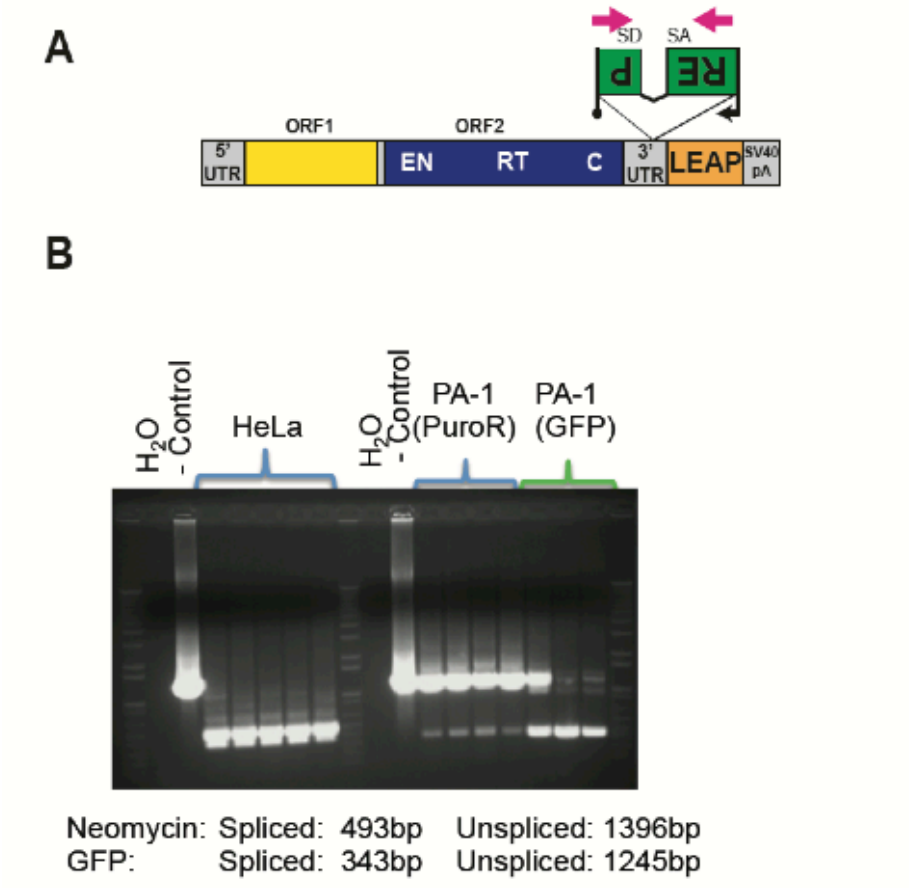


Figure 2.9: Reporter cassette PCR verifying L1 retrotransposition (Supporting Figure 2.1B).

Figure 2.10: pc-39-C characterization (Supporting Figure 2.1D).

A) Details of three PC39 insertions: PC39 is utilized as a positive control during library preparation techniques. Instead of randomly shearing gDNA from PC39 we double digest gDNA with *NdeI* and *PacI*. Pc-39-C and pc-39-B insertions contain a 3' flanking *NdeI* restriction site while pc-39-A contains a 3' flanking *PacI* site. We then prepare this digested DNA for adapter ligation and subject the library prep to the same capture and amplification techniques we used to discover our *de novo* L1 insertions. Shown here is the known poly(A) tail length, expected flanking gDNA length, and location of the restriction sites. The GRCh37/hg19 coordinates indicate the integration site of the insertion in the genome.

B) Three independent engineered L1 insertions in PC39: Following library capture and amplification techniques we observe three bands on a 0.75% agarose gel that represent distinct L1 integration events (pc-39-C ~ 1150bp; pc-39-B ~580bp; pc-39-A ~330bp). Pc-39-A and pc-39-B were previously characterized in Garcia-Perez *et al*. (2010).

C) Characterization of pc-39-C insertion: With our capture technique we identified a third, previously uncharacterized insertion, pc-39-C. Pc-39-C is located on chr19: 13627881 (GRCh37/hg19) on the reverse strand. The L1 is truncated at the 5' end and includes the last 100bp of ORF2p, and is flanked by 18bp TSD, with a 33bp poly(A) tail. The pre-integration site sequence is shown as well as the sequence surrounding the L1 integration. Red nucleotides and red triangle represent where initial EN cleavage occurred. Black triangle represents location of second-strand cleavage.

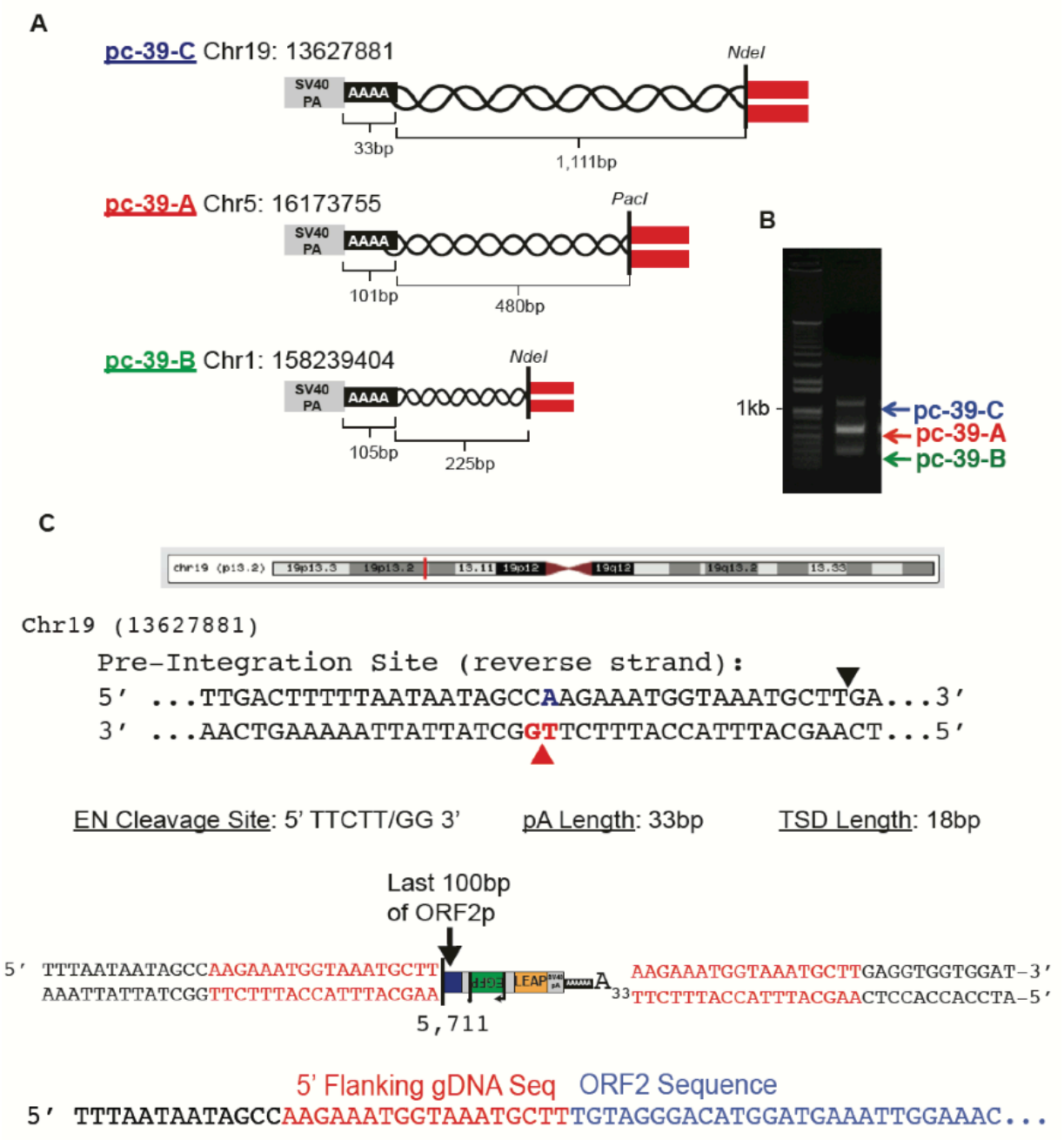


Figure 2.10: pc-39-C characterization (Supporting Figure 2.1D).

Figure 2.11: L1 insertions are located within AT-rich regions of the genome (Supporting Figure 2.1H).

A) Engineered L1 insertions are found in AT-rich regions of the genome: Plots displaying different spanning base-pair window sizes of examined sequence surrounding insertion sites (x-axis) versus the percentage of GC sequence found in the sequence (y-axis). All engineered L1 insertions display the same trend regardless of cell-type.

B) Insertions are located in T-rich sequence in the genome: Similar to Figure 2.8A, we examined the sequence surrounding insertion sites. For 100bp upstream and downstream of the L1 EN consensus cleavage site we examined the nucleotide ratio for each base-pair. In general the sequence surrounding insertions is T-rich. The proportion of T-rich sequence preference transitions to a steep increase at about 25bp upstream of the insertion site, which may be due to our requirement of insertions containing at least a 15bp poly(A) stretch. We are observing the complement sequence in this figure, which equates to a poly-T stretch.

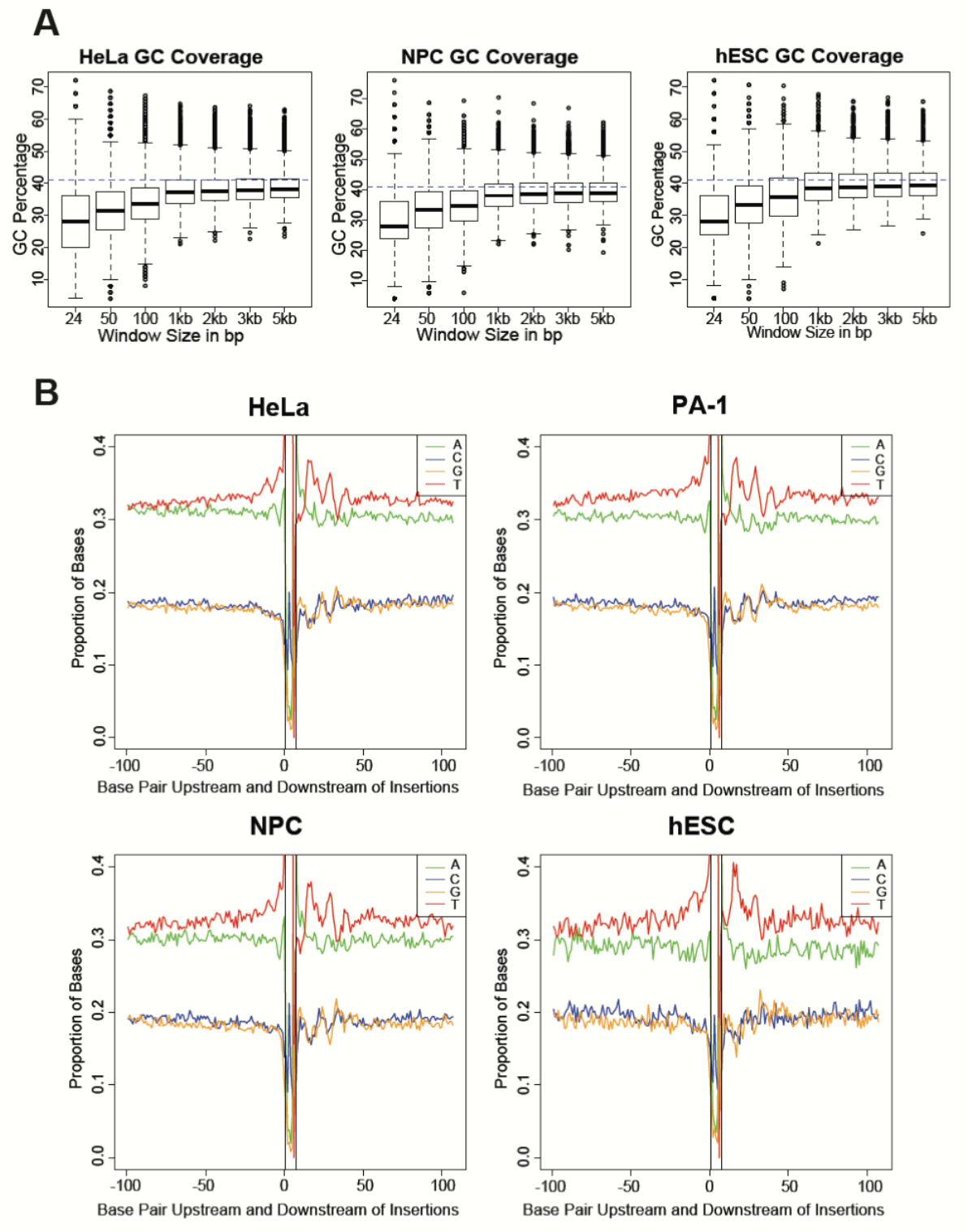


Figure 2.11: L1 insertions are located within AT-rich regions of the genome (Supporting Figure 2.1H).

Figure 2.12: L1 EN cleavage site is degenerate (Supporting Figure 2.2).

A) L1 EN site is influenced primarily by a 7mer sequence: Capturing 25bp upstream and 25bp downstream of insertion sites for a logo plot reveals that a 7mer sequence (5'-TTTTT/AA) primarily influences the EN logo. We thus examined 7mer sequences in the genome when creating our weighted random model.

B) Description of the term 'Hamming Distance': Hamming Distance refers to the number of nucleotide mismatches of a sequence as compared to some standard sequence. In this instance the sequence we want to match is the perfect L1 EN consensus cleavage site of 5'-TTTTT/AA. Thus, the 'Hamming Distance' of this sequence is 0. Any 7mer variant that has one mismatch in any position has a 'Hamming Distance' of 1, a 7mer with two mismatches has a 'Hamming Distance' of 2, etc..

C) Majority of L1 insertions occur at L1 EN consensus cleavage site with at least one or more mismatches: This plot shows increasing 'Hamming Distance' on the x-axis with the proportion of total insertions (y-axis) per sample.. Regardless of cell type, all datasets follow the same trend showing that about 10% of total insertions directly match the perfect EN consensus cleavage site, while more than half of the insertions contain 1 or 2 mismatches in this sequence.

D) *EN site drives towards AT-rich sequences in genome:* When randomly picking 10,000 sites in the human genome (left plot) regardless of the window span of sequence surrounding the site, the GC content stays relatively constant near 40%. For our weighted random model (right plot), when examining picked sites from one iteration of our model, we observe the same trend we see with our engineered L1 insertion datasets. This suggests that the variants of the EN consensus cleavage site are found within AT-rich regions of the genome and the EN consensus cleavage site is what drives L1 towards AT-rich regions of the genome.

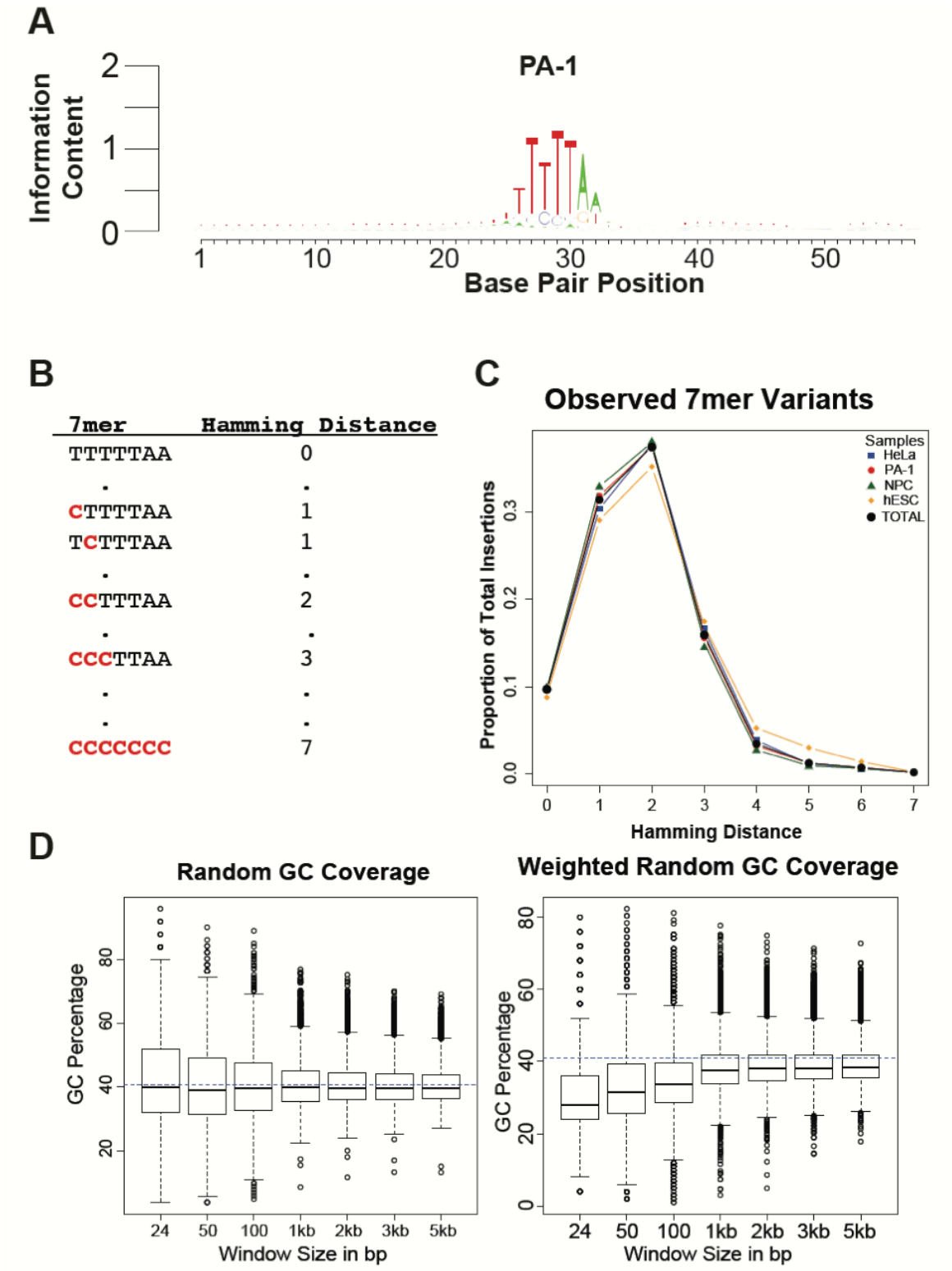


Figure 2.12: L1 EN cleavage site is degenerate (Supporting Figure 2.2).

Figure 2.13: LINE-1 is dispersed throughout the genome (Supporting Figure 2.3).

A) Insertion counts correlated with chromosome size: Chromosomes are plotted by the smallest chromosome to the largest from left to right along the x-axis. Insertion counts mapped to each chromosome are shown on the y-axis. Boxplots show the distribution observed from 10,000 iterations of the weighted random dataset. Colored points show the actual observed counts from the corresponding sample. Spearman's rho correlation for chromosome size and insertion counts for each sample: HeLa: 0.948, p-value: 2.859×10^{-6}; NPC: 0.927, p-value: 3.231×10^{-6}; hESC: 0.932, p-value: 3.165×10^{-6}. The X chromosome is a significant outlier for NPC and hESC insertions, while chromosome 5 is a significant outlier in HeLa (Bonferonni corrected p-value from linear model outlier test – HeLa: 0.0049; NPC: 0.0028; hESC: 0.0063).

B) L1 insertions are more prevalent in non-expressed regions of the genome: RNA-seq data from each cell type measured expressed regions of the genome. There are less insertions (colored squares) found within expressed regions (> 0.3 FPKM) of the genome than expected from the weighted random dataset (boxplots). We observe a significant depletion of insertions in expressed regions of the genome as compared to the weighted random datasets for HeLa, PA-1, and NPC insertions datasets (χ^2 p-values: 1.776×10^{-12}, $< 2.2 \times 10^{-16}$, $< 2.2 \times 10^{16}$ respectively).

C) L1 insertions are not preferentially located within highly expressed regions of the genome: Expressed regions of the genome (> 0.3 FPKM) were divided into 30 different bins based upon expression levels of lowest to highest (Table 2.10 and Table 2.11). For each bin, we determined the number of observed L1 retrotransposition events as compared to the weighted random dataset of 10,000 iterations (box plots). We observe as many L1 insertions as expected by the weighted random dataset in HeLa and hESCs. In NPCs we observe more insertions than expected at low expression in bin 2 (χ^2 p-value < 0.05). We observe more L1 insertions in PA-1s at lower expressed regions of the genome (bin 1 χ^2 p-value: < 0.05; bin2 χ^2 p-value: < 0.0001; bin 4 χ^2 p-value: $< 1 \times 10^{-6}$). We also observe significantly less PA-1 L1 insertions in more highly expressed regions of the genome (bin 23 χ^2 p-value: < 0.0001; bin 25 χ^2 p-value: < 0.05).

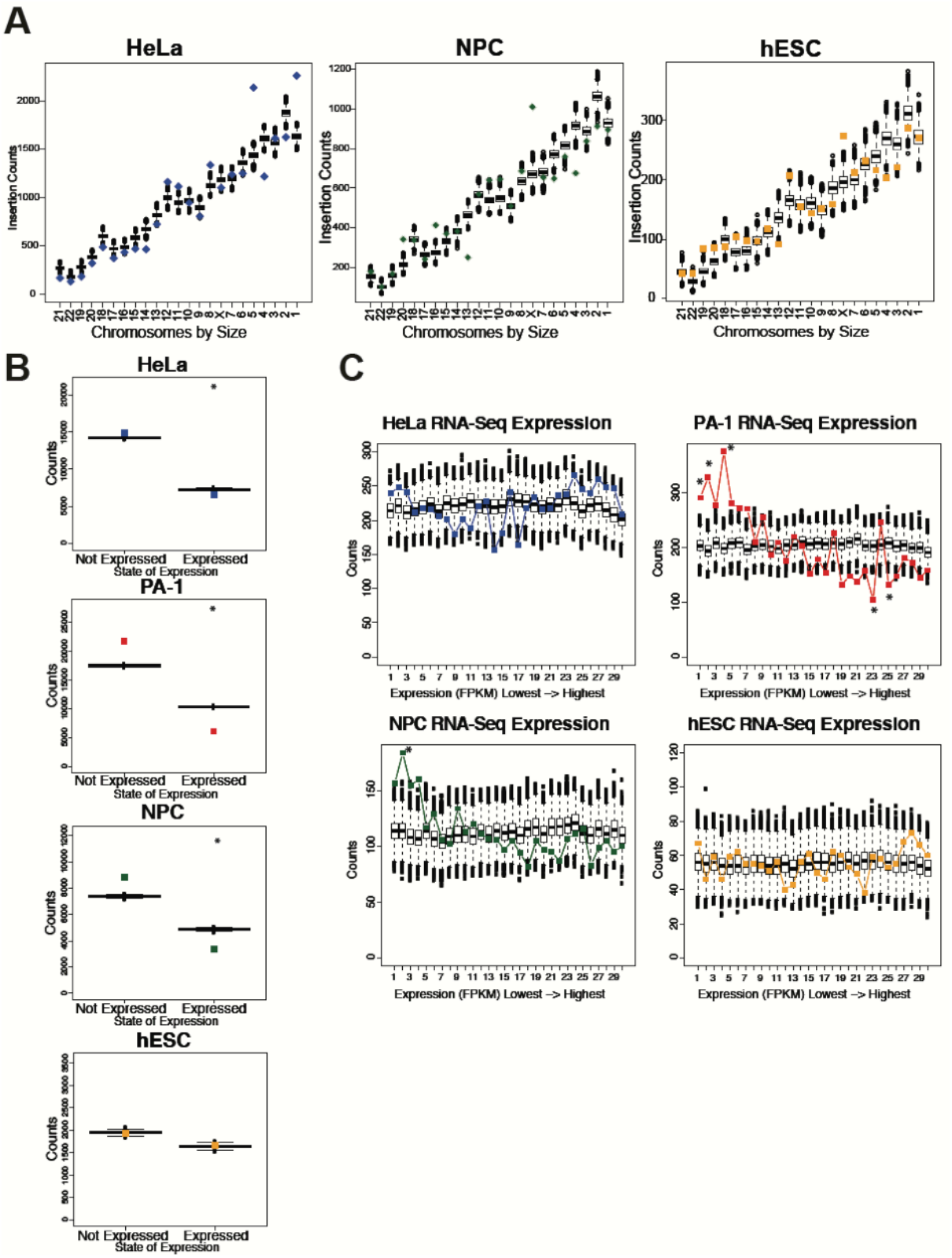


Figure 2.13: LINE-1 is dispersed throughout the genome (Supporting Figure 2.3).

Figure 2.14: Engineered insertions are dispersed throughout the human genome (Supporting Figure 2.3B).

A) Engineered HeLa insertions are dispersed throughout the human genome: Horizontal blue lines represent locations of mapped HeLa insertions throughout the human genome.

B) Engineered NPC insertions are dispersed throughout the human genome: Horizontal green lines represent locations of mapped engineered NPC insertions.

C) Engineered hESC insertions are dispersed throughout the human genome: Horizontal orange lines represent mapped locations of hESC insertions.

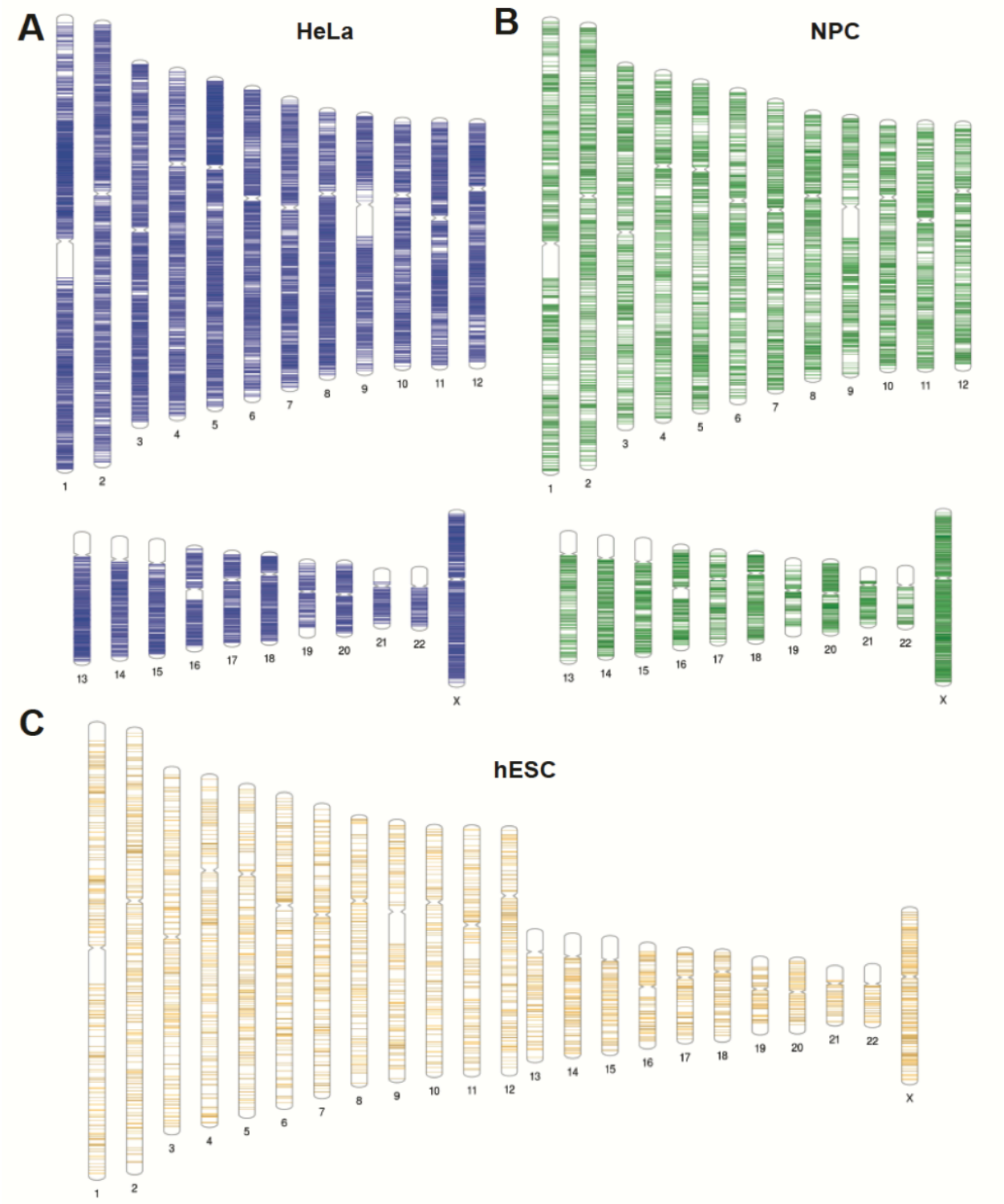


Figure 2.14: Engineered insertions are dispersed throughout the human genome (Supporting Figure 2.3B).

Figure 2.15: Endogenous LINE-1s in the human genome.

A) Young L1 insertions are within AT-rich regions of the genome: For every LINE-1 and L1Hs sequence identified by RepeatMasker in the GRCh37/hg19 human genome reference, we examined different windows sizes upstream and downstream of the annotated integration site. In general, younger, L1Hs sequences are in AT-rich regions as compared to older L1s.

B) Counts of L1 events in the genome are correlated with chromosome size: Endogenous LINE-1 insertions found in the human genome were plotted based on insertion count (y-axis) and chromosome size (x-axis). Chromosomes were plotted left to right by size. In general more insertions are found on larger chromosomes. As observed with our insertion dataset there is a spike of insertions on the X chromosome.

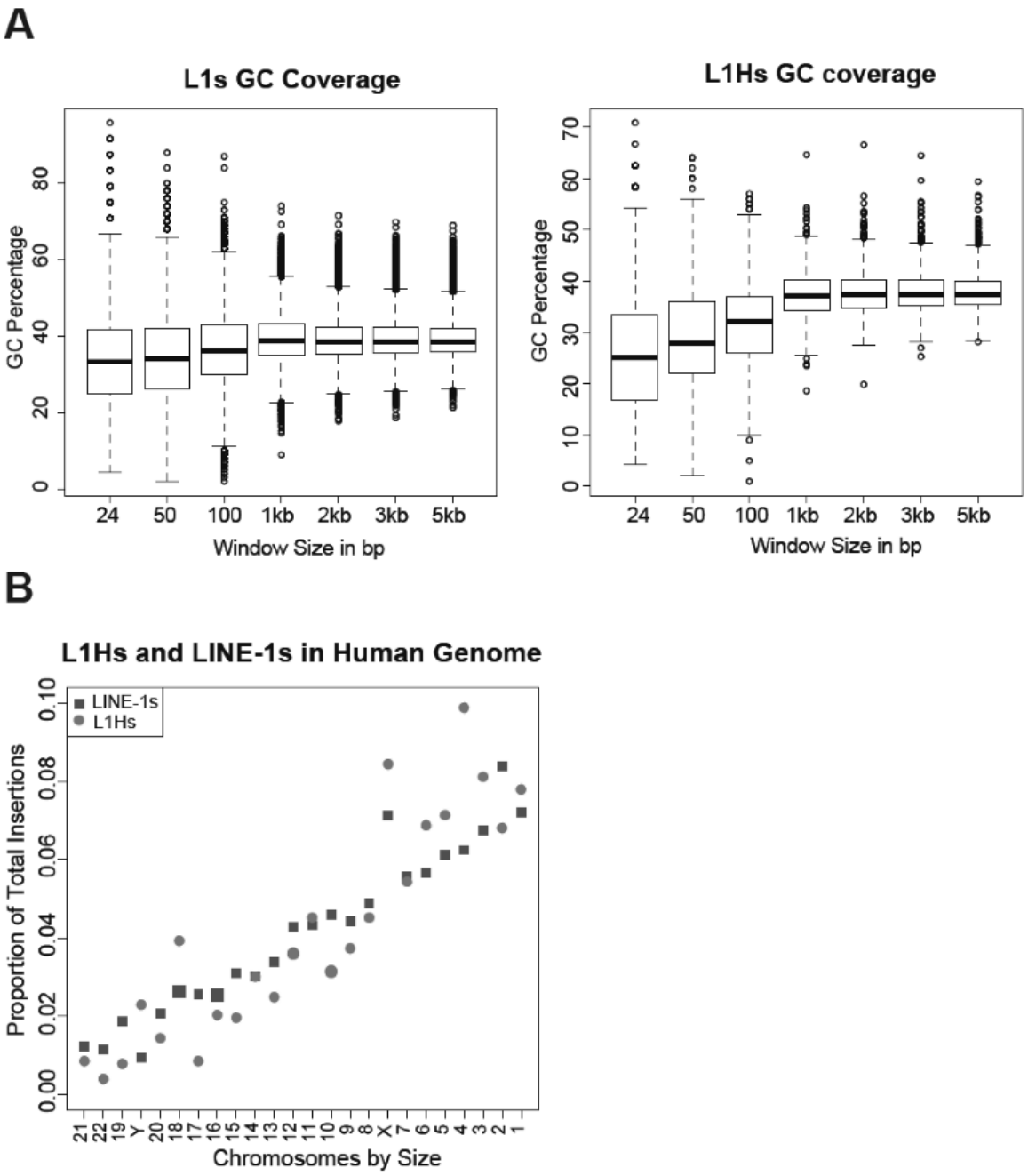


Figure 2.15: Endogenous LINE-1s in the human genome.

Figure 2.16: LINE-1 does not target transcribed regions of the genome (Supporting Figure 2.4).

A) Higher rate of transcription negatively influences integration sites: Transcribed regions of the genome were divided into 30 roughly equally sized bins (Table 2.9). Boxplots represent the range of observations from the 10,000 iterations of the weighted random datasets. In the left graph, blue squares represent the observed HeLa insertion counts in each transcriptional bin. From left to right, the rate of transcription increases in each bin. In the graph on the right, red squares represent the observed PA-1 insertion counts in each bin. Asterisks mark statistically significant differences observed between the sample insertion counts and the weighted random dataset (HeLa: bin 2 χ^2 p-value: < 0.05; PA-1: bin 1 χ^2 p-value < 0.01; bin 2 χ^2 p-value < 0.0001; bin 3 χ^2 p-value < 1 x 10^{-6}; bin 4 χ^2 p-value < 1 x 10^{-5}; bin 26 χ^2 p-value < 0.01; bin 28 χ^2 p-value < 0.01; bin 29 χ^2 p-value < 1 x 10^{-6}; bin 30 χ^2 p-value < 1 x 10^{-5}).

B) Calculating transcription bias throughout the genome: Depicted is a screenshot of the MIBrowser depicting Bru-seq data for HeLa cells on chr1. Bias is calculated as the RPKM expression of the top strand minus the bottom strand of a region in the genome over the sum total expression of both strands. The gene *CCNL2* (leftmost red rectangle) is expressed from the bottom strand and since there is no other gene present expressed from the top strand, the genome transcription Bias is -1. For the region of the genome containing *RP3-758J18* (leftmost green rectangle) expressed on the top strand and *MRPL20* and *RN7SL657P*, both expressed from the bottom strand, the Bias is -0.62 as there is more transcription from the bottom strand in the genome. Finally, since there is only transcription from the top strand where *RP4-758J18.13* is located, the bias is 1.

C) Interpreting CDF plots of transcription bias: Drawn here are hypothetical cumulative distribution function (CDF) plots for transcription bias data when looking at L1 insertions in which EN cleavage occurred on the top (top plot) or bottom strands (bottom plot) in the genome. Gray lines represent expectations from the uncorrected weighted model simulated data expectations, and the red line represents the corrected model distribution. If the L1 EN strictly prefers cleavage on the coding strand, we expect to observe an exaggerated trend in the data as shown by the dotted navy line, but if the EN strongly favors cleaving the noncoding strand then we expect to observe the dotted light blue line distribution.

D) L1 EN shows a slight preference for cleavage of coding strand during transcription: L1 insertions were first subdivided into two groups; insertions in which EN cleavage occurred on the top strand of the genome, or insertions in which EN cleavage occurred on the bottom strand of the genome. For each group we plotted the cumulative distribution function (CDF) for insertions based on Transcription Bias values (x-axis). HeLa insertions show a significant (Kolmogorov-Smirnov bootstrap test p-value < 0.05) trend towards favoring EN cleavage on the coding strand. Only PA-1 insertions in which EN cleavage occurred on the top strand show a significant preference (Kolmogorov-Smirnov boot strap test p-value < 0.01) for EN cleavage on the coding strand.

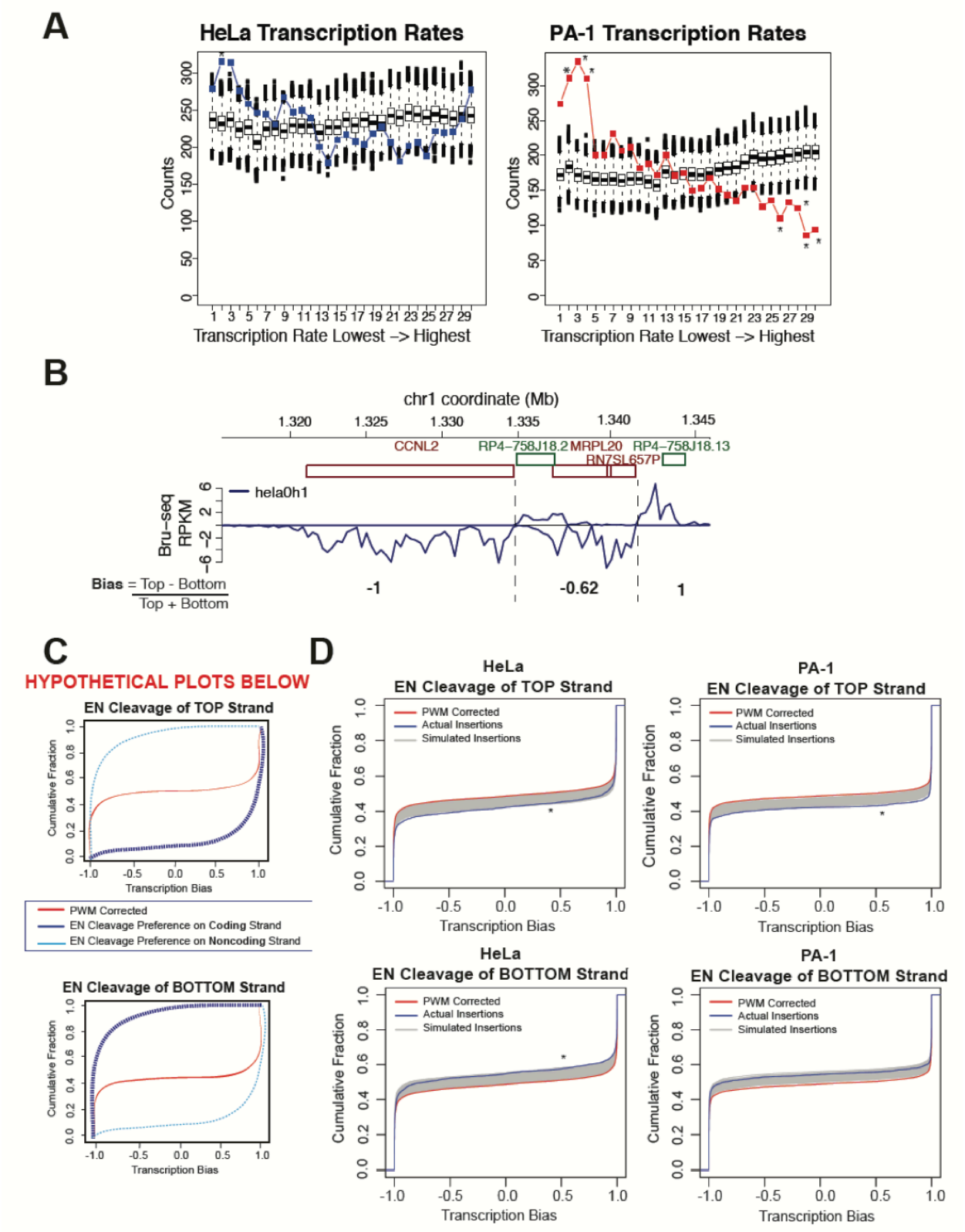


Figure 2.16: LINE-1 does not target transcribed regions of the genome (Supporting Figure 2.4).

Figure 2.17: MLV integration preference in the human genome (Supporting Information for Figure 2.5).

A) MLV integration sites show enrichment in promoter and transcribed regions of the genome: We randomly selected 27,777, MLV integration events in K562 from LaFave *et al.* (2014), a count equal to the number of observed PA-1 L1 insertion events. This analysis shows that given a sample size equivalent to that of our PA-1 integration events for a dataset with chromatin state enrichment, we should be able to detect such trends.

B) Distinct MLV 'hotspots' in the human genome: Using the same dataset in 2.16A, the downgraded MLV integration dataset containing the same number of events as in our PA-1 L1 integration dataset, plotting MLV integration spots (black horizontal lines) on the chromosomal ideogram shows distinct clustering of deep black, indicating 'hotspot' regions of MLV integration.

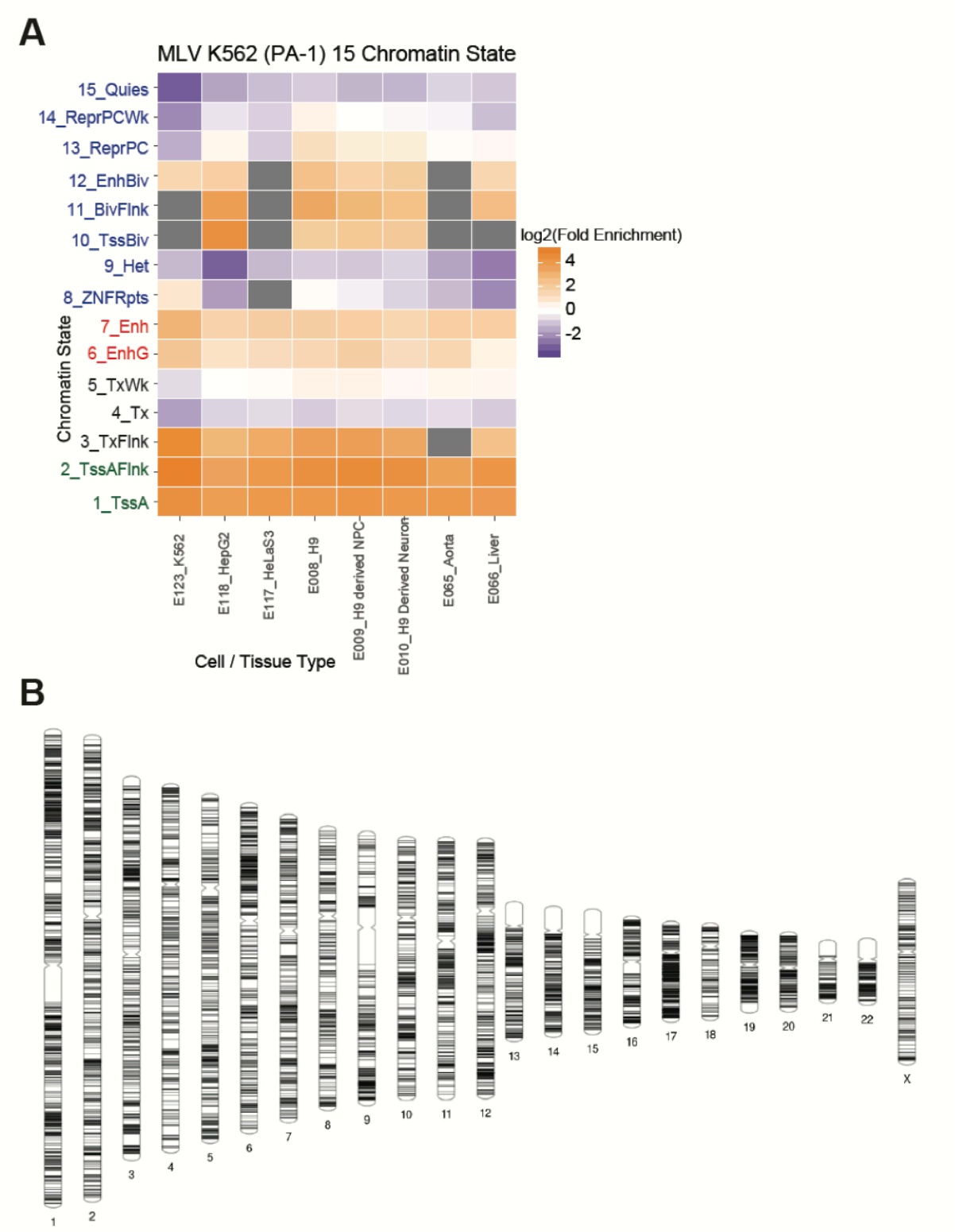


Figure 2.17: MLV integration preference in the human genome (Supporting Information for Figure 2.5).

