Influence of the genomic context on integration site selection by human L1 retrotransposons

Tania Sultana

UNIVERSITÉ CÔTE D'AZUR

École doctorale Sciences de la Vie et de la Santé

Unité de recherche : Retrotransposon and Genome Plasticity

Thèse de doctorat

Présentée en vue de l'obtention du

grade de docteur en Molecular and Cellular Interaction

de

l'UNIVERSITE COTE D'AZUR

par

Tania Sultana

Influence of the genomic context on integration site selection by human L1 retrotransposons

Dirigée par Dr. Gaël Cristofari

Soutenue le 12 December, 2016

Devant le jury composé de :

Bernard	Mari	Directeur de Recherche, CNRS	Président
Cristina	Vieira	Professor, Université Lyon 1	Rapporteur
Vincent	Parissi	Directeur de Recherche, CNRS	Rapporteur
Gaël	Cristofari	Directeur de Recherche, INSERM	Directeur de thèse

Dissertation

Influence of the genomic context on integration site selection by human L1 retrotransposons

By

Tania Sultana

Thesis submitted to Université Côte d'Azur, Sciences de la Vie et de la Santé,

Discipline: Molecular and Cellular Interaction

In fulfilment of the requirements

For the degree of Doctor of Philosophy

December, 2016

Doctoral Committee:

Dr. Bernard Mari, President

Professor Cristina Vieira, Examiner

Dr. Vincent Parissi, Examiner

Dr. Gaël Cristofari, Director of thesis

UNIVERSITÉ CÔTE D'AZUR

To my parents,

Anowara Begum,

And Kamal Bhuiyan

Acknowledgement

I would like to express my gratitude to everybody who supported me during my journey through the bumpy road of doctoral study in abroad. I am thankful to my supervisor, Dr. Gael Cristofari for his guidance and positive attitude. I would like to thank Mme Cristina Vieira, Mr Vincent Parissi, and Mr Bernard Mari for their consent to be part of my doctoral committee. I would like to thank my collaborators, Dr. Olivier Siol, Dr. Nicolas Gilbert, and Dr. Dominic Van Essen for their contribution to my thesis project. I must thank my coauthors in the review article, Alessia Zamborlini, Pascale Lesage, and my supervisor Gael Cristofari for their immense contribution till the last minutes. I am thankful to Olivier Cuvier, Vincent Calcagno, Marc Bailly-Bechet, and Sudip Shaha for their advice on data analysis.

I am greatly thankful to Department of Biochemistry and Molecular Biology of University of Dhaka to encourage me to go abroad for higher study and to allow me to leave the station for a long period. I thank my teachers throughout my education life who believed in my potential and encouraged me to continue. My sincere gratitude goes to my former supervisors, Dr. Laila N Islam, Dr. Firdausi Qadri, and Dr. Edward T Ryan. I sincerely thank my current and former colleagues Alaullah Sheikh, Abu Sayed, Richelle C Charles, and Claude Philippe, who introduced various techniques to me. I am indebted to 'Erasmus-Mundus Mobility from Asia' (EMMA 2013) to fund my study in France, especially Francine Diener and Julie Guillaumat, for their support in the time of need.

I owe many thanks to my friends and colleagues in France for their support in joys and sorrows. I would like to thank Claude, the cool, handsome fatherly figure, Aurelien, the all-rounder, Ramona, my empathetic co-passenger in this journey, and Julian, the traveler. I would like to thank the previous and current lab members who maintained a nice environment in the lab, Monika, George, Javi, Clement, Sebastien, Ashfaq, Natacha, Pilvi, Nadira, Baptiste, and Paula. My gratitude goes to my friends whose presence in my life have been short but intense, Laura-Shiva, Ivana, Dulce, Anke, Gordana, Vivien and Florent. I would like to thank member of the Gilson and the Liti team to maintain the 8th floor warm and friendly. I specially thank Mateo, Jia-xing, and Olivier to rescue me from the scripting loops, and Ben for his humor and to accompany me in the floor late night and in almost every Christmas holidays! I must thank my officemates, Nadir, Alex, Marie-Jo, Aurelien, and Ramona, for maintaining a light environment and for their support during the preparation of this thesis. All the people in Pasteur tower whose smile washed away my fatigue, will stay in my memory. The Bengali people in and around Nice, Kiran, Govinda, Kartik, Sakhwat, Aditi, Ashish, Riyad and Rini, I thank them to bring a flavor of home in my life from time to time.

My deepest gratitude goes to my parents, whom I dedicate this dissertation, for their decision to educate their daughter, and to believe in her potential; to my brother, Kamruzzaman Shohag, and my sister, Tanjina Sultana Onti, who are the two buffers of my life. I am grateful to my husband, Kaiser Shams, the only Bangladeshi men I know to leave home and career to support his wife. He, along with his family members, have set an example that a married Bangladeshi woman can have every bit of freedom. I want to say

sorry to my family members for not seeing them and not being with them when they needed me.

List of abbreviations

APE	Apurinic/Apyrimidinic endonuclease
ATLAS	Amplification typing of L1 active subfamilies
bp	Base pair
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
EN	Endonuclease
L1HS	Human-specific L1
L1/LINE1	Long Interspersed elements
LTR	Long terminal repeat
MGE	Mobile genetic element
MYA	Million year ago
NGS	Next generation sequencing
nt	Nucleotide
ORF	Open reading frame
PCR	Polymerase chain reaction
PIC	Pre-integration complex
RNA Pol II	RNA polymerase II
NPC	Nuclear pore complex
RC-L1	Retrotransposition-competent L1
RLE	Restriction-like endonuclease
RNA	Ribonucleic acid
RNase H	Ribonuclease H
RNP	Ribonucleoprotein particle
RT	Reverse transcriptase
SINE1	Short Interspersed elements
SVA	SINE-VNTR-Alu
TPRT	Target-Primed Reverse Transcription
TSD	Target site duplication
UTR	Untranslated region
VLP	Virus-like particle
TE	Transposable element

Table of Contents

Table of figures

Résumé

Les rétrotransposons sont des éléments génétiques mobiles qui se répliquent avec un intermédiaire ARN et une étape de transcription inverse. Les longs éléments nucléaires intercalés (LINE-1 ou L1 pour long "Interspersed Nuclear Elements") constituent la seule famille de rétrotransposons capables de se répliquer de manière autonome chez l'homme. Bien que la plupart des copies soient défectueuses dû à l'accumulation de mutations, le génome de chaque individu contient environ 100 copies de L1 actives. Elles contribuent à la dynamique du génome humain actuel. Le site d'intégration d'un L1 dans le génome entraîne directement un changement génétique et détermine le devenir de la copie intégrée. Ainsi, l'analyse des sites d'intégration et de leur environnement dans le génome est capitale pour comprendre l'évolution du génome humain, sa plasticité somatique dans le cancer et le vieillissement, et les interactions hôte-parasite. Plutôt que d'étudier les L1 endogènes qui ont été soumis à la pression sélective de l'évolution, nous avons choisi les sites d'intégration *de novo* de L1 exogène obtenus en transfectant un plasmide comportant un élément L1 actif dans des cellules HeLa S3. Puis, nous avons cartographié les insertions *de novo* dans le génome humain avec une haute résolution (au nucléotide près) grâce à une méthode de séquençage avec une grande profondeur, appelée ATLAS-seq. Finalement, les insertions *de novo* ont été analysées pour leur proximité avec un grand nombre d'élément génétique. Nous avons trouvé que les éléments L1 s'intègrent préférentiellement dans des régions de la chromatine faiblement exprimées et renfermant des activateurs faibles. Nous avons aussi trouvé plusieurs positions sensibles "hotspots" avec des intégrations récurrentes des L1. Nos résultats indiquent que la distribution des insertions de L1 *de novo* n'est pas aléatoire, que ce soit à l'échelle locale ou à plus petite échelle. Ainsi nous avons tracé le chemin pour identifier les facteurs cellulaires potentiels responsables du ciblage des insertions de L1.