Figure 2.18: L1 is not enriched in any specific chromatin state (Supporting Figure 2.5).

We compared our insertions datasets to the Roadmap Epigenomic Consortiums' Hidden Markov Modeled 18 chromatin states. The most comparable cell type is shown in the leftmost column for each heat map. Each box shows enrichment for the insertion dataset across the different cell types. For visual comparison of strong enrichment we included the MLV integration events in K562 cells (LaFave et al., 2014). In general, even with the 18 chromatin state dataset, L1 insertions are not enriched in any specific chromatin state.

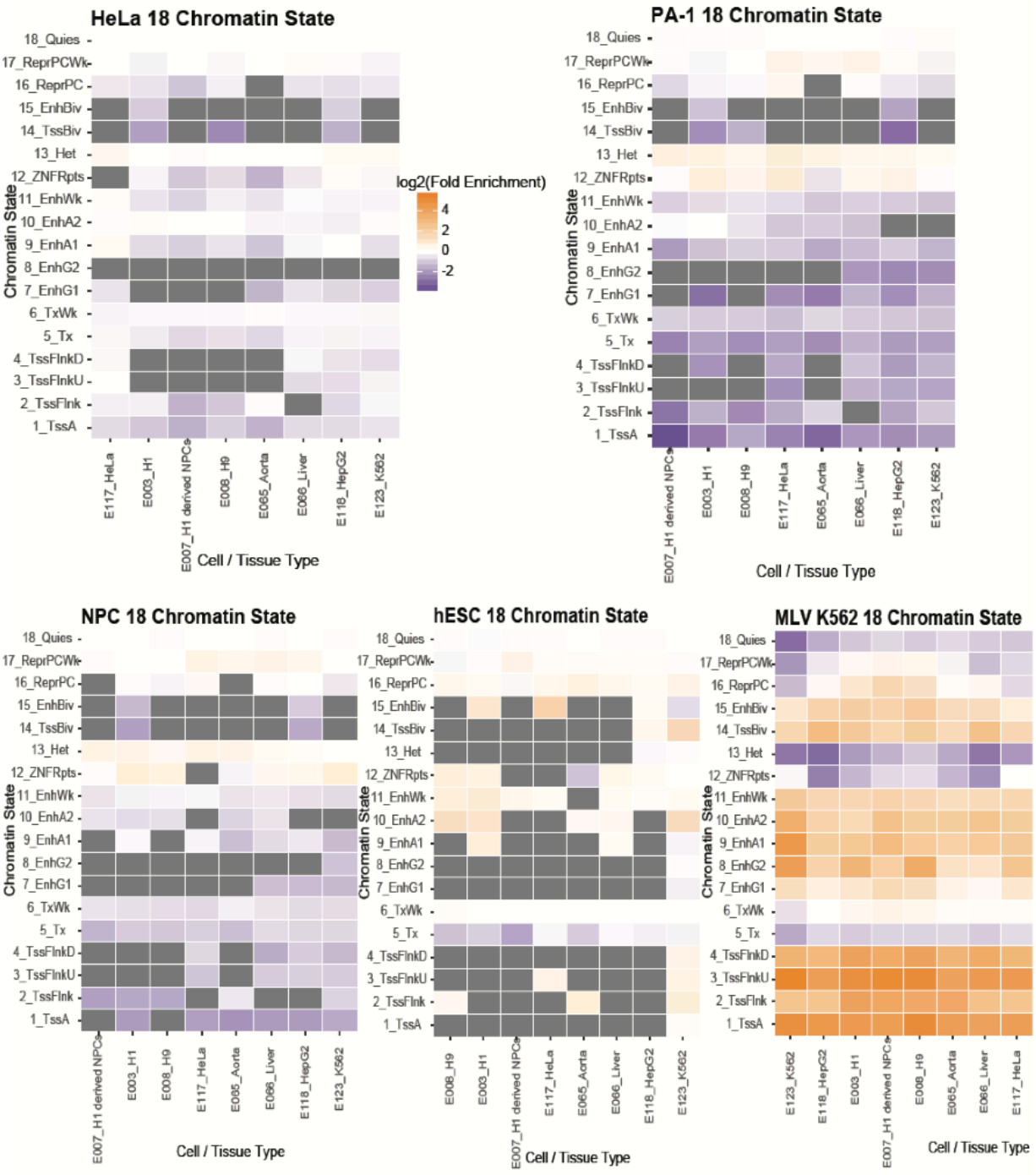


Figure 2.18: L1 is not enriched in any specific chromatin state (Supporting Figure 2.5).

Figure 2.19: Replication influences L1 integration in the human genome (Supplemental Information for Figure 2.6).

A) EN cleavage preference for lagging strand template cleavage: For all four cell types, insertions were subdivided into two different groups, insertions in which the EN cleaved the bottom strand and insertions in which the EN cleaved the top strand. Plots for bottom strand cleaved insertions that are shifted towards the left show preference for cleavage on the lagging strand template. Plots for top strand cleavage which are shifted towards the right, show a preference for cleavage of the lagging strand template. All plots except hESC insertions cleaved on the top strand show a significant shift towards L1 EN cleavage favoring the lagging strand template. While these plots show significant difference from the corrected model, insertion sets are not showing an overly exaggerated bias towards lagging strand template cleavage (Kolmogorov-Smirnov boot strap test P-values: HeLa Bottom: $< 1x10^{-6}$; HeLa Top: < 0.05; PA-1 Bottom and Top: $< 1x10^{-6}$; NPC Bottom and Top: < 0.001; hESC Bottom: < 0.05).

B) L1 EN shows slight preference for cleavage at replication fork initiation sites in HeLa and hESC:. Cumulative distribution function plots are shown for insertions based on replication fork direction (RFD) slope. A negative slope value indicates a preference for replication fork termination in the genome, while a positive slope indicates a preference for replication fork initiation. While all insertion datasets significantly differ from the corrected model distribution, the trend observed display a preference for areas of the genome between initiation and termination of forks in both PA-1s and NPCs. HeLa and hESCs insertions display less or near similar amounts of insertions near termination of forks, and then less insertions in the 'between' phase, and more insertions climbing towards replication fork initiation sites than the corrected model. Kolmogorov-Smirnov bootstrap test P-values: HeLa: < 0.01; PA-1: $< 1x10^{-16}$; NPC: < 0.01; hESC: < 0.05).

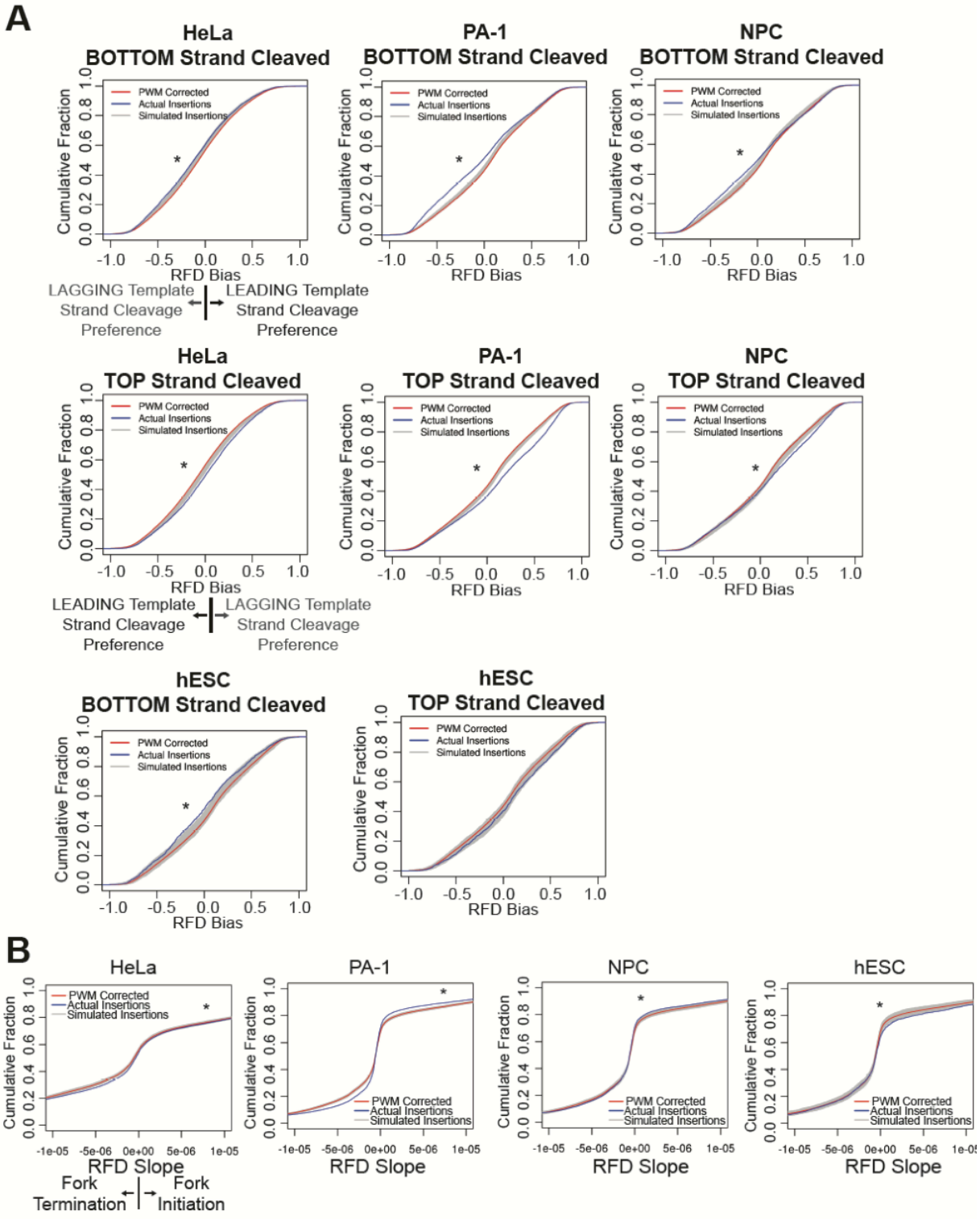


Figure 2.19: Replication influences L1 integration in the human genome (Supplemental Information for Figure 2.6).

Figure 2.20: Verifying calling algorithm.

To determine the accuracy of our calling algorithm for defining reads at base pair resolution, we randomly selected 100,000 ungapped, locations in the genome. The length of the sequence followed the distribution of CCS read lengths we observed with our actual sequenced data (Figure 2.7D). We then attached a 5' poly(A) tract to these reads that mimicked the length of poly(A) tails observed in our L1 insertion dataset (Figures 2.1G and 2.7F). These random sequences were mapped to the GRCh37/hg19 genome reference with Bowtie2 under the same conditions of our insertion data (Figure 2.7A) and then we determined the best mapped read location if possible. Further refinement of the site to base pair resolution was completed using the Smith Waterman algorithm. We accurately called 97,748 of these 100,000 reads, but were unable to identify the best mapping location for 2,136 of the reads, and we incorrectly called 116 of the reads.

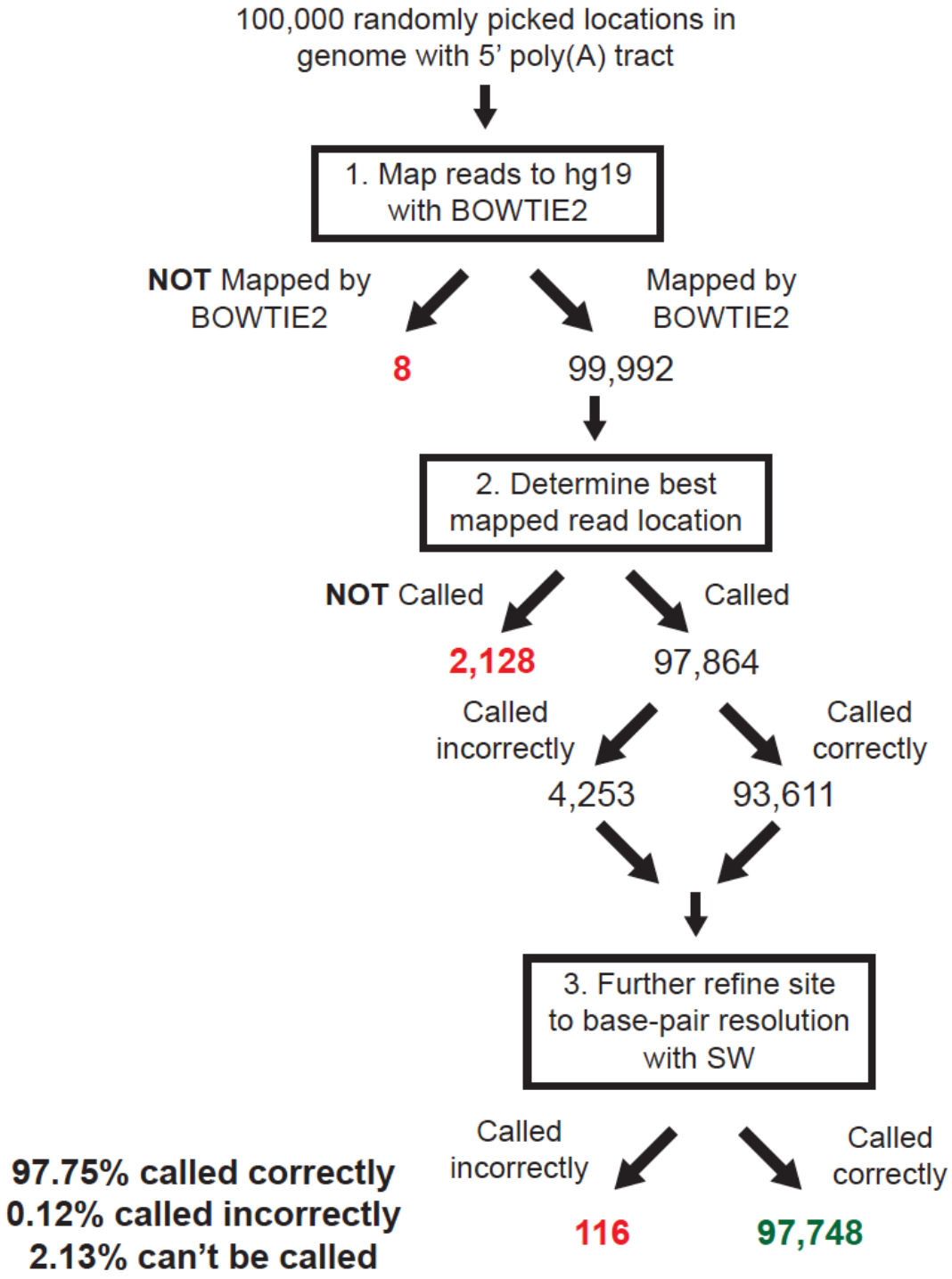


Figure 2.20: Verifying calling algorithm.

Figure 2.21: Generation, alignment, and filtering scheme for RNA-seq data.

For each cell type we collected two biological replicates of total RNA. A Ribo-Zero rRNA Removal Kit was performed to remove cytoplasmic ribosomal RNA. Samples were then library prepped with the Illumina TruSeq Stranded RNA-seq Kit. One set of biological replicates was sequenced with the Illumina HiSeq at 100bp paired-end reads, while the other biological set was Illumina HiSeq sequenced at 125bp paired-end reads. Following sequencing, reads were aligned to the GRCh37/hg19 genome reference with Tophat. Reads were then assembled with Cufflinks to get transcripts and isoforms. Then Cuffmerge was used to create the final transcription assembly using the ENSEMBL assembly as a guide. Cuffquant then quantified gene and transcript expression in terms of FPKM. We then merged biological replicates and normalized all the samples onto the same scale for cross comparison with cuffnorm.

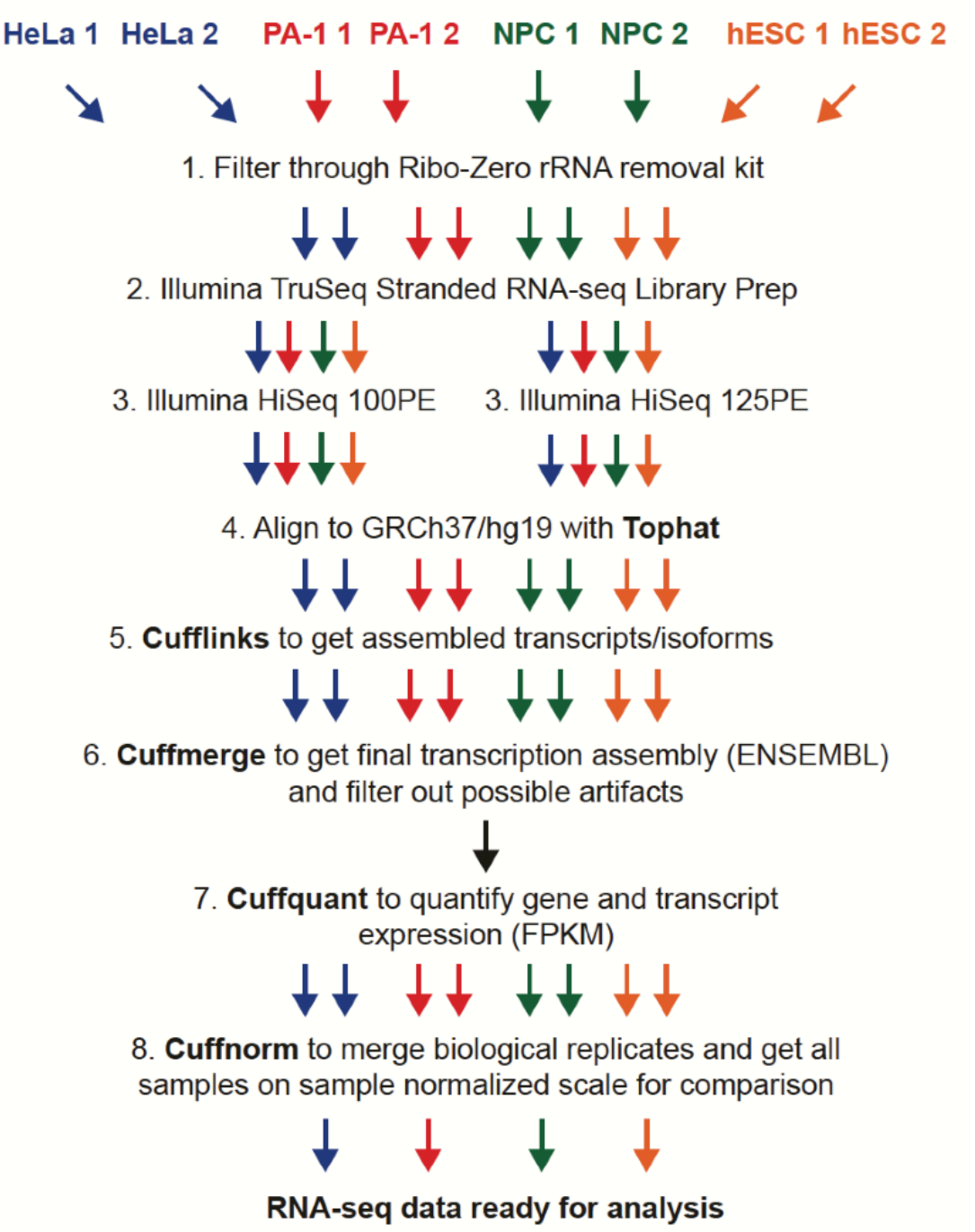


Figure 2.21: Generation, alignment, and filtering scheme for RNA-seq data.

Figure 2.22: Slight enrichment in enhancers is not associated with low GC content.

HeLa insertions appear to be somewhat enriched in enhancer states and bivalent enhancers, while hESC insertions also appear enriched in enhancer states in the Encode's 15 chromatin state (Figure 2.5). We were curious to explore if these states are correlated with being AT-rich. In general this enhancer and bivalent enhancer states are not the most AT-rich. Similarly, in Encode's 18 chromatin state analysis HeLa insertions appear enriched in enhancer states (EnhA1, EnhA2, EnhWk) as well, while hESC insertions appear enriched in some enhancer states (EnhA2, EnhWk) and repeats. These states are near the genome average GC content, suggesting that any slight enrichment in these states is not due to the states being overly AT-rich.

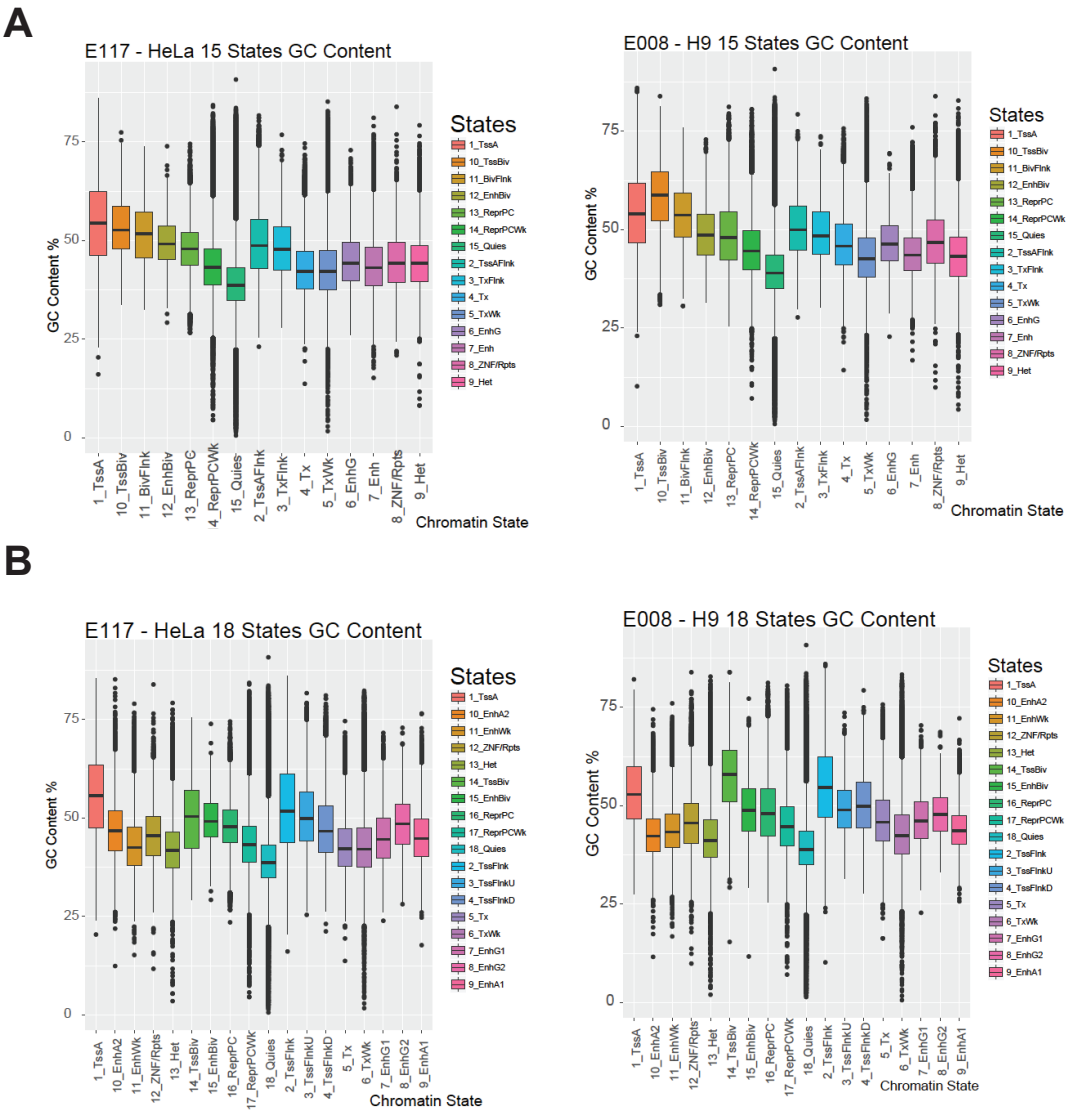


Figure 2.22: Slight enrichment in enhancers is not associated with low GC content.

Figure 2.23: Super enhancers and typical enhancers are not highly enriched for L1 insertions.

A) HeLa and hESC insertions are not enriched in super and typical enhancers: Gray boxes represent less than 30 insertions found within a given condition, and which may result in false enrichment or depletion measurements. There was no H9 hESC dataset of super and typical enhancers available so we compared our hESC insertions to the H1 hESC dataset of super and typical enhancers and find that our hESC insertions are 2.48 times enriched in typical enhancers. This enrichment is minimal when compared to enrichment levels observed with MLV integration events. Furthermore, we observe little enrichment of HeLa insertions in super and typical enhancers. Thus, hESC and HeLa insertion are not preferentially integrating within super and typical enhancers.

B) Super and Typical Enhancers are as GC-rich as average genomic levels: We examined the average GC-content of super and typical enhancers to determine if these regions of the genome are in overly AT-rich regions of the genome. Super and typical enhancers of H1 hESC and HeLa cells are slightly above average genomic GC content levels.

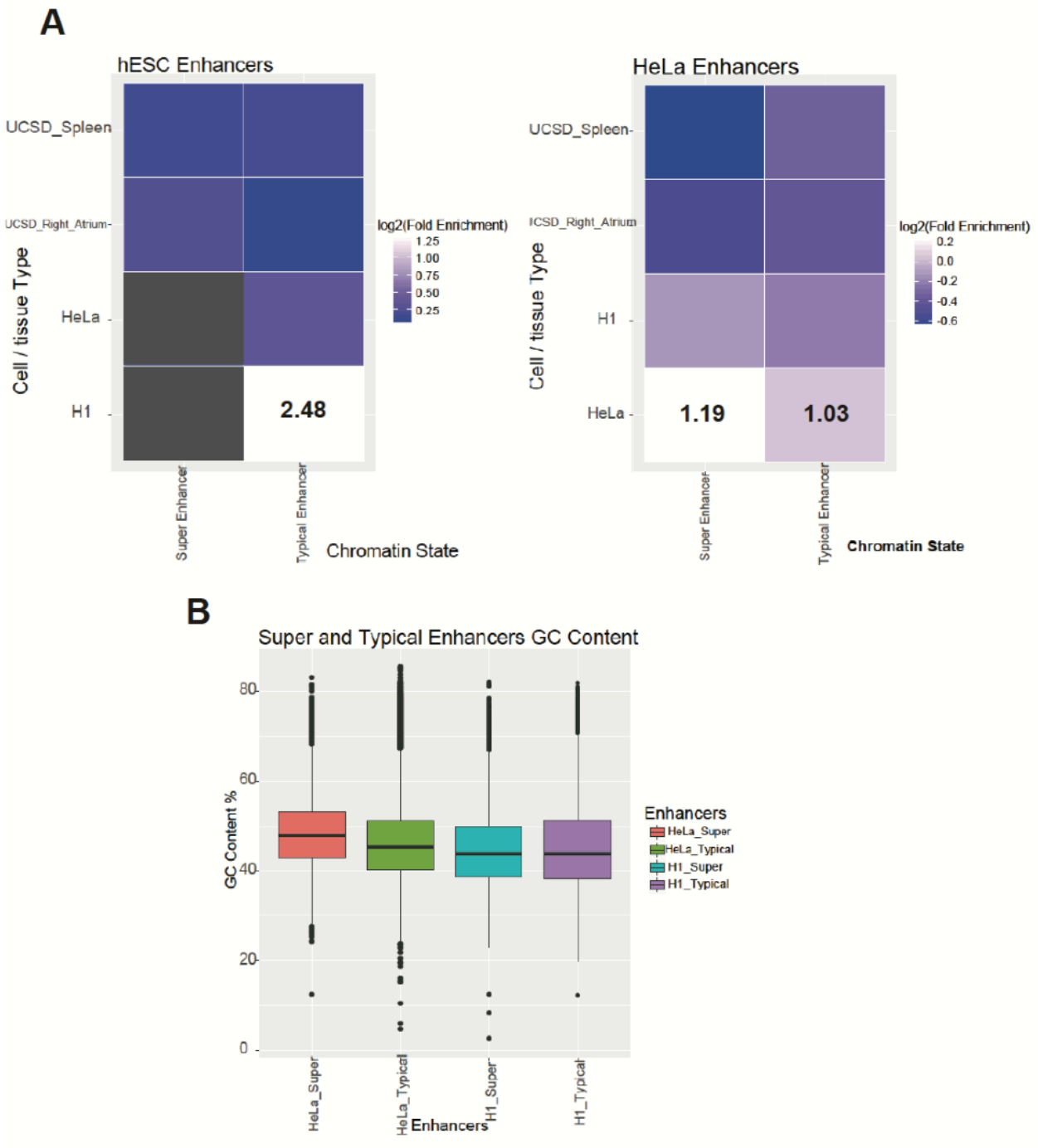


Figure 2.23: Super enhancers and typical enhancers are not highly enriched for L1 insertions.

Sample Name	Date/s Submitted for PacBio Sequencing	Selection /Screen	Proportion of Total Final Insertions
H1012	10.16.12; 2.27.13	Neo	19.14%
H1014	7.15.13	Neo	12.76%
H1015	7.15.13	Neo	2.13%
H1017	7.15.13	Neo	15.58%
H1019	11.13.13	Neo	46.96%
HJM101	7.15.13	Neo	2.50%
HGFP4 (LRE3)	11.26.13	GFP	0.93%

Table 2.1: Independent HeLa samples and contribution to final dataset.

Seven independent retrotransposition assays were performed and after completion of the assay gDNA was collected from all. Independent sample names are found in column 1, date/s in which the sample was submitted for PacBio sequencing is marked in column 2, selection or screening performed on the sample is listed in column 3 and the proportion of total insertions of the final dataset that this sample comprised is found in column 4. All samples except HGFP4 were transfected with pJM101/L1.3 and retrotransposition events were selected in the presence of Neomycin. HGFP4 was transfected with pCEP4/LRE3-*mEGFPI.*

Sample Name	Date/s Submitted for PacBio Sequencing	Selection /Screen	Proportion of Total Final Insertions
PL5	7.17.13	Puro	0.03%
PL6	7.17.13	Puro	0.51%
PL7	7.17.13	Puro	0.06%
PL8	7.17.13	Puro	1.35%
PGFP	7.30.13	GFP	27.78%
PGFP2	7.30.13	GFP	2.21%
PGFP3	8.2.13; 7.30.13	GFP	1.90%
PGFP4	11.13.13	GFP	5.70%
PGFP6	11.13.13	GFP	25.54%
PGFP7	11.13.13	GFP	9.87%
PGFP8	11.26.13	GFP	20.06%
PGFP9	11.26.13	GFP	4.99%

Table 2.2: Independent PA-1 samples and contribution to final dataset.

Independent sample names are found in column 1, date/s in which the sample was submitted for PacBio sequencing is marked in column 2, selection or screening performed on the sample is listed in column 3 and the proportion of total insertions of the final dataset that this sample comprised is found in column 4. All PA-1 samples were transfected with pCEP4/LRE3-*mEGFPI*. Samples PL5, PL6, PL7, and PL8 were selected for transfected cells in the presence of puromycin as the pCEP4/LRE3-*mEGFPI* contains a puromycin resistance. The rest of the PA-1 samples were screened for GFP positive cells that represent *de novo* L1 retrotransposition events.

Sample Name	Date/s Submitted for PacBio Sequencing	Selection/ Screen	Proportion of Total Final Insertions
NPC	2.20.14	Puro	1.87%
NPC1	5.7.14	Puro	3.16%
NPC-RA	11.18.13	Differentiated	4.10%
NPC-RA2	2.20.14	Differentiated	0.70%
NPC5	10.2.14	Puro	0.72%
NPC74	10.2.14	Puro	0.43%
NPC125	5.27.15	Puro	26.81%
NPC172	5.14.15	Puro	43.14%
NPC183	5.14.15	Puro	19.07%

Table 2.3: Independent NPC samples and contribution to final dataset.

Independent sample names are found in column 1, date/s in which the sample was submitted for PacBio sequencing is marked in column 2, selection or screening performed on the sample is listed in column 3 and the proportion of total insertions of the final dataset that this sample comprised is found in column 4. NPCs were transfected with pCEP99/UB-LRE3-*mEGFPI* and transfected cells were selected in the presence of puromycin. NPC-RA and NPC-RA2 are NPC samples that were transfected and treated with retinoic acid leading to differentiation of the NPCs.

Sample Name	Date/s Submitted for PacBio Sequencing	Selection/ Screen	Proportion of Total Final Insertions
hESC1	5.27.14	Neo	13.86%
hESC2	5.27.14	Neo	6.75%
hESC4	12.6.13	Neo	4.18%
hESC5	12.9.13	Neo	3.36%
hESC96	10.2.14	Neo	28.42%
hESC03	1.14.16	Neo	2.96%
hESC06	12.17.15	Neo	8.79%
hESC09	1.22.16	Neo	3.57%
hESC10	1.14.16	Neo	2.87%
hESC15	10.3.14	Neo	25.24%

Table 2.4: Independent hESC samples and contribution to final dataset.

Independent sample names are found in column 1, date/s in which the sample was submitted for PacBio sequencing is marked in column 2, selection or screening performed on the sample is listed in column 3 and the proportion of total insertions of the final dataset that this sample comprised is found in column 4. All hESC samples were transfected with pKUB102/L1.3-sv+ and selected for retrotransposition events in the presence of neomycin.

Table 2.5: Top 20 weighted 7mers for uncorrected and corrected models.

In the corrected model (right) observed 7mer frequencies are corrected for the frequency of the 7mer in the human genome. This correction is not made in the uncorrected model (left). The final top 20 model weights of the uncorrected and top 21 model weights of corrected model are shown. In the uncorrected model the perfect EN cleavage is given the highest weight of 1, while in the corrected model a variant of the sequence, 5'-TTTTCAA is given the highest weight of 1. Since 5'-TTTTT/AA is the most prevalent 7mer in the human genome, when 'correcting' 7mer frequencies with the frequency of the 7mer in the genome, the perfect EN consensus cleavage site becomes the 21st highest weighted 7mer.

Table 2.5: Top 20 weighted 7mers for uncorrected and corrected models.

Uncorrected Model		Corrected Model	
7mer	Weight	7mer	Weight
TTTTTAA	1	TTTTCAA	1
TTCTTAA	0.574	TTCTTAA	0.951
TTTTTAT	0.324	TTTCTAA	0.667
TTTCTAA	0.314	CTTTCAA	0.658
TTTTCAA	0.286	TTTTCAT	0.478
TTCTTAT	0.258	TTCTTAT	0.463
TTTTAAA	0.247	ATTTCAA	0.395
TTTTTGA	0.222	CTCTTAA	0.372
CTTTTAA	0.218	ATCTTAA	0.368
TCTTTAA	0.195	TTTCTAT	0.364
ATTTTAA	0.192	TTTTCGA	0.359
TTTCTAT	0.172	TTTTAAA	0.347
TCTTTAT	0.163	GTCTTAA	0.311
TTTTCAT	0.153	CTTTCAT	0.287
GTTTTAA	0.151	CTTCTAA	0.283
ATCTTAA	0.132	TTCTTAC	0.269
TTTTTGT	0.128	TTCTTGA	0.262
CTCTTAA	0.120	TCTTTAA	0.242
TTTTTCT	0.118	TTTCTAC	0.235
TTCTTGA	0.114	CTTTAAA	0.220
		TTTTTAA	0.219

Sample	Ultra-Conserved	Ultra-Conserved Noncoding	Transposon Free Regions	Total Insertions
Endogenous L1s in Human Genome				
L1s	0	1	500	951,780
L1Hs	0	0	0	1,544
Engineered L1 Insertions				
HeLa	4	11	453	21,497
PA-1	4	13	496	27,777
NPC	2	3	199	12,223
hESC	1	2	134	3,582
Engineered L1 Insertions in FANCD2 Deficient Cells				
PD20FD2 + L1.3	0	2	76	4,372
L1.3	0	7	339	18,124
L1.3/D205A	0	1	21	1,514

Table 2.6: L1 insertions in conserved regions of the human genome.

We identified endogenous LINE-1s and endogenous L1Hs sequences in the human genome with RepeatMasker and determined how many of these insertions are found within ultra-conserved (Bejerano et al., 2004), ultra-conserved noncoding (McCole et al., 2014), and transposon free regions (Simons et al., 2006) of the genome (columns). Additionally we examined whether we identified any engineered L1 insertions within these areas of the genome (rows). Engineered L1 insertions were identified in all three examined regions of the genome while endogenous L1s are absent or barely identified within these regions of the genome. Indeed this indicates that these regions of the genome are initially accessible for integration to occur, and over time selective forces remove insertions.

Accession #	Broad/ Narrow Peak	Genomic Coverage	% Total Insertions	# Insertions	Weighted Random		
					Min	Median	Max
hESC							
ENCFF495NIW	Broad	1.27%	3.38%	121	7	23	42
ENCFF552URI	Narrow	0.61%	2.04%	73	1	10	23
HeLa							
ENCFF792CTF	Broad	2.11%	3.62%	779	260	327	402
ENCFF346KEV	Narrow	0.98%	1.81%	389	87	131	177

Table 2.7: Majority of L1 insertions are not within DNase I hypersensitive sites in the genome.

We explored both broad and narrow peak (column 2) DNase I hypersensitive site datasets from ENCODE (column 1) and determined the genomic coverage of the DNase I sites in the genome (column 3) and the proportion (column 4) and number of insertions (column 5) we identified in each dataset as compared to the weighted random dataset min (column 6), median (column 7) and observed max (column 8). While there are certainly more HeLa and hESC insertions within DNase I hypersensitive sites than expected based upon the presence of the EN consensus cleavage site, a small portion of the total insertions are located within these sites. Thus, DNase I hypersensitive sites are not primarily targeted L1 integration sites in the human genome. This observation correlates nicely with the Encode 15 and 18 Chromatin State analysis performed.

Biological Replicate 1; 190bp target insert size, 100PE

Sample	# Reads	% Perfect Index Reads	% >= Q30	Mean Quality Score
$HeLa_{JVM}$	91,023,978	98.82	89.6	35.43
PA-1	56,386,346	98.44	89.3	35.36
H9	94,446,704	98.86	89.32	35.37
NPC	107,465,106	98.45	90.65	35.78

Biological Replicate 2; 190bp target insert size, 125PE

Sample	# Reads	% Perfect Index Reads	% >= Q30	Mean Quality Score
$HeLa_{JVM}$	86,945,121	98.68	84.9	34.48
PA-1	68,422,258	97.71	85.29	34.57
H9	70,103,961	98.01	85.94	34.69
NPC	84,313,027	98.26	85.15	34.66

Table 2.8: Details of two biological replicate RNA-seq runs.

The number of Illumina HiSeq reads (column 2) obtained for each cell type (column 1), and the percentage of those reads that contained perfect index reads (column 3), as well as the percentage of reads that contained a quality score above 30 (column 4) and the mean quality score of all the reads (column 5) is shown in the tables. Each table represents the summary output for each set of biological replicates.

HeLa (hela0h1) Bru-seq RPKM > 0.022		
Bin	RPKM Max Value of Bin	Genomic bps Coverage in bin
1	0.024781606	34080481
2	0.02818962	34086000
3	0.03241205	34111130
4	0.03848944	34067000
5	0.047122262	34087723
6	0.079979682	34081777
7	0.107857	34008729
8	0.129406256	34095000
9	0.15020497	33883000
10	0.1752554	33999000
11	0.21069716	34123000
12	0.253964101	34094158
13	0.38129427	34065512
14	0.45586647	34069000
15	0.522449	34097000
16	0.574950644	34095000
17	0.637818	34183000
18	0.696108	34107727
19	0.7608266	34087000
20	0.831118902	34078000
21	0.91805	34091000
22	1.184538924	34158000
23	1.319869	34087000
24	1.467176754	34133000
25	1.642058919	34102000
26	1.8294586	34091000
27	2.124498636	34091000
28	2.52164266	34090000
29	3.58425	34117000
30	2768.753	34213000

PA-1 (pa10h1) Bru-seq RPKM > 0.024		
Bin	RPKM Max Value of Bin	Genomic bps in bin
1	0.0285259	35176137
2	0.0323187	35166174
3	0.037394081	35187892
4	0.044066909	35215468
5	0.0580608	35209000
6	0.09536488	35197951
7	0.120273362	35174000
8	0.140583099	35198935
9	0.162625292	35171126
10	0.185052869	35193000
11	0.2166588	35152000
12	0.3094766	35189000
13	0.382980925	35172000
14	0.43955916	35220960
15	0.503519	35172000
16	0.564641	35251000
17	0.61669445	35190000
18	0.677942	35191000
19	0.734481888	35213000
20	0.803848065	35214000
21	0.899497099	35194000
22	1.210994053	35184000
23	1.3515564	35174000
24	1.501154	35149000
25	1.67440847	35150000
26	1.869549112	35195000
27	2.13061279	35252000
28	2.478356	35185000
29	3.14265245	35184000
30	3539.517	35162000

Table 2.9: Transcription bin thresholds (Supporting Figure 2.16A).

The left table provides the range of RPKM thresholds of Bru-seq data for the HeLa (hela0h1) sample. The right table provides the range of RPKM threshold of Bru-seq data for the PA-1 (PA10h1) sample. The first column indicates the bin number as on the corresponding graph in Figure 2.16A. The second column indicates the maximum RPKM threshold of the bin. The minimum threshold would be above the previous bin's threshold. The third column indicates the number of base pairs in the genome that lie within this binned RPKM threshold. Rows indicate the different bins.

	HeLa			PA-1	
Bin	FPKM Max Value of Bin	Genomic bps in bin	Bin	FPKM Max Value of Bin	Genomic bps in bin
1	0.499896	33094138	1	0.45842034	35934011
2	0.759631	33099809	2	0.746659396	35642246
3	1.279752415	33062286	3	1.101606055	36149574
4	1.712661	33075321	4	1.607600998	36153537
5	2.356952596	33181107	5	2.079394	35956776
6	2.966676343	33209054	6	2.673266718	35998435
7	3.788838638	32945089	7	3.235764195	36018154
8	4.622322347	33160888	8	4.016335018	35919118
9	5.589120453	33120856	9	4.728476001	35976914
10	6.487361619	33103467	10	5.633175866	35992354
11	7.489477415	33056034	11	6.653097026	35937464
12	8.505062	33136218	12	7.676101892	35908051
13	9.641971238	33115496	13	8.688693047	35949061
14	11.05339728	33123585	14	9.95716434	35944878
15	12.65272247	33102850	15	11.19146983	35784585
16	13.87764743	33005912	16	12.41333181	36085877
17	15.33025967	33135676	17	13.68018476	36384758
18	17.18377703	33105114	18	14.998743	36001246
19	19.38723312	33119129	19	16.95349403	35964887
20	21.70851967	33186671	20	19.3292167	35931436
21	24.62518039	33118869	21	21.62713492	35984487
22	28.7845	33097011	22	24.45868869	35856917
23	32.50899315	33100540	23	28.27377022	35743233
24	37.74997663	33121347	24	32.3982754	35950149
25	45.59394941	33166090	25	38.3289205	35898208
26	56.49152152	33127500	26	48.14190217	35911705
27	73.60984483	33065727	27	61.89892384	35933491
28	105.5027538	33116965	28	84.4693149	35956411
29	220.9436737	33099818	29	175.4542595	35914709
30	1036720	32873081	30	358363	35689901

Table 2.10: RNA-seq bin thresholds (Supporting Figure 2.13C).