Abstract

Long Interspersed Nuclear Elements (L1 retrotransposons) have been actively shaping the human genome. A considerable fraction of the human genome originates from L1 activity. Besides their own replication, L1 mobilize other non-autonomous non-LTR retrotransposon RNAs and occasionally some cellular mRNAs. L1-mediated insertions cause various rearrangements at the site of integration, which may contribute to the genome dynamics and sometimes can be pathogenic. To date, 124 cases of L1 mediated-integration have been reported to cause diverse set of diseases, which include a number of epithelial cancers. The consequences of L1-mediated insertions are directly dependent on the nature of insertion sites. Hence, knowing the preferred sites of L1 integration will shed light on human genome evolution and host-L1 interactions. L1's choice of integration site is partly contributed by the flexible sequence preference of L1 endonuclease (5' TTTT/A 3'), which nicks the target genomic DNA where L1 gets integrated. However, given the abundance of such favorable sites in the genome, a relatively dispersed genomic distribution of L1 is expected, which is in contrast to the observations that specific chromosomal regions seem to be particularly susceptible to the L1 machinery and behave as hotspots for L1-mediated retrotransposition. To date, two genomic regions (*c-myc* and *NF1*) and six genomic positions have been reported to be highly permissive towards L1-mediated retrotransposition. Hence, we were interested to learn the integration site preference by L1. Since, endogenous L1 copies are subjected to selective pressure over the evolutionary time, to study L1 preferred sites, we generated novel L1 insertions. We induced L1 retrotransposition in HeLa S3 cells from a plasmid borne active L1 carrying an antibiotic resistance reporter gene at its 3' end and recovered insertions from the cells surviving the G418 selection. Using an adapted in-house pipeline called ATLAS-SEQ, we selectively amplified L1-genome junctions which were sequenced by Ion Torrent sequencing, and sequencing reads were mapped to reference genome to located insertions in single nucleotide resolution. Altogether we rescued 1136 *de novo* L1 insertions from 24 libraries. D*e novo* insertion sites were examined for their proximity towards a large number of genomic features. HeLa S3-specific genomic feature data were obtained from the ENCODE consortium. We found that distribution of *de novo* L1

insertions are non-random both in their local and regional preferences. L1 preferentially integrated in the lowly-expressed chromatin and weak enhancers. We detected several hotspots of recurrent L1 insertions, factors responsible for such recurrent insertions require further evaluation. Our results pave the way to identify potential cellular factors responsible for the targeting of L1 insertions.

Introduction

1. Transposable elements shape the genomes of living organisms

1.1. DNA mobility was first evidenced by Barbara McClintock in 1950

In 1950, Barbara McClintock, the first to coin the notion of mobile genetic elements in the genome, published the presence of 'controlling elements' in maize which can move from one location to another in the genome and can regulate genes nearby the site of transposition (McClintock 1950). She showed the rise of new mutable loci due to the transposition of two loci, Ac and Ds, in the genome, which she called the ‘controlling elements’. Her fellow scientists were skeptical of her ideas since genes were widely accepted to be stable and fixed on the chromosomes at that time. In the next few years, event of transposition in the genome were confirmed by other independent studies. Nevertheless, the ability of these mobile genetic elements (MGEs) to control the genes, which Barbara emphasized in her earlier findings, was not accepted yet. The ground-breaking model of the regulatory operon by Jacob-Monod in 1960 convinced the scientists that genes can be regulated by other genomic segments which he named ‘operon’ (Jacob et al. 2005), although he opposed the concept of gene regulation by mobile genetic elements. Within few years, the discovery of insertion sequence (IS) elements in bacteria demonstrated the plasticity of the prokaryotic genomes resulting from transposition events (Adhya and Shapiro 1969; Shapiro and Adhya 1969). In the same year, Britten and Davidson proposed a model for the regulatory mechanisms in cells of higher organisms (Britten and Davidson 1969). Their model already included a role of MGEs in higher order gene regulation. Since MGEs were not known to have positive contribution to the genome, repeated DNA sequences originated from the MGEs were considered as junk of the genome (Ohno 1972). Soon after, in 1978, researchers found additional evidences supporting the impact of TEs in the genome, for example the antibiotic resistance genes in bacteria and phage or the mating-type switch loci in yeast (Kaulfers et al. 1978; Bukhari and Froshauer 1978; Kushner et al. 1979). To date, MGEs have been found in almost all species including humans, with variable occupancy levels, structures and consequences. A striking example observed in plants is shown in Figure 1-1.The impact of transposable elements on the human genome will be particularly detailed in section 4.

Here mapped, we analyzed novel integration sites of the Long Interspersed Elements-type 1 (LINE-1 or L1), which is the only autonomously-active type of TE in humans, to determine if it exhibits preferred integration site in the genome. To introduce my research, I will first give an overview of the diversity of preferences in relation with their structure, their replicative strategy, and their genomic distribution across species in chapter one. Actively replicating TEs in humans will be described in chapter two. The structure and retrotransposition mechanism of L1 will be detailed in chapter three. The consequences of human L1 integration in health and disease will be discussed in chapter four. In chapter five, I present a recently submitted review article, which review our current understanding of integration site selection by TEs and retroviruses in eukaryotes. In chapter six, I introduce the goal of our study, discuss the rationales, and the experimental approach chosen to tackle this problem. Finally, I describe our results.

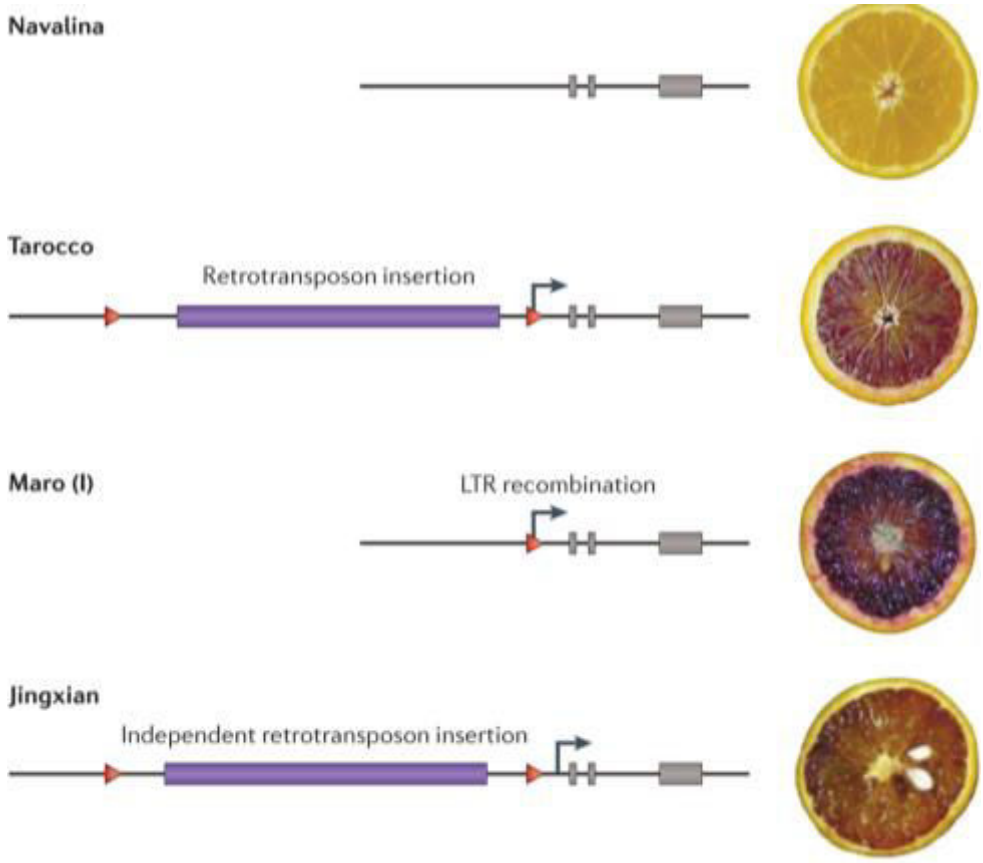

Figure 1-1. Integration of mobile genetic elements can cause genetic and phenotypic variations within species.

The native Ruby gene in the Navalina orange shows limited expression in the fruit flesh. The insertion of Rider, a long terminal repeat (LTR) retrotransposon upstream of the Ruby gene results in fruit variants with novel traits, e.g., cold-inducible Ruby expression in each of the three variants, expression of fruit flesh color in Tarocco and Maro(I) variants, and tissue specificity in Jingxian variant. Ruby exons are depicted here as grey boxes and upstream of them the LTRs flanking the retrotransposons are depicted as red triangles. From (Butelli et al. 2012; Lisch 2012).

1.2. Mobile genetic elements are diverse in structure and mechanism of mobilization

Transposable elements (TEs) can be either 'autonomous', possessing all the elements essential for transposition, and 'non-autonomous', requiring assistance from the machinery of autonomous retrotransposons. Nevertheless, no TE is strictly autonomous,

rather interacts with host factors during or at least for part of its life cycle. Classification of MGEs can be done in many ways, the basic classification is based on the nature of their transposition intermediates.

1.2.1. DNA transposons mobilize by a cut-and-paste mechanism

DNA transposons predominate in bacteria, but are also found in fungi, plants, fish and some mammals. This group of transposons does not require an RNA intermediate. DNA transposons transpose by a 'cut and paste' mechanism where the transposon-encoded transposase excise the element from its original location and help it to insert to a new location. Except for bacterial rolling-circle transposon family or when transposition is coupled to host genome replication, DNA transposons generally do not increase their total number in the genome (Curcio and Derbyshire 2003). Based on the structural variability of the catalytic domain, transposases vary in their molecular mechanisms, but in general, they recognize the short inverted repeat (IR) on both ends of the DNA transposon to excise it out of its original/donor site (Figure 1.3). Transposases are bound to transposon DNA ends until they reach the target DNA. The cleavage of the two strands at the target site are staggered, resulting in a target-site duplication (TSD) of a size typical of 4–8 bp. The Ac/Ds transposition system discovered by McClintock, is a DNA transposon. Two of the most widely studied DNA transposons used for genomic manipulation experiments and as gene therapy tools in mammals are Sleeping Beauty, a resurrected fish DNA transposon, and Piggyback, a cabbage looper moth transposon. DNA transposons occupy 3% of the human genome but none of them present any evidence of recent activity (Lander et al. 2001).