The minimum FPKM threshold is 0.3. The left table provides the range of FPKM thresholds of RNA-seq data for HeLa cells, and the right table provides the FPKM thresholds for PA-1 cells. The first column indicates the bin number as on the corresponding graph in Figure 2.13C. The second column indicates the maximum FPKM threshold for the given bin. The third column shows the number of genomic base pairs that lie within the range of that bin. Rows indicate the different bins.

NPC		
Bin	FPKM Max Value of Bin	Genomic bps in bin
1	0.459709	38102677
2	0.664583	38115969
3	1.014338781	38131130
4	1.458215479	37959868
5	2.02636	38097369
6	2.782311842	38098060
7	3.61710834	38148378
8	4.525517692	38121636
9	5.41829	37860522
10	6.375903015	38146209
11	7.420902807	38029917
12	8.41578093	38102873
13	9.5762255	38224419
14	10.7959509	38131137
15	12.06247006	38047218
16	13.5318039	38195782
17	15.14749386	37980459
18	16.81769472	38173138
19	18.62876403	38194791
20	21.01211626	37967876
21	23.46602993	37735702
22	26.39131809	38146580
23	30.1539223	38121118
24	33.87719404	38062299
25	40.6624096	38074312
26	51.22123669	37923409
27	64.99839231	38213587
28	93.5016082	38034312
29	216.4132872	38429031
30	1145440	38880418

hESC		
Bin	FPKM Max Value of Bin	Genomic bps in bin
1	0.431395542	43997452
2	0.630456722	43967342
3	0.870761	43947446
4	1.164081041	43886877
5	1.547452239	43930214
6	1.993006924	44128554
7	2.536438385	44085027
8	3.1386812	43916130
9	3.748654263	44040110
10	4.411615	44127543
11	5.18653511	43969049
12	6.088857721	43976535
13	7.01823	43968868
14	7.96448552	44118426
15	9.228033328	43996549
16	10.2992229	44048143
17	11.54539791	44053523
18	12.97345806	44020200
19	14.66226967	43986962
20	16.36990668	44001391
21	18.49413867	43909177
22	20.72052294	43927822
23	23.56412733	44073132
24	27.108519	43969626
25	31.8956	43938978
26	39.5475551	43983553
27	50.61908632	44035490
28	72.74626068	44077849
29	155.329	43975599
30	768380.5868	43674446

Table 2.11: RNA-seq bin thresholds (Supporting Figure 2.13C).

The min FPKM threshold is 0.3. The left table provides the range of FPKM thresholds of RNA-seq data for NPC cells, and the right table provides the FPKM thresholds for hESC cells. The first column indicates the bin number as on the corresponding graph in Figure 2.13C. The second column indicates the max FPKM threshold for the given bin. The third column shows the number of genomic base pairs that lie within the range of that bin. Rows indicate the different bins.

PA-1 insertion chr 20: 52111326: + poly-A tail: 22bp, EN: 5'-TACTT/AT

Sanger Sequence:

GTTCGAAATCGATAAGCTTGGATCCCCCGACCTCGAGGGGGGAGGCCGGCAAGG
CCGGATCCAGACACGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATG
CAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCAT
TATAAGCTGCAATAAACAAGTTAACAACAAAAAAAAAAAAAAAAAAAAAAAAAAAGTAT
ATTATTTACATTTTTATACATGATAACTCTTGCCTTTGTGTTGAAAAAAAAAAGTCTCT
TTTTTTTCCCCCACTCAGCAGTTATTGGAAATAGACTGTTCCCATCTGAAACCGTATC
GTAATTTGCATCAGGAAACCCAACTGCTGACATTGAGGACCTGGGTGTGTTCAATTA
TGATTTTGCTGGAGGCTGTCCCTCATTTTAATGCTGCAGCTATTGAACCACCTTCCT
GAAACCTAGCTGATACGGAATAGCAGAGACATGCCTCTCAACACCATTAGCTTT

CCS read IDs:
200442:1012058-200442:1013625-200442:1069288-205247:1009755-
205247:1061356-205247:1140523
*CCS reads predicted 27, 28, 33bp poly-A tail

PA-1 insertion chr 7: 36637839: - poly-A tail: 65bp, EN: 5'-ATTTC/AG

Sanger Sequence:

GTTCGAAATCGATAAGCTTGGATCCCCCGACCTCGAGGGGGGAGGCCGGCAAGGC
CGGATCCAGACATGATAAGATACATTGATGAGTTTGGACAAACCACAACTAGAATGCA
GTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTATTGCTTTATTTGTAACCATTAT
AAGCTGCAATAAACAAGTTAACAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAGAAATCCTAATGGGGATGATTCCA
AGACTAACTAGTTTCTGAGATACCAGCAAGGATTCCAGGAATCTCCCAGGTAGCTGA
GTCCCATTTCTTCAGATTAATCCACAGAGAACTGAGCTTGGAGAGTGTTAGCAACAT
GGAGATCCCACTTCCCAAGACAGGATTGCCTGCCGTTTTCCCACCTACCCTACCCCT
CCTCTCGACATCCCCTCTTCCGCCACTGAAGACAAGAAGTTCATCA

CCS read IDs:
220819:1107411-220819:1130961-220819:1134327-220819:1145347-
220819:1150259-220819:1153085
*CCS poly A call 33bp

Table 2.12: LINE-1 insertions validated by independent PCRs.

Sanger Sequence read show results of independent validation PCRs. The L1 sequence (5' to 3') is shown in black, including the poly(A) tail, and the flanking 3' gDNA is shown in blue. The listed PacBio CCS read IDs that support this insertion are shown and the poly(A) tail length as called from the CCS reads is listed next to the asterisk.

Table 2.12: LINE-1 insertions validated by independent PCRs (continued).

hESC insertion chr 2: 219316114: + poly-A tail: 75bp; EN: 5'-TTTTA/AT

Sanger Sequence:

GTTCGAAATCGATAAGCTTGGATCCAGACATGATAAGATACATTGATGAGTTTGGAC
AAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTAT
TGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAATAAAAAGGCCCTCCAACCACCCTAAATGGGATAACTAAGAGTATCTACTGC
AGTCATTTCAGAGGACAGAGAAGGAAAATATTTTAATTTGCTTTAATATAACCTCTTTT
CAGTAGATCACAAATGAGTTTACAAACTACTTTTTTTTCTCTTTAATTTAGGTGTTTGC
AGATAATTTTCATTATACAACTATATCCGTAGCTGTATGTGTGTATAGTTACATAATGGT
AACTACACACGATACAGAAGAA

CCS read IDs:
074805:1013371-074805:1019169-074805:1047928-074805:1048884-074805:104921
2-074805:1052260-074805:1052356-074805:1063792-074805:1065551-074805:1070
666-074805:1073764-074805:1078680-074805:1079798-074805:1084430-074805:10
88679-074805:1088837-074805:1088842-074805:1089043-074805:1089657-074805:
1092387-074805:1092424-074805:1094205-074805:1099218-074805:1105077-07480
5:1113647-074805:1123708-074805:1128160-074805:1138547-074805:1146615-0748
05:1161206-074805:1025494-074805:1029657-074805:1131242-074805:1106586
*CCS poly-A tail call 71bp

hESC insertion chr 12: 86198537: + poly-A tail: 68bp; EN: 5'-GTTCT/AA

Sanger Sequence:

GTTCGAAATCGATAAGCTTGGATCCAGACATGATAAGATACATTGATGAGTTTGGAC
AAACCACAACTAGAATGCAGTGAAAAAAATGCTTTATTTGTGAAATTTGTGATGCTAT
TGCTTTATTTGTAACCATTATAAGCTGCAATAAACAAGTTAACAACAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAGAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AGAACTGATACCAGTGTCCGAGTCTGTGTCATTTGTGTAATTGCTAAATACTCTTTGA
GTAAAGGGAGGAATCTCCCCATTGCTACTGGGAACCTCAGATTCCTTCGCTCTGTT
TTCCTTTAACTGGCACCCATCTTTGTTGCTAATGTGAAGTGAATTGAATTCCTTGGC
CAGGAGTTCATATTCTTTTGCTTTCATCTGAAGCAATGAGTCACTGTATTTAATCTCT
TTCTGGATGCCACTCAAATGAGAGTGAATTTTCAAACCAGCTTTCATGCTTTTCTCC
AAATCACACTTAACACTCTCTAAATTAGAGCTTTCCAGTTCACTTGCAGCTTCCCCTT
CCGCATCTTCATTTATATCAATGCAAACACTTTTTACCTCTTTTTCTATTTCAGCAGAG
AGCTTATCAATGAGTATTCGGTAATATTTCAGTCGTTCTTCCAGCTGTTCAATTCCAT

CCS read IDs:
134902:1023409-134902:1080825-134902:1115290-215905:1036361-215905:106688
5-215905:1081644-215905:1092223-215905:1107408-215905:1111220-215905:11149
53-215905:1146158
*CCS call poly A-tail call 71,76,77,78

Primer Name	Primer Sequence 5’ to 3’
LEAP	GTTCGAAATCGATAAGCTTGGATCC
PGFP1096288	CAAAGCTAATGGTGTTGAGAGGC
PGFP8:7:36637839_OUTER	TTGAAACTTACAAGAATGATATGG
PGFP8:7:36637839_INNER	TGATGAACTTCTTGTCTTCAGTGGC
hESC15:2:219316114_OUTER	TTCTTCTGTATCGTGTGTAGTTACC
hESC15:2:219316115_INNER	TGCAGTAGATACTCTTAGTTATCCC
hESC96:12:86198536_INNER	ATGGAATTGAACAGCTGGAAGAACG
hESC96:12:86198536_OUTER	ACATTAGTTCATCTGATCATTTCCC

Table 2.13: Independent validation PCR primer sequences.

Column one indicates the primer name used in independent validation PCRs and column 2 lists the primer sequence 5’ to 3’.

Sample	HeLa	PA-1	NPC	hESC
SINES	1411 (6.56%)	1679 (6.04%)	804 (6.58%)	350 (9.77%)
LINES	5033 (23.41%)	7373 (26.54%)	3122 (25.54%)	771 (21.52%)

Table 2.14: Engineered L1 insertions in repetitive sequences in the human genome.

Column 1 indicates the repetitive sequences in the human genome that were tested for integration either, SINES or LINES. Coordinates for repetitive sequences are those provided by RepeatMasker. For each sample, we then counted the number and proportion of total insertions (provided in parentheses) found within known SINE and LINE sequences in the GRCh37/hg19 reference.

Table 2.15: NPC DAVID results against the human genome (highest stringency).

Listed are the DAVID Results when testing all genic NPC insertions (UTRs, exonic, or intronic) as compared to the human genome. Column 1 lists the Annotation Clusters. Column 2 shows the enrichment score for a cluster as well as the key words. Column 3 indicates the number of insertions within each group. Column 4 shows the p-value of a modified Fisher's exact test (EASE score), and the multiple testing corrected Benjamini p-value is shown in column 5.

Table 2.15: NPC DAVID results against the human genome (highest stringency).

Annotation Cluster 1	**Enrichment Score: 13.16**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: Fibronectin type-III 3	39	2.4E-15	4.7E-12
UP_SEQ_FEATURE	domain: Fibronectin type-III 1	49	3.6E-14	5.2E-11
UP_SEQ_FEATURE	domain: Fibronectin type-III 2	48	1.2E-13	1.1E-10
SMART	FN3	54	2.2E-12	9.9E-10
Annotation Cluster 2	**Enrichment Score 12.73**	**Count**	**P_Value**	**Bejamini**
UP_KEYWORDS	ATP-binding	246	5.8E-15	3.8E-13
UP_KEYWORDS	Nucleotide-binding	294	1.3E-13	7.4E-12
GOTERM_MF_DIRECT	ATP binding	257	8.7E-12	1.3E-08
Annotation Cluster 3	**Enrichment Score 8.27**	**Count**	**P_Value**	**Bejamini**
UP_KEYWORDS	Transmembrane	747	1.1E-10	4.9E-09
UP_KEYWORDS	Transembrane helix	744	1.5E-10	5.6E-09
UP_SEQ_FEATURE	transmembrane region	668	3.0E-08	1.5E-05
GOTERM_CC_DIRECT	integral component of membrane	678	1.7E-06	1.6E-04
Annotation Cluster 4	**Enrichment Score: 7.29**	**Count**	**P_Value**	**Bejamini**
INTERPRO	Protein kinase, catalytic domain	102	7.8E-10	2.7E-07
UP_KEYWORDS	Serine/threonine-protein kinase	84	2.3E-09	6.0E-08
UP_SEQ_FEATURE	domain: Protein kinase	97	4.5E-09	2.9E-06
INTERPRO	Protein kinase-like domain	106	5.6E-09	1.5E-06
INTERPRO	Protein kinase, ATP binding site	80	5.5E-08	1.4E-05
UP_SEQ_FEATURE	binding site: ATP	105	6.2E-08	2.1E-05
GOTERM_MF_DIRECT	protein serine/threonine kinase activity	80	6.5E-08	3.3E-05
INTERPRO	Serine/threonine-protein kinase, active site	66	6.8E-07	1.3E-04
SMART	S TKc	79	2.7E-06	2.1E-04
UP_SEQ_FEATURE	active site: Proton acceptor	113	6.2E-06	1.1E-03
Annotation Cluster 5	**Enrichment Score: 5.93**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: AGC-kinase C-terminal	23	4.3E-08	2.1E-05
INTERPRO	AGC-kinase, C-terminal	22	3.1E-07	6.3E-05
SMART	S_TK_X	18	1.2E-04	5.0E-03
Annotation Cluster 6	**Enrichment Score 5.68**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: PH	56	2.3E-07	7.2E-05
INTERPRO	Pleckstrin homology domain	59	1.1E-06	1.8E-04
SMART	PH	59	3.7E-05	1.9E-03
Annotation Cluster 7	**Enrichment Score 5.66**	**Count**	**P_Value**	**Bejamini**
UP_KEYWORDS	SH3 domain	51	1.6E-07	2.9E-06
INTERPRO	Src homology-3 domain	50	1.7E-06	2.6E-04
UP_SEQ_FEATURE	domain:SH3	42	2.4E-06	5.3E-04
SMART	SH3	49	3.6E-05	2.0E-03
Annotation Cluster 8	**Enrichment Score 5.65**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: Ig-like C2-type 3	35	8.5E-07	2.2E-04
UP_SEQ_FEATURE	domain: Ig-like C2-type 1	46	3.4E-06	7.2E-04
UP_SEQ_FEATURE	domain: Ig-like C2-type 2	46	3.9E-06	8.0E-04

Table 2.15 (continued)

Annotation Cluster 9	Enrichment Score 5.59	Count	P_Value	Bejamini
UP_SEQ_FEATURE	domain: Fibronectin type-III 6	17	5.2E-08	2.0E-05
UP_SEQ_FEATURE	domain: Fibronectin type-III 8	13	4.3E-06	8.5E-04
UP_SEQ_FEATURE	domain: Fibronectin type-III 7	13	4.3E-06	8.5E-04
UP_SEQ_FEATURE	domain: Fibronectin type-III 9	10	4.7E-05	6.9E-03
Annotation Cluster 10	**Enrichment Score: 5.19**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: Laminin G-like 4	11	4.9E-07	1.4E-04
UP_SEQ_FEATURE	domain: Laminin G-like 3	12	2.0E-06	4.5E-04
UP_SEQ_FEATURE	domain: Laminin G-like 1	13	4.2E-05	6.5E-03
UP_SEQ_FEATURE	domain: Laminin G-like 2	13	4.2E-05	6.5E-03
Annotation Cluster 11	**Enrichment Score: 5.09**	**Count**	**P_Value**	**Bejamini**
GOTERM_MF_DIRECT	ionotropic glutamate receptor activity	11	6.9E-07	2.6E-04
INTERPRO	Ionotropic glutamate receptor	11	6.1E-06	7.6E-04
INTERPRO	Glutamate receptor, L-glutamate/glycine-binding	11	6.1E-06	7.6E-04
INTERPRO	NMDA receptor	11	6.1E-06	7.6E-04
GOTERM_MF_DIRECT	extracellular-glutamate-gated ion channel activity	11	7.2E-06	1.8E-03
SMART	PBPe	11	2.3E-05	1.5E-03
SMART	SM00918	11	2.3E-05	1.5E-03
GOTERM_BP_DIRECT	ionotropic glutamate receptor signaling pathway	12	3.1E-05	1.6E-02
Annotation Cluster 12	**Enrichment Score: 4.65**	**Count**	**P_Value**	**Bejamini**
INTERPRO	P-type ATPase, cytoplasmic domain N	15	1.8E-05	1.8E-03
INTERPRO	Cation-transporting P-type ATPase	15	1.8E-05	1.8E-03
INTERPRO	P-type ATPase, phosphorylation site	15	1.8E-05	1.8E-03
INTERPRO	P-type ATPase, A domain	15	1.8E-05	1.8E-03
UP_SEQ_FEATURE	active site:4-aspartylphosphate intermediate	15	5.7E-05	8.1E-03

Table 2.15 (continued)

Annotation Cluster 13	**Enrichment Score: 4.46**	**Count**	**P_Value**	**Bejamini**
INTERPRO	Dynein heavy chain, P-loop conaining D4 domain	10	7.8E-06	9.1E-04
INTERPRO	Dynein heavy chain, colied coil stalk	10	1.6E-05	1.7E-03
INTERPRO	Dynein heavy chain, domain-2	10	1.6E-05	1.7E-03
INTERPRO	Dynein heavy chain domain	10	1.6E-05	1.7E-03
INTERPRO	Dynein heavy chain	10	1.6E-05	1.7E-03
UP_SEQ_FEATURE	region of interest: AAA 6	9	1.7E-05	2.9E-03
UP_SEQ_FEATURE	region of interest: AAA 5	9	3.6E-05	5.9E-03
UP_SEQ_FEATURE	region of interest: AAA 4	9	3.6E-05	5.9E-03
UP_SEQ_FEATURE	region of interest: Stalk	9	3.6E-05	5.9E-03
UP_SEQ_FEATURE	region of interest: AAA 1	9	3.6E-05	5.9E-03
UP_SEQ_FEATURE	region of interest: AAA 3	9	3.6E-05	5.9E-03
UP_SEQ_FEATURE	region of interest: AAA 2	9	3.6E-05	5.9E-03
UP_SEQ_FEATURE	region of interest: Stem	8	3.4E-04	3.6E-02
UP_KEYWORDS	Dynein	12	5.0E-04	5.0E-03
Annotation Cluster 14	**Enrichment Score: 3.86**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: FERM	17	4.6E-05	6.8E-03
INTERPRO	Band 4.1 domain	17	9.9E-05	8.7E-03
INTERPRO	FERM central domain	17	9.9E-05	8.7E-03
INTERPRO	FERM domain	17	9.9E-05	8.7E-03
INTERPRO	FERM/acyl-coA-binding protein, 3-helical bundle	16	2.8E-04	2.0E-02
SMART	B41	17	5.3E-04	1.9E-02
Annotation Cluster 15	**Enrichment Score: 3.83**	**Count**	**P_Value**	**Bejamini**
INTERPRO	PDZ domain	36	4.4E-05	4.1E-03
UP_SEQ_FEATURE	domain:PDZ	27	2.2E-04	2.5E-02
SMART	PDZ	36	3.3E-04	1.3E-02
Annotation Cluster 16	**Enrichment Score: 3.51**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: EGF-like 7	12	1.1E-04	1.4E-02
UP_SEQ_FEATURE	domain: EGF-like 6	15	1.8E-04	2.2E-02
UP_SEQ_FEATURE	domain: EGF-like 5	13	1.5E-03	1.2E-01
Annotation Cluster 17	**Enrichment Score: 3.45**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: Collagen-like 1	9	2.2E-04	2.4E-02
UP_SEQ_FEATURE	domain: Collagen-like 2	9	2.2E-04	2.4E-02
UP_SEQ_FEATURE	domain: Collagen-like 3	7	9.4E-04	8.3E-02
Annotation Cluster 18	**Enrichment Score: 3.36**	**Count**	**P_Value**	**Bejamini**
UP_SEQ_FEATURE	domain: Fibronectin type-III 12	7	2.6E-04	2.8E-02
UP_SEQ_FEATURE	domain: Fibronectin type-III 11	7	5.2E-05	5.1E-02
UP_SEQ_FEATURE	domain: Fibronectin type-III 10	7	5.2E-04	5.1E-02
UP_SEQ_FEATURE	domain: Fibronectin type-III 13	7	5.2E-04	5.1E-02

References

Aken, B.L., Achuthan, P., Akanni, W., Amode, M.R., Bernsdorff, F., Bhai, J., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P.*, et al.* (2017). Ensembl 2017. Nucleic Acids Res *45*, D635-D642.

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

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

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M.*, et al.* (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Basame, S., Wai-lun Li, P., Howard, G., Branciforte, D., Keller, D., and Martin, S.L. (2006). Spatial assembly and RNA binding stoichiometry of a LINE-1 protein essential for retrotransposition. J Mol Biol *357*, 351-357.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159-1170.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

Bennett, E.A., Keller, H., Mills, R.E., Schmidt, S., Moran, J.V., Weichenrieder, O., and Devine, S.E. (2008). Active Alu retrotransposons in the human genome. Genome Res *18*, 1875-1883.

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

Bickel, P.J., Boley, N., Brown, J.B., Huang, H.Y., and Zhang, N.R. (2010). Subsampling Methods for Genomic Inference. Ann Appl Stat *4*, 1660-1697.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491-500.

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

Buzdin, A., Gogvadze, E., Kovalskaya, E., Volchkov, P., Ustyugova, S., Illarionova, A., Fushan, A., Vinogradova, T., and Sverdlov, E. (2003). The human genome contains many types of chimeric retrogenes generated through in vivo RNA recombination. Nucleic Acids Res *31*, 4385-4390.

Buzdin, A., Ustyugova, S., Gogvadze, E., Vinogradova, T., Lebedev, Y., and Sverdlov, E. (2002). A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3' terminus of l1. Genomics *80*, 402-406.

Christensen, S.M., and Eickbush, T.M. (2005). R2 target-primed reverse transcription: ordered cleavage and polymerization steps by protein subunits asymmetrically bound to the target DNA. Mol Cell Biol *25*, 6617-28.

Conley, M.E. (2005). Two Independent Retrotransposon Insertions at the Same Site Within the Coding Region of BTK. Hum Mutat.

Cost, G.J., and Boeke, J.D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry *37*, 18081-18093.

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

Cost, G.J., Golding, A., Schlissel, M.S., and Boeke, J.D. (2001). Target DNA chromatinization modulates nicking by L1 endonuclease. Nucleic Acids Res *29*, 573-577.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., and Gage, F.H. (2009). L1 retrotransposition in human neural progenitor cells. Nature *460*, 1127-1131.

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

Dimitrieva, S., and Bucher, P. (2013). UCNEbase--a database of ultraconserved non-coding elements and genomic regulatory blocks. Nucleic Acids Res *41*, D101-109.

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

Dombroski, B.A., Scott, A.F., and Kazazian, H.H., Jr. (1993). Two additional potential retrotransposons isolated from a human L1 subfamily that contains an active retrotransposable element. Proc Natl Acad Sci U S A *90*, 6513-6517.

Doucet, A.J., Hulme, A.E., Sahinovic, E., Kulpa, D.A., Moldovan, J.B., Kopera, H.C., Athanikar, J.N., Hasnaoui, M., Bucheton, A., Moran, J.V.*, et al.* (2010). Characterization of LINE-1 ribonucleoprotein particles. PLoS Genet *6*.

Doucet-O'Hare, T.T., Rodic, N., Sharma, R., Darbari, I., Abril, G., Choi, J.A., Young Ahn, J., Cheng, Y., Anders, R.A., Burns, K.H.*, et al.* (2015). LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. Proc Natl Acad Sci U S A *112*, E4894-4900.

Ergun, S., Buschmann, C., Heukeshoven, J., Dammann, K., Schnieders, F., Lauke, H., Chalajour, F., Kilic, N., Stratling, W.H., and Schumann, G.G. (2004). Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. J Biol Chem *279*, 27753-27763.

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

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A.*, et al.* (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483-496.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J.*, et al.* (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49-59.

Evrony, G.D., Lee, E., Park, P.J., and Walsh, C.A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. Elife *5*.

Ewing, A.D., Gacita, A., Wood, L.D., Ma, F., Xing, D., Kim, M.S., Manda, S.S., Abril, G., Pereira, G., Makohon-Moore, A.*, et al.* (2015). Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res *25*, 1536-1545.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res *20*, 1262-1270.

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

Freeman, J.D., Goodchild, N.L., and Mager, D.L. (1994). A modified indicator gene for selection of retrotransposition events in mammalian cells. Biotechniques *17*, 46, 48-49, 52.

Garcia-Perez, J.L., Marchetto, M.C.N., Muotri, A.R., Coufal, N.G., Gage, F.H., O'Shea, K.S., and Moran, J.V. (2007). LINE-1 retrotransposition in human embryonic stem cells. Hum Mol Genet *16*, 1569-1577.

Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P.*, et al.* (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature *466*, 769-773.

Gartler, S.M., and Riggs, A.D. (1983). Mammalian X-chromosome inactivation. Annu Rev Genet *17*, 155-190.

Gasior, S.L., Preston, G., Hedges, D.J., Gilbert, N., Moran, J.V., and Deininger, P.L. (2007). Characterization of pre-insertion loci of de novo L1 insertions. Gene *390*, 190-198.

Gilbert, N., Lutz, S., Morrish, T.A., and Moran, J.V. (2005). Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol *25*, 7780-7795.

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

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

Hancks, D.C., Goodier, J.L., Mandal, P.K., Cheung, L.E., Kazazian, H.H. Jr. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. Hum Mol Genet *20*, 3386-3400.

Hancks, D.C., and Kazazian, H.H., Jr. (2016). Roles for retrotransposon insertions in human disease. Mob DNA *7*, 9.

Helman, E., Lawrence, M.S., Stewart, C., Sougnez, C., Getz, G., and Meyerson, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome Res *24*, 1053-1063.

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

Huang, C.R., Schneider, A.M., Lu, Y., Niranjan, T., Shen, P., Robinson, M.A., Steranka, J.P., Valle, D., Civin, C.I., Wang, T.*, et al.* (2010). Mobile interspersed repeats are major structural variants in the human genome. Cell *141*, 1171-1182.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res *37*, 1-13.

Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Inoue, H., Nojima, H., and Okayama, H. (1990). High efficiency transformation of Escherichia coli with plasmids. Gene *96*, 23-28.

Iskow, R.C., McCabe, M.T., Mills, R.E., Torene, S., Pittard, W.S., Neuwald, A.F., Van Meir, E.G., Vertino, P.M., and Devine, S.E. (2010). Natural mutagenesis of human genomes by endogenous retrotransposons. Cell *141*, 1253-1261.

Januszyk, K., Li, P.W., Villareal, V., Branciforte, D., Wu, H., Xie, Y., Feigon, J., Loo, J.A., Martin, S.L., and Clubb, R.T. (2007). Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. J Biol Chem *282*, 24893-24904.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. PNAS *94, 1872-7.*

Kazazian, H.H., Jr., and Moran, J.V. (2017). Mobile DNA in Health and Disease. N Engl J Med *377*, 361-370.

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

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

Khazina, E., and Weichenrieder, O. (2009). Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. Proc Natl Acad Sci U S A *106*, 731-736.

Kimberland, M.L., Divoky, V., Prchal, J., Schwahn, U., Berger, W., and Kazazian, H.H., Jr. (1999). Full-length human L1 insertions retain the capacity for high frequency retrotransposition in cultured cells. Hum Mol Genet *8*, 1557-1560.

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

Kopera, H.C., Moldovan, J.B., Morrish, T.A., Garcia-Perez, J.L., and Moran, J.V. (2011). Similarities between long interspersed element-1 (LINE-1) reverse transcriptase and telomerase. Proc Natl Acad Sci U S A *108*, 20345-20350.

Kulpa, D.A., and Moran, J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol *13*, 655-660.

LaFave, M.C., Varshney, G.K., Gildea, D.E., Wolfsberg, T.G., Baxevanis, A.D., and Burgess, S.M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res *42*, 4257-4269.

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

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.

Lee, E. (2012). Landscape of Somatic Retrotransposition in Human Cancers. Science *337*, 967-970.

Levin, H.L., and Moran, J.V. (2011). Dynamic interactions between transposable elements and their hosts. Nat Rev Genet *12*, 615-627.

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

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

Lyon, M.F. (1998). X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet *80*, 133-137.

Macia, A., Munoz-Lopez, M., Cortes, J.L., Hastings, R.K., Morell, S., Lucena-Aguilar, G., Marchal, J.A., Badge, R.M., and Garcia-Perez, J.L. (2011). Epigenetic control of retrotransposon expression in human embryonic stem cells. Molecular and cellular biology *31*, 300-316.

Maita, N., Anzai, T., Aoyagi, H., Mizuno, H., and Fujiwara, H. (2004). Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. J Biol Chem *279*, 41067-41076.

Martin, F., Maranon, C., Olivares, M., Alonso, C., Lopez, M.C., (1995). Characterization of a non-long terminal repeat retrotransposon cDNA (L1Tc) from Trypanaosoma cruzi: homology of the first ORF with the ape family of DNA repair enzymes. J Mol Biol. *247*, 49-59

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

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

McCole, R.B., Fonseka, C.Y., Koren, A., and Wu, C.T. (2014). Abnormal dosage of ultraconserved elements is highly disfavored in healthy cells but not cancer cells. PLoS Genet *10*, e1004646.

Melander, A., Olsson, J., Lindberg, G., Salzman, A., Howard, T., Stang, P., Lydick, E., Emslie-Smith, A., Boyle, D.I., Evans, J.M.*, et al.* (1999). 35th Annual Meeting of the European Association for the Study of Diabetes : Brussels, Belgium, 28 September-2 October 1999. Diabetologia *42*, A1-A330.

Miki, Y. (1992). Disruption of the APC Gene by a Retrotransposal Insertion of L1 Sequence in a Colon Cancer. Cancer Research *52*, 643-645.

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

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

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

Muotri, A.R., Marchetto, M.C., Coufal, N.G., Oefner, R., Yeo, G., Nakashima, K., and Gage, F.H. (2010). L1 retrotransposition in neurons is modulated by MeCP2. Nature *468*, 443-446.

Ostertag, E.M., DeBerardinis, R.J., Goodier, J.L., Zhang, Y., Yang, N., Gerton, G.L., and Kazazian, H.H., Jr. (2002). A mouse model of human L1 retrotransposition. Nat Genet *32*, 655-660.

Ostertag, E.M., Prak, E.T., DeBerardinis, R.J., Moran, J.V., and Kazazian, H.H., Jr. (2000). Determination of L1 retrotransposition kinetics in cultured cells. Nucleic Acids Res *28*, 1418-1423.

Paulsen, M.T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E.A., Magnuson, B., Wilson, T.E., and Ljungman, M. (2014). Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. Methods *67*, 45-54.

Paulsen, M.T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E.A., Tsan, Y.C., Chang, C.W., Tarrier, B., Washburn, J.G., Lyons, R.*, et al.* (2013). Coordinated regulation of synthesis and

stability of RNA during the acute TNF-induced proinflammatory response. Proc Natl Acad Sci U S A *110*, 2240-2245.

Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.L., and Hyrien, O. (2016). Replication landscape of the human genome. Nat Commun *7*, 10208.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Lower, J., Stratling, W.H., Lower, R., Schumann, G.G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nuc Acids Res *40*, 1666-1683.

Richardson, S.R., Gerdes, P., Gerhardt, D.J., Sanchez-Luque, F.J., Bodea, G.O., Munoz-Lopez, M., Jesuadian, J.S., Kempen, M.H.C., Carreira, P.E., Jeddeloh, J.A*., et al.* (2017). Heritable L1 retrotransposition in the mouse primordial germline and early embryo. Genome Res.

Riggs, A.D. (1990). Marsupials and Mechanisms of X-Chromosome Inactivation. Aust J Zool *37*, 419-441.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J*., et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., and Pachter, L. (2011). Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol *12*, R22.

Rodic, N., Steranka, J.P., Makohon-Moore, A., Moyer, A., Shen, P., Sharma, R., Kohutek, Z.A., Huang, C.R., Ahn, D., Mita, P*., et al.* (2015). Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. Nat Med *21*, 1060-1064.

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

Scott, A.F., Schmeckpeper, B.J., Abdelrazik, M., Comey, C.T., O'Hara, B., Rossiter, J.P., Cooley, T., Heath, P., Smith, K.D., and Margolet, L. (1987). Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. Genomics *1*, 113-125.

Scott, E.C., and Devine, S.E. (2017). The Role of Somatic L1 Retrotransposition in Human Cancers. Viruses *9*.

Scott, E.C., Gardner, E.J., Masood, A., Chuang, N.T., Vertino, P.M., and Devine, S.E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res *26*, 745-755.

Simons, C., Pheasant, M., Makunin, I.V., and Mattick, J.S. (2006). Transposon-free regions in mammalian genomes. Genome Res *16*, 164-172.

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

Speek, M. (2001). Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. Mol Cell Biol *21*, 1973-1985.

Sultana, T., Zamborlini, A., Cristofari, G., and Lesage, P. (2017). Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet *18*, 292-308.

Swergold, G.D. (1990). Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol *10*, 6718-6729.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327-338.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol *3*, research0052.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol *28*, 511-515.

Tubio, J.M., Li, Y., Ju, Y.S., Martincorena, I., Cooke, S.L., Tojo, M., Gundem, G., Pipinikas, C.P., Zamora, J., Raine, K.*, et al.* (2014). Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science *345*, 1251343.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sanchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M.*, et al.* (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228-239.

van den Hurk, J.A.J.M., Meij, I.C., Seleme, M.d.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S.*, et al.* (2007). L1 retrotransposition can occur early in human embryonic development. Hum Mol Genet *16*, 1587-1592.

Vidaud, D., Vidaud, M., Bahnak, B.R., Siguret, V., Gispert Sanchez, S., Laurian, Y., Meyer, D., Goossens, M., and Lavergne, J.M. (1993). Haemophilia B due to a de novo insertion of a human-specific Alu subfamily member within the coding region of the factor IX gene. Eur J Hum Genet *1*, 30-36.

Watanabe, K., Ueno, M., Kamiya, D., Nishiyama, A., Matsumura, M., Wataya, T., Takahashi, J.B., Nishikawa, S., Nishikawa, S., Muguruma, K.*, et al.* (2007). A ROCK inhibitor permits survival of dissociated human embryonic stem cells. Nat Biotechnol *25*, 681-686.

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

Wei, W., Morrish, T.A., Alisch, R.S., and Moran, J.V. (2000). A Transient Assay Reveals That Cultured Human Cells Can Accommodate Multiple LINE-1 Retrotransposition Events. Analytical Biochemistry *284*, 435-438.

Wimmer, K., Callens, T., Wernstedt, A., and Messiaen, L. (2011). The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. PLoS Genet *7*, e1002371.

Wissing, S., Munoz-Lopez, M., Macia, A., Yang, Z., Montano, M., Collins, W., Garcia-Perez, J.L., Moran, J.V., and Greene, W.C. (2012). Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. Hum Mol Genet *21*, 208-218.

Wulff, K., Gazda, H., Schroder, W., Robicka-Milewska, R., and Herrmann, F.H. (2000). Identification of a novel large F9 gene mutation-an insertion of an Alu repeated DNA element in exon e of the factor 9 gene. Hum Mutat *15*, 299.

Yirmiya, N., Sigman, M., and Freeman, B.J. (1994). Comparison between diagnostic instruments for identifying high-functioning children with autism. J Autism Dev Disord *24*, 281-291.

Zhang, A., Dong, B., Doucet, A.J., Moldovan, J.B., Moran, J.V., and Silverman, R.H. (2014). RNase L restricts the mobility of engineered retrotransposons in cultured human cells. Nucleic Acids Res *42*, 3803-3820.

Zhang, Y., Romanish, M.T., and Mager, D.L. (2011). Distributions of transposable elements reveal hazardous zones in mammalian introns. PLoS Comput Biol *7*, e1002046.

Chapter 3

The Influence of the ATR and FANCD2 DNA Repair Proteins on L1 Integration Preferences in the Human Genome

Dr. Huira Kopera, in the laboratory of Dr. John V. Moran, performed the ATR knockdown, GFP knockdown, and L1 retrotransposition experiments. Mr. Cesar Lopez-Ruiz, in the laboratory of Dr. José L. García-Perez, performed FANCD2-deficient L1 retrotransposition assays. I characterized the resultant L1 retrotransposition events and conducted the data analysis on all the samples presented in this Chapter.

Overview

Previous studies have demonstrated that human LINE-1s lacking endonuclease (L1 EN) activity can retrotranspose by an endonuclease-independent mechanism (L1 ENi) in cultured cells that contain mutations in genes that render the non-homologous end joining (NHEJ) pathway of DNA repair and p53 inactive. As a result, L1 ENi retrotransposition events are hypothesized to integrate at endogenous DNA lesions, which include dysfunctional telomeres. Here, we explore whether the disruption of DNA repair pathways other than NHEJ influence L1 integration in the human genome. First, we knocked down the Ataxia telangiectasia mutated and Rad3-Related (ATR) protein in HeLa cells, and then tested whether ATR knockdown affects the integration preference of engineered L1s. Second, we explored whether the Fanconi anemia (FA) complex influences L1 integration. Defects in the FA pathway lead to the accumulation of inter-strand DNA crosslinks (ICLs) that can interfere with DNA replication. If left unrepaired, the resultant DNA lesions can be processed into double-strand DNA breaks, which may provide a pathway for enhanced ENi L1 retrotransposition. Third, we also examined whether missense mutations in a proliferating cell nuclear antigen (PCNA) interacting domain (PIP) affect L1 integration preferences. Our findings suggest that different DNA repair factors influence L1 integration in the retrotransposition process. FA proteins can

influence initial integration, while ATR and possibly interaction with PCNA appears to influence the later stages of reverse transcription and integration of the L1.

Introduction

A Brief Background about ATR

Below I provide a brief background about ATR biology and the DNA repair response. Notably, the DNA repair response has been thoroughly reviewed by Shiotani and Zou (Shiotani and Zou, 2009), and later by Awasthi, Foiani, and Kumar (Awasthi et al., 2015).

The Ataxia telangiectasia mutated and Rad3-related protein (ATR) is a guardian of the genome and responds to a broad spectrum of DNA damage, including single-strand endonucleotyic nicks, double-strand DNA breaks, and DNA damage that interferes with DNA replication (Zou, 2007). ATR primarily operates during the S and G2 phases of the cell cycle (Jazayeri et al., 2006) and plays a critical role in stabilizing the genome during DNA replication (Brown and Baltimore, 2003). Additionally, ATR plays a key role in the intra-S and G2/M cellular DNA checkpoint pathways (Abraham, 2001). ATR is an essential protein and ATR-deficiency leads to cell autonomous lethality within a few cell divisions (Brown and Baltimore, 2000).

ATR recognizes structures that arise as a result of DNA damage or the inhibition of DNA replication [*e.g.*, single-strand DNA (ssDNA) and double-stranded DNA (dsDNA)/single strand DNA junctions (Shiotani and Zou, 2009)]. For example an accumulation of ssDNA can occur at replication forks when DNA polymerase and DNA helicase activities become discordant (Byun et al., 2005; Walter and Newport, 2000). Additionally, it is believed that common fragile sites represent unreplicated ssDNA regions of DNA that are caused by stalled or collapsed replication forks that are a consequence of DNA replication inhibition (Casper et al., 2002). Several DNA repair mechanisms (*e.g.*, nucleotide excision repair, mismatch repair, and long-patch base excision repair) can also induce ssDNA gaps (Friedberg, 2003; Lopes et al., 2006), as can the resection of double strand DNA breaks by exonucleases and/or endonucelases (Lee et al., 1998).

The replication protein A (RPA) complex initially recognizes ssDNA that arises during DNA damage (Figure 3.1). The ATR-interacting protein (ATRIP), which interacts with ATR, then binds to RPA-coated ssDNA substrates, allowing the localization of the ATR-ATRIP complex to sites of DNA damage (Ball et al., 2005; Namiki and Zou, 2006; Zou and Elledge, 2003). The full activation of the ATR-ATRIP DNA checkpoint response then requires additional ATR regulators, including RAD9, RAD1, and HUS1 (which are components of the 9-1-1 complex), and RAD17. RAD17 recruits the 9-1-1 complexes to 5' dsDNA/ssDNA junctions that arise upon the resection of double-stranded DNA breaks that can arise as a consequence of stalled DNA replication forks (Ellison and Stillman, 2003; Majka et al., 2006; Zou and Elledge, 2003). The recruitment of the ATR-ATRIP and 9-1-1 complexes then stimulates ATR kinase activity at DNA damage sites (Kumagai et al., 2006).

In contrast to the situation by which ATR responds to DNA damage arising during DNA replication, the activation of ATR at double-stranded DNA breaks depends on another kinase, the Ataxia telangiectasia mutated protein (ATM) (Jazayeri et al., 2006; Myers and Cortez, 2006). The MRN complex, which consists of meiotic recombination protein 11 (MRE11), RAD50, and Nijmegen breakage syndrome protein 1 (NBS1) is one of the first lines of cellular defense that recognizes double-strand DNA breaks. The MRN complex then recruits and stimulates ATM kinase activity at double-stranded DNA breaks at all phases in the cell cycle (Berkovich et al., 2007; Falck et al., 2005; Kitagawa et al., 2004; Lee and Paull, 2005; You et al., 2005). ATM activation then leads to the activation of exonuclease and endonuclease activities that function in double-strand break resection, resulting in the formation of long regions of single-strand DNAs at the initial double strand break site in genomic DNA (Jazayeri et al., 2006). The resultant single-strand DNA then serves as a substrate for ATR recruitment and its subsequent activation (Shiotani and Zou, 2009).

ATR is a kinase and phosphorylates a number of proteins that play critical roles in the ATR DNA damage checkpoint response (*e.g.*, RPA, ATRIP, RAD17, members of the 9-1-1 complex, and claspin). The ATR DNA damage checkpoint response ultimately leads to cell cycle arrest to ensure that the damaged DNA can undergo DNA

repair (Shiotani and Zou, 2009). ATR can also phosphorylate claspin, a protein thought to interact with DNA replication forks, in response to DNA damage. Claspin phosphorylation initiates a signal transduction cascade, where ATR phosphorylates the S-phase checkpoint 1 kinase (CHk1), allowing its dissociation from chromatin (Kumagai et al., 2004). The phosphorylated form of the Chk1 kinase can then phosphorylate downstream effector proteins that play critical roles in the G1/S and G2/M cell cycle transitions (Shiotani and Zou, 2009)

Previous work in the Moran lab by Dr. Huira Kopera, using retrotransposition assays in HeLa cells, demonstrated that ATR knockdown results in 3.5- to 10-fold increase in retrotransposition of an engineered wildtype L1 when compared to control HeLa cells. Examination of several integration events suggests that ATR knockdown results in longer (*e.g.*, increased length) L1 insertions. Furthermore, ATR knockdown does not lead to increased L1 ENi retrotransposition. These data suggest that ATR knockdown does not lead to substrates that could be utilized as a primer for reverse transcription by L1 RT during ENi retrotransposition.

In collaboration with Dr. Kopera, we tested whether ATR knockdown affects the integration preferences of engineered human L1s. We hypothesized that ATR knockdown compromises DNA damage sensing within a cell, which may alter the L1 integration profiles in the genome. Alternatively, ATR may not affect initial integration, but instead may respond to retrotransposition intermediates generated during TPRT, thereby leading to altered structures of integrated L1s when compared to ATR-proficient cells. To discriminate between these possibilities, we sought to identify whether there were differences in *de novo* engineered L1 integration preferences in ATR-proficient *vs*. ATR-deficient cells. More specifically, we sought to determine if the DNA damage resulting from the absence of ATR could specifically influence L1 integration. We further hypothesized that the knockdown of ATR may render fragile sites, regions associated with stalled replication forks, or replication origins ‘hot spots’ for L1 integration. In this chapter, I address L1 integration in the human genome when ATR is knocked down, and the potential genomic instability that L1 utilizes for integration.