1.2.2. Retrotransposons replicate in the genome by a copy-and-paste mechanism

Retrotransposons are a group of TEs that replicates in the genome by a copy-paste mechanism. This means that the actual fragment of mobile DNA is not altered. Instead it is transcribed into an intermediate RNA copy, whose reverse transcribed DNA gets integrated into a new location after a reverse transcription step. Apart from this basic property, retrotransposons can vary by their structure (Figure 1.2) and by their mechanism of

transposition (Figure 1.3). The two main classes differ by the presence of long terminal repeat (LTR) at their extremities, and are thus called LTR and non-LTR retrotransposons.

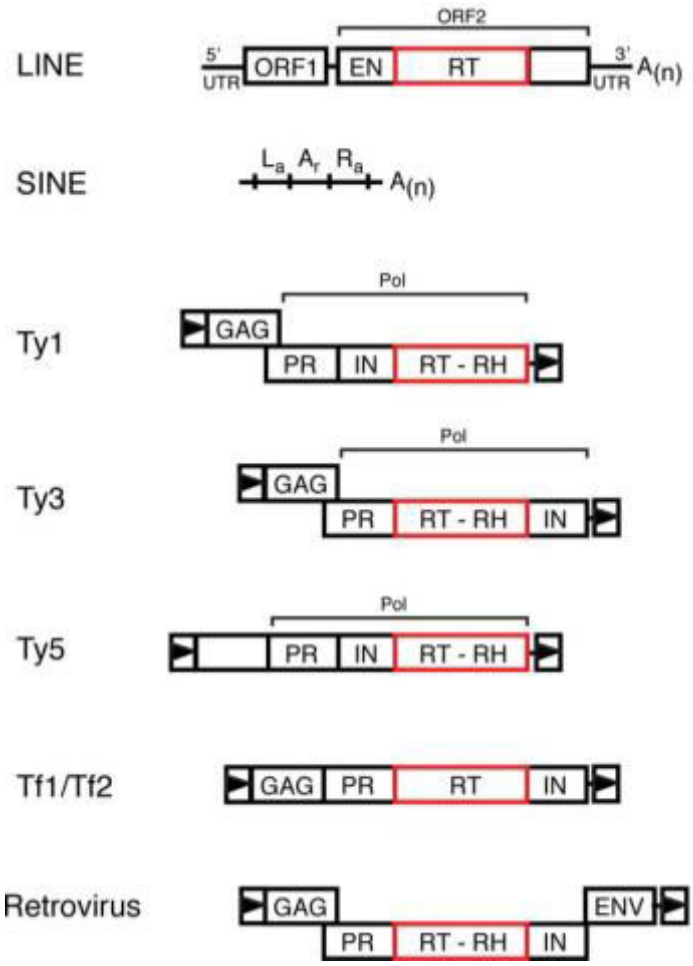

Figure 1-2 Retrotransposon architectures.

Structure (RNA) of some retrotransposons prototypes with the reverse transcriptase (RT) sequence in red (not to scale), from top to bottom: LINE, SINE, Ty1, Ty3, Ty5, Tf1/Tf2, retrovirus. Rectangles represent protein-coding sequences. Coding sequences are as follows: EN, endonuclease; GAG, gag protein; PR, protease; IN, integrase; RH, ribonuclease H domain; Pol, polymerase domain; ENV, envelope protein. UTR, untranslated region; A(n), poly(A) tail; La, left-arm region; Ar, adenosine-rich region; Ra, right-arm region; LTRs, boxed triangles. From (Beauregard et al. 2008).

1.2.2.A. LTR-containing retroelements include LTR-retrotransposons, retroviruses and endogenous retroviruses

LTR-retrotransposons are particularly abundant in eukaryotes, specially in plants where they are the dominating group of transposons. They contain open reading frames (ORFs) that minimally encode Gag and Pol proteins, and are flanked by direct long terminal repeats (LTR) on each end (Figure 1-2). The gag gene encodes the structural components of the VLP. The pol gene encodes a polyprotein with multiple protein domains and catalytic structures (protease, integrase, reverse transcriptase, and RNase H) and is further processed into individual mature proteins by the enzymatic activity of the protease. The reverse transcription of an LTR-retrotransposon RNA occurs in cytoplasmic particles called 'virus-like particles' (VLPs) (Figure 1.3) using host tRNA as primer. Within the VLP, the reverse transcriptase (RT) synthesizes a short cDNA from the 5' end of LTR retrotransposon RNA. Upon completion of transcription, this cDNA is transferred to the 3'end of the same or a second RNA copy which is used as a template for rest of the cDNA synthesis. RNase H degrades most of the RNA in the RNA/DNA hetero-duplex except the relatively resistant poly-purine tracts. These poly-purine tracts act as primers on the cDNA strand to synthesize the second DNA strand. A second strand transfer allows to complete LTR ends synthesis (reviewed in (Hughes 2015)). An integrase-homodimer bind to each end of the dsDNA. DNA together with the bound integrase tetramer is called intasome or integration complex. Integration complex escorts and integrates the DNA in the new genomic target site. LTR-retrotransposons have been extensively studied in Saccharomyces cerevisiae (Baker's yeast) and Drosophila melanogaster (fruit fly). According to the sequence similarity of RT among retrotransposons and the order of the protein domains in the Pol gene, LTR-retrotransposons have been classified in two groups: Ty1/Copia and Ty3/Gypsy. Ty1, Ty2, Ty4 and Ty5 yeast LTR-retrotransposons fall in the Ty1/Copia group and Ty3 LTR-retrotransposons fall in the Ty3/Gypsy group (Xiong and Eickbush 1990) (reviewed in (Eickbush and Jamburuthugoda 2008)). In *Schizosaccharomyces pombe* (fission yeast), *Tf1* and *Tf2* are members of the *Ty3/Gypsy* family.

- Retroviruses originate from LTR-retrotransposons

Retroviruses share a common evolutionary ancestry with LTR-retrotransposons and are assumed to originate from LTR-retrotransposons. One major difference between LTR-retrotransposons and retroviruses is that retroviruses have acquired an envelope gene (env) over the evolutionary period (Figure 1.2) (Finnegan 1983; Temin 1980). The env genes of some of the retroviral families have been traced back to their original viral source (Malik et al. 2000). For example, the origin of the env in gypsy, cer, and tas retroviruses has been tracked to Baculoviridae, Phlebovirus, and Herpesviridae genus of DNA viruses respectively while env gene from mammalian retroviruses has been captured from RNA viruses (Terzian et al. 2001; 2000; Malik et al. 2000).

The *env* gene encodes a surface transmembrane glycoprotein allows budding of viral particles from host cell membranes and binding to host receptors exposed on the cell membrane of target cells to permit cellular entry. The acquisition of envelope genes rendered retroviruses infectious by allowing them to pass from one cell to another or by cell-to-cell contact. Consequently, while LTR-retrotransposons depend solely on vertical transmission, retroviruses are capable of horizontal transmissions. Notably, integrase is conserved among retroviruses and LTR-retrotransposons. The viral diploid RNA genome is reverse transcribed by RT to a linear double stranded viral DNA (vDNA) molecule flanked by LTR at both ends (Shimotohno et al. 1980; Ju and Skalka 1980). Like LTR-retrotransposons, vDNA along with some viral and host cellular proteins, notably viral integrase, forms the pre-integration complexes (PICs) (Bowerman et al. 1989; Wei et al. 1997). When the PICs arrive in their site of integration, integrase processes 3’ end of vDNA and inserts it into the host cell genomic DNA. The integrated viral DNA, also known as provirus, replicate along with the host DNA and can act as a reservoir for future infections. It is functionally equivalent to the LTR-retrotransposons found in genomes. There are seven genus of retroviruses: α-retrovirus through ε-retrovirus, lentivirus, and spumavirus. Gammaretroviruses and spumaviruses are phylogenically more closely related to each other than to lentiviruses (Weiss 2006).

- ### Endogenous retroviruses originate from the ancient retroviral infections of the germline genome

Retroviral infections of the germline from the ancient times can be transmitted to the next generations and accumulated in some genomes (Coffin et al. 1997) (reviewed in (Stoye 2012)). These permanently integrated retroviruses are called endogenous retroviruses (ERVs). Phylogenetic studies of Gypsy retrovirus in eight species of Drosophila has revealed that both vertical and horizontal transmissions were involved in the evolution of insect endogenous retroviruses (Terzian et al. 2000). In human, human endogenous retroviruses (HERVs) comprise 8% of the genome (Lander et al. 2001) but none of them are fully replication competent due to accumulation of mutations although non-infectious HERV particles can be produced in particular circumstances (Grow et al. 2015).These HERVs resemble known exogenous retroviruses—Class I HERVs are most homologous to the gammaretroviruses, Class II HERVs to betaretroviruses, and Class III HERVs to spumaviruses (Medstrand et al. 2002).

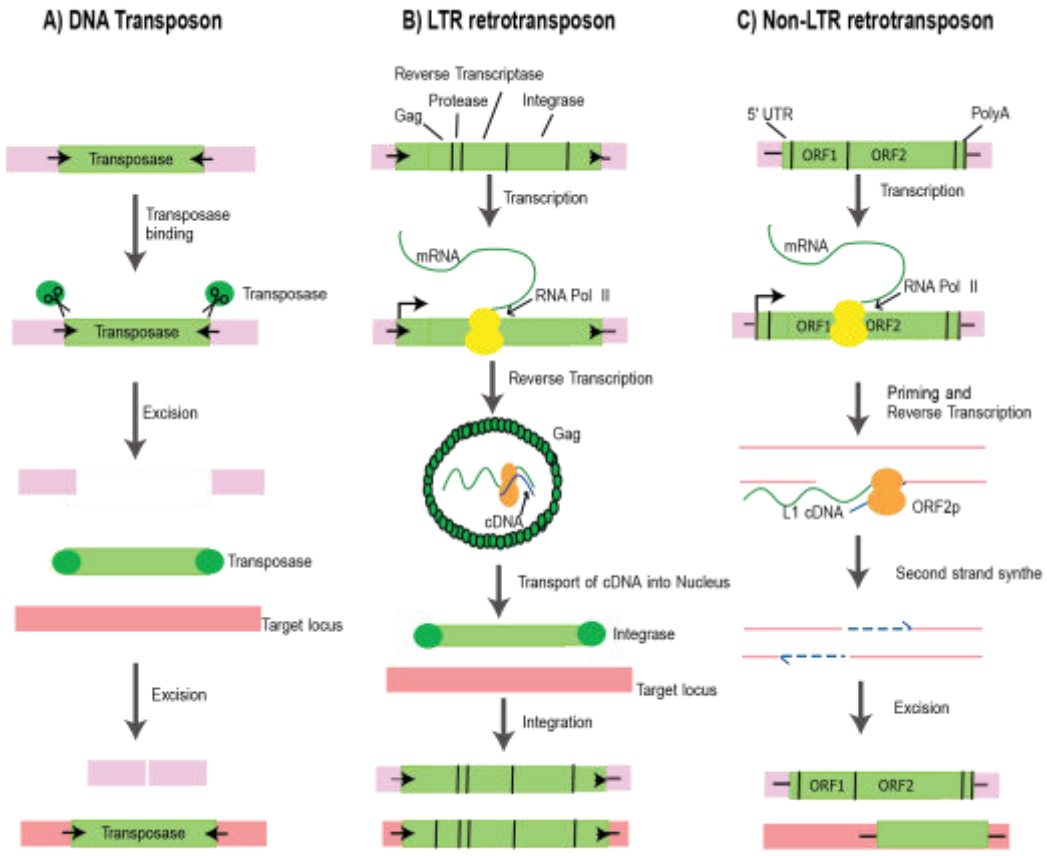

Figure 1-3 Replication models for the three major classes of transposable elements.

TEs are presented as light green bars, terminal inverted repeats in DNA transposon and LTR in LTR-retrotransposon as black arrows, transposase and integrase proteins as green circles, transposon donor sites as light pink bars, novel integration sites as dark pink bars, target site duplications (TSD) as black horizontal lines, RNA polymerase II in yellow color, Gag proteins as dark green circles, reverse transcriptase protein in orange shape and color, RNA transcript of TEs as green waves and their reverse complement cDNAs as blue waves. Adapted from (Levin and Moran 2011).

1.2.2.B. Non-LTR retrotransposons vary in their endonuclease domains

Non-LTR retrotransposons contain no long-terminal repeat and are the likely ancestor of the LTR-retrotransposons (Figure 1.2). In contrast to the LTR retrotransposons, which in most cases, use the host tRNA, to prime reverse transcription (Ke et al. 1999), non-LTR retrotransposons use host genomic DNA ends at the target site to initiate the reverse transcription (Figure 1.3). This process was first detailed for the R2 element in silkworm and is called Target-Primed Reverse Transcription (TPRT) (Luan et al. 1993) Non-LTR

retrotransposons fall into either of the two classes: i) the RLE-encoding elements, and ii) the APE-encoding elements. RLE-encoding elements contain a Restriction enzyme-Like Endonuclease domain in the C-terminus of the single open reading frame (ORF) while the APE-encoding elements contain a APurinic/apyrimidinic Endonuclease domain in one of its two ORFs (Yang et al. 1999; Kapitonov et al. 2009; Feng et al. 1996). The single ORF of RLE elements is necessary for replication and contains an RT and an EN domain. For the APE-encoding elements, the first ORF encodes a protein with nucleic acid binding and chaperone activity, and the second ORF encodes both endonuclease and reverse transcriptase activities. Most of the RLE encoding elements, for example, R2 elements from insects and arthropods, are sequence specific, while only a small subset of the APE-encoding elements is site specific or show weak specificity for target sites (Fujiwara 2015). One major difference between these two classes is that the APE domain directly contributes to the sequence specificity of the target site, while in the RLE encoding elements, sequence specificity comes from the DNA binding motif rather than the RLE domain itself. Many of the well-defined non-LTR retrotransposons are APE encoding, for example, L1 elements from mammals, TRAS/SART and R1 elements from Bombyx mori, TART-HetA-TAHRE elements from Drosophila melanogaster, and TRE5-A from Dictyostelium discoideum. The structure and replication mechanism of human L1, a non-LTR retrotransposon, will be detailed in section 3.

1.2.2.C. Retroelements share a common ancestor with RNA viruses

Retrotransposons use a reverse transcriptase (RT) activity to replicate. Phylogenetic analysis of the RT domains provided information about the origin and divergence of retroelements. Eickbush team built a phylogenetic tree of RT including 82 retroelements from various species and RNA polymerases from RNA viruses to infer the origin and evolution of the retroelements. Their study concluded that RNA viruses and retroelements share a common ancestor. The progenitor elements did not have the LTR, which was acquired later in the evolutionary period. Both the LTR and non-LTR ancestral elements contained gag and pol genes (Xiong and Eickbush 1988; 1990). Non-LTR elements are as old as eukaryotes (Malik et al. 1999). It is proposed that while invading the nucleus of primitive eukaryotes, some mobile group II introns have lost the RT ORF and became splicosomal

introns; others lost the intron RNA structure and become non-LTR retrotransposons (Robart and Zimmerly 2005). Alternately, some studies in *Drosophila* and as well as in other higher order species suggest an evolutionary link between the telomerase complex and the reverse transcriptase domain of retrotransposons (Pardue and DeBaryshe 2003).

Building a phylogenetic tree rooted by the RT sequence of group II introns revealed upto eleven distinct non-LTR retrotransposon clades. The oldest three clades of non-LTR elements (CRE, R2, R4) were sequence-specific by virtue of a restriction enzyme-like endonuclease (RLE) domain located downstream of the RT domain (Malik et al. 1999). Eight clades including L1 and R1, evolved from these three original clades by the acquisition of an apurinic-apyrimidic endonuclease-like (APE-like) domain upstream of the RT domain with broader specificity (Malik et al. 1999). Four of the APE-like domain containing clades which include R1, later acquired RNase H domain downstream of the RT domain (Malik et al. 1999).

1.3. Transposable elements differ in their genomic distribution

1.3.1. Host-TE interactions over an evolutionary time results in non-random distribution of TEs

TEs exhibit highly diverse genomic distribution. The variable distributions of mobile elements in the contemporary genome arise both from their eventual integration preferences and from a variety of selective pressures. Indeed, deleterious insertions will be lost and beneficial insertions will be maintained over an evolutionary period. TEs' selection of sites for integration, and the hosts' strategy to minimize TE-mediated damage, collectively presents the pattern of TE distribution we observe in the genome (Martin and Bushman 2001; Han et al. 2006; Brady et al. 2009; Kazazian 2004). Indeed, few studies comparing the de *novo* versus the fixed insertions, or the younger versus the older insertions evidenced the differences in insertion distribution (Brady et al. 2009; Ovchinnikov et al. 2001; Barr et al. 2005). Comparison between the patterns of *de novo* and the fixed insertion of human endogenous retrovirus (HERV-K) showed that the *de novo* insertions were slightly enriched in transcription units, gene-rich regions, and near the

histone marks associated with the active transcription units and the regulatory regions while the fixed insertions were found preferentially outside the transcription units (Brady et al. 2009). Orientation of TEs also contribute to the post integration elimination frequency. For example, HERV-K fixed insertions, which were in the same transcriptional orientation relative to the host gene were prone to elimination to ensure minimum disruption of host mRNA synthesis. In contrast, novel insertions within transcription units showed no such orientation bias (Brady et al. 2009). A similar difference in insertion features and orientation bias was found for avian sarcoma-leukosis virus (ASLV) insertions in chicken cells (Barr et al. 2005). Youngest HERV-K elements in the human genome showed a distribution intermediate between the *de novo* integration sites and the older fixed HERV-Ks confirming the changes in genomic distribution of TE over time.