A Brief Background on Fanconi Anemia

Another DNA damage repair pathway involves the Fanconi Anemia (FA) family of proteins. Twenty-two FA complementation groups have been identified thus far (subtypes A, B, C, D1, D2, E, F, G, I, J, L, M, N, O, P, Q, R, S, T, U, V, and W) (Bogliolo et al., 2013; Bogliolo and Surralles, 2015; de Winter et al., 2000a; de Winter et al., 2000b; de Winter et al., 1998; Deans and West, 2011; Dorsman et al., 2007; Foe et al., 1996; Howlett et al., 2002; Huang et al., 2006; Knies et al., 2017; Levitus et al., 2005; Levran et al., 2005; Litman et al., 2005; Meetei et al., 2004; Meetei et al., 2005; Mosedale et al., 2005; Pickering et al., 2013; Reid et al., 2007; Sims et al., 2007; Smogorzewska et al., 2007; Strathdee et al., 1992; Timmers et al., 2001; Xia et al., 2007). Mutations in genes comprising the Fanconi anemia (FA) pathway can result in both autosomal recessive and X-linked recessive forms of the disease (Sasaki, 1975).

Clinically, FA patients exhibit specific developmental abnormalities, hematological defects, and have increased susceptibilities to cancers (Moldovan and D'Andrea, 2009). A hallmark of FA is hypersensitivity to a class of DNA damaging agents [*e.g.*, mitomycin C (MMC), diepoxybutane, and cisplatin (D'Andrea and Grompe, 2003)] that create inter-strand DNA crosslinks (ICLs) that, if unrepaired, can lead to the accumulation of stalled DNA replication forks that effectively block both DNA replication and RNA transcription (Scharer, 2005). These toxic ICLs are removed primarily during DNA replication (Akkari et al., 2000).

The FA pathway acts to remove ICLs by coordinating the actions of several different DNA repair pathways (*e.g.*, nucleotide excision repair, translesion synthesis, and homologous recombination (Niedzwiedz et al., 2004). Moreover, clastogenic agents [*e.g.*, ultraviolet (UV) and ionizing radiation (IR)], chemical treatments that block DNA replication (*e.g.*, hydroxyurea), and ICLs that spontaneously arise during the process of DNA replication can activate the FA pathway (Dunn et al., 2006; Garcia-Higuera et al., 2001; Howlett et al., 2002; Taniguchi et al., 2002). Most of the resultant DNA lesions arising during replication can be repaired using conventional DNA repair pathways; however, ICL repair relies upon the FA pathway.

When DNA replication becomes blocked at an ICL, a double-strand break is generated by endonucleases (De Silva et al., 2000; Hanada et al., 2006). The resultant double-strand DNA break leads to the uncoupling of one sister chromatid from the other. Incisions occur in the phosphodiester backbone on both sides of the ICL (Niedernhofer et al., 2004), the remaining cross-linked base undergoes processing, and specialized DNA polymerases can then continue replication by bypassing the resultant lesion (Niedzwiedz et al., 2004). The replication fork can then be re-established, which likely involves sister-chromatid mediated homologous recombination.

The mono-ubiquitination of FANCD2 is a key event in FA activation. Following exposure to DNA lesions that arise during DNA replication, FANCM and FAAP24 recruit the FA core complex (which consists of the FA proteins A, B, C, E, F, G, and L) to DNA lesions (Garcia-Higuera et al., 1999; Medhurst et al., 2001; Meetei et al., 2004; Meetei et al., 2003). Recruitment of the FA core complex leads to the mono-ubiquitination of the downstream effector proteins FANCD2 and FANCI. Mono-ubiquinated FANCD2 and FANCI then localize with FANCD1 at stalled replication forks with other DNA repair proteins, which include RAD51, BRCA1, and the proliferating cell nuclear antigen DNA polymerase processivity factor, PCNA (Garcia-Higuera et al., 2001; Howlett et al., 2005; Hussain et al., 2004; Smogorzewska et al., 2007; Taniguchi et al., 2002). Once DNA repair is completed, the FANCD2 protein undergoes de-ubiquitination and the cell cycle resumes (D'Andrea and Pellman, 1998).

Interestingly, ATR is the main upstream regulator of the FA pathway (Friedel et al., 2009). Following the inhibition of DNA synthesis or replication (also known as replication stress), ATR phosphorylates FANCD2, which leads to FANCD2 mono-ubiquitination (Andreassen et al., 2004; Ho et al., 2006; Howlett et al., 2005; Pichierri and Rosselli, 2004). Furthermore, since mutations in the FANCD2 PCNA protein interacting (PIP)-box disrupt FANCD2 mono-ubiquitination, PCNA also may play a role in FANCD2 mono-ubiquitination (Howlett et al., 2009).

An immortalized fibroblast cell line, called PD20F (Coriell Institute GM16633, NA16633), was generated from a 7 year-old male FA patient (Whitney et al., 1995). This patient harbors a maternally inherited FANCD2 allele containing an S126G amino

acid, which promotes mis-splicing, leading, to the inclusion of 13bp from intron 5 into the FANCD2 mRNA (Timmers et al., 2001). This 13bp insertion generates a frameshift mutation that is predicted to result in a severely truncated FANCD2 protein of 180 amino acids. The paternally inherited FANCD2 allele contains a missense change, R1236H. These inherited changes result in markedly diminished FANCD2 protein levels (Timmers et al., 2001).

Utilizing this PD20F immortalized fibroblast cell line, Mr. Cesar Lopez-Ruiz in Dr. José García-Perez lab, performed L1 retrotransposition assays in PDF20F and complemented PD20F cells using three different engineered L1s: wild-type L1.3, an L1.3/D205A EN mutant, and an L1.3/PIP6 mutant. I then performed our L1 capture techniques to determine where the *de novo* engineered L1 integration events occur in these cells. Given the role of FANCD2 in DNA repair and the prevention of stalled replication forks, we hypothesized that the resultant engineered L1 insertions may be driven towards stalled replication forks. Thus, we were eager to explore if there were any potential differences in L1 insertion site profiles in the FANCD2-deficient PD20F versus the complemented FANCD2-proficient cell lines.

While endonuclease independent (ENi) retrotransposition rarely occurs in most tissue culture cell types, cell types with inactivated NHEJ and p53 functions exhibit abundant ENi retrotransposition (Coufal et al., 2011; Morrish et al., 2007; Morrish et al., 2002). Curiously the L1.3/D205A EN mutant retrotransposes at 23% of wild-type L1.3 levels in PD20F, but does not exhibit appreciable retrotransposition in complemented PD20F cells. These data suggest that PD20F cells may contain unrepaired DNA lesions that facilitate L1 ENi retrotransposition. As such, we also wanted to explore if there were any differences in insertion site profiles in the FANCD2-deficient PD20F between the wildtype and EN-deficient engineered L1s.

PCNA and its Role During LINE-1 Retrotransposition

PCNA is a DNA sliding clamp that is an essential co-factor for DNA polymerases during replication (O'Donnell et al., 2013; Siddiqui et al., 2013). PCNA tethers polymerases to DNA and dramatically increases their processivity (Choe and Moldovan,

2017). PCNA lacks enzymatic activity and participates in several DNA repair processes by recruiting various enzymes to DNA. Most proteins interact with PCNA through a canonical PCNA-interacting protein (PIP) box. L1 contains a PIP box at residues 407-415 in the ORF2p (Taylor et al., 2013). There are several hypotheses about how PCNA interaction with L1 may be involved in L1 retrotransposition. One hypothesis is that PCNA may support L1 retrotransposition by recruiting L1 ORF2p onto genomic DNA until a preferred L1 EN target site is recognized to initiate TPRT. Another possibility is that PCNA may act as a processivity factor for L1 RT. Finally, it is possible that ORF2p recruits PCNA to help repair nicks and gaps at junctions between L1 and the host DNA that arise during TPRT. We sought to determine if the loss of PCNA binding affects L1 EN consensus cleavage site recognition and L1 integration preferences.

An L1.3 PIP mutant was created, L1.3/PIP6, which contains a YY414,415 AA mutation located within a PIP-box, which is located between the ORF2p EN and RT domains at amino acids 407-415 (Taylor et al., 2013). Curiously, this L1.3/PIP6 is severely compromised for retrotransposition in HeLa cells, as well as the NHEJ-deficient Chinese Hamster Ovary cell line XR-1, but readily retrotransposes in the FANCD2-deficient cell line PD20F, and the FANCA deficient Chinese Hamster cell line VH4 (unpublished data from Dr. José García-Perez). Intriguingly, a previous study observed decreased ORF2p/PCNA co-purification in HEK293T cells, and a decrease in the retrotransposition efficiency of a L1.3/PIP6 double mutant when expressed from a synthetic L1 in HeLa and HEK293T cells (Taylor et al., 2013). Additionally, PCNA has been implicated in recruiting RNase H2 to RNA/DNA hybrids in genomic DNA (Bubeck et al., 2011). This ability of PCNA is curious, as L1 does not encode a recognizable RNase H activity, even though RNase H activity might be required to degrade the original L1 RNA template strand from the L1 RNA/cDNA duplex that occurs during first-strand L1 cDNA synthesis (Richardson et al., 2014) Indeed, such an RNase H activity could potentially facilitate L1 second-strand L1 cDNA synthesis. Indeed, PCNA may recruit an RNase H activity that functions in L1 integration. Thus, we were interested to test whether the lack of interaction with PCNA affected the integration preferences of the L1.3/PIP6 mutant in PD20F cells.

Since both ATR and FANCD2 are involved in DNA repair pathways, data from these retrotransposition conditions are presented together in this chapter. Curiously, since ATR is a regulator of the FA pathway, we were interested to compare L1 insertions between ATR and FANCD2-deficient conditions to test whether these proteins play different roles in L1 integration. In Chapter 2, we observed that wildtype L1s intersperse throughout the genome regardless of cell type used in our studies. In this chapter we investigate whether we observe a shift in L1 insertion preference when DNA repair processes are perturbed in the cell.

Results

Previous Work: Generation of L1 Insertions

Dr. Huira Kopera performed L1 retrotransposition assays to determine whether ATR knockdown influences L1 integration preferences in the human genome. The retrotransposition indicator cassette *mblastI*, was used to tag the 3' untranslated region (UTR) of a human L1 (pJJ101/L1.3) (Kopera et al., 2011). The resultant construct contains an engineered retrotransposition competent L1 (L1.3) (Dombroski et al., 1994; Sassaman et al., 1997) sequence with a blasticidin deaminase indicator cassette in its 3'UTR (Morrish et al., 2002). Only cells that harbor a *de novo* L1 retrotransposition event will become resistant to blasticidin drug selection. An advantage to the blasticidin retrotransposition cassette is that blasticidin kills cells quickly allowing us to capture ATR knockdown and L1 retrotransposition simultaneously.

HeLa cells were co-transfected with a pJJ101/L1.3 expression constructs and small interfering RNAs that target either ATR (siATR) or GFP (siGFP), which should not affect HeLa cell viability of L1 retrotransposition (Figure 3.2). As a negative control for retrotransposition, an L1 RT mutant (D702A; pJJ105/L1.3) was transfected in parallel. Cells were selected for *de novo* engineered L1 retrotransposition events in the presence of blasticidin three days post-transfection. At completion of the retrotransposition assay, cells were fixed, and were counted to determine the relative L1 retrotransposition efficiency. We observed wildtype L1 retrotransposition events under both ATR and GFP knockdown conditions. Genomic DNA was collected from cells ten days post-

transfection and sequencing libraries were created specifically capturing amplified *de novo* engineered L1 events.

Mr. Cesar Lopez-Ruiz, in Dr. José García-Perez's laboratory, performed the wildtype and mutant L1 retrotransposition assays in the immortalized fibroblast PD20F cell line. Briefly, a wildtype L1 construct, pJJ101/L1.3, and mutant derivatives (pJJD205A/L1.3 and pJJ101/L1.3-YY414,415AA) were transfected into PD20F cells and retrotransposition assays were performed as noted above. All of the constructs contain the retrotransposition-competent L1.3 tagged with the *mblastI* retrotransposition indicator cassette in the L1 3'UTR. An L1 EN mutant (pJJD205A/L1.3), an L1 PIP mutant (pJJ101/L1.3-YY414, 415AA), and an L1 RT mutant (pJJ105/L1.3) were also assayed for retrotransposition in FANCD2-deficient PD20F cells and FANCD2 complemented PD20F cells, which contain a full-length FANCD2 cDNA that was delivered into cells by a retroviral expression vector (Pulsipher et al., 1998). Wildtype L1 retrotransposition occurs in both FANCD2 complemented and FANCD2-deficient PD20F cells. The L1 EN and PIP mutant retrotransposition events are only observed in FANCD2-deficient PD20F cells.

FANCD2-deficient and FANCD2-proficient cells were transfected with the wildtype and mutant L1 constructs noted above. Five days post transfection cells were selected in the presence of blasticidin. A week after selection, cells were isolated, gDNA was collected, and sequencing libraries were assembled specifically capturing *de novo* engineered L1 events.

Capture of L1 insertions in ATR-deficient and FANCD2-deficient Cell Lines

Engineered L1 integration capture procedures were performed as described in Chapter 2. Briefly, genomic DNAs were randomly sheared to 3kb and adapter sequences described were ligated onto the end-repaired, dA-tailed genomic DNA. The dual-biotinylated LEAP primer was used to linearly amplify the 3' end of *de novo* L1 retrotransposition events and their associated 3' flanking sequence. Following biotin capture on streptavidin bead, nested PCR was performed with oligonucleotide primers complementary to the SV40pA signal primer sequence and 3' adapter sequence.

Captured products were characterized using Pacific Bio Science (PacBio) single molecule real-time (SMRT) circular consensus sequence (CCS) sequencing and subjected to the same filtering and analysis schemes described in Chapter 2 (Figure 2.7A). The average CCS read length obtained for insertions was ~600bp (Figure 3.3B).

We identified 3,815 *de novo* engineered L1 insertions in siATR knockdown experiments and 474 *de novo* engineered L1 insertions in siGFP knockdown experiments. In PD20F FANCD2-deficient retrotransposition assays, we identified 18,124 wildtype L1.3, 1,514 L1.3/D205A EN-mutant, and 13,162 L1.3/PIP6 mutant integration events. We also identified an additional 4,372 wildtype L1.3 insertions from the FANCD2-complemented PD20F cell line.

The presence of more than one independent CCS for a single insertion site indicates that at least two independent molecules for the same insertion were identified by PacBio sequencing, which gives us greater confidence that we are observing a *bona fide de novo* engineered L1 integration site as opposed to a PCR artifact. At least two or more independent L1 insertion CCS reads were identified in the siATR knockdown experiments (14.78% of the total insertions) and siGFP knockdown experiments (38.82% of the total insertions) (Figure 3.3C). Moreover, at least two or more independent CCS reads containing L1 insertions were identified in the following cell lines: PD20F FANCD2-deficient cells [wildtype L1.3 (14.61% of the total insertions); L1.3/D205A (51.65% of the total insertions); and L1.3/PIP6 (27.12% of the total insertions)] and FANCD2 complemented cells [wildtype L1 (20.8% of the total insertions)]. Notably, as in Chapter 2, the L1 insertions contained long poly(A) tails (Figure 3.3A), occurred within AT-rich regions of the genome (Figure 3.4) (Gasior et al., 2007; Lander et al., 2001), and integrated at an L1 EN consensus cleavage site (Figure 3.5). Importantly, the L1 consensus target site sequence from L1 EN mutants differs from those of the other L1 constructs (Figure 3.5C) (see below for more discussion). Thus, we have generated a robust engineered L1 insertion dataset from these cell lines.

Engineered Insertions are Located Throughout the Genome

As in Chapter 2, we compared our L1 insertion dataset to 10,000 simulations of the weighted random dataset model that uses the L1 EN degenerate consensus cleavage site to randomly select L1 integration site in the human genome. The wildtype L1 insertions from ATR and GFP knockdown experiments are distributed throughout the genome (Figure 3.6) and, in general, positively correlate with chromosome size (Spearman's rho in the siATR and siGFP experiments is 0.93 and 0.91, respectively). As in Chapter 2, we once again observed a significant increase in the numbers of L1 insertion sites on chromosomes 1 and 5 in the siATR knockdown experiments, likely because there are more than two copies of these chromosomes in HeLa cells. A slight increase, although still within expected boundaries of the weighted random dataset, of insertions on chromosomes 1 and 5 is also seen in the siGFP dataset; however, the smaller difference likely reflects the smaller number of L1 integrations characterized in this experiment (3815 siATR knockdown *vs.* 474 siGFP knockdown L1 insertions, respectively).

The L1 insertions in the FANCD2-deficient cell line are also distributed throughout the genome (Figure 3.6) and positively correlate with chromosome size (Spearman's rho ranges between 0.89 and 0.90). Curiously, all four retrotransposition conditions in PD20F cells show an increase of insertions on chromosomes 8, 11, and 20 when compared to the weighted random dataset. However, none of these large chromosomal insertion counts are significant when testing for outliers using a linear regression model with respect to chromosome size and insertion counts. To our knowledge, PD20F cells do not exhibit chromosome abnormalities, although many FA-deficient cells have an increased proportion of cells with 4N DNA content (Kaiser et al., 1982; Kubbies et al., 1985). However, there is no reason to believe that this would result in an increase in insertions only on these three specific chromosomes. This observed phenomenon appears to be PD20F cell type specific and is observed in both FANCD2-deficient and FANCD2-proficient cellular conditions. Finally, we plotted L1 insertions on a chromosomal ideogram, where each horizontal line represents the genomic location

of an integration event (Figure 3.7). The L1 insertions are clearly distributed across chromosomes and do not show specific 'hot spot' regions of integration preference.

L1 Integration into L1 EN Consensus Cleavage Site Dependent Upon L1 EN Activity

We next analyzed the L1 integration sites in greater detail to determine whether ATR knockdown alters L1 integration preference. In ATR and GFP knockdown cells, the wildtype L1 retrotransposition events integrate into the 7bp L1 EN consensus cleavage site (5'-TTTTT/AA-3') (Figure 3.5A). Correcting the L1 EN consensus cleavage sites for the proportion of 7mers found in the human genome still displays the distinct degenerate L1 EN consensus cleavage site (Figure 3.5B). In FANCD2-proficient and FANCD2-deficient cells, wildtype L1 retrotransposition events integrate into a degenerate L1 EN consensus cleavage site (Figure 3.5C). In FANCD2-deficient cells, L1 PIP mutant retrotransposition events also integrate into a degenerate L1 EN consensus cleavage site (Figure 3.5C). Thus, deficiencies in ATR and FANCD2 do not appear to alter wildtype or PIP mutant L1 integration preferences (Figure 3.5B).

In contrast to the above data, L1 ENi retrotransposition events in FANCD2-deficient cells display a shift in integration within L1 target site sequence variants. We observed a decreased preference for a T in the 2nd and 5th position of the 7bp L1 EN consensus sequence, and an even greater decreased preference for an A in the 6th and 7th position (Figure 2.7B). This shift in target site preference is consistent with the hypothesis that ENi is occurring at endogenous lesions in genomic DNA.

Genes are not Preferential L1 Integration Sites

We next addressed whether the knockdown of ATR or the absence of FANCD2 alters the ability of engineered L1s to retrotranspose into genes identified in UCSC build GRCh37/hg19 of the human genome. Consistent with the data reported in Chapter 2, we did not identify significant depletions or enrichments of any of our L1 insertion datasets within exons when compared to our random weighted model (Figure 3.8A). The L1 retrotransposition events from ATR knockdown and GFP knockdown cells also show no significant depletion or enrichment of insertions within introns of genes when compared to the weighted random dataset (Figure 3.8B).

Of the PD20F sample insertion datasets, the L1.3/PIP6 insertions show a significant depletion of insertions in introns (Pearson's χ^2 test, p-value 1.792×10^{-5}). The wildtype L1 retrotransposition events in FANCD2 complemented PD20F also show a significant depletion of intronic insertions (Pearson's χ^2 test, p-value 0.04724). Notably, these datasets contain the greatest number of L1 insertions examined in the chapter, suggesting that large insertion datasets are needed to observe a depletion of insertions in introns. Additionally, we did not observe a significant excess of L1 insertions in the antisense orientation of genes for any of the analyzed datasets. However, as reported in Chapter 2, there is a slight preference for antisense integration into genes, which likely reflects the increased prevalence of the L1 EN consensus cleavage site in the sense strand of genes (Figure 3.8C).

L1 Integration in ATR-deficient Cells is Independent of Gene Transcription and Expression

We next explored whether ATR knockdown influences L1 integration towards expressed regions of the genome. During transcription, the DNA coding strand becomes untethered from its complementary strand, potentially becoming an accessible target for L1 EN cleavage. We found that transcription status does not have a significant effect on the distribution of L1 insertions in ATR knockdown HeLa cells (Figures 3.9A).

We next examined L1 retrotransposition events from ATR and GFP knockdown with HeLa cell Bru-seq transcription data (Paulsen et al., 2014; Paulsen et al., 2013) and RNA-seq expression data. Consistent with the data reported in Chapter 2, we did not observe an enrichment or depletion of L1 insertions in transcribed regions of the genome when compared to the weighted random model (Figures 3.9A, 3.9B, 3.10A, and 3.10B). Moreover, the rate of transcription does not significantly alter L1 integration preferences (Figure 3.9C). However, as observed in Chapter 2 (Figure 2.4D), there appears to be a slight underlying preference towards EN cleavage on the coding strand in the genome (Figure 3.9D). Notably, our approach may have pitfalls if ATR knockdown significantly alters the HeLa cell transcription profile and results in global changes in the assignments of expressed and non-expressed genes. That being stated, ATR knockdown does not appear to significantly alter L1 integration profiles.

L1 Integration in ATR Knockdown Cells is Independent of Replication

When DNA polymerase activity and DNA helicase activity become discordant, increased amounts of ssDNA, an ATR substrate, can be generated at stalled DNA replication forks. Thus, we next determined whether ATR knockdown altered L1 integration preferences into lagging or leading strand DNA replication templates. We did not observe statistically significant changes in the numbers of L1 insertions favoring the lagging strand template when comparing our wildtype L1 insertion dataset with Okazaki sequencing data from HeLa cells (Petryk et al., 2016) (Figure 3.11A and 3.11B). Similarly, we did not observe an L1 integration bias towards replication fork initiation or termination sites (Figure 3.11C). Once again, our approach may have pitfalls if ATR knockdown affects replication fork initiation, directionality, or termination in a significant portion of the genome. That being stated, the preliminary data presented here suggests that ATR knockdown does not lead to replication biased changes in L1 integration preferences.

ENi Integration Events Display Replication Bias Favoring Okazaki Fragments as ENi Substrates

We next compared the insertion profiles of wildtype L1 retrotransposition events generated in FANCD2-deficient cells with the Okazaki sequencing data from lymphoblastoid cells and HeLa cells. Comparisons with both cell types showed the same trends, but in this chapter we only show results from comparisons with lymphoblastoid cell OK-seq data. Consistent with data obtained from wildtype L1 retrotransposition events reported in Chapter 2 (Figure 2.6B), wildtype L1s show a slight preference for L1 EN cleaving the lagging strand template in FANCD2-deficient cells (Figures 3.11A and 3.11B; bottom strand cleavage Ks.bootstrap p-value $< 1 \times 10^{-6}$; top strand cleavage Ks.bootstrap p-value < 0.001). A similar trend also was observed for wildtype L1 insertions in FANCD2-proficient cell lines (Figures 3.11A and 3.11B).

As opposed to wildtype L1s, endonuclease-deficient L1s likely exploit endogenous DNA lesions to mediate their insertion (Morrish et al., 2007; Morrish et al., 2002). Indeed, Okazaki fragments present at DNA replication forks during lagging

strand DNA synthesis contain a 3'-OH group, which may provide a suitable substrate for the L1 RT to initiate TPRT in the absence of L1 EN activity. Thus, we next compared the insertion profiles of ENi L1 retrotransposition events generated in FANCD2-deficient cells with the Okazaki sequencing data. As opposed to the trends observed for wildtype L1s (Figures 2.6B, 3.11A and 3.11B), the ENi retrotransposition integration events show an opposite integration preference (Figures 3.11A and 3.11B; Kolmogorov-Smirnov boot strap test bottom and top strand analysis p-values < 0.01). Together, the above data suggest that wildtype L1s may show a slight preference for insertion into the lagging strand template.

PCNA is Not Required for L1 Integration into Replication Forks

Since PCNA is an essential co-factor for DNA replication, PCNA may be involved in guiding L1 to replication forks to direct integration. Thus, we hypothesized that the L1.3/PIP6 PIP-box mutant may block interactions with PCNA and alter L1 integration preferences within the genome. To test this hypothesis, we compared the L1 insertion datasets in FANCD2-deficient cells to lymphoblastoid Okazaki sequencing data. We observed a slight preference for L1 EN cleavage on the lagging strand template for the L1.3/PIP6 PIP-box mutant (cleavage on top strand Ks.bootstrap test p-value < 0.01; cleavage on bottom strand Ks.bootstrap test p-value < 0.05) (Figures 3.11A and 3.11B). Notably, this trend is the same as that observed for wildtype L1s in FANCD2-deficient cells, as well as the L1 insertions reported in Chapter 2. This data suggests that PCNA is not involved in guiding L1 EN cleavage to a specific strand in replication forks.

Finally, we wanted to test whether the L1.3/PIP6 mutant preferentially integrates into replication fork origins or replication fork termination sites. While PD20F L1.3/PIP6 insertions significantly differ from the weighted random dataset (Ks.bootstrap test p-value <0.01), most insertions are found within regions of the genome that are replicated by overlapping extending forks, and not overwhelmingly in replication origins or sites of termination (Figure 3.11C). Intriguingly, wildtype L1 insertions in FANCD2-deficient cells are depleted in replication termination sites, while L1.3/PIP6 PIP-box mutant insertions are depleted in both replication origins and replication termination sites. Thus,

comparisons of these data suggest that PCNA may possibly play a role in guiding wildtype L1s to replication fork origins.

ATR Knockdown does not Drive L1 Integration to Known Fragile Sites in the Human Genome

Common chromosomal fragile sites are unstable genomic regions that break under replication stress. Aphidicolin, an inhibitor of DNA replication, induces common fragile sites . Fungtammasan *et al.* (2012) published a set of defined aphidicolin-induced common fragile sites, as well as a set of non-fragile sites. Since it is believed that fragile sites represent unreplicated ssDNA regions that are caused by stalled or collapsed DNA replication forks, and ssDNA regions are substrates for ATR activation, we hypothesized we might observe an increase of L1 integration into stalled replication forks in the absence of ATR. Thus, we next asked whether L1 insertions in ATR knockdown cells are enriched within common fragile sites when compared to simulated L1 insertions generated using our weighted random dataset (Table 3.1). In general, we observed that ATR knockdown neither significantly increases L1 integration into fragile site regions, nor does ATR knockdown significantly decrease L1 integration into non-fragile sites of the genome. Notably, these comparisons are limited in their ability to determine if fragile sites induced specifically by ATR knockdown are L1 integration preferences, as these data sets represent common fragile sites induced by aphidicolon. Furthermore, these data sets will not allow us to determine if preferential integration occurs at random fragile sites induced by ATR knockdown.

FANCD2-deficient Cells do not Harbor More L1 Integration Sites in Common Fragile Sites

Since FANCD2 is involved in the repair of ICLs, which typically arise during DNA replication, we next hypothesized that FANCD2-deficient cells may lead to an increase in stalled replication forks that arise at common fragile sites (Casper et al., 2002). Thus, we tested whether L1 insertions in FANCD2-deficient cells are enriched within common fragile sites when compared to simulated L1 insertions generated using our weighted random dataset. In general, though we did observe a slight over-representation of

wildtype L1 insertions into non-fragile sites in FANCD2-proficient cells as compared to our simulated L1 integration sites generated from our weighted random model (Table 3.1). We did not observe an enrichment of wildtype, EN mutant, or L1.3/PIP6 insertions into common fragile sites in FANCD2-deficient cells (Table 3.1). However, the L1.3/PIP6 insertions exhibit an increase in non-fragile sites as well. Thus, we did not observe an increase of L1 integration sites into common fragile sites, but these results cannot rule out L1 integration preference in fragile sites induced by stalled replication forks in FANCD2-deficient cells. Aphidicolin induced common fragile sites may differ in location from FANCD2-deficient fragile sites.

Discussion

ATR Affects L1 Integration Post-EN Cleavage

Our data suggests that ATR knockdown does not significantly affect L1 integration preference. Intriguingly, data generated by Dr. Huira Kopera indicated that ATR knockdown leads to increased L1 retrotransposition in HeLa cells. Moreover, the resultant integration events tend to be 'longer' in length, are often flanked by long target site duplications, and have a higher incidence of genomic deletions at the L1 integration site. How these structural alterations arise requires further study.

The activation of ATR at double-strand DNA breaks is reliant upon ATM (Jazayeri et al., 2006; Myers and Cortez, 2006). Interestingly, ATM-deficient mice and NPCs also exhibit slightly elevated levels of L1 retrotransposition (Coufal et al., 2011). Moreover, L1 retrotransposition events in ATM-deficient NPCs appear to be longer in length when compared to ATM-proficient controls (Coufal et al., 2011). Thus, both ATM and ATR may play a role in combating L1 retrotransposition.

Given the above data, we hypothesize that ATR and ATM may recognize DNA structures that arise during TPRT to inhibit L1 retrotransposition. Potential substrates include: (1) single-strand L1 cDNAs; (2) single-strand L1 cDNA/dsDNA junctions; or (3) single nicks or single-strand DNA substrates that arise on the 'top' DNA strand during TPRT. The activation of ATR and/or ATM at these structures may then lead to the recruitment of DNA repair proteins that recognize TPRT as DNA damage intermediates,

thereby inhibiting retrotransposition. Consistent with the idea that TPRT intermediates may be recognized as sites of DNA damage, it is notable that the overexpression of components of the nucleotide excision repair pathway strongly inhibits engineered L1 retrotransposition in cultured human cells (Servant et al., 2017).

The L1 PIP-box is Not Required for Directing ORF2p to L1 EN Target Site

FANCD2-deficient L1.3/PIP6 insertions display a long poly(A) tail (Figure 3.3A), are found in AT-rich regions of the genome (Figure 3.4), and contain the L1 EN consensus cleavage site (Figure 3.5C). This data suggests that interaction with PCNA is not required for L1s to integrate into L1 EN consensus cleavage sites throughout the genome. This further implies that L1 ORF2p/PCNA interactions would likely be important for L1 retrotransposition after the initiation of L1 EN cleavage of genome DNA during TPRT.

Since PCNA plays an important role in DNA replication, we hypothesized that the interaction of L1 ORF2p with PCNA may be involved in guiding L1 to replication forks. Our data does not suggest any profound discrepancies in L1 integration preference at replication forks. We observe the same trend in data among wildtype L1s and PIP6—they both show a slight preference at cleaving the lagging strand template of DNA replication forks to mediate L1 integration (Figures 3.11A and 3.11B). Similar trends were also observed for wildtype L1 insertions in FANCD2-deficient cells (Figure 3.11A and 3.11B), as well as in other cell types examined in Chapter 2 (Figures 2.6 and 2.7).

We observed suggestive evidence that PCNA may be involved in directing L1 integration towards DNA replication origins (Figure 3.11C). This data is by no means conclusive, but certainly is worthy of follow up experimentation. Indeed, comparing the L1 PIP mutant insertion profiles with OK-seq data generated from the PD20F FA-deficient cell line, could yield a definitive conclusion as to whether PCNA guides L1 to replication fork origins. Given this suggestive evidence that PCNA may guide L1 to replication fork origins, it will also be interesting to test whether PCNA guides endonuclease-deficient L1s to stalled replication forks. Indeed, the examination of the

L1 insertion profiles of PIP-box/EN- double mutants could critically test if PCNA plays a role in ENi L1 retrotransposition.

Finally, it is possible that PCNA plays an important role during TPRT. For example ORF2p/PCNA interactions could potentially increase the processivity of the L1 RT activity or could recruit other proteins [*e.g.*, RNase H2 (Bubeck et al., 2011) or DNA ligase] to sites of TPRT, thereby facilitating the completion of L1 integration. Indeed, examining the structures of L1 PIP-box retrotransposition events may provide insight about the roles of PCNA during L1 integration.

FANCD2-deficiency Does Not Influence Wildtype L1 Integration Preference

The wildtype L1 integration events in FANCD2-deficient cells show the same general integration preferences as the wildtype insertion datasets analyzed in Chapter 2. These data are consistent with previous studies from our lab (Morrish et al., 2007; Morrish et al., 2002) and suggest that wildtype L1s integrate via canonical TPRT even under cellular conditions that promote ENi L1 retrotransposition. An additional means to test this hypothesis would be to characterize wildtype L1 integration sites in FANCD2-deficient cells to determine if the resultant L1 integration events exhibit structural hallmarks indicative of canonical TPRT or ENi retrotransposition (*e.g.* 5' and 3' truncations, lack TSDs, deletions at integration site).

Finally, since FANCD2 is primarily involved in repairing ICLs that arise throughout the genome during DNA replication, the data presented here cannot directly assess whether wildtype L1s preferentially target unrepaired ICLs as integration substrates. Thus, while we did not observe a significant skewing of wildtype or L1.3/PIP6 integration events in FANCD2-deficient cells, we cannot rule out the possibility that L1 integration preferences in these cells may be influenced by unresolved ICLs in the genome.

Endonuclease-independent Retrotransposition Can Utilize Replication Forks for Integration

Our observations in FANCD2-deficient cells strongly suggest that stalled replication forks or structures arising during the repair of stalled replication forks may provide integration substrates for ENi L1 retrotransposition. The resultant ENi L1 retrotransposition integration events also differ, subtly but significantly, from other integration data sets with respect to their target integration site sequence (Figure 3.5C). The ENi L1 integration sites are still enriched for the first five T residues, but show a distinct decrease in the last two A residues in the 7bp target site sequence. Moreover, examination of the nucleotides surrounding the EN mutant L1 integration target site show a distinct peak in T rich sequence up to 25bp upstream of the integration site as compared to other insertion datasets (Figure 3.4A). These data suggest that interactions between the L1 poly(A) tail and single-strand, T-rich 3' overhangs are important structures to initiate ENi L1 retrotransposition (Kulpa et al., 2006). The decrease in the preference of the last two A residues, which are critical for L1 EN target site (Figure 3.5C), provide additional support that ENi L1 retrotransposition events are using endogenous lesions as DNA repair substrates.

Future studies are needed to determine the nature of the integration substrates used to accommodate ENi L1 retrotransposition in FANCD2-deficient [and for that matter XRCC4-deficient cells (Morrish et al., 2002)]. Stalled replication forks may be processed to double-strand break repair intermediates that are favorable substrates of ENi L1 retrotransposition. Alternatively, and consistent with the data presented in Figures 3.11A and 3.11B, it is possible that the 3'OH groups present at Okazaki fragments could be used as primers to initiate ENi L1 retrotransposition. Indeed, in certain respects, the structures at DNA replication forks and/or at double-strand breaks that arise from collapsed DNA replication forks may be similar to those at dysfunctional telomeres, which can serve as substrates for ENi L1 retrotransposition (Morrish et al., 2007; Morrish et al., 2002). Thus, our data suggest a new mechanism by which L1s lacking endonuclease activity utilize properties of the replication cycle to integrate within the genome.

Materials and Methods

Plasmids were purified using a Qiagen Plasmid Midi kit (#12143).

Plasmids

pJJ101/L1.3 : This plasmid was described previously (Kopera et al., 2011). It contains a full-length retrotransposition-competent L1 element, L1.3 (Sassaman et al., 1997), tagged with a *mblastI* retrotransposition indicator cassette in the 3'UTR and subcloned in pCEP4 (Invitrogen).

pJJD205A/L1.3 : This plasmid was described previously (Kopera et al., 2011). This construct is similar to pJJ101/L1.3 but contains a D205A mutation in the ORF2p EN domain.

pJJ101/L1.3-YY414,415AA : is a derivative of JJ101/L1.3 that contains a double missense mutations in the PIP domain of L1-ORF2p (YY at position 414 and 415 mutated to A).

pJJ105/L1.3 : This plasmid was described previously (Kopera et al., 2011). This construct is identical to pJJ101/L1.3 but contains a missense mutation (D702A) in the ORF2p RT domain.

pCEP/GFP: has been described previously (Alisch et al., 2006). It consists of a pCEP4 backbone (Invitrogen/Life Technologies; Carlsbad, CA #V04450) that contains the coding sequence of the humanized Renilla green fluorescent protein (hrGFP) from phrGFP-C (Stratagene) driven by a CMV promoter.

Cells culture and transfection of ATR and GFP knockdown HeLa cells

HeLa cells were grown in a humidified incubator at 37°C and 7% CO_2. HeLa cells were grown in DMEM (Life Technologies #11960051) with 10% FBS, 1× Pen Strep Glutamine (100 U/mL penicillin, 100 µg/mL streptomycin, and 292 µg/mL glutamine; PSG), and 1mM NaPyruvate. All DNA transfections were performed with FuGene6 (Promega #E2692) according to manufacturer's directions. Briefly, $1x10^6$ HeLa cells were plated in a 10 cm tissue culture dish. The following day, 18 µg of pJJ101/L1.3 LINE-1 plasmid DNA was incubated with 90 µL of 10 µM GFP or ATR siRNA (Dharmacon, #P-002048-

01-20 and #M-003202-05-0005, respectively) and 60 µl of DuoFect (Dharmacon, #T-2010-03) in 1 ml of OptiMem for 20 minutes at room temperature before adding the transfection mix to the cells. After 16-18 hours, transfections were stopped by changing the media (day 1 post-transfection). Three days post-transfection, retrotransposition events were selected under 10 mg/ml blasticidin. Media was changed again 6 days post-transfection. Ten days post-transfection, cells were harvested for genomic DNA purification. Genomic DNA was isolated and purified using the Qiagen Blood and Cell Culture DNA Midi kit (Qiagen #13343).

PD20F Cell culture

All reagents were purchased from GIBCO-Life Technologies unless otherwise indicated. PD20F (FANCD2 mutant cells) and complemented PD20FD2 cells (FANCD2 mutant cells complemented with a retroviral vector containing the human FANCD2 cDNA) were grown using Dulbecco's Modified Eagle Medium (DMEM) high glucose, GlutaMAX, supplemented with 10% fetal bovine serum (FBS), and Penicillin-Streptomycin (10,000 U/mL). Cells were passaged by standard trypsinization (using a 0.05% stock). The absence of *Mycoplasma spp*. in cultured cells was confirmed at least once a month by a PCR-based assay (Minerva) and STR-genotyping confirmed the identity of the cell lines (LorGen, Granada, Spain).

Transfection of PD20F cells and Retrotransposition assays

All retrotransposition assays performed in PD20F and FANCD2 complemented PD20F were performed in 4 biological replicate reactions. PD20F and PD20FD2 cells were transfected using Fugene6 (Promega #E2692) using manufacturer instructions. Briefly, $8x10^4$ cells were plated per 100mm culture plates and transfected 16h later using 10µl of Fugene6 and 4µg of each plasmid DNA using OptiMEM (Invitrogen #31985062) following the manufacturer instructions. 24h later, fresh media was added and cells were cultured for the next 4 days, changing the media every other day. Five days after transfection cells were selected with 2µg/ml Blasticidin-S (Invitrogen) for the following 7 days, changing the media every other day. After the selection process, blast-resistant foci were harvested by trypsinization and genomic DNA extracted. Transfection

efficiency controls were included, co-transfecting cells in parallel with a GFP expression vector and determining the percentage of GFP-expressing cells 48h post-transfection by FACS.

Genomic DNA isolation

Genomic DNAs from PD20F retrotransposition assays were purified using a DNeasy Blood & Tissue Mini Kit (Qiagen #69504) following the manufacturer's instructions.

Acknowledgements

Thanks to Dr. Huira Kopera who performed the siATR and siGFP retrotransposition assays in HeLa cells. Dr. Kopera also isolated gDNA from retrotransposition assays and provided text for the materials and methods. Dr. Kopera provided helpful reviews on ATR background and engaged in enlightening discussions regarding these experiments. Thanks to Mr. Cesar Lopez-Ruiz for performing the retrotransposition assays in PD20F cell line as well as isolating and sending gDNA after completion of assays. Thanks to Dr. José García-Perez for this opportunity to identify engineered L1 insertion events from the PD20F retrotransposition assays, as well as providing pertinent text for the materials and methods.

Figure 3.1 Recognition of DNA damage by ATR

This figure was adapted and modified from Shiotani and Zou et al. 2009. DNA damage that results in single-strand DNA (ssDNA) either from stalled replication forks, or resection of DNA becomes a target for replication protein A (RPA). Ataxia telangiectasia mutated and Rad3-Related (ATR) and ATR-interacting protein (ATRIP) are then recruited to the RPA-ssDNA complex. Independently, RAD17 is also recruited to the RPA-ssDNA complex. Then the 9-1-1 complex, consisting of Rad9, Rad1, and HUS1, is recruited to the site of damage along with TopBP1. At this point ATR becomes activated to stabilize the genome during this DNA damage.

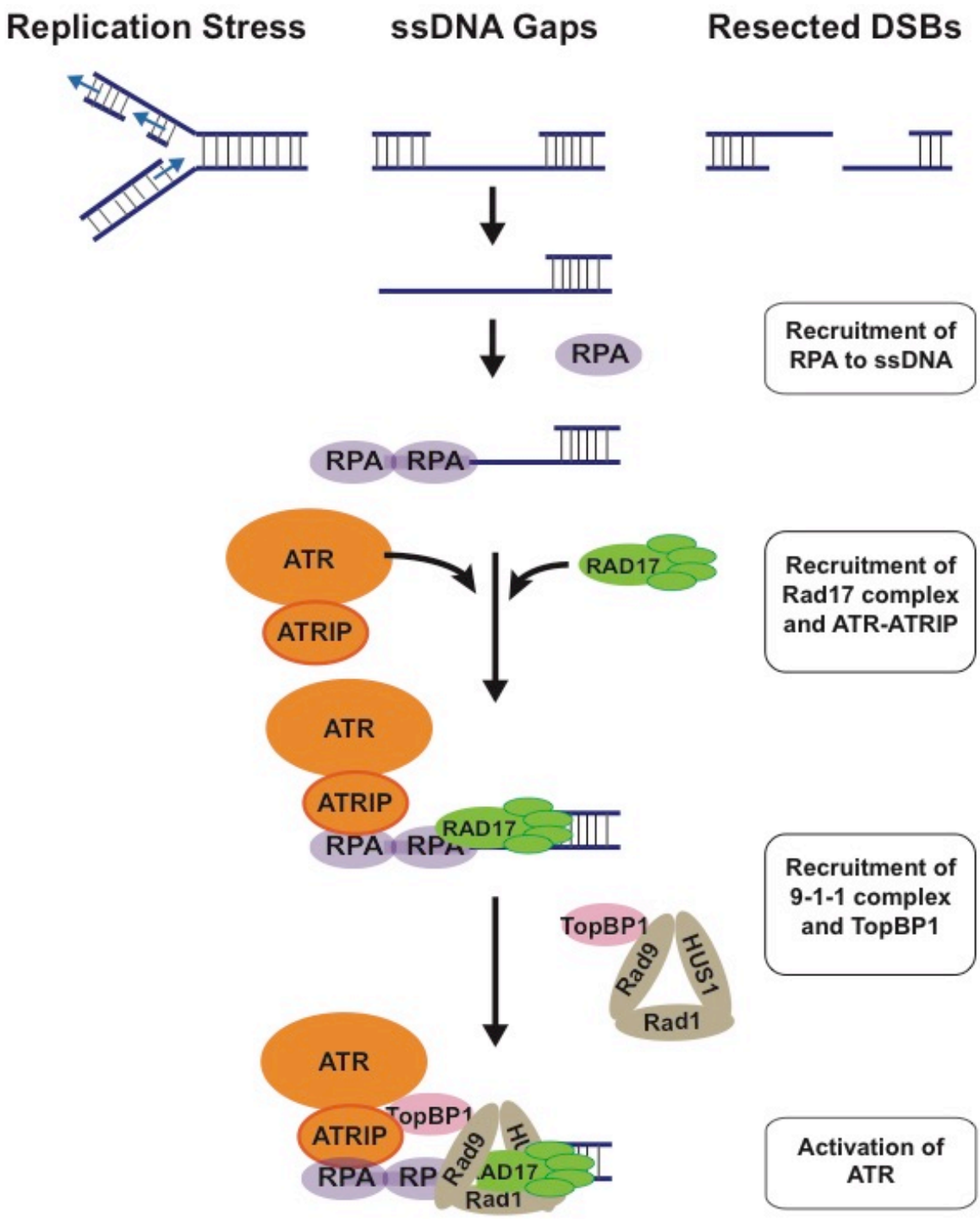


Figure 3.1 Recognition of DNA damage by ATR

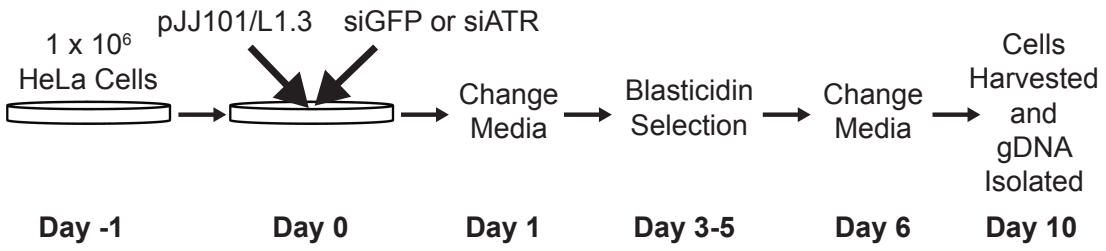


Figure 3.2 Schematic of siRNA knockdown retrotransposition assay in HeLa cells

On Day -1, $1x10^6$ HeLa cells are plated in a 10cm tissue culture dish. Twenty-four hours later on Day 0, the HeLa cells are co-transfected with pJJ101/L1.3 and siRNA, either siGFP or siATR. Twenty-four hours post-transfection media is changed. Three days after transfection selection with blasticidin begins for the next 48 hours. On day 6 the media is changed, and cells are harvested and gDNA isolated on Day 10.

Figure 3.3 Engineered insertions display known LINE-1 insertion characteristics

A) *Engineered insertions display long poly(A) tails:* Histograms of poly(A) tail lengths in base pairs (x-axis) for PD20F insertions complemented with FANCD2 and transfected with pJJ101/L1.3 (labeled 'PD20F : PD20FD2 + L1.3'), PD20F cells transfected with pJJ101/L1.3 (labeled 'PD20F : L1.3'), PD20F cells transfected with L1.3 EN mutant pJJD205A/L1.3 (labeled 'PD20F : L1.3/D205A'), and PD20F cells transfected with L1.3 PIP-box mutant pJJ101/L1.3-YY414,415AA (labeled 'PD20F : L1.3/PIP6'). Histogram of poly(A) tail lengths for HeLa retrotransposition assays involving wild type L1.3 transfected with pJJ101/L1.3 and siATR knockdown (labeled 'siATR'; right top plot) or siGFP knockdown (labeled 'siGFP'; right bottom plot). These insertions had the representative longer poly(A) tail length due to the SV40 polyadenylation signal present in the engineered L1 constructs.

B) *PacBio Sequencing of Insertions results in long CCS Reads:* Histogram of PacBio CCS read lengths in base pairs (x-axis) acquired for insertions from the respective labeled transfection conditions.

C) *Independent PacBio CCS reads support insertions:* A plot of the proportion of insertions that are supported by 1 or more independent PacBio CCS reads. Number of independent CCS reads supporting a single insertion is show on the x-axis, while proportion of total insertions is on the y-axis.

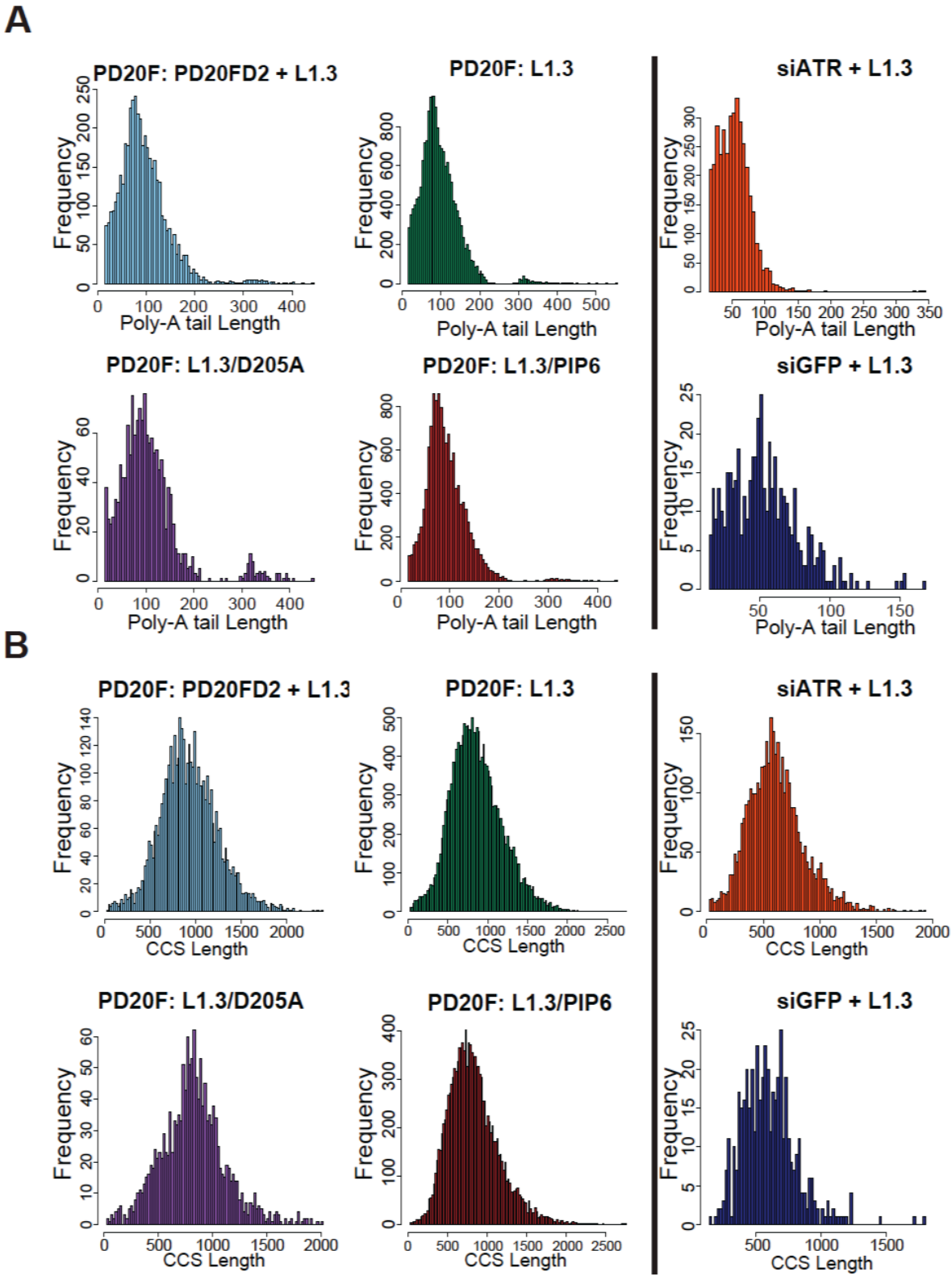


Figure 3.3 Engineered insertions display known LINE-1 insertion characteristics

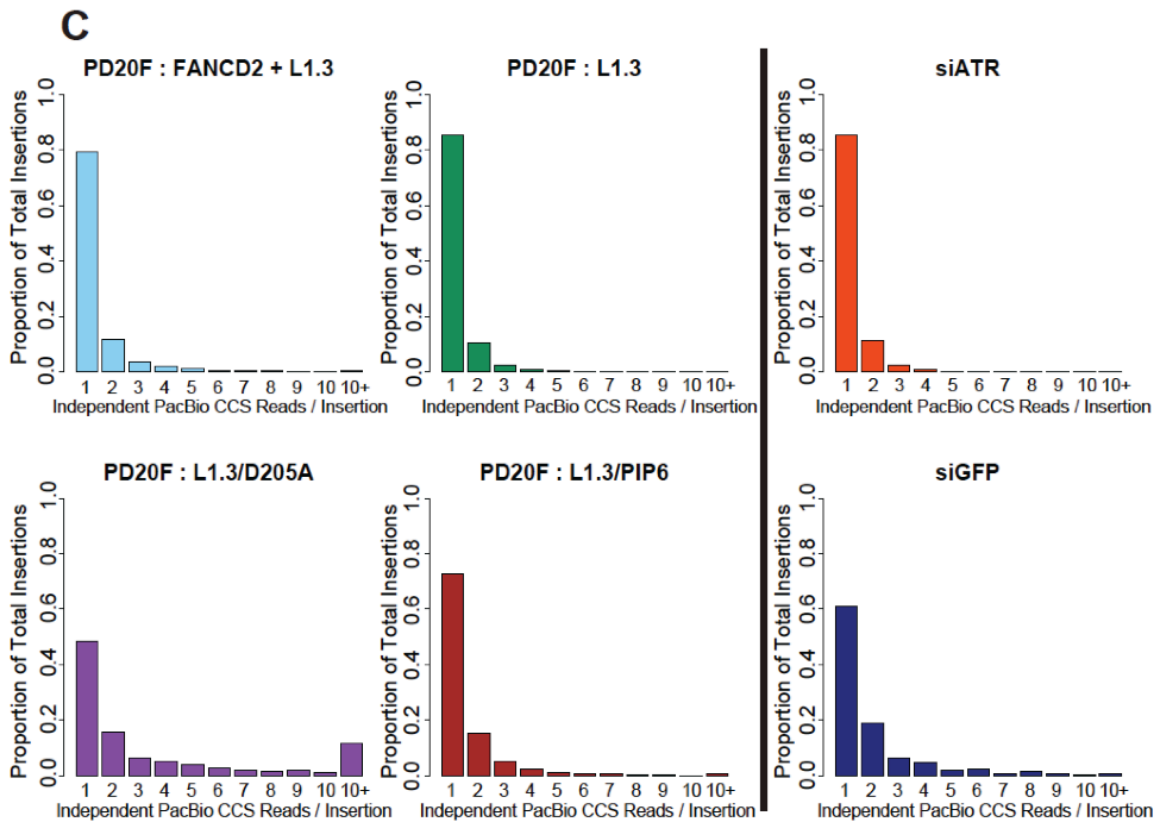


Figure 3.3 Engineered insertions display known LINE-1 insertion characteristics

Figure 3.4 L1 insertions are located within A-T rich regions of the human genome.

A) *Insertions are located in T-rich sequencings in the genome:* Plotting nucleotide frequency of 100bp upstream and downstream of insertions from the corresponding 7bp EN consensus cleavage site reveals the sequence directly surrounding integration sites are primarily T-rich. Green represents Adenine, blue represents Cytosine, orange represents Guanine and red represents Thymine nucleotides.

B) *Engineered L1 insertions are found in AT-rich regions of the genome:* Different lengths (x-axis) of genomic sequence upstream and downstream of L1 insertion sites were examined for GC content (y-axis). Boxplots represent the range of distributions of GC content observed for each respective window size. The box represents the middle 50% of the data, while the line in the middle of the box represents the median observed observation. Whiskers represent the ranges for the bottom 25% and top 25% of the data excluding outliers. The blue dotted line represents the average human GC content of 40.94%.

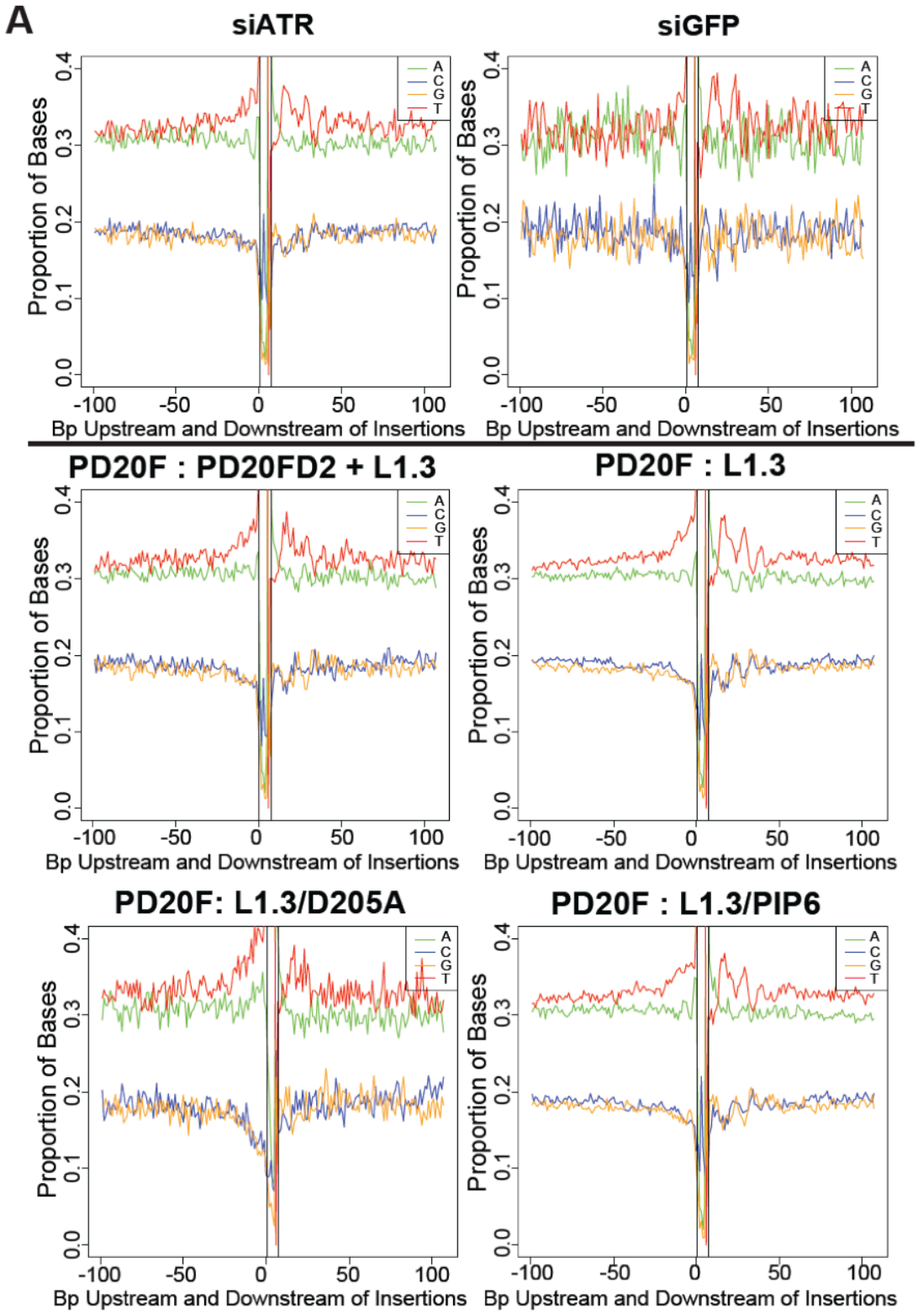


Figure 3.4 L1 insertions are located within A-T rich regions of the human genome.

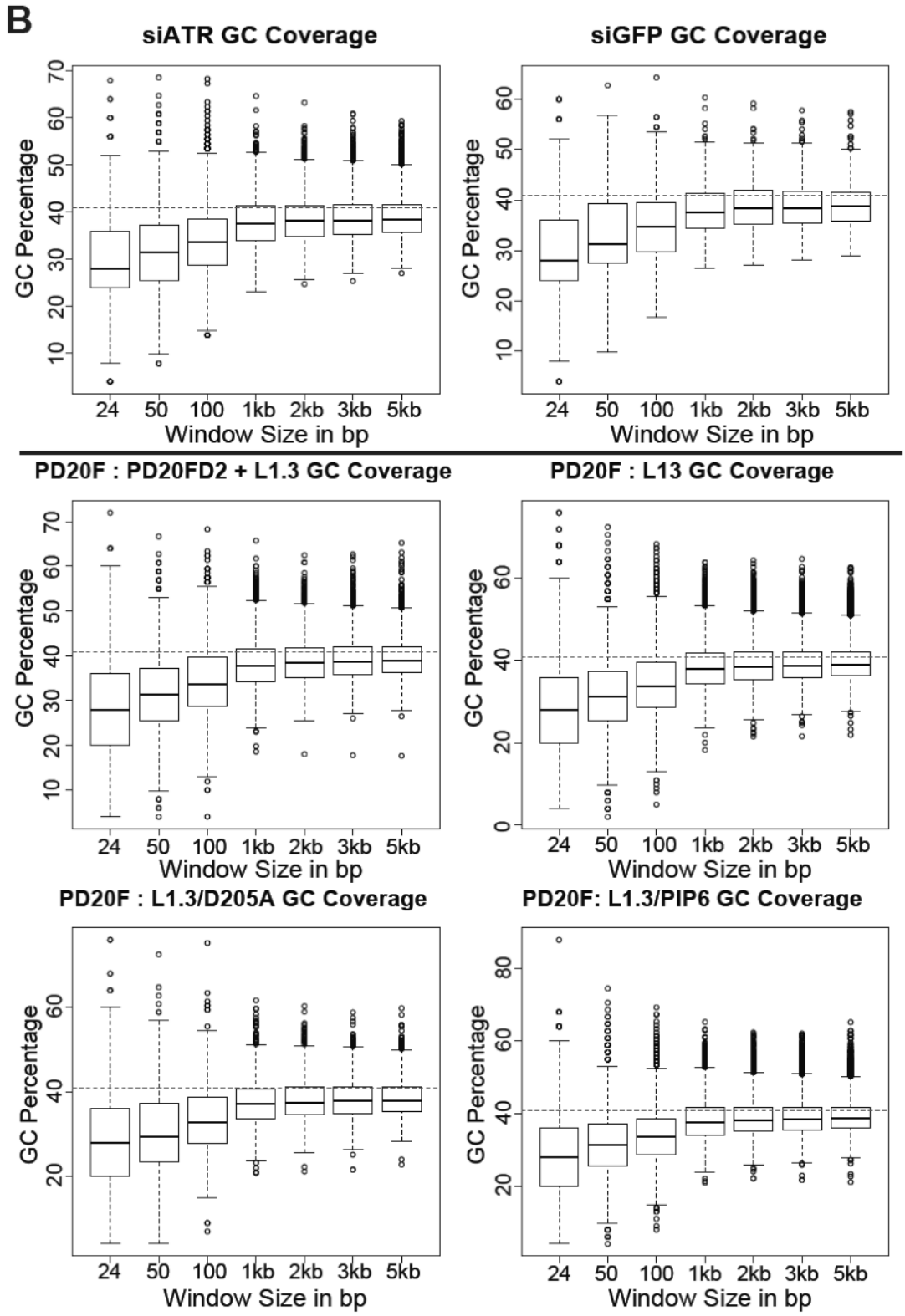


Figure 3.4 L1 insertions are located within A-T rich regions of the human genome.

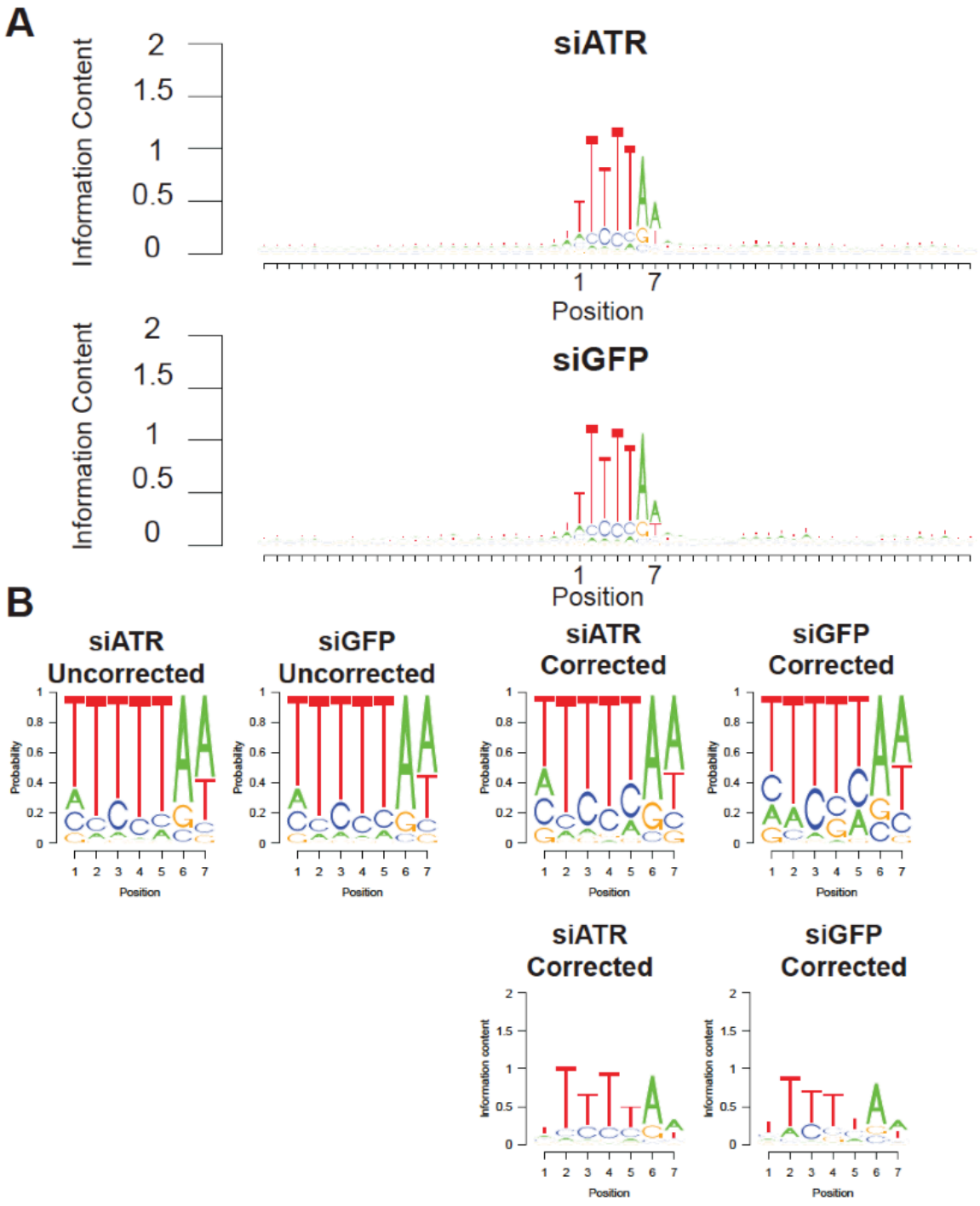


Figure 3.5 Engineered L1 insertions display endonuclease consensus cleavage site

C

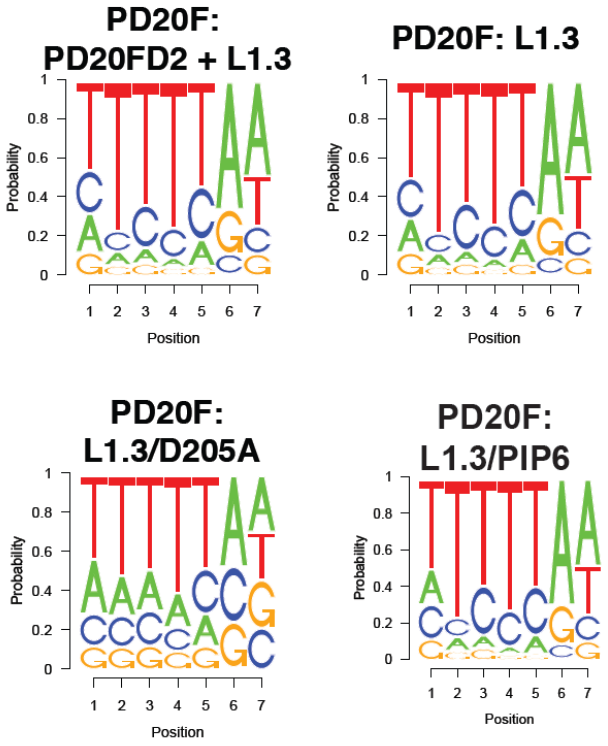


Figure 3.5 Engineered L1 insertions display endonuclease consensus cleavage site

A) *L1 EN site is influenced primarily by a 7mer sequence:* We analyzed twenty-five base pairs upstream and downstream of the 7bp EN consensus cleavage site for insertions from siATR and siGFP retrotransposition assays and created a logo plot. These insertions display the expected 7bp EN degenerate consensus cleavage site.

B) *L1 EN consensus cleavage site observed in ATR knockdown experiments:* Logo plots of the specific 7bp EN consensus sequence was further examined in siATR and siGFP samples. Whether nucleotides were corrected for genomic frequency of EN sites in the genome (labeled 'Corrected') or uncorrected (labeled 'Uncorrected'), the strong 5'-TTTTT/AA EN cleavage is observed.

C) *L1 EN consensus cleavage site observed at FANCD2-deficient integration events:* Logo plots of FANCD2-deficient EN consensus cleavage sites.

Figure 3.6 L1 insertions are located throughout the genome

We compared insertion counts (y-axis) on chromosomes (x-axis) in respective PD20F insertion assay conditions (top plots) and siATR (bottom left plot) or siGFP (bottom right plot) knockdown retrotransposition assays. Boxplots represent the range of observations observed from the 10,000 iterations of the weighted random dataset. Each colored point represents the observed counts of insertions on each respective chromosome for each corresponding dataset.

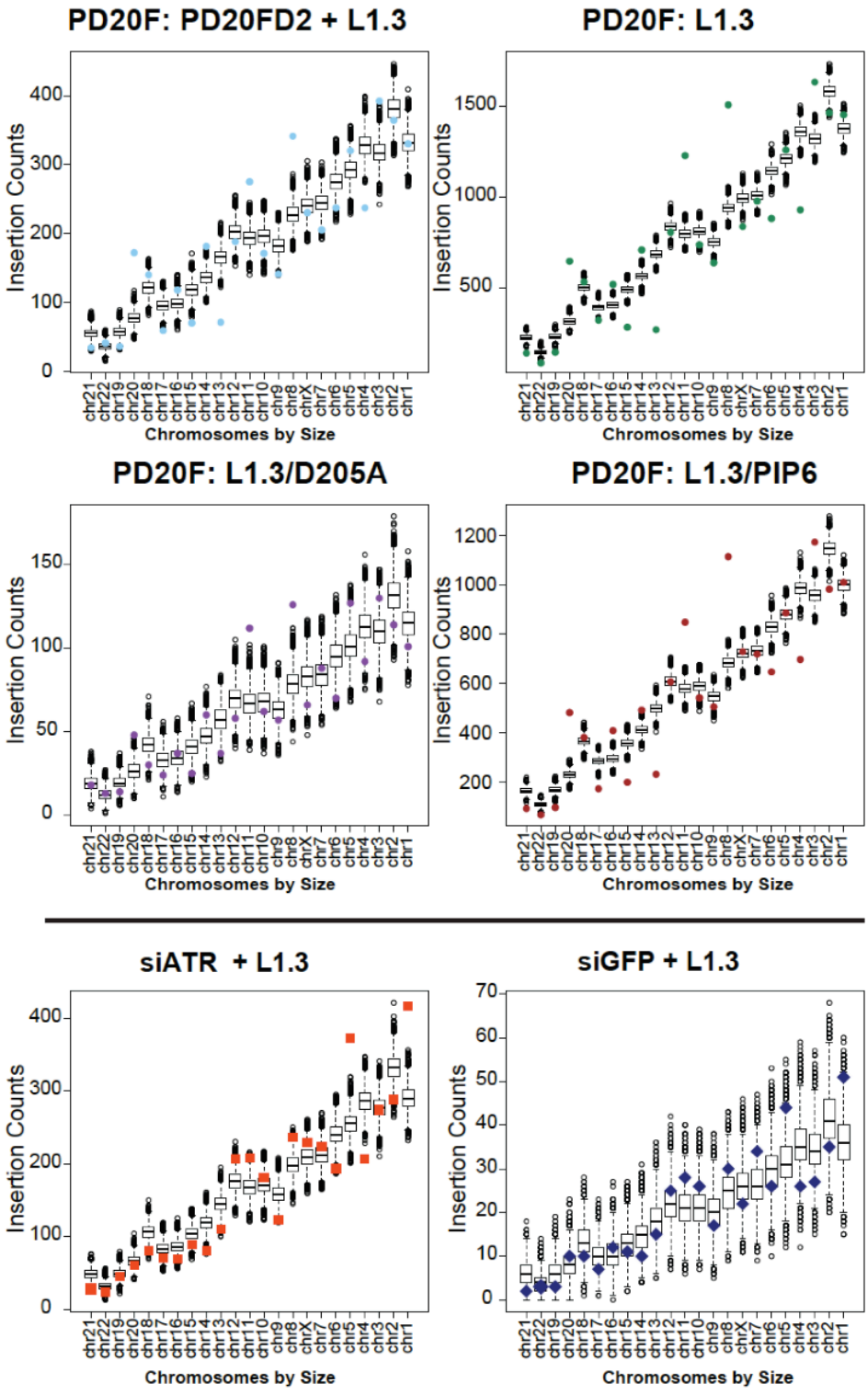


Figure 3.6 L1 insertions are located throughout the genome

Figure 3.7 L1 insertions are interspersed across chromosomes

Depicted are chromosomal ideograms for all the insertions of the given sample listed above. Each horizontal line represents a single insertion site on the corresponding location on the given chromosome. Insertions are dispersed across chromosomes. PD20F is a male cell line so insertions are found on chromosome Y, while siATR and siGFP experiments were performed in the female HeLa cell line.

Figure 3.7 L1 insertions are interspersed across chromosomes

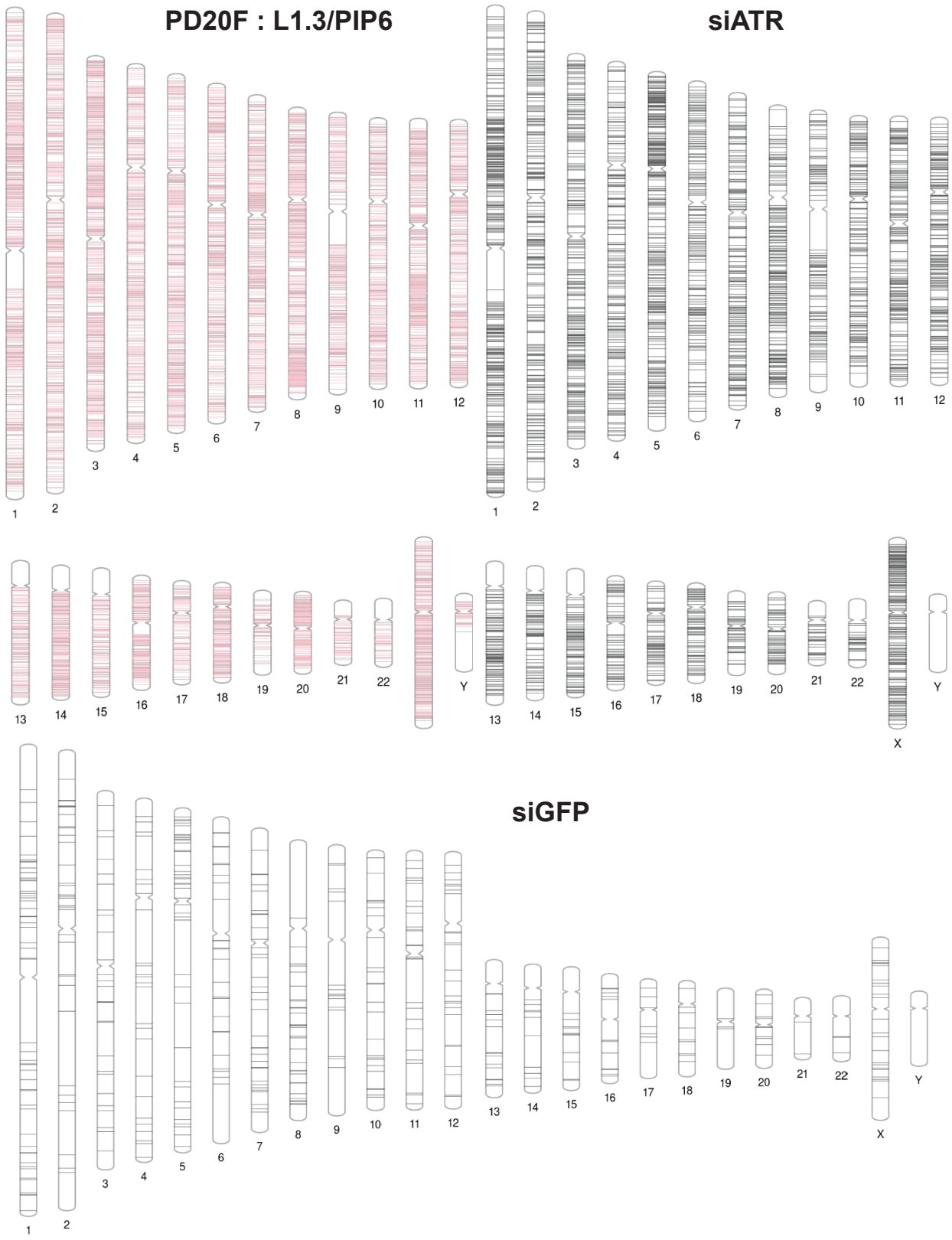


Figure 3.7 L1 insertions are interspersed across chromosomes

Figure 3.8 L1 inserts into genes

A) *L1 insertions are located within exons:* We explored the amount of insertions that are located within the UCSC genome browser's annotation of exons of genes. Both insertion datasets (colored dot) fall within the expected range of exonic insertions as compared to the weighted random dataset (boxplots).

B) *L1 integration can occur at introns:* We plotted the expected weighted random distribution of insertions within UCSC defined introns (boxplots) vs. the observed insertion counts (colored dots). PD20F complemented FANCD2, L1.3 insertions and L1.3/PIP6 insertions show significantly less intronic insertions (proportion test p-values 0.047, $1.792e^{-05}$ respectively).

C) *Antisense genic insertion preference driven by L1 EN consensus cleavage site:* We plotted the ratio of antisense to sense insertions found within exons and introns as defined by the UCSC genome browser annotation. In general, both sets of insertions (colored dots) fall within the expected range as compared to the weighted random dataset (boxplots).

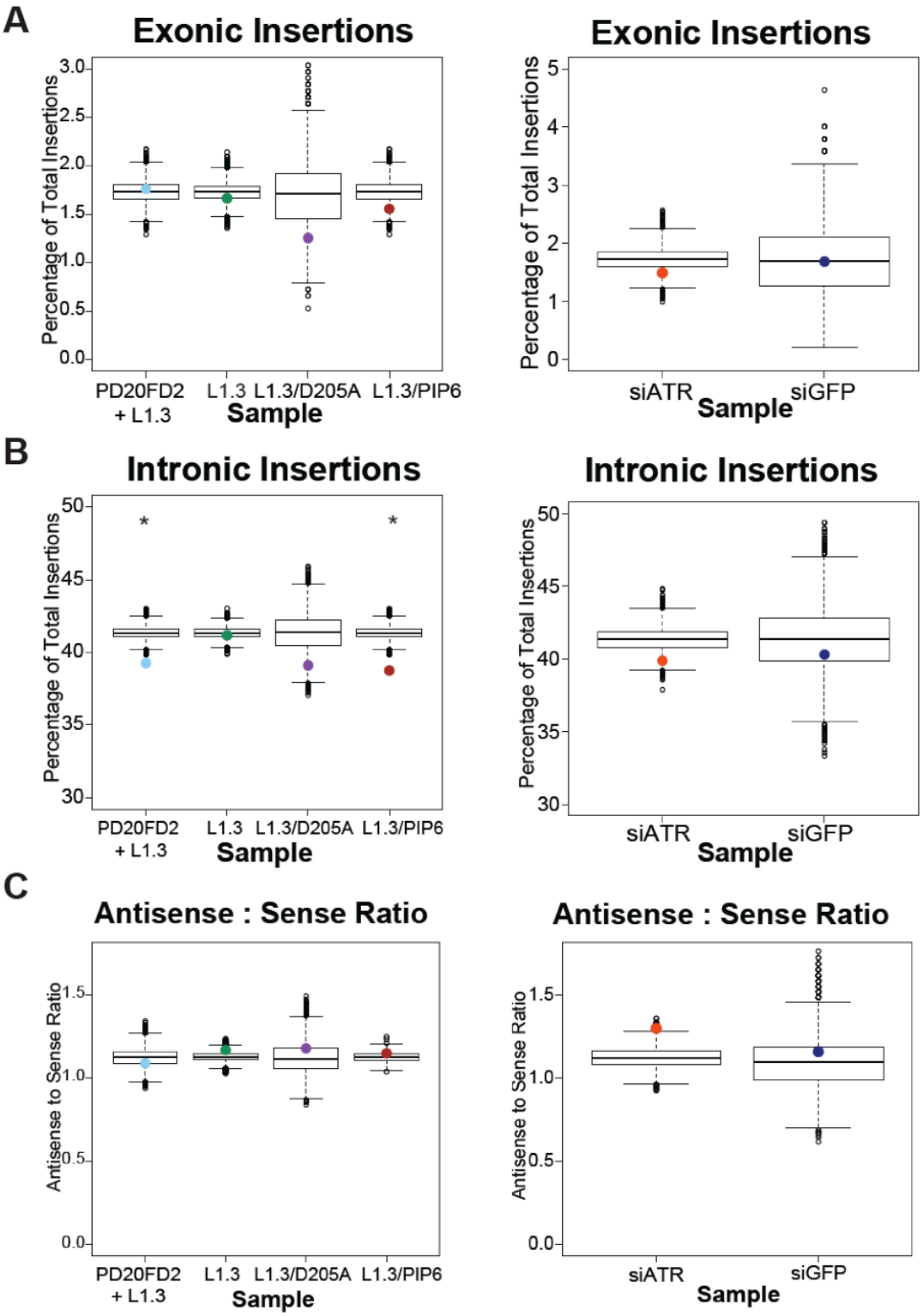


Figure 3.8 L1 inserts into genes

Figure 3.9 Transcribed regions of genome are not preferential L1 integration sites

A) *Transcription is not preferentially targeted by L1 EN:* Utilizing Bru-seq data performed on HeLa cells we defined the genome into transcribed regions (RPKM > 0.025) vs. not transcribed regions. We then plotted the counts of insertions (colored squares) that fell within each category as compared to the weighted random datasets (box plots).

B) *Transcription does not influence L1 integration preference:* Cumulative Distribution Function (CDF) plots of siATR and siGFP insertions with respect to transcription levels (x-axis) as determined either by HeLa segments Bru-seq data (top plots) or HeLa 1kb binned Bru-seq data (bottom plots). Red line represents the corrected weighted random model, blue line indicates the actual observed insertion dataset, and grey lines represent the weighted random model simulations.

C) *Rate of transcription has no influence on L1 integration:* Of transcribed regions in the HeLa genome we divided these regions into 30 distinct bins, from lowest transcription (left most of x-axis) to highest transcription (right most of x-axis). Regardless of transcription rate (x-axis) we observed the expected number of insertions (y-axis) as based on the weighted random model (box plot). Observed insertion counts from datasets are shown by connected colored squares.

D) *L1 EN shows no inherit preference for cleavage of coding or noncoding strand during transcription:* Transcription Bias for each region in the genome is calculated as the absolute value of the rate of transcription on the top strand minus the bottom strand over the cumulative (top and bottom strand) transcription level in the region. We then distinguished transcription bias into 11 distinct bins from 0 to 1, where 0 indicates no strand transcription preference, and 1 indicates that one strand in the genome (top or bottom) is always biased towards being transcribed. For each bin we then counted the proportion of insertions in which EN cleavage occurred on the coding strand. Boxplots represent the distribution observed from the weighted random dataset and colored connected squares represent observed insertions from the corresponding dataset.