1.3.2. Analysis of novel integration sites reveals TE-specific favored genomic sites

With the advances in sequencing technologies and the availability of annotated genomic features in the reference genomes, integration site selectivity has been evidenced for a number of TEs in the past decade. While some TEs favor integration into specific genomic regions or features, others show more dispersed pattern of insertions. Besides their primary choice of sites, some TEs also show secondary preferences for alternative features in response to physiological stimuli (Dai et al. 2007). Additionally, a microfeature within a preferred macrofeature can participate in integration site selection. For example, Ty5 integrates into heterochromatin, but more specifically in nucleosome free regions and open sites within the heterochromatin (Baller et al. 2011). Despite their complexity and diversity of the mechanism to integrate in their favored site, an overview of integration site specific TEs is presented below and is reviewed in section 5.1

1.3.2.A. Transposable elements enriched in or near gene-rich regions.

Many TEs integrate into gene-rich regions, although most of the events occur in sites that prevent disruption of ORFs. For example, the P element, a DNA transposon from *D. melanogaster* avoids disruption of ORFs by integrating within 500bp upstream of transcription start sites (Bellen et al. 2011). Ty1, Ty2, Ty3 and Ty4 yeast LTR-

retrotransposons integrate within element-specific window upstream of RNA polymerase III transcripts, namely tDNA genes, while Tf1 and Tf2 yeast LTR-retrotransposons integrate upstream of RNA polymerase II transcripts. Non-LTR retrotransposons are also known to integrate in gene-rich regions. The Dictyostelium discoideum non-LTR retrotransposon, TRE5A preferentially integrates ~48 bp upstream of the tRNA genes, whereas TRE3A integrates downstream of tRNA genes (Siol et al. 2006; 2011; Winckler et al. 2002). Likewise, few non-LTR insect retrotransposons from R1 and R2 clades integrate into 18S and 28S rDNA locus (reviewed in (Fujiwara 2015)). Certain retroviruses also exhibit preferential integration in gene-rich regions. For example, HIV-1 preferentially integrates into intronic regions of highly transcribed genes (Schroder et al. 2002; Singh et al. 2015), whereas murine leukaemia viruses (MLVs) shows a strong integration preference near certain regulatory sequences, for e.g, strong enhancers, promoters and transcriptional start sites (LaFave et al. 2014; Mitchell et al. 2004; Ciuffi 2008).

1.3.2.B. Transposable elements enriched in telomeric regions.

A number of non-LTR retrotransposons integrate specifically in telomeres. TRAS1 and SART1 from the R1 clade in silkworm integrate into the 'TTAGG' repeats of the telomeres and are involved in a telomerase-independent telomere maintenance pathway. Het-A, TART, TAHRE (Telomere-associated and HeT-A related element) in Drosophila are located at the extreme ends of the telomeres (Biessmann and Mason 2003; Pardue and DeBaryshe 2000; Rashkova et al. 2002b; 2002a). Het-A and TART are non-autonomous elements and complement each other for successful retrotransposition. HeT-A lacks reverse transcriptase, but retrotranspose by recruiting the TART reverse transcriptase. Likewise, TART recruits HeT-A gag protein (Casacuberta and Pardue 2005; Pardue et al. 2005) to access the target sequence. In certain condition, human L1 also integrates at telomeres, for example, in cells lacking functional p53 and non-homologous end-joining pathway (NHEJ) which can naturally arise in cancer (Morrish et al. 2002).

1.3.2.C. Transposable elements enriched in heterochromatin.

Some transposons target heterochromatin which contains relatively few genes. Chromoviruses, which are members of *Ty3/Gypsy* LTR-retrotransposons, contain a chromodomain in their integrase domain. Chromodomains are involved in chromatin remodeling by binding methylated histones (Eissenberg 2001; Nielsen et al. 2001). Chromoviruses integrate in the heterochromatin of eukaryotes from fungi to vertebrates (Gao et al. 2008; Malik et al. 1999). Fusion of the chromodomain from fungal MAGGY chromovirus with the Tf1 integrase is sufficient to redirect Tf1 to heterochromatin (Gao et al. 2008). The Ty5 LTR-retrotransposon also integrates preferentially into heterochromatin in S. cerevisiae. Approximately 75% of Ty5 integration events occur within the telomeric and sub-telomeric heterochromatin while the rest integrates in easily accessed sites in open chromatin (Baller et al. 2011).

1.3.2.D. Transposable elements dispersed across the genome

Finally, many of the TEs do not show any identified site-selectivity and rather integrates in a pattern close to random, such as *Sleeping Beauty* (a resurrected DNA transposon), or the avian sarcoma leukosis virus (ASLV) (an alpha-retrovirus) (Gogol-Döring et al. 2016; Mitchell et al. 2004; Narezkina et al. 2004). There are other TEs for which indirect evidence supports non-random integration but for which site-specificity has not been properly investigated yet, for example, human L1 and Alu non-LTR retrotransposons (see section 6). TEs which integrate randomly in the genome have been manipulated to be used as gene delivery vehicles for functional genomics study and for clinical gene therapy. Retroviral vectors have been widely used for these purposes due to their high delivery efficiency, and long term-stable expression of the delivered transgenes. However, recent studies have revealed that many of the retroviruses used for gene delivery are biased for particular genomic sites and can cause serious damage by integrating the transgene into genomic sites of cellular importance (Hacein-Bey-Abina et al. 2003). Hence, choice of vector influences the extent of damage due to insertional mutagenesis. Genomic safe harbors for transgene integration are genomic locations where landing of a transgene will be the least damaging to the host. These sites are often located far away from the coding-, non-coding- and regulatory

sequences. TEs, which do not show integration site-specificity, possess relatively lower chances of landing in regions affecting genes and thereby has the potential to be used as safer gene delivery tools. DNA transposons from the mariner superfamily, including the Sleeping Beauty, are potential gene delivery vectors under study due to their minimal target site requirements, integration in wide range of hosts irrespective to the tissue types (Claeys Bouuaert and Chalmers 2010). Of course the safest situation should be site-specific integration in safe harbors.

2. A limited number of TE families are actively replicating in modern humans

The initial analysis and sequencing of human genome in 2001 revealed unpredicted information on genome composition (Lander et al. 2001). 45% of the human genome is composed of TEs. The protein coding sequences occupies only 2% of the genome whereas 45% of the human genome is occupied by repeat elements derived from the activities of mobile genetic elements (Figure 2-1). DNA transposons represent 3% of the genome, LTR-retrotransposons 8%, and non-LTR retrotransposons 34%. Among the non-LTR retrotransposons, LINEs and SINEs comprise 21% and 13% respectively. LINE-1 is the only autonomously active TE family in the contemporary human genome. Other retrotransposons, for e.g., Alu and SVA elements are also active and employ the LINE-1 machinery for mobilization. Thereby, in this section we will focus on LINE-1 elements.

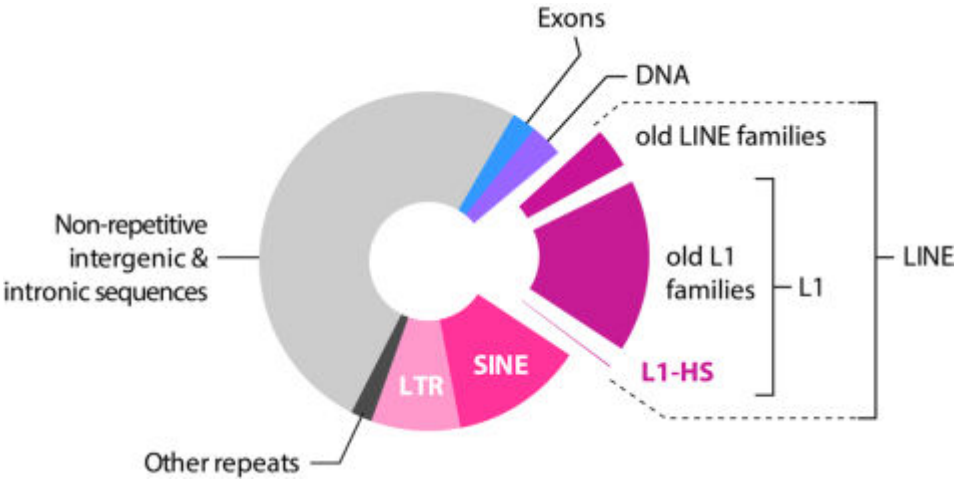

Figure 2-1 Transposable element content of the human genome.

Half of our genome is occupied by repeat elements. Human specific L1 (L1HS) forms a tiny fraction of genome and solely contribute to the total pool of retrotransposition activity. Adapted from (Lander et al. 2001).