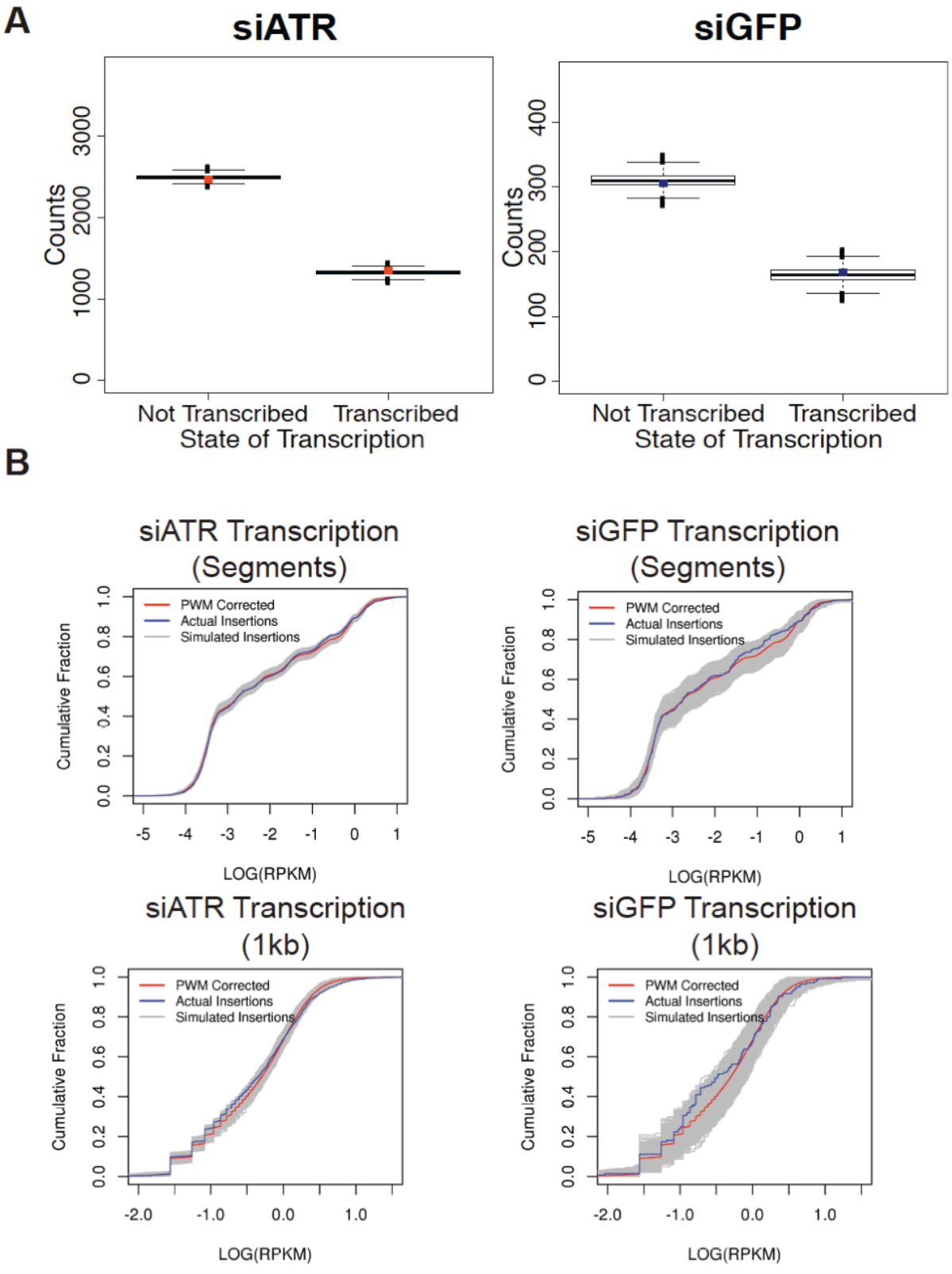


Figure 3.9 Transcribed regions of genome are not preferential L1 integration sites

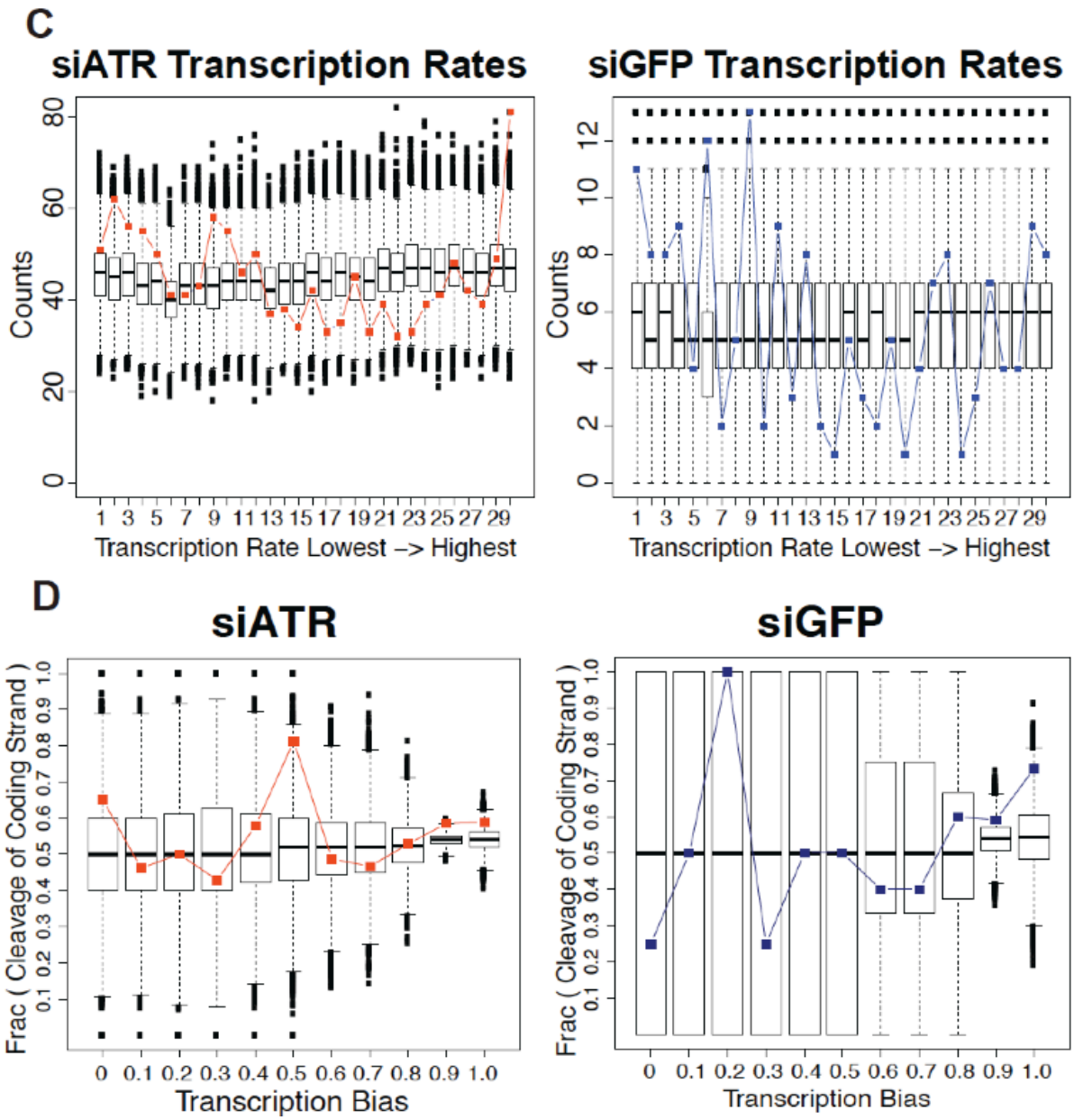


Figure 3.9 Transcribed regions of genome are not preferential L1 integration sites

Figure 3.10 Gene expression determined from RNA-seq does not influence L1 integration

A) *RNA-seq expression does not influence L1 integration preference:* Using RNA-seq data we divided the genome into two distinct bins, those regions of the genome expressed (FPKM > 0.3) and those regions not expressed. Weighted random dataset distributions is shown by the box plots and observed insertion counts are shown by the colored squares. RNA-seq expression from HeLa cells shows that insertions in ATR (left plot) and GFP (right plot) knockdown experiments are within expected ranges of expressed regions of the genome as compared to the weighted random model.

B) *Rate of expression does not influence integration:* Expressed regions of the genome (FPKM > 0.3) were divided into 30 distinct bins from lowest expression rates (left most of x-axis on plot) to the highest expressed rate (right most of x-axis). Box plots represent the distribution observed from the weighted random dataset and colored connected squares represent the observed insertion counts for the respective dataset.

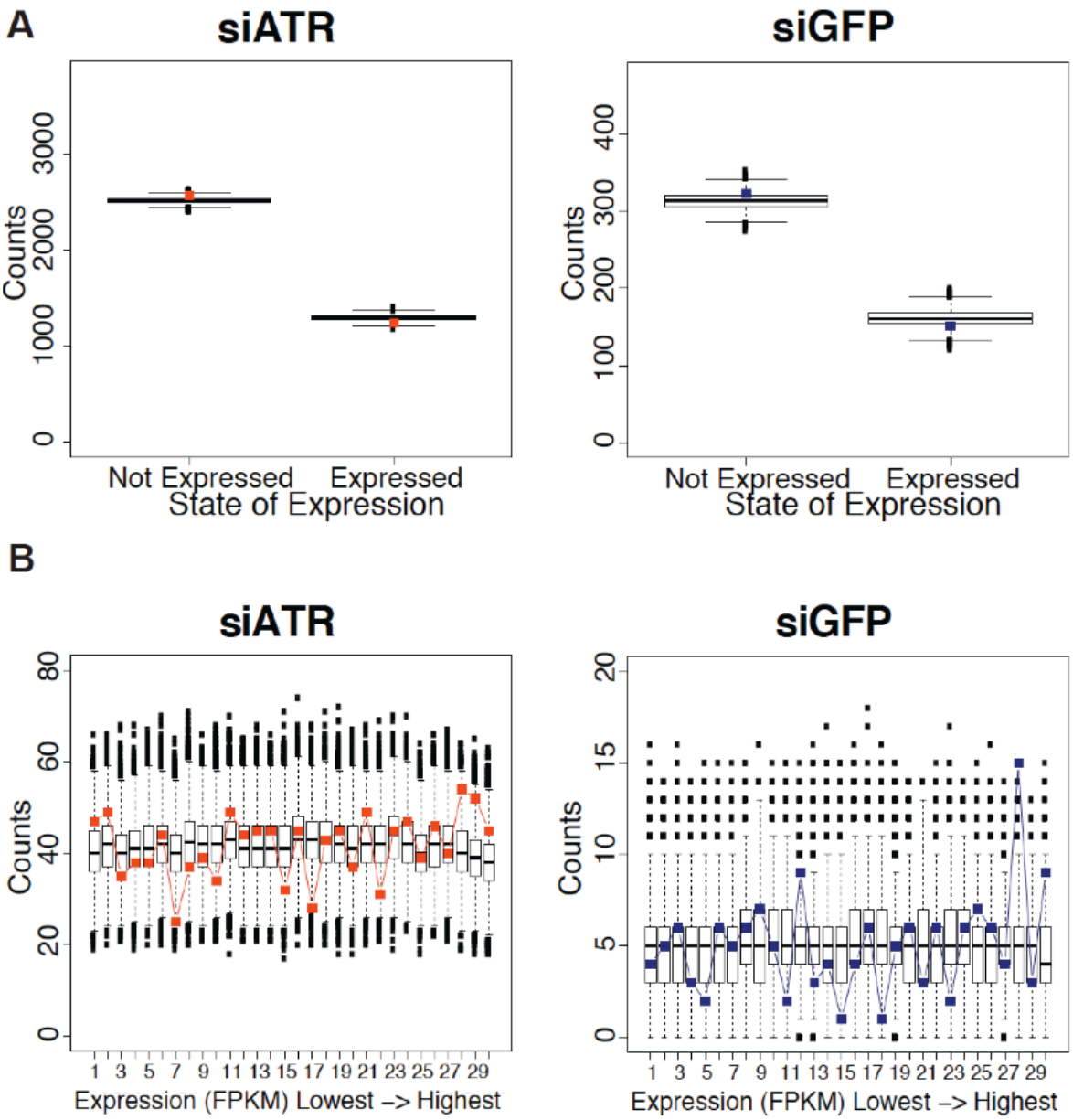


Figure 3.10 Gene expression determined from RNA-seq does not influence L1 integration

Figure 3.11 Replication has no influence on L1 integration

A) *L1 EN shows slight preference for cleavage of lagging strand template during replication:* We used the previously published HeLa and lymphoblastoid OK-seq data from Petryk *et al.* 2016 to determine if L1 EN preferentially cleaves the lagging strand template. HeLa OK-seq data was used for siATR and siGFP L1.3 insertion datasets and L1.3/PIP6 insertions in FANCD2-deficient cells was compared to the lymphoblastoid OK-seq data. Replication Fork Direction was previously described and it's value can be used to determine the leading or lagging strand template during DNA replication. We took the absolute value of this measurement and divided the genome into 11 bins (x-axis). For each bin we then determined the proportion of insertions found in each bin in which the EN cleavage occurred on the lagging strand template during replication (y-axis). Boxplots represent the observed weighted random dataset distributions and colored connected squares represent the actual observed insertion data.

B) *L1 EN shows slight preference for cleavage of lagging strand template during replication (an alternative plot):* We divided insertion datasets based upon whether L1 EN cleaved the top strand (+) or bottom strand (-) in the genome. We then plotted these insertions with respect to replication fork direction (RFD) bias. For insertions in which the EN cleaved the bottom (-) strand for integration into the genome a negative RFD bias value indicates lagging strand template cleavage is preferred whereas a positive RFD bias value indicated leading strand template cleavage preference. Red line represents the corrected weighted random model, blue line indicates the actual observed insertion dataset, and grey lines represent the weighted random model simulations. CDF plots of insertions in which EN cleaved the bottom (top plots) or top strand (bottom plots) during replication. FANCD2-deficient PIP6 mutant insertions significantly differ from the weighted random model when insertions are cleaved on the top strand (Kolmogorov-Smirnov bootstrap test p-value < 0.05) as well as when insertions are cleaved on the bottom strand in the genome (Kolmogorov-Smirnov bootstrap test p-value < 0.01). Both wildtype L1.3 insertions datasets in PD20F, including complemented with PD20FD2, show a trend favoring EN cleavage on the leading template strand. The D205A EN mutant in PD20F shows the opposite trend. Since the L1.3/D205A lacks EN activity this data suggest that the Okazaki fragments are utilized for priming during reverse transcription (Kolmogorov-Smirnov boot strap test P-values: PD20F L1.3 Bottom: < $1x10^{-6}$; PD20F L1.3 Top: < 0.001; PD20F L1.3/D205A Bottom and Top: < 0.01).

C) *L1 EN shows no strong preference for replication initiation or termination sites in the genome:* Replication fork direction slope was plotted with respect to the cumulative fraction of insertions with that value. A negative RFD slope value indicates fork termination while a positive RFD slope value indicates fork initiation. Red line represents the corrected weighted random model, blue line indicates the actual observed insertion dataset, and grey lines represent the weighted random model simulations. FANCD2-deficient L1.3/PIP6 mutant insertions significantly differ from the corrected weighted random model (Kolmogorov-Smirnov bootstrap test p-value < 0.01). FANCD2-deficient L1.3 insertion dataset differs significantly from the corrected weighted random model (Kolmogorov-Smirnov bootstrap test p-value < 0.05).

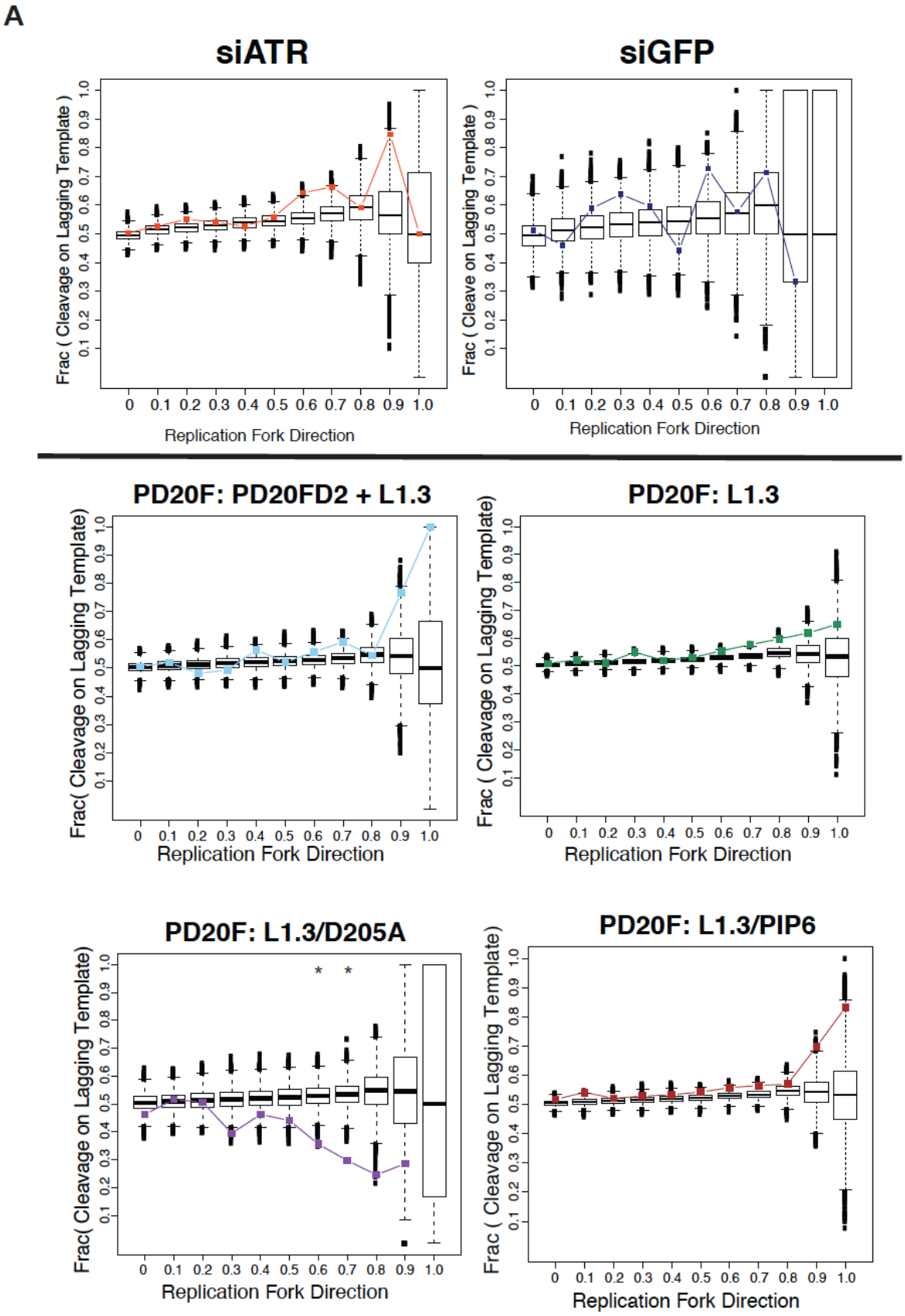


Figure 3.11 Replication has no influence on L1 integration

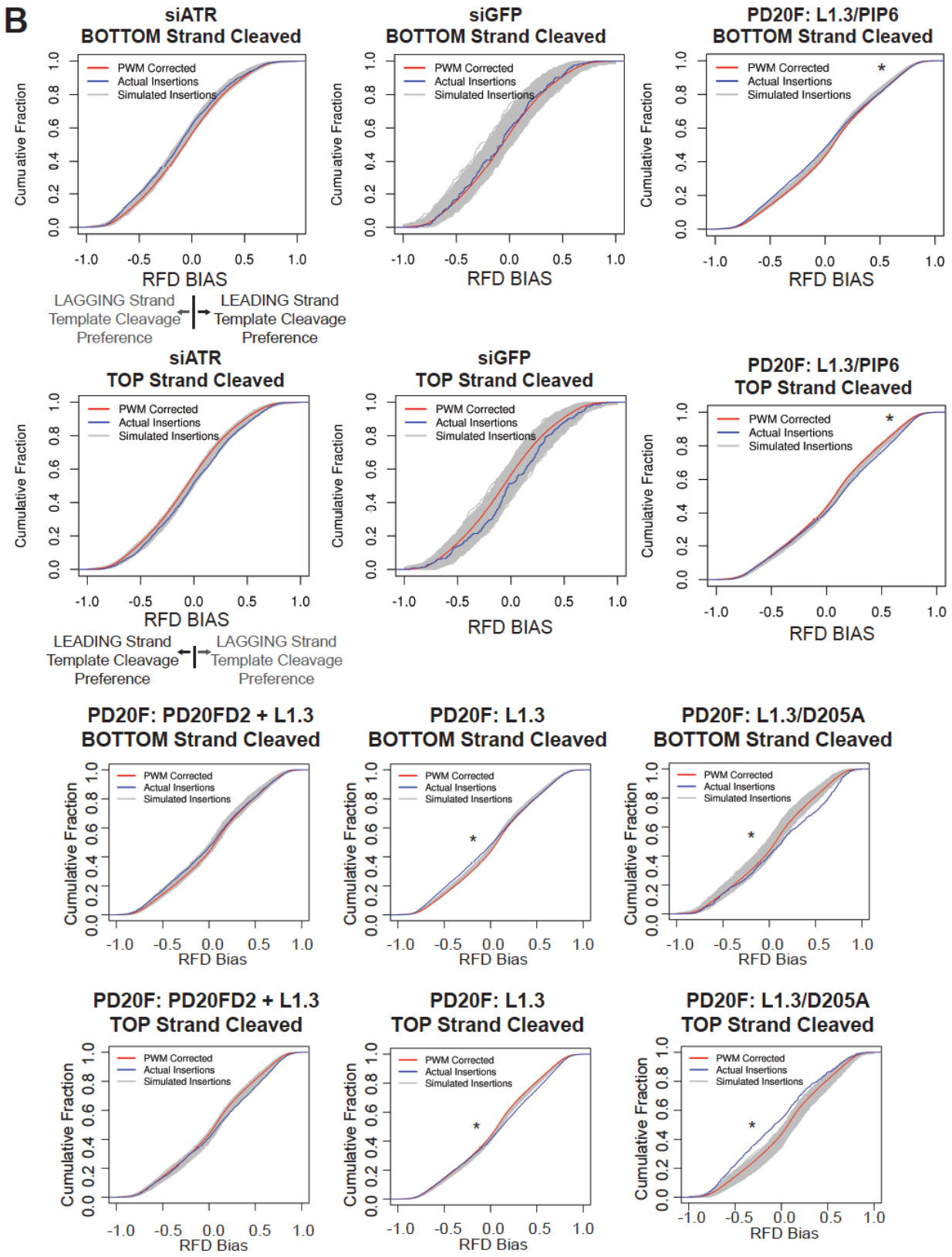


Figure 3.11 Replication has no influence on L1 integration

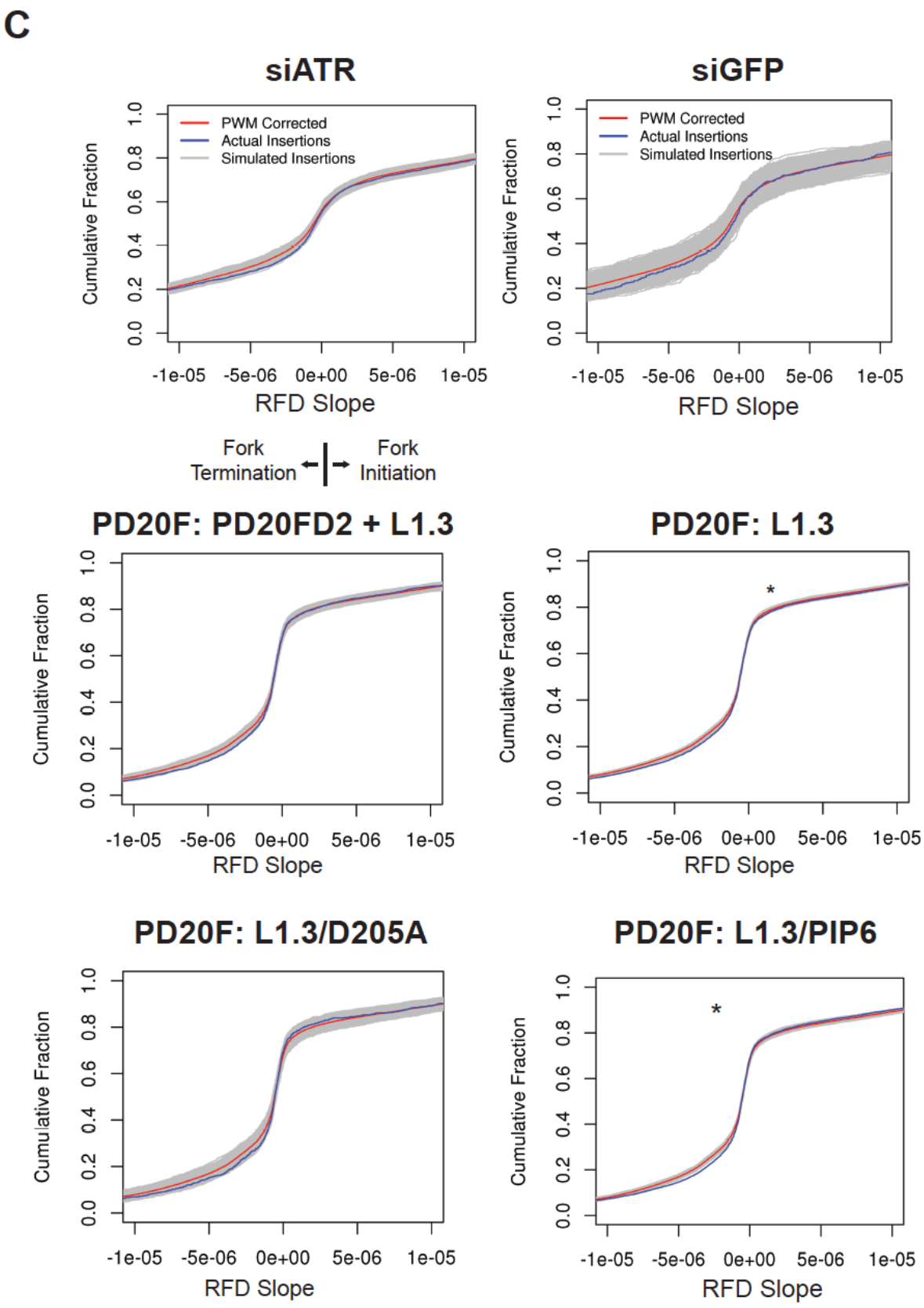
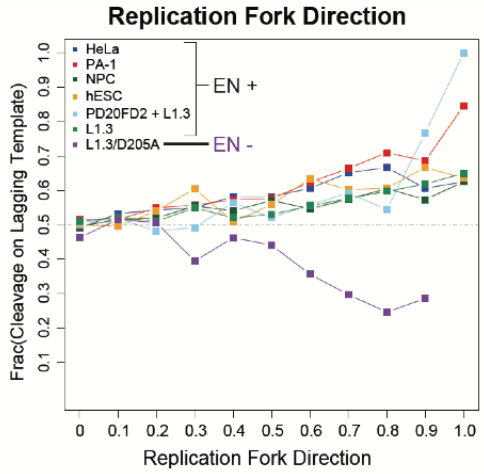


Figure 3.11 Replication has no influence on L1 integration



Sample	Proportion of Total Insertions Cleaved on Lagging Template
HeLa	55.51%
PA-1	57.67%
NPC	54.08%
hESC	55.17%
PD20FD2 + L1.3	52.88%
L1.3	53.62%
L1.3/D205A	42.39%

Figure 3.12 The L1 endonuclease domain influences cleavage and L1 integration preference in replication forks.

In this figure we plotted all the engineered insertion samples (including those from Chapter 2) with respect to the proportion of insertions cleaved on the lagging strand template (y-axis) versus the replication fork direction (x-axis). Each color represents a different sample (see legend on top left of plot). All insertion datasets generated from an engineered L1 with wildtype endonuclease activity show a preference for cleavage on the lagging strand template during replication, as more than 50% of the total insertions show this preference (table to right of plot). The L1.3/DD205A EN- mutant insertions only cleave the lagging strand template for 42.39% of the total insertions. This strong difference in trend between the two L1s, wildtype EN and mutant EN, suggest a difference in integration mechanisms.

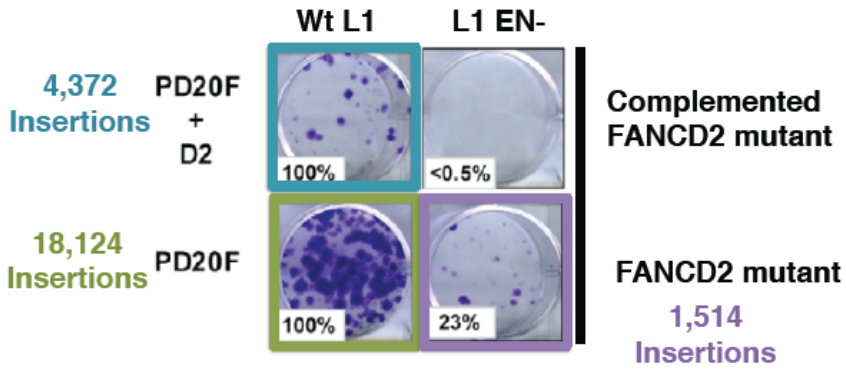


Figure 3.13 FANCD2-deficient cells, PD20F, support wildtype and ENi retrotransposition.

FANCD2 deficient (bottom row) or FANCD2 complemented (top row) were transfected with a wildtype L1 (left column) or L1 EN- mutant (right column) and following completion of the retrotransposition assay colonies were fixed and stained. Endonuclease mutant L1 is capable of 'jumping' in FANCD2 deficient cells to levels 23% of wildtype levels.

	Sample	Weighted Random				Sample	Weighted Random		
Data	siATR	Min	Median	Max	Data	siGFP	Min	Median	Max
Fragile Sites	681 (17.85%)	531	625	720	Fragile Sites	78 (16.46%)	49	77	108
Non-fragile Sites	1470 (38.53%)	1390	1521	1630	Non-fragile Sites	184 (38.82%)	144	189	224

	Sample	Weighted Random				Sample	Weighted Random		
Data	PD20FD2 + L1.3	Min	Median	Max	Data	L1.3	Min	Median	Max
Fragile Sites	654 (14.96%)	623	715	806	Fragile Sites	2906 (16.03%)	2782	2968	3151
Non-fragile Sites	1864 (42.63%)	1623	1745	1845	Non-fragile Sites	7431 (41.00%)	6974	7229	7476

	Sample	Weighted Random				Sample	Weighted Random		
Data	L1.3/ D205A	Min	Median	Max	Data	L1.3/PIP6	Min	Median	Max
Fragile Sites	233 (15.39%)	189	248	318	Fragile Sites	2064 (15.68%)	2004	2156	2313
Non-fragile Sites	605 (39.96%)	528	603	677	Non-fragile Sites	5519 (41.93%)	5031	5248	5474

Table 3.1 Insertions in fragile and non-fragile sites

Column 1 indicates the published dataset and their respective genomic locations of fragile and non-fragile sites. There are two different fragile sites datasets, the second published dataset is from Mrasket *et al.* 2010 as indicated, and the first Fragile Sites dataset as well as last Non-Fragile Sites dataset is from Fungtammasan *et al.* 2012. Column 2 indicates the number of insertions (percentage of total insertions given in parenthesis) for the given sample in the listed published data set. Columns 3-5 give the min, median, and max of expected insertions of the weighted random dataset within the corresponding listed dataset. In general we observe the expected amount of insertions in fragile sites. We do not observe an increase of insertions in fragile sites, nor a decrease in non-fragile sites in the siATR sample expected.

References

Abraham, R.T. (2001). Cell cycle checkpoint signaling through the ATM and ATR kinases. Genes Dev *15*, 2177-2196.

Akkari, Y.M., Bateman, R.L., Reifsteck, C.A., Olson, S.B., and Grompe, M. (2000). DNA replication is required To elicit cellular responses to psoralen-induced DNA interstrand cross-links. Mol Cell Biol *20*, 8283-8289.

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

Andreassen, P.R., D'Andrea, A.D., and Taniguchi, T. (2004). ATR couples FANCD2 monoubiquitination to the DNA-damage response. Genes Dev *18*, 1958-1963.

Awasthi, P., Foiani, M., and Kumar, A. (2015). ATM and ATR signaling at a glance. J Cell Sci *128*, 4255-4262.

Ball, H.L., Myers, J.S., and Cortez, D. (2005). ATRIP binding to replication protein A-single-stranded DNA promotes ATR-ATRIP localization but is dispensable for Chk1 phosphorylation. Mol Biol Cell *16*, 2372-2381.

Berkovich, E., Monnat, R.J., Jr., and Kastan, M.B. (2007). Roles of ATM and NBS1 in chromatin structure modulation and DNA double-strand break repair. Nat Cell Biol *9*, 683-690.

Bogliolo, M., Schuster, B., Stoepker, C., Derkunt, B., Su, Y., Raams, A., Trujillo, J.P., Minguillon, J., Ramirez, M.J., Pujol, R.*, et al.* (2013). Mutations in ERCC4, encoding the DNA-repair endonuclease XPF, cause Fanconi anemia. Am J Hum Genet *92*, 800-806.

Bogliolo, M., and Surralles, J. (2015). Fanconi anemia: a model disease for studies on human genetics and advanced therapeutics. Curr Opin Genet Dev *33*, 32-40.

Brown, E.J., and Baltimore, D. (2000). ATR disruption leads to chromosomal fragmentation and early embryonic lethality. Genes Dev *14*, 397-402.

Brown, E.J., and Baltimore, D. (2003). Essential and dispensable roles of ATR in cell cycle arrest and genome maintenance. Genes Dev *17*, 615-628.

Bubeck, D., Reijns, M.A., Graham, S.C., Astell, K.R., Jones, E.Y., and Jackson, A.P. (2011). PCNA directs type 2 RNase H activity on DNA replication and repair substrates. Nucleic Acids Res *39*, 3652-3666.

Byun, T.S., Pacek, M., Yee, M.C., Walter, J.C., and Cimprich, K.A. (2005). Functional uncoupling of MCM helicase and DNA polymerase activities activates the ATR-dependent checkpoint. Genes Dev *19*, 1040-1052.

Casper, A.M., Nghiem, P., Arlt, M.F., and Glover, T.W. (2002). ATR regulates fragile site stability. Cell *111*, 779-789.

Choe, K.N., and Moldovan, G.L. (2017). Forging Ahead through Darkness: PCNA, Still the Principal Conductor at the Replication Fork. Mol Cell *65*, 380-392.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Marchetto, M.C., Muotri, A.R., Mu, Y., Carson, C.T., Macia, A., Moran, J.V., and Gage, F.H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc Natl Acad Sci U S A *108*, 20382-20387.

D'Andrea, A., and Pellman, D. (1998). Deubiquitinating enzymes: a new class of biological regulators. Crit Rev Biochem Mol Biol *33*, 337-352.

D'Andrea, A.D., and Grompe, M. (2003). The Fanconi anaemia/BRCA pathway. Nat Rev Cancer *3*, 23-34.

De Silva, I.U., McHugh, P.J., Clingen, P.H., and Hartley, J.A. (2000). Defining the roles of nucleotide excision repair and recombination in the repair of DNA interstrand cross-links in mammalian cells. Mol Cell Biol *20*, 7980-7990.

de Winter, J.P., Leveille, F., van Berkel, C.G., Rooimans, M.A., van Der Weel, L., Steltenpool, J., Demuth, I., Morgan, N.V., Alon, N., Bosnoyan-Collins, L.*, et al.* (2000a). Isolation of a cDNA representing the Fanconi anemia complementation group E gene. Am J Hum Genet *67*, 1306-1308.

de Winter, J.P., van der Weel, L., de Groot, J., Stone, S., Waisfisz, Q., Arwert, F., Scheper, R.J., Kruyt, F.A., Hoatlin, M.E., and Joenje, H. (2000b). The Fanconi anemia protein FANCF forms a nuclear complex with FANCA, FANCC and FANCG. Hum Mol Genet *9*, 2665-2674.

de Winter, J.P., Waisfisz, Q., Rooimans, M.A., van Berkel, C.G., Bosnoyan-Collins, L., Alon, N., Carreau, M., Bender, O., Demuth, I., Schindler, D.*, et al.* (1998). The Fanconi anaemia group G gene FANCG is identical with XRCC9. Nat Genet *20*, 281-283.

Deans, A.J., and West, S.C. (2011). DNA interstrand crosslink repair and cancer. Nat Rev Cancer *11*, 467-480.

Dombroski, B.A., Feng, Q., Mathias, S.L., Sassaman, D.M., Scott, A.F., Kazazian, H.H., Jr., and Boeke, J.D. (1994). An in vivo assay for the reverse transcriptase of human retrotransposon L1 in Saccharomyces cerevisiae. Mol Cell Biol *14*, 4485-4492.

Dorsman, J.C., Levitus, M., Rockx, D., Rooimans, M.A., Oostra, A.B., Haitjema, A., Bakker, S.T., Steltenpool, J., Schuler, D., Mohan, S.*, et al.* (2007). Identification of the Fanconi anemia complementation group I gene, FANCI. Cell Oncol *29*, 211-218.

Dunn, J., Potter, M., Rees, A., and Runger, T.M. (2006). Activation of the Fanconi anemia/BRCA pathway and recombination repair in the cellular response to solar ultraviolet light. Cancer Res *66*, 11140-11147.

Ellison, V., and Stillman, B. (2003). Biochemical characterization of DNA damage checkpoint complexes: clamp loader and clamp complexes with specificity for 5' recessed DNA. PLoS Biol *1*, E33.

Falck, J., Coates, J., and Jackson, S.P. (2005). Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. Nature *434*, 605-611.

Foe, J.R., Rooimans, M.A., Bosnoyan-Collins, L., Alon, N., Wijker, M., Parker, L., Lightfoot, J., Carreau, M., Callen, D.F., Savoia, A.*, et al.* (1996). Expression cloning of a cDNA for the major Fanconi anaemia gene, FAA. Nat Genet *14*, 488.

Friedberg, E.C. (2003). DNA damage and repair. Nature *421*, 436-440.

Friedel, A.M., Pike, B.L., and Gasser, S.M. (2009). ATR/Mec1: coordinating fork stability and repair. Curr Opin Cell Biol *21*, 237-244.

Garcia-Higuera, I., Kuang, Y., Naf, D., Wasik, J., and D'Andrea, A.D. (1999). Fanconi anemia proteins FANCA, FANCC, and FANCG/XRCC9 interact in a functional nuclear complex. Mol Cell Biol *19*, 4866-4873.

Garcia-Higuera, I., Taniguchi, T., Ganesan, S., Meyn, M.S., Timmers, C., Hejna, J., Grompe, M., and D'Andrea, A.D. (2001). Interaction of the Fanconi anemia proteins and BRCA1 in a common pathway. Mol Cell *7*, 249-262.

Gasior, S.L., Preston, G., Hedges, D.J., Gilbert, N., Moran, J.V., and Deininger, P.L. (2007). Characterization of pre-insertion loci of de novo L1 insertions. Gene *390*, 190-198.

Hanada, K., Budzowska, M., Modesti, M., Maas, A., Wyman, C., Essers, J., and Kanaar, R. (2006). The structure-specific endonuclease Mus81-Eme1 promotes conversion of interstrand DNA crosslinks into double-strands breaks. EMBO J *25*, 4921-4932.

Ho, G.P., Margossian, S., Taniguchi, T., and D'Andrea, A.D. (2006). Phosphorylation of FANCD2 on two novel sites is required for mitomycin C resistance. Mol Cell Biol *26*, 7005-7015.

Howlett, N.G., Harney, J.A., Rego, M.A., Kolling, F.W.t., and Glover, T.W. (2009). Functional interaction between the Fanconi Anemia D2 protein and proliferating cell nuclear antigen (PCNA) via a conserved putative PCNA interaction motif. J Biol Chem *284*, 28935-28942.

Howlett, N.G., Taniguchi, T., Durkin, S.G., D'Andrea, A.D., and Glover, T.W. (2005). The Fanconi anemia pathway is required for the DNA replication stress response and for the regulation of common fragile site stability. Hum Mol Genet *14*, 693-701.

Howlett, N.G., Taniguchi, T., Olson, S., Cox, B., Waisfisz, Q., De Die-Smulders, C., Persky, N., Grompe, M., Joenje, H., Pals, G.*, et al.* (2002). Biallelic inactivation of BRCA2 in Fanconi anemia. Science *297*, 606-609.

Huang, T.T., Nijman, S.M., Mirchandani, K.D., Galardy, P.J., Cohn, M.A., Haas, W., Gygi, S.P., Ploegh, H.L., Bernards, R., and D'Andrea, A.D. (2006). Regulation of monoubiquitinated PCNA by DUB autocleavage. Nat Cell Biol *8*, 339-347.

Hussain, S., Wilson, J.B., Medhurst, A.L., Hejna, J., Witt, E., Ananth, S., Davies, A., Masson, J.Y., Moses, R., West, S.C.*, et al.* (2004). Direct interaction of FANCD2 with BRCA2 in DNA damage response pathways. Hum Mol Genet *13*, 1241-1248.

Jazayeri, A., Falck, J., Lukas, C., Bartek, J., Smith, G.C., Lukas, J., and Jackson, S.P. (2006). ATM- and cell cycle-dependent regulation of ATR in response to DNA double-strand breaks. Nat Cell Biol *8*, 37-45.

Kaiser, T.N., Lojewski, A., Dougherty, C., Juergens, L., Sahar, E., and Latt, S.A. (1982). Flow cytometric characterization of the response of Fanconi's anemia cells to mitomycin C treatment. Cytometry *2*, 291-297.

Kitagawa, R., Bakkenist, C.J., McKinnon, P.J., and Kastan, M.B. (2004). Phosphorylation of SMC1 is a critical downstream event in the ATM-NBS1-BRCA1 pathway. Genes Dev *18*, 1423-1438.

Knies, K., Inano, S., Ramirez, M.J., Ishiai, M., Surralles, J., Takata, M., and Schindler, D. (2017). Biallelic mutations in the ubiquitin ligase RFWD3 cause Fanconi anemia. J Clin Invest *127*, 3013-3027.

Kopera, H.C., Moldovan, J.B., Morrish, T.A., Garcia-Perez, J.L., and Moran, J.V. (2011). Similarities between long interspersed element-1 (LINE-1) reverse transcriptase and telomerase. Proc Natl Acad Sci U S A *108*, 20345-20350.

Kubbies, M., Schindler, D., Hoehn, H., Schinzel, A., and Rabinovitch, P.S. (1985). Endogenous blockage and delay of the chromosome cycle despite normal recruitment and growth phase explain poor proliferation and frequent edomitosis in Fanconi anemia cells. Am J Hum Genet *37*, 1022-1030.

Kulpa D.A., and Moran J.V. (2006). Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol *13*, 655-60.

Kumagai, A., Kim, S.M., and Dunphy, W.G. (2004). Claspin and the activated form of ATR-ATRIP collaborate in the activation of Chk1. J Biol Chem *279*, 49599-49608.

Kumagai, A., Lee, J., Yoo, H.Y., and Dunphy, W.G. (2006). TopBP1 activates the ATR-ATRIP complex. Cell *124*, 943-955.

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

Lee, J.H., and Paull, T.T. (2005). ATM activation by DNA double-strand breaks through the Mre11-Rad50-Nbs1 complex. Science *308*, 551-554.

Lee, S.E., Moore, J.K., Holmes, A., Umezu, K., Kolodner, R.D., and Haber, J.E. (1998). Saccharomyces Ku70, mre11/rad50 and RPA proteins regulate adaptation to G2/M arrest after DNA damage. Cell *94*, 399-409.

Levitus, M., Waisfisz, Q., Godthelp, B.C., de Vries, Y., Hussain, S., Wiegant, W.W., Elghalbzouri-Maghrani, E., Steltenpool, J., Rooimans, M.A., Pals, G.*, et al.* (2005). The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. Nat Genet *37*, 934-935.

Levran, O., Attwooll, C., Henry, R.T., Milton, K.L., Neveling, K., Rio, P., Batish, S.D., Kalb, R., Velleuer, E., Barral, S.*, et al.* (2005). The BRCA1-interacting helicase BRIP1 is deficient in Fanconi anemia. Nat Genet *37*, 931-933.

Litman, R., Peng, M., Jin, Z., Zhang, F., Zhang, J., Powell, S., Andreassen, P.R., and Cantor, S.B. (2005). BACH1 is critical for homologous recombination and appears to be the Fanconi anemia gene product FANCJ. Cancer Cell *8*, 255-265.

Lopes, M., Foiani, M., and Sogo, J.M. (2006). Multiple mechanisms control chromosome integrity after replication fork uncoupling and restart at irreparable UV lesions. Mol Cell *21*, 15-27.

Majka, J., Binz, S.K., Wold, M.S., and Burgers, P.M. (2006). Replication protein A directs loading of the DNA damage checkpoint clamp to 5'-DNA junctions. J Biol Chem *281*, 27855-27861.

Medhurst, A.L., Huber, P.A., Waisfisz, Q., de Winter, J.P., and Mathew, C.G. (2001). Direct interactions of the five known Fanconi anaemia proteins suggest a common functional pathway. Hum Mol Genet *10*, 423-429.

Meetei, A.R., Levitus, M., Xue, Y., Medhurst, A.L., Zwaan, M., Ling, C., Rooimans, M.A., Bier, P., Hoatlin, M., Pals, G.*, et al.* (2004). X-linked inheritance of Fanconi anemia complementation group B. Nat Genet *36*, 1219-1224.

Meetei, A.R., Medhurst, A.L., Ling, C., Xue, Y., Singh, T.R., Bier, P., Steltenpool, J., Stone, S., Dokal, I., Mathew, C.G.*, et al.* (2005). A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. Nat Genet *37*, 958-963.

Meetei, A.R., Sechi, S., Wallisch, M., Yang, D., Young, M.K., Joenje, H., Hoatlin, M.E., and Wang, W. (2003). A multiprotein nuclear complex connects Fanconi anemia and Bloom syndrome. Mol Cell Biol *23*, 3417-3426.

Melander, A., Olsson, J., Lindberg, G., Salzman, A., Howard, T., Stang, P., Lydick, E., Emslie-Smith, A., Boyle, D.I., Evans, J.M.*, et al.* (1999). 35th Annual Meeting of the European Association for the Study of Diabetes : Brussels, Belgium, 28 September-2 October 1999. Diabetologia *42*, A1-A330.

Moldovan, G.L., and D'Andrea, A.D. (2009). How the fanconi anemia pathway guards the genome. Annu Rev Genet *43*, 223-249.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. (2007). Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature *446*, 208-212.