2.1. L1HS predominantly mobilizes sequences in *cis*

2.1.1. Waves of L1 amplification contributed to primate genome evolution

LINE-1 retrotransposons have been amplifying in mammalian genomes for more than 160 million years (Burton et al. 1986; Smit et al. 1995). L1 sequences accumulate mutations in a neutral rate, thereby older sequences are proportionately more divergent from the active L1 consensus sequence compared to the younger ones (Voliva et al. 1984; Boissinot et al. 2000; Lee et al. 2007). Sequence comparison between individual genomic L1 sequences in the contemporary genome and a consensus sequence derived from modern-active LINE-1s have unearthed the age of 21 primate-specific L1 subfamilies, termed as L1PA1 to L1PA16 and L1PB1 to L1PB4, where an increase in the number of the terms denotes an increase in the age (Smit et al. 1995; Khan et al. 2006). The most prolific families are L1PA8 to L1PA3, which amplified 40 to 12 million years ago (MYA) (Khan et al. 2006). The human specific L1 subfamily, L1HS, also known as the L1PA1 family, and emerged only ˜4 millions of years ago (MYA), sometimes after the divergence between humans and chimpanzees (6 MYA). Recent studies suggest that host defense proteins have evolved in parallel to the evolution of L1 families to protect the genome from the mutagenic effects. Restriction host factors are often specifically active against a given L1 subfamily or a group of them, but are unable to counteract the mobility of other sub families (Castro-Diaz et al. 2014; Jacobs et al. 2014). As a consequence, over the evolutionary time, one L1 sub family has been replaced by another, wave after wave. During a certain period, only one dominant family was mostly active, (Boissinot et al. 2000) whereas closely related families coexisted (if they had different 5'UTR) for a short period until one of them finally took over (Khan et al. 2006; Boissinot et al. 2000; Cabot et al. 1997; Casavant and Hardies 1994).

2.1.2. L1s are the only source of retrotransposition machinery in human genome

Currently, only a set of very few L1HS belonging to L1-Ta (transcribed, subset 'a') and pre-Ta subfamily is transcriptionally active. Approximately 400 L1 elements in the human genome falls in pre-Ta category/subfamily (Salem et al. 2003) which contains a diagnostic ACG trinucleotide at positions 5930-5932 and a G nucleotide (position 6015) in their 3'

untranslated region (3'UTR) (see Figure 2-2) (Boissinot et al. 2000). The group L1-Ta is the most recent L1 subfamily in the human genome and contains a diagnostic 5'-ACA-3' trinucleotide (position 5930-5932). This subfamily has further evolved into two branches, Ta-0 and Ta-1, and has approximately 520 members in our genome (Myers et al. 2002; Boissinot et al. 2000). The Ta-0 subfamily is older than the Ta-1. The Ta-1 subset contains the largest number of active L1s accounting for around half of the Ta family, followed by the Ta-0 and the pre-Ta subfamilies (Brouha et al. 2002; Boissinot et al. 2000; Beck et al. 2011; Sassaman et al. 1997). Out of the 459 L1-Ta elements analyzed in the reference human genome, 192 belong to the Ta-1 subset, 137 to the Ta-0 subset. The subset for the remaining 130 elements is either indistinguishable on account of truncations or rearrangements of the diagnostic nucleotides, or they fall in an intermediate subset between Ta-0 and Ta-1 (Myers et al. 2002). Ta-1 differentiated into two groups, Ta-1nd (no deletion of G at nucleotide position '74') and Ta-1d (see Figure 2-2). The youngest subset of Ta-1, Ta-1d, arose about 1.4 MYA and accounts for approximately two thirds of the Ta-1 subfamily (Boissinot et al. 2000).

Among the 500,000 L1 sequences in the current human genome (Lander et al. 2001), only a variable set of 80 to 100 elements are full length and potentially retrotransposition competent due to 5' truncations and to the accumulation of other alterations in the L1 body (Beck et al. 2010; Brouha et al. 2003). Retrotransposition-competent L1s are the only source of transposition events in the current genome. Depending on the method of analysis, the estimated rate of inheritable L1 retrotransposition events in humans varies between 1 in 20 to 1 in 200 births (Ewing et al. 2015; Cordaux et al. 2006; Xing et al. 2009).

L1 encoded proteins preferentially mobilize their own mRNA, a phenomenon known as cis preference (Esnault et al. 2000; Wei et al. 2001). Besides their autonomous activity, L1 proteins can also occasionally act in trans to mobilize non-autonomous non-LTR retrotransposons (e.g., human Alu and SVA elements) (Raiz et al. 2012; Dewannieux et al. 2003; Hancks et al. 2012; 2011) and cellular mRNAs leading to processed pseudogene (retropseudogenes) formation (Wei et al. 2001; Esnault et al. 2000). Typical hallmarks of L1-mediated retrotransposition includes target site duplications (TSD), 5'-truncations, 5'-end inversions, poly(A) tail of variable length, and absence of introns. Each L1-mobilized

element, including Alu elements, are flanked by the direct repeats of variable length at the integrated site created by staggered cuts generated by the EN during TPRT (Cost et al. 2002).

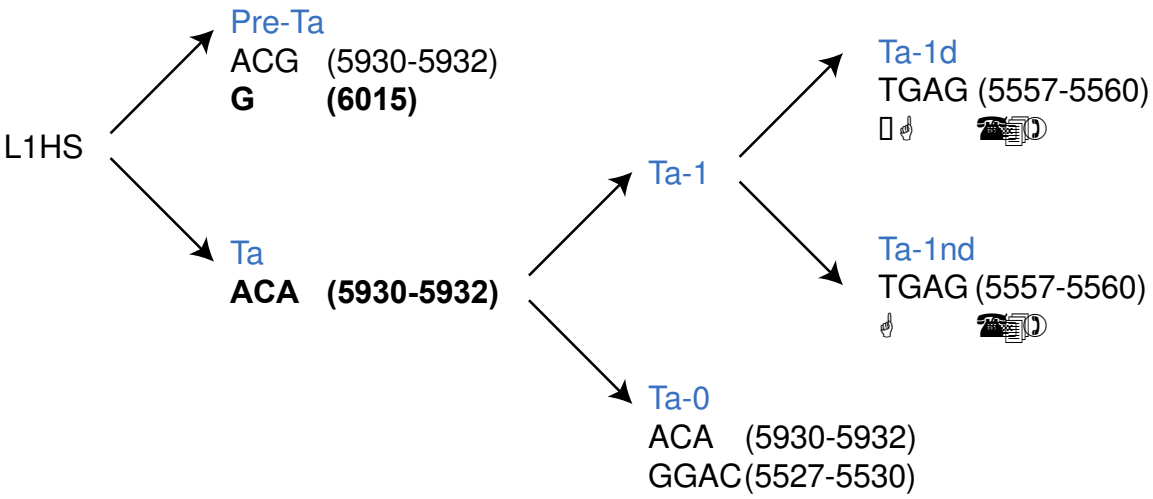

Figure 2-2. Classification of active L1HS subfamilies and the position of their diagnostic nucleotides.

Diagnostic nucleotides are presented below the L1HS subfamily, and their positions are in parentheses. Nucleotides in bold represent diagnostic nucleotides that are also found in derived younger subfamilies. From (Boissinot et al. 2000).

2.2. Alu and SVA are repeated non-coding sequences mobilized by L1 in *trans*

2.2.1. Alu is the most abundant TE family in humans and hijacks L1 proteins for mobilization

With more than 1 million copies in the human genome, Alu elements are the most abundant retrotransposons by copy number, occupying 11% of the genome (Lander et al. 2001). Alu elements retrotranspose more frequently compared to other TE in humans, with an estimated retrotransposition rate of one event in every twenty human newborns (Cordaux et al. 2006; Xing et al. 2009; Huang et al. 2010; Li et al. 2001). Alu elements originate from the cellular 7SL RNA, which is part of the signal recognition particle. It appeared ~65 MYA, followed by a duplication and by a deletion of the central 7SL-specific

sequence (Ullu and Tschudi 1984; Ullu et al. 1982; Ullu and Weiner 1984; Quentin 1992a; 1992b; Jurka and Zuckerkandl 1991). Thereby, Alu elements consist of a bipartite structure, where the left and right monomers are highly similar, and are separated by a central A-rich region (Figure 2-2). The left monomer that is the 5' half of an Alu element, contains an RNA polymerase III promoter (A and B boxes) which is lacking in the right monomer (Chu et al. 1995). The right terminus of the right monomer consists a poly(A) tail of variable length, but lacks conventional RNA polymerase III termination signal. This allows polymerase III to bypass the signal and the transcript includes a unique flanking genomic sequence until an RNA polymerase III termination signal (a stretch of four to six consecutive thymidine) is encountered (Chu et al. 1995). Hence, each Alu mRNA is unique and varies in length. In respect to their age, Alu elements can be classified into three major subfamilies: AluJ, AluS and AluY (from oldest to youngest). The active Alu 'core elements' in the contemporary human genome is comprised of all elements from AluY subfamily and most of the elements from AluS subfamily (Bennett et al. 2008).