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

Mosedale, G., Niedzwiedz, W., Alpi, A., Perrina, F., Pereira-Leal, J.B., Johnson, M., Langevin, F., Pace, P., and Patel, K.J. (2005). The vertebrate Hef ortholog is a component of the Fanconi anemia tumor-suppressor pathway. Nat Struct Mol Biol *12*, 763-771.

Myers, J.S., and Cortez, D. (2006). Rapid activation of ATR by ionizing radiation requires ATM and Mre11. J Biol Chem *281*, 9346-9350.

Namiki, Y., and Zou, L. (2006). ATRIP associates with replication protein A-coated ssDNA through multiple interactions. Proc Natl Acad Sci U S A *103*, 580-585.

Niedernhofer, L.J., Odijk, H., Budzowska, M., van Drunen, E., Maas, A., Theil, A.F., de Wit, J., Jaspers, N.G., Beverloo, H.B., Hoeijmakers, J.H.*, et al.* (2004). The structure-specific endonuclease Ercc1-Xpf is required to resolve DNA interstrand cross-link-induced double-strand breaks. Mol Cell Biol *24*, 5776-5787.

Niedzwiedz, W., Mosedale, G., Johnson, M., Ong, C.Y., Pace, P., and Patel, K.J. (2004). The Fanconi anaemia gene FANCC promotes homologous recombination and error-prone DNA repair. Mol Cell *15*, 607-620.

O'Donnell, M., Langston, L., and Stillman, B. (2013). Principles and concepts of DNA replication in bacteria, archaea, and eukarya. Cold Spring Harb Perspect Biol *5*.

Paulsen, M.T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E.A., Magnuson, B., Wilson, T.E., and Ljungman, M. (2014). Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA. Methods *67*, 45-54.

Paulsen, M.T., Veloso, A., Prasad, J., Bedi, K., Ljungman, E.A., Tsan, Y.C., Chang, C.W., Tarrier, B., Washburn, J.G., Lyons, R.*, et al.* (2013). Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response. Proc Natl Acad Sci U S A *110*, 2240-2245.

Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.L., and Hyrien, O. (2016). Replication landscape of the human genome. Nat Commun *7*, 10208.

Pichierri, P., and Rosselli, F. (2004). The DNA crosslink-induced S-phase checkpoint depends on ATR-CHK1 and ATR-NBS1-FANCD2 pathways. EMBO J *23*, 1178-1187.

Pickering, A., Zhang, J., Panneerselvam, J., and Fei, P. (2013). Advances in the understanding of the Fanconi anemia tumor suppressor pathway. Cancer Biol Ther *14*, 1089-1091.

Pulsipher, M., Kupfer, G.M., Naf, D., Suliman, A., Lee, J.S., Jakobs, P., Grompe, M., Joenje, H., Sieff, C., Guinan, E.*, et al.* (1998). Subtyping analysis of Fanconi anemia by immunoblotting and retroviral gene transfer. Mol Med *4*, 468-479.

Reid, S., Schindler, D., Hanenberg, H., Barker, K., Hanks, S., Kalb, R., Neveling, K., Kelly, P., Seal, S., Freund, M.*, et al.* (2007). Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. Nat Genet *39*, 162-164.

Richardson, S.R., Narvaiza, I., Planegger, R.A., Weitzman, M.D., and Moran, J.V. (2014). APOBEC3A deaminates transiently exposed single-strand DNA during LINE-1 retrotransposition. Elife *3*, e02008.

Sasaki, M.S. (1975). Is Fanconi's anaemia defective in a process essential to the repair of DNA cross links? Nature *257*, 501-503.

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

Scharer, O.D. (2005). DNA interstrand crosslinks: natural and drug-induced DNA adducts that induce unique cellular responses. Chembiochem *6*, 27-32.

Servant, G., Streva, V.A., Derbes, R.S., Wijetunge, M.I., Neeland, M., White, T.B., Belancio, V.P., Roy-Engel, A.M., and Deininger, P.L. (2017). The Nucleotide Excision Repair Pathway Limits L1 Retrotransposition. Genetics *205*, 139-153.

Shiotani, B., and Zou, L. (2009). ATR signaling at a glance. J Cell Sci *122*, 301-304.
Siddiqui, K., On, K.F., and Diffley, J.F. (2013). Regulating DNA replication in eukarya. Cold Spring Harb Perspect Biol *5*.

Sims, A.E., Spiteri, E., Sims, R.J., 3rd, Arita, A.G., Lach, F.P., Landers, T., Wurm, M., Freund, M., Neveling, K., Hanenberg, H.*, et al.* (2007). FANCI is a second monoubiquitinated member of the Fanconi anemia pathway. Nat Struct Mol Biol *14*, 564-567.

Smogorzewska, A., Matsuoka, S., Vinciguerra, P., McDonald, E.R., 3rd, Hurov, K.E., Luo, J., Ballif, B.A., Gygi, S.P., Hofmann, K., D'Andrea, A.D.*, et al.* (2007). Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. Cell *129*, 289-301.

Strathdee, C.A., Gavish, H., Shannon, W.R., and Buchwald, M. (1992). Cloning of cDNAs for Fanconi's anaemia by functional complementation. Nature *358*, 434.

Taniguchi, T., Garcia-Higuera, I., Andreassen, P.R., Gregory, R.C., Grompe, M., and D'Andrea, A.D. (2002). S-phase-specific interaction of the Fanconi anemia protein, FANCD2, with BRCA1 and RAD51. Blood *100*, 2414-2420.

Taylor, M.S., LaCava, J., Mita, P., Molloy, K.R., Huang, C.R., Li, D., Adney, E.M., Jiang, H., Burns, K.H., Chait, B.T.*, et al.* (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell *155*, 1034-1048.

Timmers, C., Taniguchi, T., Hejna, J., Reifsteck, C., Lucas, L., Bruun, D., Thayer, M., Cox, B., Olson, S., D'Andrea, A.D.*, et al.* (2001). Positional cloning of a novel Fanconi anemia gene, FANCD2. Mol Cell *7*, 241-248.

Walter, J., and Newport, J. (2000). Initiation of eukaryotic DNA replication: origin unwinding and sequential chromatin association of Cdc45, RPA, and DNA polymerase alpha. Mol Cell *5*, 617-627.

Whitney, M., Thayer, M., Reifsteck, C., Olson, S., Smith, L., Jakobs, P.M., Leach, R., Naylor, S., Joenje, H., and Grompe, M. (1995). Microcell mediated chromosome transfer maps the Fanconi anaemia group D gene to chromosome 3p. Nat Genet *11*, 341-343.

Xia, B., Dorsman, J.C., Ameziane, N., de Vries, Y., Rooimans, M.A., Sheng, Q., Pals, G., Errami, A., Gluckman, E., Llera, J.*, et al.* (2007). Fanconi anemia is associated with a defect in the BRCA2 partner PALB2. Nat Genet *39*, 159-161.

You, Z., Chahwan, C., Bailis, J., Hunter, T., and Russell, P. (2005). ATM activation and its recruitment to damaged DNA require binding to the C terminus of Nbs1. Mol Cell Biol *25*, 5363-5379.

Zou, L. (2007). Single- and double-stranded DNA: building a trigger of ATR-mediated DNA damage response. Genes Dev *21*, 879-885.

Zou, L., and Elledge, S.J. (2003). Sensing DNA damage through ATRIP recognition of RPA-ssDNA complexes. Science *300*, 1542-1548.

Chapter 4

Conclusions

Overview

My dissertation research has focused on identifying LINE-1 insertion preferences in the human genome. In chapter two, I examined tens of thousands of engineered LINE-1 insertions from four different cell lines that are proxies for *in vivo* cell types known to accommodate endogenous *de novo* L1 retrotransposition events. I discovered that the L1 EN is the principal factor dictating retrotransposition events into degenerate L1 EN consensus cleavage sites in AT-rich regions of the genome. Notably, the L1 EN consensus sequence is enriched on the sense strand of genes. I further demonstrate that gene expression, DNA replication status, and epigenetic features only exert minimal effects on L1 integration preferences. Thus, L1 EN is the principal determinant driving L1 integration throughout the human genome.

In Chapter 3, I explored the effects of various DNA repair processes on L1 integration preference. Specifically, I examined whether two DNA repair proteins, ATR and FANCD2, influence L1 integration. My data, in collaboration with Dr. Huira Kopera, strongly suggest that ATR senses TPRT intermediates; knocking down ATR expression does not significantly affect L1 integration preferences. By comparison, FANCD2-deficient cells accommodate high levels of ENi L1 retrotransposition. The integration preference of wildtype L1s is not significantly affected in FANCD2-deficient cells; however, the available data suggest that L1 endonuclease-deficient mutants can target replication intermediates, perhaps using the 3'OH group present at Okazaki fragments or resultant double-strand DNA breaks present at collapsed DNA replication forks to initiate ENi retrotransposition. Interestingly, it is proposed that other endonuclease-

deficient non-LTR retrotransposons (*e.g.*, some mobile group II introns), can retrotranspose by a similar mechanism.

In Chapter 3, I further explored whether mutations that disrupt the L1 PCNA-interacting protein (PIP) domain hinder the ability of ORF2p to interact with PCNA, influence L1 integration. Interestingly, PIP mutants display similar integration preferences as wildtype L1s.

Finally, in collaboration with Dr. Thomas Widmann, Mr. Alejandro Roldán, Mrs. Sarah Emery, and Dr. Weichen Zhou I have demonstrated that the capture technique used to characterize engineered L1 insertions in Chapters 2 and 3 can be easily modified to capture *de* novo engineered zebrafish L2 integration sites (Appendix A), as well as endogenous L1Hs integration sites in the human genome (Appendix B). Below, I discuss the significance of the data presented in this dissertation and suggest possible future directions for subsequent research.

LINE-1 Endonuclease Drives LINE-1 Integration Preference

Identification of Authentic Engineered L1 Integration Events

In Chapter 2, I generated a large data set of 65,079 engineered L1 retrotransposition events in various human tissue culture cell lines. Our first concern was to verify that our capture based technique and bioinformatics pipeline led to the identification of *bona fide* L1 integration events. We verified that our insertions end in a poly(A) tail (Figures 2.1G and 2.7F), integrate into a previously identified degenerate L1 EN consensus cleavage site (Figures 2.1I and 2.7G), and occurred within sequences in the genome with a high local AT-content (Figures 2.1H and 2.7F). Since our insertions displayed these known L1 structural hallmarks, we were confident that our dataset contained authentic *de novo* L1 retrotransposition events (Gilbert et al., 2002; Lander et al., 2001; Morrish et al., 2002; Symer et al., 2002; Szak et al., 2002). Moreover, the presence of poly(A) tails reassured us that PacBio CCS sequencing technology allows us to sequence through poly(A) rich sequences, which often present a technical challenge for experiments using Illumina-based sequencing technology. Finally, many of our insertions are supported by more than one independent CCS read, which ranged

from 4.7% of HeLa cells insertions to ~60% of hESC insertions (Figures 2.1H and 2.7E). As expected, the number of duplicate reads reflects the number of independent retrotransposition events examined in our experiments. Thus, I generated a robust dataset of authentic, *de novo* L1 engineered integration events.

A Closer Examination of the LINE-1 Endonuclease

We hypothesized the L1 EN must be a determinant in L1 integration, as previous studies indicated that the vast majority of engineered L1 insertions occur at an L1 EN degenerate consensus cleavage site (5'-TTTT/A-3') (Gilbert et al., 2002; Jurka et al., 1997; Morrish et al., 2002; Symer et al., 2002; Szak et al., 2002). While the L1 EN cleavage site is typically described as a 5bp sequence, we found that that the L1 EN consensus cleavage is actually contributed by 7bp located at the integration site (5'-TTTTT/AA-3') (Figures 2.1I and 2.7G). We discovered that each nucleotide within the 7mer sequence is not independent (Figure 2.2B). The first position of the EN cleavage site is independent from the next four base pairs which is considered as one unit, which are then independent from the last two base pairs considered as another unit of the cleavage site (This creates a 1-4-2 model). We conclude that the L1 EN does not cut DNA in a random manner, further confirming that a simple random model, which treats all possible sites in the genome equally, cannot mimic L1 integration sites.

In depth examination of the L1 EN consensus cleavage site revealed that a large majority (44.92%) of our L1 insertions often contain a cytosine residue instead of a thymidine residue in the T-rich sequence (*e.g.*, 5'-TTCTT/AA-3') (Figure 2.2C). It remains possible that L1 EN, which is classified as, and displays sequence similarities to, apurinic/apyrimidinic-like endonucleases (APEs), recognizes cytidine residues within the T-rich sequence to cleave at the target site. Indeed, previous work suggests that the human L1 EN crystal structure is most closely related to human apurinic/apyrimidinic endonuclease 1 (APE1), which cleaves DNA at apurinic and apyrimidinic sites of DNA damage (Mol et al., 2000; Weichenrieder et al., 2004). The dispersed cytosines found throughout the L1 EN cleavage site may be recognized by the L1 EN enzymatically and initiate the cleavage reaction, as in similar respects to other APEs.

I also studied the possible influence of the spatial configuration of nucleotides at the L1 EN cleavage. Specifically, as a consequence of local sequence-dependent unwinding of the helix and structural flexibility, the 5'-T_nA_n-3' junction normally results in a wider minor groove structure (Cost and Boeke, 1998), where base stacking is minimal (Mack et al., 2001; Stefl et al., 2004). It is thought that the TpA junction in the minor grove of DNA is sensed, contacted, and widened by the insertion of a hairpin loop protruding from the L1-EN protein surface. Thus, target recognition may involve the accommodation of the 3' adenine residue in an extra-helical conformation in a pocket of L1 EN (Weichenrieder et al., 2004). Another study also suggested that the L1 EN recognizes structural features present at the DNA target sequence, such as the structural flexibility of the DNA at the 5'-poly(T)/A-3' junction, rather than specific nucleotides in the target sequence (Repanas et al., 2007). We hoped our model would be able to capture this aspect of L1 EN cleavage specificity.

We created a weighted random model that accounted for the fact that L1 EN drives L1 integration into degenerate L1 EN consensus cleavage sequences. Since the L1 EN cleavage site is degenerate (*e.g.*, 5'-TTTT/A-3' and variants of that sequence, such as 5'-TTTA/A-3', 5'-TTCT/A-3', *etc.*) we used our empirical data to construct a model that captures this variability. As such, we used all possible 7mer sequence variants (12,288) in our modeling; 97% of our total insertions are found within 6% of these variants (Figure 2.2D). For L1 EN cleavage sites observed by three or more insertions, we used our observed frequencies of the corresponding EN cleavage site. For L1 EN cleavage sites observed by fewer than three insertions, we used our 1-4-2 model to generate a position probability matrix and calculate frequencies for the integration sites (Figure 2.2E). These frequencies were then weighted by the most commonly observed EN consensus cleavage site which happened to be 5'-TTTTT/AA-3'. Our weighted random model picked as many positions in the genome as total insertions in our dataset. To capture the variability of L1 EN integration sites, we repeated the L1 EN weighted integration picking process 10,000 times.

To confirm the validity of our model, we asked if the simulated L1 EN sites display the same variability in the L1 EN consensus cleavage site observed in our

actual data (Figure 2.2F). The sequences surrounding the simulated sites are located in local AT-rich sequences of the genome (Figure 2.12D), just as we observed with our actual data set (Figures 2.1H and 2.11), which is consistent with the idea that L1 EN drives L1 integration sites into AT-rich regions of the genome. Notably, as opposed to previous studies, our model accounts for known L1 insertion preferences. Previous studies only analyzed a limited number of retrotransposition events and tend to use old L1 and/or young L1Hs sequences within the genome as benchmarks for comparisons to their empirical data (Baillie et al., 2011; Ovchinnikov et al., 2001). Making comparisons of this nature does not account for selective forces that have influence the distribution of L1s over evolutionary time, thereby skewing the genomic L1 distribution.

LINE-1 Integrates Throughout the Genome

Now that we had generated a model that assumes that L1 EN is the principal factor in directing L1 integration at a degenerate consensus cleavage site, we were ready to examine whether other genomic features might influence L1 integration. In essence our model controls for the known L1 EN univariate, and we now wanted to explore other variables, such as specific genomic features, which may also influence L1 integration, resulting in a possible multivariable model for L1 integration. As observed previously, we saw a significant increase of engineered L1 insertions on the X chromosome in PA-1, NPCs, hESC, but not HeLa cells (Figures 2.3A, 2.3B, 2.14A, and 2.15). There are longstanding debates in the field as to whether L1s prefer to insert on the X-chromosome, or whether the X-chromosome simply accumulates greater numbers of L1 insertions over evolutionary time, since unique portions of the X-chromosome cannot undergo meiotic recombination in males (Langley et al., 1988; Wichman et al., 1992). Indeed, some have suggested that the increase of L1 insertions on the X chromosome plays a functional role in X-inactivation (Bailey et al., 2000; Gartler and Riggs, 1983; Lyon, 1998; Riggs, 1990). All of our cell types examined are female. Thus, I speculate the observed increase of engineered L1 insertions on the X-chromosome may be due to specific, but not yet studied, epigenetic modifications that influence L1 integration. The hypothesis follows that the X-chromosomes in HeLa cells would lack such features.

We did not find a preference for L1 integration within the exons or introns of genes (Figures 2.3C and 2.3D). We did observe a slight antisense preference for L1 insertions within genes in HeLa, PA-1, and hESC-derived NPCs; however, careful analyses revealed that this is due to an increase in the presence of L1 EN cleavage sites on the sense strand of genes (Figure 2.3F). Importantly, these data differ from observations of genomic L1s, which show a profound accumulation in the antisense orientation of genes (Smit, 1999). Thus, Darwinian selective forces likely influence the accumulation of genomic L1s over evolutionary time, perhaps because alleles containing L1s in the sense orientation are detrimental and selectively removed from the population.

Intriguingly, L1 insertions in hESCs show a significant antisense insertion bias into genes. We speculate that this observed enrichment could occur if L1 into the sense orientation are more detrimental in hESCs than in other cell types. For example, this scenario could result if sense strand L1 insertions promote hESC differentiation. These somewhat paradoxical findings warrant further examination.

Several reports have claimed that L1 preferentially targets expressed genes, especially in neurons (Upton et al., 2015). By comparison, we did not observe any preference for L1 integration within expressed or transcribed regions of the genome by comparing L1 insertion profiles to both Bru-seq and RNA-seq datasets generated from representative cell lines used in our study (Figures 2.4B, 2.4C, 2.13B, and 2.13C). It is formally possible that L1 insertion preferences within neurons differ from those in cultured cells. However, it is more likely that the single cell WGA amplification used to detect *de novo* L1 integration events in neurons are plagued by high false positive rates (*e.g.*, due to a high rate of chimeric sequencing artifacts) and only a handful of insertions were verified by orthogonal validation experiments (Evrony et al., 2016). If so, the L1 insertion datasets in neurons may not be appropriate for assessing L1 integration preferences. Given the extensive validation of our datasets, we believe the L1 insertion preferences in neurons warrants re-examination.

We observed that engineered L1s readily insert into ‘transposon-free’ regions of the genome (Table 2.6) (Bejerano et al., 2004; Simons et al., 2006), as well as into

protein codon exons. These results differ significantly from genomic L1s, where endogenous L1s are either absent or severely depleted from these regions. Thus, we conclude that 'transposon-free' regions of the genome and protein codon exons are accessible L1 integration sites. It stands to reason that alleles containing these detrimental insertions would be subject to strong negative selective pressures and would be removed from the population. Indeed, these data are consistent with the idea that many disease-producing L1 insertions represent evolutionary dead ends. For example, individuals presenting with Duchenne muscular dystrophy due to an L1-mediated retrotransposition event will sadly seldom father children. Our data also clearly demonstrate the ability of the cultured cell L1 retrotransposition assay to 'capture' mutagenic insertions and underscores our assertion that our data set represents a relatively unbiased set of L1 integration events in the human genome.

Since we observed a depletion of L1 integration sites in expressed and transcribed regions of the genome, we asked whether L1 preferentially integrates within heterochromatic regions of the genome. We compared our L1 integration data set to the 15 and 18 chromatin state data generated by the NIH roadmap epigenomics mapping consortium (Roadmap Epigenomics et al., 2015). Notably, we did not observe a significant enrichment of engineered L1 integration sites within any of the modeled chromatin states (Figures 2.5 and 2.19), though we did observe a depletion of L1 insertions into certain regions, such as genes (Figure 2.3C and 2.3D).

As a control for our analyses, we reanalyzed a previously published MLV retrovirus integration data set and confirmed that MLV exhibits preference for integrating into transcriptional start sites, enhancers, and active promoters (LaFave et al., 2014). Indeed, downsizing the MLV integration site datasets to reflect the numbers of L1 integration events examined in our study still revealed an appreciable enrichment of MLV integration sites in transcriptional start sites, enhancers, and active promoters, (Figure 2.17). These findings support the assertion that our sample sizes could have detected a correlation between the 15 to 18 chromatin states examined by the NIH roadmap epigenomics mapping consortium and L1 integration preferences. Finally, the above findings are consistent with previous studies from our laboratory and support the

hypothesis that the epigenetic silencing of engineered L1 retrotransposition events in PA-1 cells either during or immediately after their insertion is due to an active mechanism and cannot simply be explained by the insertion of L1 into heterochromatic regions of the genome (Garcia-Perez et al., 2010).

We compared our insertion data set to previously published Okazaki fragment sequencing (Petryk et al., 2016), and found that wildtype L1s do not exhibit a strong integration preference into DNA replication initiation sites, DNA replication termination sites, or template or lagging strand DNA templates. Okazaki fragment DNA sequencing (OK-seq), involves capture of Okazaki fragments during DNA replication and characterization by high-throughput DNA sequencing. We downloaded the previously published OK-seq data performed on HeLas and a lymphoblastoid cell line (Petryk et al., 2016). OK-seq involves capture of Okazaki fragments during DNA replication and characterization by high-throughput DNA sequencing. We did find a slight L1 integration preference into the lagging strand template in each of our cell types (Figures 2.6B and 2.19A). Importantly, and in agreement with previous studies, these data further suggest that a proportion of engineered L1 retrotransposition events likely occur during the S-phase of the cell cycle (Kubo et al., 2006; Morrish et al., 2002) (also see below).

We readily admit that it would be ideal to perform OK-seq on a representative group of cell lines used in our studies; however, such an analysis is expensive, time-consuming, and labor intensive. Moreover, unlike RNA expression data, previous studies suggest that DNA replication timing is reasonably correlated between different cell types (Petryk et al., 2016), thereby justifying our approach. That being stated, we acknowledge that cell type specific peculiarities in replication timing will be missed in our analyses.

There are other caveats and potential limitations to our study. First, we cannot rule out the possibility that genomic features other than the presence of a degenerate L1 EN cleavage site influence L1 integration preferences to some extent. Although the size of our dataset has provided an unprecedented opportunity to examine L1 insertion preferences in a statistically robust manner, the dataset size also allows us to observe statistically significant deviations from our weighted random model that have a small

effect size on L1 integration preference. Second, it remains possible that the use of engineered L1s containing selectable (*i.e.*, neomycin) or screenable (*i.e.*, green fluorescence protein) retrotransposition indicator cassettes lead to an ascertainment bias in our data set. If so, we would expect to see an enrichment of L1 integration events in open chromatin and/or expressed genes; however, we did not observe such enrichments. Notwithstanding these caveats, we conclude that L1 EN is the principal factor that drives L1 integration preferences in the genome.

FANCD2-deficient Cells Display ENi and PCNA-independent Retrotransposition

ENi Retrotransposition in FANCD2-deficient Cells

In collaboration with Dr. José L. Garcia-Perez, we characterized a smaller subset of engineered wildtype L1 retrotransposition events in both FANCD2-deficient and FANCD2-proficient cells. As in Chapter 2, we did not observe significant deviations in L1 integration preferences from those generated in simulations using our weighted random model with respect to the L1 chromosomal distribution or L1 insertions into genes. Similar results also were obtained using an L1 containing a missense mutant that blocks the ability of L1 ORF2p to interact with PCNA. Once again, these data again support the hypothesis that L1 EN is the driving force for L1 integration in these cell lines.

Previous data from our laboratory demonstrated that L1s could bypass the requirement of L1 EN and undergo endonuclease-independent retrotransposition in Chinese Hamster Ovary cells and human cell lines that are compromised for the non-homologous end joining (NHEJ) pathway of DNA repair and lack p53 activity (Coufal et al., 2011; Morrish et al., 2007; Morrish et al., 2002). Indeed, dysfunctional telomeres could be used as ENi integration substrates in DNA protein kinase catalytic subunit-deficient CHO cells (Morrish et al., 2007). Thus, through either diffusion or by associating with host factors, endonuclease-deficient L1s are able to localize to genomic sites of disrepair to initiate ENi retrotransposition.

Intriguingly, ENi L1 retrotransposition also occurs in a subset of Fanconi anemia-deficient cells. The examination of a cohort of ENi L1 retrotransposition events in

FANCD2-deficient cells suggests that endonuclease-deficient L1s can target replication intermediates, perhaps by using a 3'OH group present at Okazaki fragments or at double-strand DNA breaks present at collapsed DNA replication forks, to initiate ENi retrotransposition (Figures 3.11A and 3.11B and 4.1). These findings are curiously similar to previous studies with an endonuclease-deficient group II intron, which suggested that it preferentially integrates into lagging strand templates, most likely utilizing Okazaki fragments to prime cDNA synthesis (Ichiyanagi et al., 2003; Ichiyanagi et al., 2002).

How some EN-deficient L1 integration events target DNA replication forks for integration requires elucidation. Indeed, it will be interesting to examine whether L1 ORF2p EN-/PIP domain double mutants exhibit similar or different integration preferences when compared to EN-deficient L1 mutants. Moreover, the nature of the lesion(s) used to initiate ENi in FANCD2-deficient cells require elucidation. A thorough characterization of the structural hallmarks associated with ENi retrotransposition events (*e.g.*, the presence or absence of TSDs and genomic alterations) may provide valuable insight into this process. Preliminary data from Dr. Garcia-Perez indicates that wildtype and ENi retrotransposition events isolated from FANCD2 cells generally contain poly(A) tails; thus, it seems likely that our capture method allows the characterization of a representative cohort of ENi retrotransposition events.

Does ENi Retrotransposition in FANCD2-deficient Cells Recognize Replication Forks for Integration?

The FA pathway is involved in the replication-dependent removal of ICLs; the failure to repair ICLs can lead to stalled replication forks (Scharer, 2005). Thus, FANCD2-deficient cells may harbor more substrates for ENi retrotransposition than FANCD2-proficient cells. Consistent with this idea, previous studies demonstrated that treating FANCD2-deficient cells with low doses of aphidicolin, which inhibits DNA polymerase α and δ and increases the appearance of common fragile sites (Glover et al., 1984), led to an increase in chromosomal gaps and breaks in metaphase chromosomes (Howlett et al., 2005). Common fragile sites are late-replicating regions of

the genome particularly susceptible to replication fork stalling or collapse, which if unrepaired leads to genomic instability (Howlett et al., 2005).

We did not observe an increase of ENi L1 retrotransposition events at common fragile sites in FANCD2-deficient cells (Table 3.1). However, if endonuclease-deficient L1s use stalled replication forks as integration substrates, it is possible that aphidicolin treatment may lead to an increase of ENi L1 retrotransposition at common fragile sites. That being stated, such experiments may be technically difficult because treating an asynchronous population of cells with aphidicolin is a blunt tool—only a small portion of cells in the population may have the lesion required for ENi L1 retrotransposition. Thus, negative results will be difficult to interpret. Moreover, it also is possible that stalled replication forks are not the actual substrates used for ENi L1 retrotransposition. Instead, ENi might occur at double-strand DNA breaks that occur as a consequence of stalled replication forks. Indeed, such double-strand DNA breaks may also serve as ENi retrotransposition integration substrates in XRCC4- and a subset of DNA-PKcs-deficient Chinese Hamster Ovary cells (Morrish et al., 2007).

Does ENi Retrotransposition Represent an Ancestral Mechanism of L1 Retrotransposition?

In sum, our results suggest that ENi L1 retrotransposition may represent an ancestral mechanism used by L1-like retrotransposons prior to the acquisition of an endonuclease domain. Under this scenario, L1-like elements were reliant on genomic features (*e.g.*, site of genomic DNA damage, replication forks, and, less frequently, dysfunctional telomeres) as sites to initiate TPRT in the absence of an endonuclease. The acquisition of an endonuclease domain then allowed L1 to exert its will on the genome, creating site-specific endonucleolytic breaks to liberate 3'-OH groups, which, in essence, allowed it to autonomously insert throughout the genome (Figure 4.2). Notably, this scenario differs to that of other retrotransposons discussed in the introductory chapter, where the acquisition of a site-specific endonuclease or interactions between an element-encoded integrase-type activity and other proteins allowed transposable elements to target genomic 'safe havens' to minimize TE damage

to the host. Indeed, we posit that the acquisition of an endonuclease, as originally implied by its name, allowed L1 to become an interspersed retrotransposon.

Does PCNA Interact with L1 post-EN Cleavage?

PCNA is an important co-factor for DNA polymerase during replication (O'Donnell et al., 2013; Siddiqui et al., 2013) and also participates in several DNA repair processes. Previous studies suggest that an L1 containing a missense mutation in the PIP domain exhibits a decrease in L1 retrotransposition efficiency (Taylor et al., 2013). Drs. Jose L. Garcia-Perez and Tomoichiro Miyoshi observed similar results—an L1 PIP mutant, PIP6, exhibited < 5% of wild-type L1 retrotransposition levels in HeLa cells. Moreover, the L1 PIP6 mutant does not phenocopy L1 EN-deficient mutants because it does not jump in the NHEJ-deficient Chinese Hamster Ovary cell line XR-1.

In Chapter 3, I found that the L1 PIP6 mutant does not affect L1 integration preferences, as the resultant insertions integrate into a degenerate L1 EN consensus cleavage site (Figures 3.4B and 3.5C) and are located throughout the genome (Figures 3.6 and 3.7). Thus, L1 EN activity is not disrupted in the PIP6 mutant, suggesting that PCNA may interact with L1 after the initiation of TPRT, perhaps by recruiting additional host factors required for the completion of L1 integration. Precedence for such interactions exist. For example, PCNA is implicated in recruiting RNase H2 to RNA/DNA hybrids in genomic DNA (Bubeck et al., 2011). It is tempting to speculate that such recruitment may play a role in TPRT (Figure 4.3).

Since PCNA also is involved in several DNA repair pathways, including long patch DNA repair synthesis, it also is possible that PCNA may aid in the completion of L1 integration. For example, LIG1 is a human DNA ligase that interacts with PCNA (Prasad et al., 1996; Srivastava et al., 1998) and is involved in the ligation of single-stranded DNA breaks during long patch repair (Caldecott, 2008). One could easily imagine that the recruitment of LIG1 plays a role in ligating the L1 cDNA to genomic DNA (Figure 4.3). Clearly, the characterization of L1 PIP6 insertions from FANCD2-deficient cells represents a first step toward addressing the above possibilities.

Does PCNA Guide L1 to Replication Forks?

We compared our L1 PIP6 mutant insertion data set in FANCD2-deficient cells with OK-seq replication data (Petryk et al., 2016) to determine if loss of PCNA binding effects integration preference in initiating or terminating replication forks. Examination of our L1 PIP6 mutant insertions shows a significant depletion of insertions at both initiating and terminating replication forks (Figure 3.11C). When we adjust for the observed depletion of wildtype L1 insertions in FANCD2-deficient cells at terminating replication forks, this suggests PCNA may guide L1 to initiating replication forks. These are very preliminary results and need further validation. Additional analysis with Okazaki sequencing directly from FANCD2-deficient cell line could help tease these results out. Although OK-seq in FANCD2-deficient cells may prove to be difficult to perform, as well as interpret, if these cells harbor stalled replication forks.

Since PCNA is a co-factor involved in DNA replication and L1 retrotransposition is believed to occur during S-phase (Kubo et al., 2006; Morrish et al., 2002), PCNA may guide L1 to replication forks. If PCNA guides L1 to replication forks, PCNA-deficient retrotransposition in FANCD2-deficient cells may occur because cells harbor an abundance of stalled replication forks, and moreover it would suggest that L1 can find such forks in abundance on its own.

Additional retrotransposition assays of a double L1 EN- and PIP-box mutant in FANCD2-deficient cells could elucidate whether the L1 EN- mutant is dependent upon PCNA to guide L1 integration. We may expect decreased retrotransposition efficiency in this double mutant if PCNA is directly involved in guiding the L1 EN mutant to its priming substrate. We believe that L1 EN mutants utilize the Okazaki fragments or downstream double-stranded substrates as primers for reverse transcription, but if the L1 can no longer be led to replication forks, what does the L1 RT use as a primer to initiate cDNA synthesis? Would this double mutant kill L1 retrotransposition entirely in FANCD2-deficient cells? If this double mutant leads to observable levels of retrotransposition, characterization of integration events and integration sites could help determine if PCNA may play a role in ENi L1 retrotransposition.

ATR Affects L1 Integration Post-EN Cleavage

Recognition and signaling of DNA damage is mediated by the ataxia telangiectasia mutated (ATM) and ATM Rad3-related (ATR) proteins, which bind to broken DNA ends (Christmann et al., 2003). Previous reports show that ATM-deficient NPCs exhibit increased wild-type L1 retrotransposition levels when compared to ATM-proficient NPCs (Coufal et al., 2011). The characterization of L1 retrotransposition events from ATM-deficient NPCs have typical L1 structural hallmarks, suggesting they integrated via canonical TPRT. However, there appear to be longer, or perhaps more L1 retrotransposition events in ATM-deficient NPCs compared to those in ATM-proficient NPCs. Thus, ATM may act to inhibit L1 retrotransposition in NPCs.

Importantly, Dr. Huira Kopera observed that L1 retrotransposition increases in HeLa cells upon ATR knockdown. Since a complete knockout of ATR leads to cell death, it is hypothesized that ATR may recognize single-strand DNA breaks or other intermediates generated during TPRT. Indeed, consistent with this model, ATR knockdown neither enhances ENi retrotransposition nor does it affect L1 integration preferences in the genome (see Chapter 3). Moreover, L1 integration events characterized from ATR-deficient cells are often longer than those characterized from ATR-proficient cells, and contain an increased frequency of intra-L1 rearrangements and/or genomic deletions at the L1 integration site. Thus, like ATM, ATR may act to inhibit L1 retrotransposition in cultured cells.

Does ATR Act at Second-Strand Cleavage?

ATR also has the potential to recognize single-strand DNA nicks generated during TPRT second strand cleavage. I speculate that second-strand cleavage could lead to an exposed single strand DNA gap that is bound by replication protein A (RPA), which subsequently recruits ATR, the ATR-interacting protein (ATRIP), and the RAD17 complex to participate in DNA repair (Figure 3.1 and Figure 4.4). In the absence of ATR, this substrate would not be subject to repair; as such, one may observe longer retrotransposition events. Similarly, if the second strand endonucleolytic nick occurs upstream (leading to a 5' overhang) of the initial EN cleavage site, or is resected, failure

to repair the resultant single strand DNA gap could lead to the genomic deletions accompanying L1 retrotransposition events in ATR-deficient cells.

Concluding Remarks

My thesis has examined L1 integration preferences in the human genome and ultimately concludes that the L1 EN is the principal determinant for L1 integration preference. Our data suggests that the degenerate L1 EN consensus cleavage site is responsible for previously observed L1 integration characteristics, such as integration events into AT-rich regions of the genome.

We generated a weighted random model that mimics L1 integration sites in degenerate L1 EN consensus cleavage sites distributed throughout the human genome. We utilized this model to determine if certain cellular conditions, cellular host factors, or alternative retrotransposition pathways may influence L1 integration preference. We suggest a new mechanism for EN-independent L1 integration utilizing Okazaki fragments or resultant double strand DNA breaks present at collapsed DNA replication forks to initiate EN-independent retrotransposition. Analysis of additional integration sites from DNA repair protein knockdown or deficiency experiments suggests evidence for cellular factors preventing L1 mobilization pre-EN cleavage (FA pathway) or post-EN cleavage (ATR). Lastly, we examined integration sites of a PCNA-interacting protein-box L1 mutant and find evidence to suggest that PCNA aids L1 integration post-EN cleavage. As demonstrated in this thesis, we have generated an invaluable toolset for the transposable element biology field to gain mechanistic insights about how additional cellular conditions, cellular host factors, or retrotransposition pathways might influence L1 integration in the human genome.

Figure 4.1 A Mechanism for wildtype L1 and L1 EN mutant integration into replication forks.

A) *Replication Fork Structure*: As depicted here the green arrows represent Okazaki fragments or the replication intermediates complementary to the lagging strand template. These intermediates may be replicated DNA primed off the Okazaki fragments but not yet ligated to form one singular strand. The orange arrow represents the leading strand in a replication fork.

B) *Proposed integration of wildtype L1 during replication*: L1 EN may preferentially cleave the lagging strand template during replication, as the strand may be more accessible than the leading strand template, which is being passed by the DNA polymerase for replication.

C) *Proposed integration of L1 EN mutant during replication*: We propose, and our data in Chapter 2 suggests, that the L1 EN mutant (blue circle with star) utilizes Okazaki fragments or replication intermediates containing a free 3'-hydroxl to prime and allow for reverse transcriptase of L1 RNA. In this model second strand cleavage would need to occur by other cellular host factors to ensure integration into the genome. An alternative possibility is that stalled replication forks lead to double strand breaks which can then be utilized as primers for RT activity and integration of L1 EN mutants.

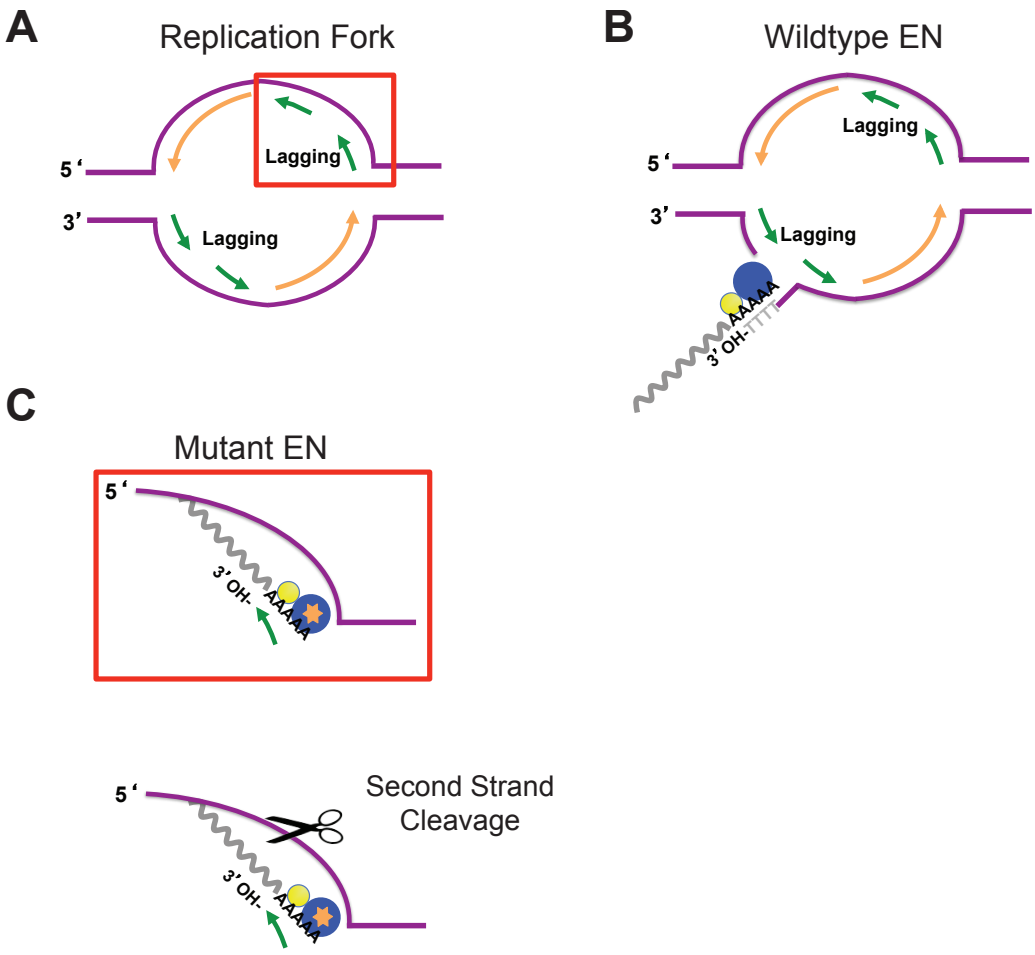


Figure 4.1 A mechanism for wildtype L1 and L1 EN mutant integration into replication forks.

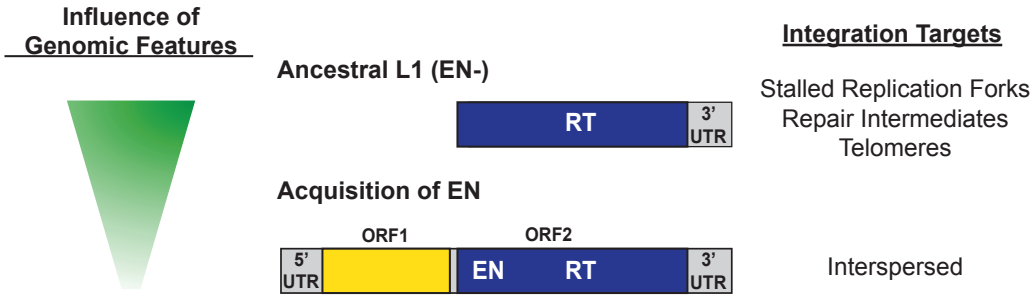


Figure 4.2 Acquisition of the L1 endonuclease has allowed L1 to integrate throughout the human genome.

The L1.3/D205A EN mutant is similar to an ancestral state of L1, which lacks endonuclease activity. This ancestral state, lacking EN activity must rely upon DNA double strand breaks created throughout the genome in order to ensure integration. DNA break opportunities can be presented during stalled replication forks, DNA repair intermediates, or at the end of chromosomes near telomeres. The integration sites of the ancestral L1 are heavily influenced and reliant upon these genomic features. Acquisition of the EN activity allowed L1 to mobilize throughout the genome, no longer completely reliant upon genomic features, and thus becoming interspersed throughout the genome.

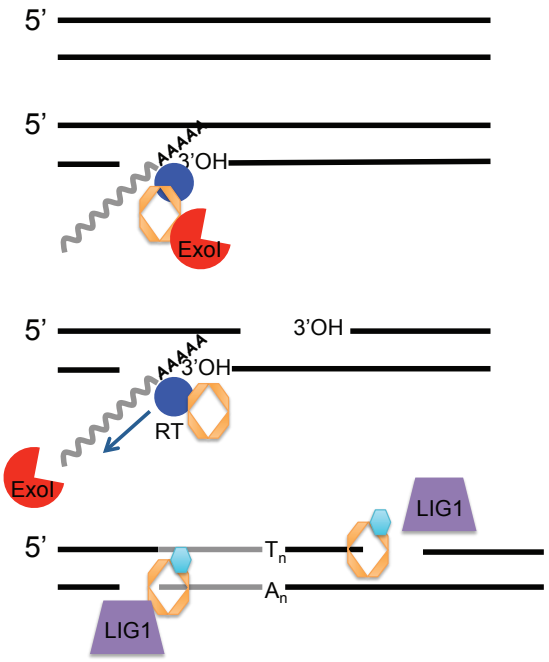


Figure 4.3 PCNA, a recruiter for cellular host RNase H or ligase to aid in L1 retrotransposition.

During TPRT the L1 EN activity of the ORF2p (shown by blue circle) cleaves at a double stranded DNA substrate. If PCNA is bound to ORF2p during TPRT it may be used to recruit Exo1 or similar proteins that contain RNase H activity. This RNase H activity would act after reverse transcriptase of the L1 RNA to cDNA, removing the L1 RNA from the RNA/DNA hybrid so that second strand synthesis can occur. Another proposed mechanism of PCNA is that it recruits ligase to ligate the remaining ssDNA breaks that flank the L1 integration site. PCNA may sit on the 3' end of the cleavage, become modified (light blue hexagon) and thus act as a signal for LIG1 to come to the site and finish retrotransposition.