Alu recruits L1 proteins for retrotransposition. The Alu RNA folds into separate structure for each monomer. For efficient transposition, Alu RNA binds to the SRP9 and SRP14 signal recognition particle heterodimer (Sarrowa et al. 1997). Two determinants of Alu activity are, first, its primary ~280bp core sequence, and second, the ability of the Alu mRNA to form a ribonucleoprotein complex (RNP) with the SRP9/14 heterodimer (Bennett et al. 2008). Alu elements do not encode any protein and use the L1 retrotransposition machinery to integrate into the target sites, this is why sometimes they are called 'a parasite's parasite' (Weiner 2002). The poly(A) tail of the Alu RNP competes for the L1 ORF2p reverse transcriptase (Doucet et al. 2015b; Boeke 1997; Dewannieux et al. 2003; Mills et al. 2007; Sinnett et al. 1991) and the efficiency of transposition is dependent on the length of poly(A) (Dewannieux et al. 2003). The sequence PolyA is the site where the reverse transcription is initiated and is shared by all L1-mobilized template RNAs (L1, Alu, SVA and cellular mRNAs) (Doucet et al. 2015b; Kajikawa and Okada 2002; Esnault et al. 2000). The Alu poly(A) tail is an internal part of Alu sequences, whereas L1 poly(A) tail is added via the canonical polyadenylation pathway. Apart from the poly(A) tail, upstream

and downstream flanking sequences of Alu progenitor sequences also influence its transcription and transposition (Ullu and Weiner 1985; Comeaux et al. 2009).

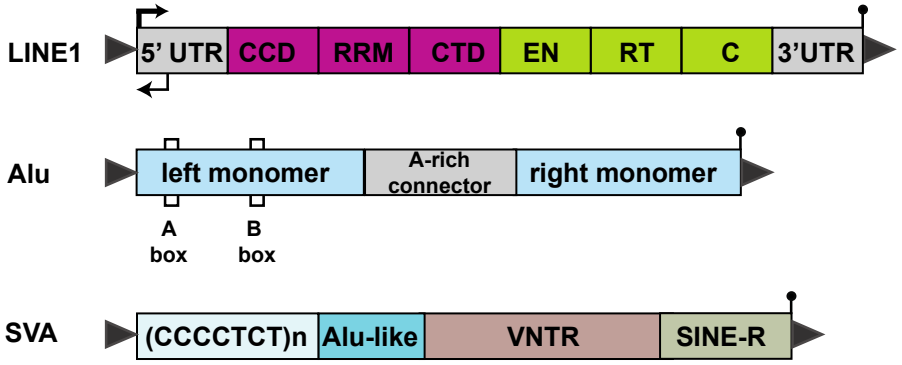

Figure 2-3. Structures of active human transposable elements.

LINE1, Alu and SVA elements are illustrated. L1 ORF1 domains are presented in magenta, L1 ORF2 domains are in light green, L1 untranslated regions are in light grey, target site duplications are in black arrowheads. UTR, untranslated region; CCD, coiled coil domain; RRM, RNA recognition motif; CTD, carboxy-terminal domain; EN, endonuclease domain, RT, reverse transcriptase domain; C, cysteine rich domain; VNTR, variable number of GC-rich **t**andem **r**epeats; lollipop, polyadenylation signal.

2.2.2. SVAs are composite non-coding sequences and show hallmarks of L1-mediated mobilization

SINE-VNTR-Alu (SVA) elements are compound repeat elements, i.e, they are composed of other repeats. They originated 25 MYA and comprise the youngest active family of mobile elements in humans. ~2700 copies SVA has been identified in the human genome which represents 0.2% of it (Wang et al. 2005). In general, an SVA element is ~2Kb and structured as follow, (5' to 3'): an array of hexameric tandem repeats (CCCTCT)n, two antisense Alu-like fragments, a variable number of GC-rich tandem repeats (VNTR), a SINE-R sequence sharing identity with the retroviral Env gene and the right LTR of HERV-K sequence, and a terminal polyA tail (Figure 2.2). The presence of a canonical poly(A) signal (AATAAA) at its end suggests that SVA transcription is RNA pol II mediated, although no internal RNA pol II promoter could be detected (Wang et al. 2005). SVA retrotransposition shows the

hallmarks of L1-mediated mobilization (Hancks et al. 2012; Raiz et al. 2012). However, some differences have been found between L1 and SVA retrotransposition. For example, transduction of 5' flanking sequences is more frequent for SVA elements (10%) as compared to L1s (Damert et al. 2009; Hancks et al. 2009).

2.3. L1-mediated processed retropseudogene formation contributes to human genome plasticity

Besides retrotransposon RNAs, L1 can also mobilize protein coding mRNAs (Esnault et al. 2000; Wei et al. 2001) and small nuclear RNAs, such as U6 (Doucet et al. 2015a). The integrated copies of the mobilized genes lack intron and promoter, and thereby are called processed pseudogenes. Like Alu and SVA elements, processed pseudogenes exhibit the regular hallmarks of L1-mediated TPRT mechanism. The human reference genome contains ~8,000 to 17,000 processed pseudogenes (Torrents et al. 2003), of which ribosomal protein genes are the most abundant (Zhang et al. 2002). Although most of the processed pseudogenes are non-functional due to the loss of regulatory sequences by 5' truncations and other rearrangements and the absence of promoter, some of them became functional and have provided new cellular function adding diversity to the genome. This has been demonstrated by the integration of a cyclophilinA pesudogene inside the TRIM5 gene in primates within the last 6My. Both of these genes are antiviral restriction factors and give protection against retroviruses through different mechanisms. Remarkably the resulting fusion protein is functional and provided new defense mechanism against exogenous viruses (Sayah et al. 2004; Malfavon-Borja et al. 2013).

3. L1 replicates by an RNA-mediated copy-and-paste mechanism

3.1. L1 is a 6kb DNA sequence

The first consensus sequence of human specific L1 element was derived from the alignment of 35 human L1 sequences by Scott et al in 1987 (Scott et al. 1987). This consensus sequence had ORFs similar to the L1 ORFs from other eukaryotic species (Scott et al. 1987). Within four years, Dombroski et al. succeeded to isolate an active full-length source L1 from chromosome 22 which was the progenitor of a truncated copy which inserted into the factor VIII gene on X chromosome, causing hemophilia A in a newborn (Dombroski et al. 1991). A prototype human L1 has a ~900 bp 5' untranslated region (UTR) with a weak promoter activity for RNA polymerase II (Swergold 1990), two open reading frames, ORF1 and ORF2 separated by a 63 bp spacer, a 3'UTR ending with a weak polyadenylation signal (Moran et al. 1999), and a long poly(A) tail of variable length (Figure 2-3). Recently, an additional ORF in the 5'UTR and in inverse orientation to L1, named ORF0, has been discovered. The encoded protein ORF0p, which is 70 amino acid long, slightly enhances L1 retrotranspostion in cultured cells if overexpressed in trans, by a mechanism yet to be revealed (Denli et al. 2015).

3.1.1. L1 5'UTR contains a bidirectional promoter

L1 is transcribed from its internal RNA pol II promoter located in the 5'UTR (Swergold 1990). The first 670 nt of the 5'UTR display promoter activity. Deletion analysis has shown that the the first 155 nt of 5'UTR contains the cis acting regulatory element essential for L1 transcription (Swergold 1990). An overlapping antisense promoter activity resides between 400-600nt of the 5'UTR and is responsible for the transcription of sequences upstream of L1 (Speek 2001; Nigumann et al. 2002). The antisense promoter is not essential for retrotransposition since the entire 5'UTR can be uploaded by a strong heterologous promoter in cell culture retrotransposition assays. The L1 sense promoter forms an initiator element with the upstream flanking genomic sequence which may influence the efficiency of L1 transcription (Lavie et al. 2004). Hence, L1 promoter strength, in part, is dependent on its integration site.

3.1.2. L1 elements contain 2 ORFs required for L1 mobility

The first L1 ORF is called ORF1, which is ~723 bp long and codes for ORF1p. ORF1p is a 241 amino acid protein (~40 kDa) (Scott et al. 1987) with nucleic acid binding (Kolosha and Martin 2003) and chaperone activities (Martin and Bushman 2001). ORF1p contains three major domains (Figure 2-3), an N-terminal poorly conserved coiled coil domain (CCD), followed by an RNA recognition domain (RRM), and a well conserved carboxy-terminal domain (CTD) (Figure 2-3). Through the interaction of leucine zippers of the N-terminal coiled coil domain, ORF1p forms homotrimers able to bind nucleic acids in a sequence independent manner (Khazina et al. 2011; Basame et al. 2006; Martin et al. 2003; Callahan et al. 2012; Naufer et al. 2016). Mutations in the conserved motifs of either of these three domains limit or abolish L1 retrotransposition efficiency suggesting the importance of each conserved motifs in L1 retrotransposition (Kulpa and Moran 2005; Basame et al. 2006). ORF1p contains four critical phospho-acceptor residues, two serines in N-terminal domain and two threonines in the RRM domain. Mutations of these amino acids inhibit L1 retrotransposition but have no significant effect on the ability of ORF1p to anneal RNA in vitro (Cook et al. 2015).