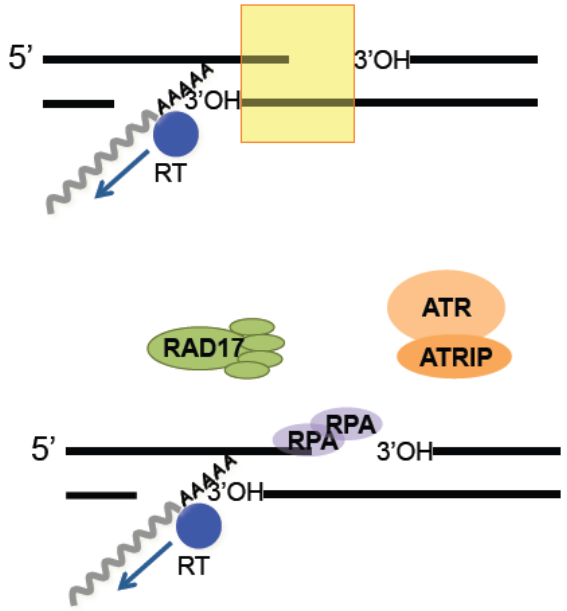


Figure 4.4 Proposed model of ATR acting at second strand cleavage during TPRT.

During TPRT, it is unclear how or when second strand cleavage occurs, but it must happen for integration to occur. At the site of second strand cleavage a ssDNA complex is formed which is the ideal substrate for ATR induced DNA damage repair. This ssDNA may become bound by RPA, which then signals ATR, ATRIP, and RAD17 binding to begin repair. ATR could be seen as a cellular host factor trying to prevent L1 integration. Knockdown of ATR may force the second strand to undergo additional repair mechanisms to resolve the break.

References

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

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M.*, et al.* (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W.J., Mattick, J.S., and Haussler, D. (2004). Ultraconserved elements in the human genome. Science *304*, 1321-1325.

Bubeck, D., Reijns, M.A., Graham, S.C., Astell, K.R., Jones, E.Y., and Jackson, A.P. (2011). PCNA directs type 2 RNase H activity on DNA replication and repair substrates. Nucleic Acids Res *39*, 3652-3666.

Caldecott, K.W. (2008). Single-strand break repair and genetic disease. Nat Rev Genet *9*, 619-631.

Christmann, M., Tomicic, M.T., Roos, W.P., and Kaina, B. (2003). Mechanisms of human DNA repair: an update. Toxicology *193*, 3-34.

Cost, G.J., and Boeke, J.D. (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. Biochemistry *37*, 18081-18093.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Marchetto, M.C., Muotri, A.R., Mu, Y., Carson, C.T., Macia, A., Moran, J.V., and Gage, F.H. (2011). Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. Proc Natl Acad Sci U S A *108*, 20382-20387.

Evrony, G.D., Lee, E., Park, P.J., Walsh, C.A. (2016). Resolving rates of mutation in the brain using single-neuron genomics. Elife *5*

Garcia-Perez, J.L., Morell, M., Scheys, J.O., Kulpa, D.A., Morell, S., Carter, C.C., Hammer, G.D., Collins, K.L., O'Shea, K.S., Menendez, P.*, et al.* (2010). Epigenetic silencing of engineered L1 retrotransposition events in human embryonic carcinoma cells. Nature *466*, 769-773.

Gartler, S.M., and Riggs, A.D. (1983). Mammalian X-chromosome inactivation. Annu Rev Genet *17*, 155-190.

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

Glover, T.W., Berger, C., Coyle, J., and Echo, B. (1984). DNA polymerase alpha inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. Hum Genet *67*, 136-142.

Howlett, N.G., Taniguchi, T., Durkin, S.G., D'Andrea, A.D., and Glover, T.W. (2005). The Fanconi anemia pathway is required for the DNA replication stress response and for the regulation of common fragile site stability. Hum Mol Genet *14*, 693-701.

Ichiyanagi, K., Beauregard, A., and Belfort, M. (2003). A bacterial group II intron favors retrotransposition into plasmid targets. Proc Natl Acad Sci U S A *100*, 15742-15747.

Ichiyanagi, K., Beauregard, A., Lawrence, S., Smith, D., Cousineau, B., and Belfort, M. (2002). Retrotransposition of the LI.LtrB group II intron proceeds predominantly via reverse splicing into DNA targets. Mol Microbiol *46*, 1259-1272.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. PNAS *94, 1872-7.*

Kubo, S., Seleme, M.C., Soifer, H.S., Perez, J.L., Moran, J.V., Kazazian, H.H., Jr., and Kasahara, N. (2006). L1 retrotransposition in nondividing and primary human somatic cells. Proc Natl Acad Sci U S A *103*, 8036-8041.

LaFave, M.C., Varshney, G.K., Gildea, D.E., Wolfsberg, T.G., Baxevanis, A.D., and Burgess, S.M. (2014). MLV integration site selection is driven by strong enhancers and active promoters. Nucleic Acids Res *42*, 4257-4269.

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

Langley, C.H., Montgomery, E., Hudson, R., Kaplan, N., and Charlesworth, B. (1988). On the role of unequal exchange in the containment of transposable element copy number. Genet Res *52*, 223-235.

Lyon, M.F. (1998). X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet *80*, 133-137.

Mack, D.R., Chiu, T.K., and Dickerson, R.E. (2001). Intrinsic bending and deformability at the T-A step of CCTTTAAAGG: a comparative analysis of T-A and A-T steps within A-tracts. J Mol Biol *312*, 1037-1049.

Mol, C.D., Izumi, T., Mitra, S., and Tainer, J.A. (2000). DNA-bound structures and mutants reveal abasic DNA binding by APE1 and DNA repair coordination [corrected]. Nature *403*, 451-456.

Morrish, T.A., Garcia-Perez, J.L., Stamato, T.D., Taccioli, G.E., Sekiguchi, J., and Moran, J.V. (2007). Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature *446*, 208-212.

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

O'Donnell, M., Langston, L., and Stillman, B. (2013). Principles and concepts of DNA replication in bacteria, archaea, and eukarya. Cold Spring Harb Perspect Biol *5*.

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

Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.L., and Hyrien, O. (2016). Replication landscape of the human genome. Nat Commun *7*, 10208.

Prasad, R., Singhal, R.K., Srivastava, D.K., Molina, J.T., Tomkinson, A.E., and Wilson, S.H. (1996). Specific interaction of DNA polymerase beta and DNA ligase I in a multiprotein base excision repair complex from bovine testis. J Biol Chem *271*, 16000-16007.

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

Riggs, A.D. (1990). Marsupials and Mechanisms of X-Chromosome Inactivation. Aust J Zool *37*, 419-441.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J.*, et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Scharer, O.D. (2005). DNA interstrand crosslinks: natural and drug-induced DNA adducts that induce unique cellular responses. Chembiochem *6*, 27-32.

Siddiqui, K., On, K.F., and Diffley, J.F. (2013). Regulating DNA replication in eukarya. Cold Spring Harb Perspect Biol *5*.

Simons, C., Pheasant, M., Makunin, I.V., and Mattick, J.S. (2006). Transposon-free regions in mammalian genomes. Genome Res *16*, 164-172.

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

Srivastava, D.K., Berg, B.J., Prasad, R., Molina, J.T., Beard, W.A., Tomkinson, A.E., and Wilson, S.H. (1998). Mammalian abasic site base excision repair. Identification of the reaction sequence and rate-determining steps. J Biol Chem *273*, 21203-21209.

Stefl, R., Wu, H., Ravindranathan, S., Sklenar, V., and Feigon, J. (2004). DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. Proc Natl Acad Sci U S A *101*, 1177-1182.

Symer, D.E., Connelly, C., Szak, S.T., Caputo, E.M., Cost, G.J., Parmigiani, G., and Boeke, J.D. (2002). Human l1 retrotransposition is associated with genetic instability in vivo. Cell *110*, 327-338.

Szak, S.T., Pickeral, O.K., Makalowski, W., Boguski, M.S., Landsman, D., and Boeke, J.D. (2002). Molecular archeology of L1 insertions in the human genome. Genome Biol *3*, research0052.

Taylor, M.S., LaCava, J., Mita, P., Molloy, K.R., Huang, C.R., Li, D., Adney, E.M., Jiang, H., Burns, K.H., Chait, B.T., *et al.* (2013). Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. Cell *155*, 1034-1048.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sanchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M., *et al.* (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228-239.

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

Wichman, H.A., Van den Bussche, R.A., Hamilton, M.J., and Baker, R.J. (1992). Transposable elements and the evolution of genome organization in mammals. Genetica *86*, 287-293.

Appendix A

Identification of *de novo* LINE-2 Insertions in Early Development of Zebrafish

Work in this appendix was done in collaboration with Dr. Thomas Widmann and Mr. Alejandro Roldán while in the laboratory of Dr. José L. Garcâ-Perez. Dr. Thomas Widmann performed retotransposition experiments in the zebrafish and isolated gDNA. I performed LINE-2 capture techniques on the retrotransposed zebrafish gDNA. Mr. Alejandro Roldán performed bioinformatics analyses to identify LINE-2 insertions.

Introduction

In collaboration with Dr. Thomas Widmann and Dr. José L. García-Perez, currently at the University of Edinburgh, we sought to capture, amplify, and sequence *de novo* engineered LINE-2 insertions from early embryonic zebrafish *Danio rerio*. The aim was to determine the location of *de novo* LINE-2 events during early embryogenesis and determine if LINE-2 preferentially integrates within the *D. rerio* genome.

Two retrotransposition-competent LINEs, ZfL2-1 and Zfl2-2, were identified in the zebrafish genome (Kapitonov and Jurka, 2003; Sugano et al., 2006). Both ZfL2-1 and ZfL2-2 are members of the L2 clade (Sugano et al., 2006). They are classified as stringent type LINEs as their encoded enzymatic machinery specifically recognizes its own RNA 3' tail during retrotransposition (Sugano et al., 2006). This RNA 3' tail is predicted to form a stem-loop structure, which may aid in recognition by the reverse transcriptase (Kajikawa et al., 2005). Upon integration these elements do not generate target site duplications and their 3' tails are composed of microsatellites; of which specific microsatellite sequences can be used to identify distinct element families (Kapitonov and Jurka, 2003). These microsatellites are not sequence targets for these

LINE-2 elements, but rather are inserted into the genome together with the element sequence (Kapitonov and Jurka, 2003). It is believed that generation of these 3' end microsatellites is a result of non-templated additions by the LINE-2 reverse transcriptase (Kapitonov and Jurka, 2003).

ZfL2-1 is approximately 5kb long and encodes two open reading frames (ORFs). ZfL2-1 ORF1 encodes a protein composed of a putative coiled-coil (CC) motif and an esterase (ES) domain, while ORF2 is composed of an endonuclease (EN) domain and reverse transcriptase (RT) domain (Figure A.1). Each ORF is required for efficient retrotransposition (Sugano et al., 2006). While the functional significance of the esterase domain, a catalytic triad composed of serine, histidine and aspartic acid residues is unknown, point mutations within the domain result in reduced retrotransposition efficiency, suggesting that this ES domain has an enhancing function during retrotransposition (Ho et al., 1997; Sugano et al., 2006). It has been hypothesized that the ES domain is important for interaction with cellular membranes, aiding in penetration of host cells (Kapitonov and Jurka, 2003; Nakamura, et al., 2012). The 3' termini of ZfL2-1 elements are composed of $(ATTGA)_n$ which follows 5'-GCTTGA and the polyadenylation signal (Kapitonov and Jurka, 2003).

ZfL2-2 on the other hand is approximately 4.2kb long and contains a repeated sequence in the 5' untranslated region (UTR). The element encodes only one open reading frame with EN and RT domains. The 3' termini of ZfL2-2 contain $(AAATGT)_n$ and they do not have any polyadenylation signal (Figure A.1).

Here we examined engineered ZfL2-1 and Zfl2-2 integration events from zebrafish embryos. We modified the capture and amplification techniques described in Chapter 2, and resulting amplicon products were sequenced by Pacific Biosciences single molecule real time (SMRT) circular consensus sequence (CCS) sequencing technology. This appendix describes the specific modifications performed to capture *de novo* engineered ZfL2-1 and Zfl2-2 integration events.

Results/Discussion

Generating LINE-2 Insertions in Zebrafish

Dr. Thomas Widmann generated ZfL2-1 and ZfL2-2 DNA containing a T7 promoter upstream to generate ZfL2-1 and ZfL2-2 mRNA transcripts to be injected into zebrafish eggs. Three or eight days post injection, zebrafish embryos were collected and genomic DNA (gDNA) was isolated by Proteinase-K and phenol extraction protocols. Transfected zebrafish gDNA was then sent to the University of Michigan where I performed capture methods and PacBio Sequencing.

Capture of de novo LINE-2 Insertions

Utilizing the same general format for capture of *de novo* insertion events in Chapter 2, methods were slightly modified to allow for LINE-2 capture. Transfected zebrafish gDNA was randomly sheared to 3kb fragments with the Covaris S220/E220 series and following the same format as in Chapter 2, libraries were created with the same adapter sequences. Libraries were subjected to a linear amplification utilizing a biotinylated primer specific to both the ZfL2-1 and ZfL2-2 sequence (Figure A.2). Following biotinylation amplification, biotinylated products were streptavidin bead captured and subjected to a nested PCR with an adapter specific primer as well as primers specifically designed for each the ZfL2-1 and ZfL2-2 sequences. Successfully amplified products were column purified. A portion of amplified products was ligated into a cloning vector and Sanger sequenced to verify the capture of *de novo* LINE-2 insertion events in the zebrafish genome. The remaining products were sent for PacBio SMRT CCS sequencing at the University of Michigan Sequence Core and sequences were sent to Dr. José García-Perez's lab for further analysis and characterization by Mr. Alejandro Roldán.

Materials and Methods

Primer sequences: All oligonucleotides used in this study were synthesized by Integrated DNA Technologies (IDT; Coralville, Iowa).

PCR Library Preparation

Top strand adapter with T overhang; purified by high-performance liquid chromatography (HPLC):
5'-GGAAGCTTGACATTCTGGATCGATCGCTGCAGGGTATAGGCGAGGACAACT-3'

Bottom strand adapter with 5' phosphorylation and 3' amino modifier; purified by high-performance liquid chromatography (HPLC):
5'-/5Phos/GTTGTCCT/3AmMO/-3'

10μM final concentrations of annealed adapters were made by incubation of both the top and bottom strand adapters at 95˚C for 5 minutes, followed by allowing the tube to naturally come to room temperature.

15μg of isolated gDNA was sheared to 3kb random fragments following protocols for the Covaris S220/E220 operating systems. Sheared gDNA was purified following QIA-quick PCR Purification Kit (Qiagen #28104). Purified sheared gDNA was end repaired following NEBNext End Repair Module (NEB #E6050). End repaired gDNA was purified following QIA-quick PCR Purification Kit (Qiagen #28104). A non-templated dAMP was then incorporated on the 3' end of the purified end repaired gDNA as outlined by NEBNext dA-Tailing Module (NEB #E6053). Following this reaction dA-tailed gDNA was subsequently purified with the MinElute PCR Purification Kit (Qiagen #28004). Annealed adapters were ligated onto the final purified DNA molecules in the following conditions: 1μg DNA to 90μM of annealed adapter (Final Adapter concentration of 4.5μM) in a 20μl total volume reaction with 1μl (200U) of T4 DNA ligase (NEB #M0202). Ligation reactions were incubated overnight at 16˚C and heat inactivated at 65˚C for 20 minutes. Samples were purified of excess linkers with QIAquick PCR purification kits (Qiagen #28104) and eluted in 50μl EB Buffer.

Uni-linear Biotinylated amplification

Biotinylated Zebrafish; purified by high-performance liquid chromatography (HPLC); 5' Dual Biotin; 18bp internal spacer;

5'-/52-Bio//iSp18/ATTTACCGTAAGTTATGTAACGCGG-3'

Linear extension reactions were performed with Roche Expand Long Range dNTP Pack PCR system. Reactions contained 500ng of template gDNA, the Manufacturer's Expand Long Range Buffer including 12.5mM $MgCl_2$, 0.25μM Biotinylated Zebrafish primer, 500μM PCR Nucleotide mix (dATP, dCTP, dGTP, dTTP at 10μM each), 3% DMSO and 3.5U of Expand Long Range Enzyme. Cycling conditions used are as follows: 94˚C for 3.5 minutes, followed by 30 cycles of 94˚C, 15s; 65˚C, 30s; 68˚C, 3 min., and a seven-minute extension at 68˚C.

Biotin Capture

The Uni-linear extension products were subsequently column purified with QIA-quick PCR Purification Kit (Qiagen #28104). Purified products were biotin captured following Dynabeads kilobaseBINDER Kit (Invitrogen #60101) for 3 hours at room temperature while rotating the tube. After capture, beads were placed on a magnet and washed twice with the Wash Buffer, and washed a final time with ddH_2O. Final biotin captured products were eluted to 30μl with ddH_2O.

Nested PCR

ZfL2-1 Insertion Primers

Adapter primer ZfL2-1: 5'-ATCGATCGCTGCAGGGTATAGG

Zebra primer 3 with ZfL2-1: 5'-ATATGGGCTATGAACTAATG

ZfL2-2 Insertion Primers

Adapter primer Zfl2-2: 5'-GCTTGACATTCTGGATCGATCGC

Zebra primer 5 with ZfL2-2: 5'-ATTAAATTCCTGCAGGTTTGGG

Each 30μl biotinylated captured product was divided amongst 3 separate PCR reactions, where 10μl is used as the starting template per PCR reaction. PCR reactions were performed with Invitrogen Taq DNA Polymerase (Invitrogen #18038042). PCR reactions were brought to 50 uL total volume with ddH_2O and contained final concentrations of 1X of the manufacturer's provided PCR buffer minus Mg, 1.5 mM $MgCl_2$, 0.4μM Adapter primer and 0.4μM zebra primer, 800μM Deoxynucleotide (dNTP) Solution Mix (NEB) (Mix initially contains dATP, dCTP, dGTP, dTTP at 10mM each), and 1U of *Taq* DNA Polymerase. Cycling conditions are as follows: 94˚C for 2 minutes, followed by 35 cycles of 94˚C, 45s; 57˚C, 30s; 68˚C, 3 min and 15s., and a seven-minute extension at 68˚C.

Final PCR products were column purified with QIA-quick PCR Purification kit (Qiagen #28104) and eluted to a final volume of 50μl with elution buffer. Samples were measured for gDNA concentration with Invitrogen's Qubit fluorometer and then sent to the University of Michigan's Sequencing Core for PacBio Single Molecule Real Time (SMRT) circular consensus sequencing.

PCR Product Characterization

Additionally, products were cloned into TA Cloning Kit Dual Promoter (pCR II) cloning vector (Invitrogen #K207020), transformed, and plasmid DNA recovered by Wizard Plus SV minipreps DNA purification system (Promega #A1460). Individual clones were then sequenced with M13 Forward and M13 Reverse primers. Resulting Sanger sequences were then blatted to the UCSC Zv9/danRer7 Zebrafish Genome Browser (http://genome.ucsc.edu/) (Kent, 2002) to determine successful capture and amplification of *de novo* L2 events and flanking 3' gDNA.

Acknowledgements

Thanks to Dr. Thomas Widmann who generated engineered L2 insertion events in zebrafish embryos and subsequently isolated gDNA. Dr. Widmann was influential in creating the zebrafish primers used in the capture techniques. Mr. Alejandro Roldán analyzed the PacBio CCS generated data to identify *de novo* engineered L2 insertion events.

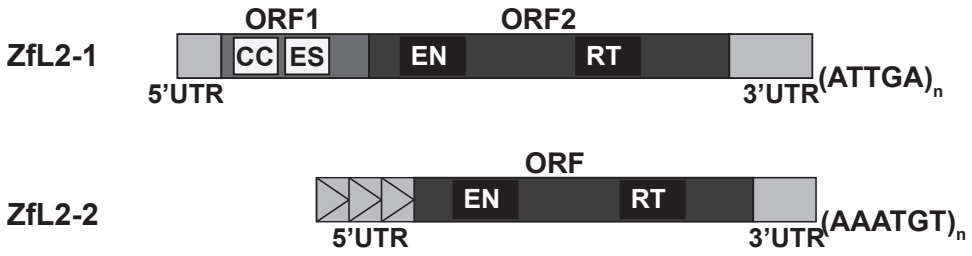


Figure A.1 Schematics of the ZfL2-1 and ZfL2-2 elements. ZfL2-1 is approximately 5kb long and encodes two open reading frame (ORF) proteins. ORF1p contains a coiled-coil (CC) domain as well as an esterase (ES) domain. ORF2p contains an endonuclease (EN) and reverse transcriptase (RT) domain. The element ends in a 3'UTR and typically insertions contain the $(ATTGA)_n$ microsatellite sequence. ZfL2-2 is approximately 4.2 kb long and contains one ORF with an EN and RT domain. The element is followed by an $(AAATGT)_n$ microsatellite sequence.

5’...gcc**ATTTACCGTAAGTTATGTAACGCGG**aactccat**ATATGGG CTATGAACTAATG**accc...3’

Figure A.2 ZfL2-1 sequence showing primer sequences. ZebrafishL2-1 sequence depicting the presence of the biotinylated zebrafish primer sequence (orange) and the Zebra primer 3 sequence (blue).

References

Ho, Y.S., Swenson, L., Derewenda, U., Serre, L., Wei, Y., Dauter, Z., Hattori, M., Adachi, T., Aoki, J., Arai, H*., et al.* (1997). Brain acetylhydrolase that inactivates platelet-activating factor is a G-protein-like trimer. Nature *385*, 89-93.

Kajikawa, M., Ichiyanagi, K., Tanaka, N., and Okada, N. (2005). Isolation and characterization of active LINE and SINEs from the eel. Mol Biol Evol *22*, 673-682.

Kapitonov, V.V., and Jurka, J. (2003). The esterase and PHD domains in CR1-like non-LTR retrotransposons. Mol Biol Evol *20*, 38-46.

Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.
Sugano, T., Kajikawa, M., and Okada, N. (2006). Isolation and characterization of retrotransposition-competent LINEs from zebrafish. Gene *365*, 74-82.

Nakamura, M., Okada, N., Kajikawa, M. (2012). Self-interaction, nucleic acid binding, and nucleic acid chaperone activities are unexpectedly retained in the unique ORF1p of zebrafish LINE. Mol Cell Biol *32*, 458-69.

Appendix B

Identifying Somatic Endogenous LINE-1 Insertions

The work presented here is in collaboration with the Michigan Brain Somatic Moscaicism Network. Mrs. Sarah Emery in Dr. Jeffrey Kidd's laboratory performed tissue culture, genomic DNA extractions, and single cell whole genome amplification experiments. Dr. Weichen Zhou in Dr. Ryan Mill's laboratory is performing analysis of the MiSeq sequencing results generated from this Appendix.

Introduction

As described in Chapter 1, L1 is actively mobilizing within our genomes today, and a number of studies have created methods to capture polymorphic or somatic L1Hs insertions. L1Hs is the currently active L1 sequence within the human genome to date (Beck et al., 2011; Brouha et al., 2003; Moran et al., 1996; Sassaman et al., 1997). A number of studies have identified somatic L1Hs events in the human brain (Baillie et al., 2011; Erwin et al., 2016; Evrony et al., 2012; Evrony et al., 2015; Upton et al., 2015). As part of a National Institute of Mental Health (NIMH) initiative to define the frequency and pattern of somatic mutations in neurotypical individuals and in schizophrenia populations, we devised a set of methods to specifically capture 3' L1Hs sequences and their 3' flanking gDNA (McConnell et al., 2017). We want to address how somatic L1Hs insertions may contribute to neuronal diversity within the neurotypical spectrum as well as in diseased brains.

Neurons are among the longest-lived cells in the body, arising from neural stem cells and progenitor cells that must undergo tens of billions of cell divisions before birth and during the first years of life in order to generate the ~80 billion neurons in the fully developed human brain (Lui et al., 2011; McConnell et al., 2017). Thus, adult neurons have the potential to accumulate several somatic mutations throughout their lifetime.

Background on Schizophrenia

Schizophrenia is a devastating chronic psychotic illness affecting 0.5-1.0% of the population worldwide. The neurodevelopmental hypothesis of schizophrenia states that abnormalities during critical early periods of brain development may trigger the later appearance of clinical symptoms (Bloom, 1993; Murray et al., 1992; Weinberger, 1987). Schizophrenia has an inherited genetic risk estimated at 64% (Lichtenstein et al., 2009). Numerous GWAS studies have led to the idenfication of one-hundred and eight genetic loci containing common alleles with minor risk associated with schizophrenia (odds ratio < 1.2) (Schizophrenia Working Group of the Psychiatric Genomics, 2014). Since somatic brain DNA variations have been found to contribute to the incidence of other neuronal diseases such as early-onset Alzheimer's disease (Beck et al., 2004), brain malformations (Evrony et al., 2012), and Sturge-Weber syndrome (Nakashima et al., 2014; Shirley et al., 2013), several studies have begun to investigate the role of L1 as a source of DNA variation in schizophrenia.

Role of L1 in Progression of Schizophrenia

In 2014 Bundo *et al.* reported significant increase of L1 DNA content, via quantitative real-time PCR, in prefrontal cortex neurons of patients with schizophrenia as compared to healthy controls. Likewise, increased L1 copy number was observed in iPS cell-derived neurons of patients with schizophrenia containing a 22q11 deletion. The 22q11 deletion is a well-defined high-risk genetic factor for schizophrenia, affecting about 1%-2% of schizophrenia patients (Karayiorgou et al., 2010). Because the 22q11 deletion results in the deletion of several genes related to schizophrenia, the L1 increase in schizophrenia patients with the 22q11 deletion is likely to modulate phenotypes of schizophrenia rather than be a direct cause. Examination of affected genes by brain-specific L1 insertions revealed overrepresentation of neuronal function-related terms such as synapse and protein phosphorylation in schizophrenia compared to controls.

Doyle *et al.* 2017 performed L1-seq (Ewing and Kazazian, 2010) on gDNA also obtained from postmortem dorsolateral prefrontal cortex neurons of schizophrenic

patients and healthy controls. An increase of intragenic novel L1 insertions was observed in schizophrenic patients as compared to cases studied. The Bundo *et al.* and Doyle *et al.* studies present evidence that schizophrenic patients may harbor L1 insertions, some of which are located in genes implicated in the pathophysiology of schizophrenia.

What remains to be determined is the functional significance of L1 insertions within genes related to schizophrenia. Furthermore, the population allele frequencies of most detected L1 insertions remain unknown making it difficult to firmly establish association of each with schizophrenia. Whether novel L1 insertions are the cause or the result of putative alterations in brain development in schizophrenia remains to be determined.

Here we present a set of methods which were created and used to capture L1Hs insertions from two female members of the CEPH pedigree 1463: NA12878 and NA12890. β-Lymphocyte cells of each individual were obtained and gDNA was isolated. After completion of the initial testing of these two different presented methods on NA12878, we will perform the more robust method on pooled gDNA from a neurotypical brain chunk, and then move on to several neurotypical and schizophrenia brain samples.

Results/Discussion

Mrs. Sarah Emery kept NA12878 and NA12890 lymphoblastoid cell lines in tissue culture and when confluent, isolated pooled gDNA and single-cell gDNA from each. Mrs. Emery also performed whole genome amplification (WGA) on single-cell gDNA isolations and I performed L1Hs capture techniques on resulting samples. Depending upon the protocol followed, either 5μg or 20μg of gDNA was randomly sheared to 1kb or 2kb fragments with the Covaris M220 series. Following shearing, gDNA was end-repaired and dA-tailed so that our designed adapters with a T overhang could be ligated onto gDNA overnight.

Revised Iskow et al. 2010 method

Final prepped libraries were subjected to two amplification cycling conditions. In the revised Iskow *et al.* 2010 method, the first PCR performed involves an L1Hs specific sequence containing the 'ACA' tri-nucleotide specific to L1Hs elements in the 3'UTR, as well as an 'outside' primer specific to the ligated adapter sequence (Figure B.1). After completion of initial amplification these first PCR products are used as a template in subsequent PCR amplification conditions. This 2nd PCR involves a downstream L1Hs primer sequence, upstream of the L1Hs poly-A signal, and an 'internal' adapter primer sequence. The Resultant products are run on a 1.2% agarose gel and fragments, ~600bp in size, are gel extracted and purified.

Biotinylated Capture Method

Biotinylated capture techniques involve a uni-linear amplification with a dual-biotinylated L1Hs specific primer. This primer is the same sequence in the revised Iskow *et al.* 2010 method but contains dual biotins and an internal 18bp spacer on the 5' end of the primer. After this uni-linear amplification, products are streptavidin bead captured and subjected to several washes. The washed captured products are PCR amplified, with the same 2nd PCR conditions as in the revised Iskow *et al*. 2010 method. Amplified products are run on a 1.2% agarose gel, size selected to ~600bp fragments, and gel extracted.

Illumina MiSeq Primer Modifications

Amplified products to be run on the Illumina MiSeq sequencing machine, should contain p5i502/03 and p7i702/03 primers in the 2nd and final PCR amplification step. These primers add the required sequence needed to complement the MiSeq sequencing platform. When running the Ilumina MiSeq sequencer, the i7 primer, Rd1 SeqPrimer (Adapter) and Rd2 SeqPrimer (L1Hs) need to be added to the appropriate wells of the MiSeq sequencing cartridge. Read 1 from the MiSeq sequencer results should contain the adapter sequence and subsequent 3' flanking gDNA of the L1 insertions, while read 2 should contain the 3' end of L1Hs sequence, including a poly-A tail, and possibly 3' flanking gDNA of the insertion site.

Materials and Methods

Adapter Design

Primer sequences: All oligonucleotides used in this study were synthesized by Integrated DNA Technologies (IDT; Coralville, Iowa).

Top strand adapter with T overhang; purified by high-performance liquid chromatography (HPLC):
5'-GGAAGCTTGACATTCTGGATCGATCGCTGCAGGGTATAGGCGAGGACAACT-3'

Bottom strand adapter with 5' phosphorylation and 3' amino modifier; purified by high-performance liquid chromatography (HPLC): 5'-/5Phos/GTTGTCCT/3AmMO/-3'

10μM final concentrations of annealed adapters were made by incubation of 10μl of 100μM top and 10μl of 100μM bottom strand adapters with 10μl 10X T4 DNA Ligase Reaction Buffer (NEB #M0202) and 70μl of ddH_2O bringing the total volume to 100μl at 95˚C for 5 minutes, followed by allowing the tube to naturally come to room temperature.

Pooled gDNA Library Preparation

Isolated gDNA, max of 20μg DNA, was sheared to 2kb in Covaris' clear miniTUBE (Covaris PN 520064) with the Covaris M220 series following the appropriate setting and protocols found at: http://covarisinc.com/resources/protocols/. Following shearing, gDNA is end repaired using NEBNext End Repair Module (NEB #E6050). End repaired fragments are then column purified with the QIAquick PCR purification kit (Qiagen #28104). Next, purified products are dA-tailed with NEBNext dA-tailing Module Protocol (NEB #E6053). Finally, products are purified once more with the QIAquick PCR purification kit (Qiagen #28104) and eluted to a max volume of 50μl.

Whole Genome Amplified Single-cell Library Preparation

Isolated gDNA, max of 5μg, is sheared to a target bp peak of ~1kb with the Covaris microTUBE-50 (Covaris PN 520166) on the Covaris M220 series. Settings for shearing

were modified as follows: Peak Incident Power (W) of 50, Duty Factor of 20%, Cycles per Burst of 200, Treatment time of 20s, Temperature at 20˚C. Following shearing, sheared gDNA is end repaired and dA-tailed in one single reaction following NEBNext Ultra End Repair/dA-Tailing Module (NEB #E7442). Following completion of this reaction, products are column purified with Qiagen's MinElute PCR column purification (Qiagen #28004) and eluted to a final volume of 10µl.

Annealing Adapters

Annealed adapters were ligated onto the final purified DNA molecules in the following conditions: Up to 1µg DNA to 90µM of annealed adapter (Final Adapter concentration of 4.5µM) in a 20µl total volume reaction with 1µl (200U) of T4 DNA ligase (NEB #M0202). Ligation reactions were incubated overnight at 16˚C and heat inactivated at 65˚C for 20 minutes. Samples were purified of excess linkers with the QIAquick PCR purification kit (Qiagen #28104) and eluted in 50µl EB Buffer. NOTE: For WGA single-cell samples, final purification is performed with Qiagen's MinElute PCR column purification (Qiagen #28004) and eluted to a final volume of 12µl.

PCR Primers

L1Hs TA Iskow: 5'- ATA CCT AAT GCT AGA TGA CAC A

L1Hs TA Biotin: 5'- /52-Bio//iSp18/ATA CCT AAT GCT AGA TGA CAC A

L1Hs Nested: 5'- CAT GGC ACA TGT ATA CAT ATG TAA CTA ACC TGC ACA ATG TG

Adapter Outside: 5'- GCT TGA CAT TCT GGA TCG ATC GC

Adapter Shifted: 5'- ATC GAT CGC TGC AGG GTA TAG G

Illumina Primers

p5i502 Adapter: 5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACA TGT CAC ATG ATC GAT CGC TGC AGG GTA TAG G

Index for Illumina MiSeq: 5'- ATGTCACATG

p5i503 Adapter: 5'- AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT TTC CGT TAT ATC GAT CGC TGC AGG GTA TAG G

Index for Illumina MiSeq: 5'- TTTCCGTTAT

p7i702 L1Hs: 5'- CAA GCA GAA GAC GGC ATA CGA GAT AGA CGA CCG AGC ACA TGT ATA CAT ATG TAA CTA ACC TGC ACA ATG TG

Index for Illumina MiSeq: 5'- TCGGTCGTCT

p7i703 L1Hs: 5'- CAA GCA GAA GAC GGC ATA CGA GAT TCG AAC CAG GGC ACA TGT ATA CAT ATG TAA CTA ACC TGC ACA ATG TG

Index for Illumina MiSeq: 5'- CCTGGTTCGA

Illumina Sequencing Primers

i7Primer (L1Hs): 5'- GGT ACA TGT GCA CAT TGT GCA GGT TAG TTA CAT ATG TAT ACA TGT GC

Rd1 SeqPrimer (Adapter): 5'- ATC GAT CGC TGC AGG GTA TAG GCG AGG ACA ACT

Rd2 SeqPrimer (L1Hs): 5'- GCA CAT GTA TAC ATA TGT AAC TAA CCT GCA CAA TGT GCA CAT GTA CCC

Rd1 SeqPrimer should be added to Pos12 and Rd2 SeqPrimer should be added to Pos14 of Illumina MiSeq wells for sequencing.

Modified Iskow et al. 2010 PCR Capture and Amplification

1st PCR Conditions:

Starting library template of 40-100ng is amplified with Invitrogen Platinum *Taq* DNA Polymerase (Invitrogen #10966018). In 50µl total volume reactions 1X manufactured PCR Buffer, -Mg, 1.5mM $MgCl_2$, 0.2mM each dNTP, 0.4µM each primer (L1Hs TA Iskow and Adapter Outside), and 2U of Platinum *Taq* DNA Polymerase were mixed. The first PCR reaction conditions were as follows: Initial denaturation at 96˚C for 2 minutes; 12 cycles: denature at 96˚C for 30s; 60˚C annealing for 1 min 30s; extension at 72˚C for 1 min 30s; followed by a final extension at 72˚C for 3 min.

2nd PCR Conditions:

Utilizing the same Invitrogen Platinum *Taq* DNA polymerase as in the 1st Iskow PCR, after completion of the 1st PCR, 5µl of the first PCR products are used as a template for

the 2nd PCR. L1Hs TA Nested and Adapter Shifted primers should be used, but if next step is sequencing on Illumina MiSeq sequencing platform then p5i502 Adapter and p7i702 L1Hs primers should be used in PCR reactions. In 50µl total volume reactions 1X manufactured PCR Buffer, -Mg, 1.5mM $MgCl_2$, 0.2mM each dNTP, 0.4µM each primer (L1 TA Nested and Adapter Shifted), and 2U of Platinum *Taq* DNA Polymerase were mixed. The 2nd PCR reaction conditions were as follows: Initial denaturation at 96˚C for 2 minutes; 20 cycles: denature at 96˚C for 30s; 60˚C annealing for 30s; extension at 72˚C for 1 min 30s; followed by a final extension at 72˚C for 5 min.

Biotin L1Hs Capture and Amplification

1st PCR conditions and Biotin Capture:

Starting library template of 40-100ng is linearly extended with Invitrogen Platinum *Taq* DNA Polymerase (Invitrogen #10966018). In 50µl total volume reactions 1X manufactured PCR Buffer, -Mg, 1.5mM $MgCl_2$, 0.2mM each dNTP, 0.4µM L1Hs TA Biotin Primer, and 2U of Platinum *Taq* DNA Polymerase were mixed. The first PCR reaction conditions were as follows: Initial denaturation at 96˚C for 3 minutes; 35 cycles: denature at 96˚C for 30s; 60˚C annealing for 1 min 30s; extension at 72˚C for 2 min; followed by a final extension at 72˚C for 3 min.

Biotinylated products were then biotin captured with Invitrogen's Dynabeads kilobaseBINDER Kit (Invitrogen #60101) for 3 hours at room temperature while rotating the tube [Note: Invitrogen's Dynabeads kilobaseBINDER kit is the only streptavidin bead kit that works successfully with this protocol]. After capture, beads were placed on a magnet and washed twice with the provided Wash Buffer, and washed a final time with ddH_2O. Final biotin captured products are brought to 30µl volume with ddH_2O.

2nd PCR Conditions:

Template for the 2nd PCR is 5µl of the biotin bead captured products. The same Invitrogen Platinum *Taq* DNA polymerase as in the 1st PCR is used again. L1Hs TA Nested and Adapter Shifted primers should be used, but if next step is sequencing on Illumina MiSeq sequencing platform then p5i502 Adapter and p7i702 L1Hs primers

should be used in PCR reactions (or p5i503 Adapter and p7i703 L1Hs can be used). In 50µl total volume reactions 1X manufactured PCR Buffer, -Mg, 1.5mM $MgCl_2$, 0.2mM each dNTP, 0.4µM each primer (L1 TA Nested and Adapter Shifted), and 2U of Platinum *Taq* DNA Polymerase were mixed. The 2nd PCR reaction conditions were as follows: Initial denaturation at 96˚C for 3 minutes; 20 cycles: denature at 96˚C for 30s; 60˚C annealing for 30s; extension at 72˚C for 2 min; followed by a final extension at 72˚C for 5 min.

Final Purification and Illumina MiSeq Sequencing:

Final PCR products were loaded and run on a 1.2% UltraPure low melting point agarose gel (Invitrogen # 16520050). Products ~500bp in size were gel extracted and purified with QIAquick Gel Extraction Kit (Qiagen #28704). Final gDNA concentrations are determined with Invitrogen's Qubit Flourometer. Sarah Emery then performed real-time PCR to determine more accurate sample concentrations before running samples on Illumina MiSeq sequencing machine with MiSeq Reagent Kit v3(600-cycle) (Illumina MS102-3003).

Acknowledgements

Mrs. Sarah Emery in Dr. Jeffrey Kidd's lab was a tremendous help on this project. She was also invaluable in helping run samples on the MiSeq machine in the Moran Laboratory. Dr. Weichen Zhou in Dr. Ryan Mill's laboratory analyzed the paired-end MiSeq sequencing data for analysis and identification of polymorphic LINE-1 insertions.

5’...gggagat**ATACCTAATGCTAGATGACACA**ttagtgggtgcagcgca
ccag**CATGGCACATGTATACATATGTAACTAACCTGCACAAT**
GTGcacatgtaccctaaaacttagagtataataaa...3’

Figure B.1 The L1Hs specific primer sequences to amplify L1Hs integration events and flanking 3’ gDNA.

The 3’ end of the L1Hs sequence is shown. Displayed in orange is the L1Hs TA primer specific to the ‘ACA’ tri-nucleotide found in L1Hs specific sequences. The L1Hs nested primer sequence created for the 2nd nested PCR amplification is displayed in navy. The red ‘g’ represents a SNP expected if successful amplification of L1Hs has occurred. A polymorphic or somatic insertion, if present, should follow this sequence with a poly-A tail and flanking 3’ gDNA.

References

Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., Fell, M.*, et al.* (2011). Somatic retrotransposition alters the genetic landscape of the human brain. Nature *479*, 534-537.

Beck, C.R., Garcia-Perez, J.L., Badge, R.M., and Moran, J.V. (2011). LINE-1 elements in structural variation and disease. Annu Rev Genomics Hum Genet *12*, 187-215.

Beck, J.A., Poulter, M., Campbell, T.A., Uphill, J.B., Adamson, G., Geddes, J.F., Revesz, T., Davis, M.B., Wood, N.W., Collinge, J.*, et al.* (2004). Somatic and germline mosaicism in sporadic early-onset Alzheimer's disease. Hum Mol Genet *13*, 1219-1224.

Bloom, F.E. (1993). Advancing a neurodevelopmental origin for schizophrenia. Arch Gen Psychiatry *50*, 224-227.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc Natl Acad Sci U S A *100*, 5280-5285.

Erwin, J.A., Paquola, A.C., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I., Butcher, C.R., Herdy, J.R.*, et al.* (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat Neurosci *19*, 1583-1591.

Evrony, G.D., Cai, X., Lee, E., Hills, L.B., Elhosary, P.C., Lehmann, H.S., Parker, J.J., Atabay, K.D., Gilmore, E.C., Poduri, A.*, et al.* (2012). Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. Cell *151*, 483-496.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J.*, et al.* (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49-59.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res *20*, 1262-1270.

Karayiorgou, M., Simon, T.J., and Gogos, J.A. (2010). 22q11.2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. Nat Rev Neurosci *11*, 402-416.

Lichtenstein, P., Yip, B.H., Bjork, C., Pawitan, Y., Cannon, T.D., Sullivan, P.F., and Hultman, C.M. (2009). Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet *373*, 234-239.

Lui, J.H., Hansen, D.V., and Kriegstein, A.R. (2011). Development and evolution of the human neocortex. Cell *146*, 18-36.

McConnell, M.J., Moran, J.V., Abyzov, A., Akbarian, S., Bae, T., Cortes-Ciriano, I., Erwin, J.A., Fasching, L., Flasch, D.A., Freed, D.*, et al.* (2017). Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network. Science *356*.

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

Murray, R.M., O'Callaghan, E., Castle, D.J., and Lewis, S.W. (1992). A neurodevelopmental approach to the classification of schizophrenia. Schizophr Bull *18*, 319-332.

Nakashima, M., Miyajima, M., Sugano, H., Iimura, Y., Kato, M., Tsurusaki, Y., Miyake, N., Saitsu, H., Arai, H., and Matsumoto, N. (2014). The somatic GNAQ mutation c.548G>A (p.R183Q) is consistently found in Sturge-Weber syndrome. J Hum Genet *59*, 691-693.

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

Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421-427.

Shirley, M.D., Tang, H., Gallione, C.J., Baugher, J.D., Frelin, L.P., Cohen, B., North, P.E., Marchuk, D.A., Comi, A.M., and Pevsner, J. (2013). Sturge-Weber syndrome and port-wine stains caused by somatic mutation in GNAQ. N Engl J Med *368*, 1971-1979.

Upton, K.R., Gerhardt, D.J., Jesuadian, J.S., Richardson, S.R., Sanchez-Luque, F.J., Bodea, G.O., Ewing, A.D., Salvador-Palomeque, C., van der Knaap, M.S., Brennan, P.M.*, et al.* (2015). Ubiquitous L1 mosaicism in hippocampal neurons. Cell *161*, 228-239.

Weinberger, D.R. (1987). Implications of normal brain development for the pathogenesis of schizophrenia. Arch Gen Psychiatry *44*, 660-669.