The second L1 ORF is called ORF2, it is 3843bp long, and codes for ORF2p, a 149 kDa protein with three domains: an N-terminal apurinic/apyrimidic endonuclease (APE) like domain (Feng et al. 1996; Cost and Boeke 1998), a reverse transcriptase domain (RT) (Mathias et al. 1991), and a C-terminal cysteine rich domain of unclear function (Fanning and Singer 1987) (Figure 2-3). The EN and RT domain play critical role in L1 retrotransposition and will be discussed in depth in section 3.3.2. Mutations in the C-terminal domain interferes with RNP formation and limit L1 retrotransposition (Moran et al. 1996). However, the biochemical role of the C-domain in L1 retrotransposition remains poorly understood.

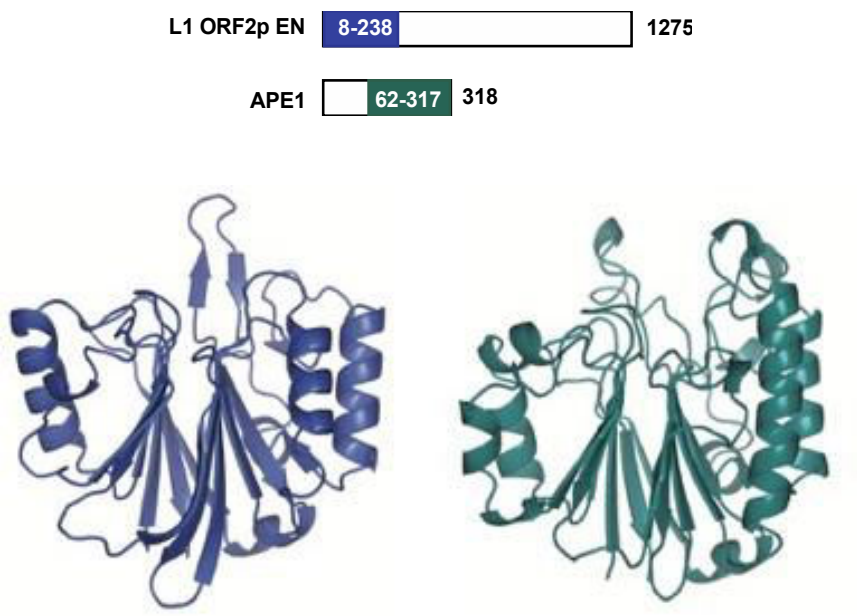

Figure 3-1. Crystal structure of the ORF2p endonuclease domain compared to the one of APE.

Bars represent full-length proteins containing phosphohydrolase domains at the colored positions. The respective structures are drawn as ribbon diagrams juxtaposed in the same orientation with the substrate binding surface on top. A common, central sandwich is surrounded by individual helices and surface loops. PDB accession codes for L1 and APE1 EN are 1vyb and 1dew respectively. From (Weichenrieder et al. 2004).

3.1.3. L1 ends with a weak polyadenylation signal

L1 has a weak transcription termination signal for RNA pol II (AATAAA) (Moran et al. 1999; 1996) in the ~200 bp of its 3'UTR. This signal is often bypassed by RNA pol II. In such cases, transcription continues until a downstream termination signal is found (Moran et al. 1999; Goodier et al. 2000). If such L1 transcripts containing non-L1 genomic sequence at their 3'end are used as template for reverse transcription, the newly generated L1 copy also carries this non-L1 sequence from the progenitor locus. Such events are called 3' transduction. Almost one third of somatic L1 retrotransposition events carry 3' transduced sequences (Tubio et al. 2014; Goodier et al. 2000). 3' transductions contribute to genomic

expansion and to shuffle protein-coding exons throughout the genome giving rise to gene duplications (Moran et al. 1999; Xing et al. 2006).

3.2. The L1 ribonucleoprotein particle represent the core of the L1 replication machinery

As few as approximately hundred L1 elements among the 500,00 in the human genome are full length and capable of retrotransposition when expressed from a plasmid with a strong promoter. Accumulation of mutations in the L1 sequence over time limits L1 ability to replicate. Thus, to persist in the genome, L1 elements must continue to replicate and expand in the genome to maintain a functional progeny.

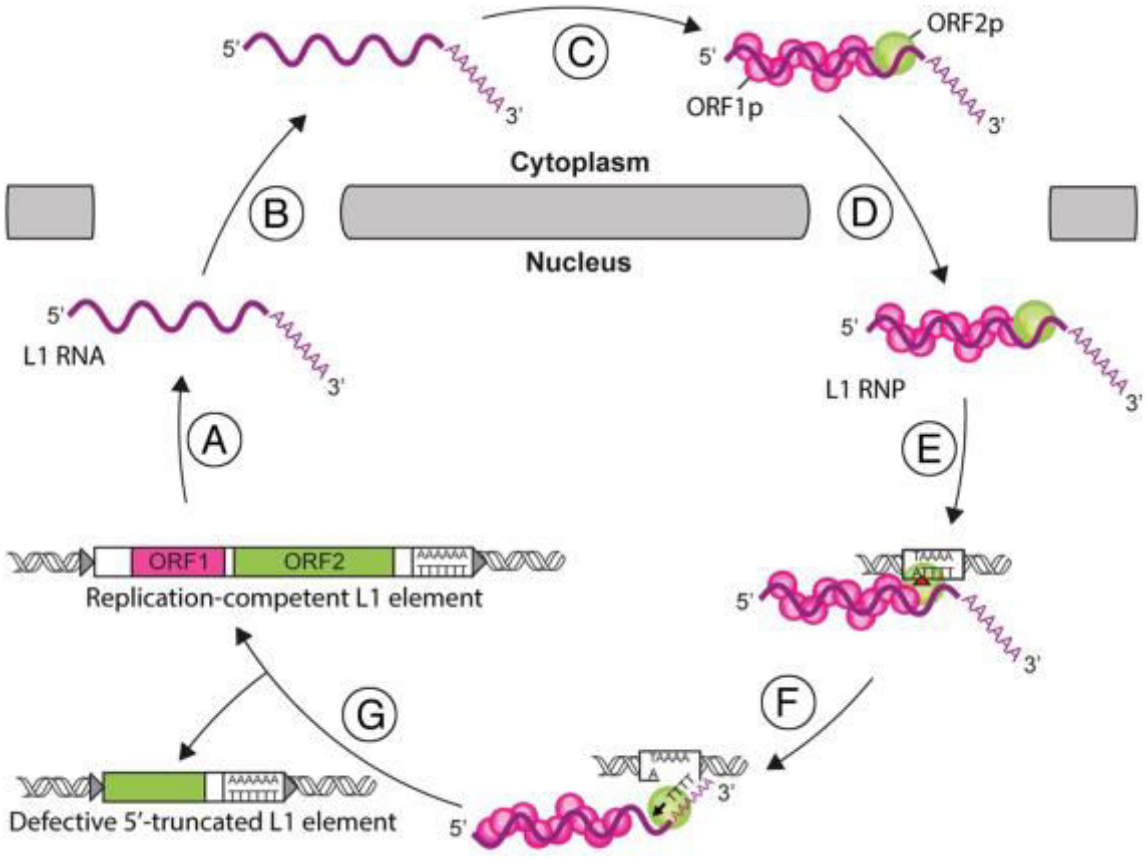

Figure 3-2. The L1 life cycle.

L1 is transcribed to a bicistronic L1 mRNA (A), which is exported to the cytoplasm (B), where ORF1p and ORF2p proteins are translated and bind to the L1 RNA to form L1 ribonucleoprotein particles (RNP) (C). The L1 RNP is imported into the nucleus (D), where

L1 ORF2p endonuclease (EN) activity nicks the first target DNA strand (red arrowhead, E) and reverse transcriptase (RT) initiates the reverse transcription of L1 RNA (black arrow, F). The final steps are not completely understood yet (G), L1 reverse transcription is often abortive and results in a 5' truncated L1 progenitor. Progenitor full-length L1 may continue to replicate if the site of integration is open for expression.

3.2.1. L1 transcription starts predominantly at +1nt position

L1 transcription was considered to be initiated at the first nt of the L1 element to produce a bicistronic L1 RNA (Swergold 1990; Minakami et al. 1992). However, a later study has shown that transcription initiation is not strictly restricted to the first nucleotide of L1 (Lavie et al. 2004). L1 transcription initiation site is variable. Transcription may even start from the flanking upstream nucleotides (Lavie et al. 2004). Several transcription factors are known to play role in LINE-1 transcription, for e.g., ying yang 1 (YY1) binds to nucleotide +13 to +21 of the L1 sequence (Minakami et al. 1992; Becker et al. 1993), SOX family transcription factors binds to two central regions within the L1 5'UTR (nt 472–477 and 572–577) (Tchénio et al. 2000), T-cell factor/lymphoid enhancer factor (TCF/LEF) binds to sequences overlapping with SOX (Kuwabara et al. 2009), p53 binds to multiple sites (Harris et al. 2009), and runt related transcription factor 3 (RUNX3) binds to nucleotides 83–101 (Yang et al. 2003). Protection of L1 upstream sequences in DNAse footprint experiments supports the notion that other transcription factors might also be involved in binding immediate L1 flanking upstream sequences and may influence L1 transcription (Mathias and Scott 1993).

3.2.2. L1 ORF1p is translated in a cap-dependent manner

L1 promoter directs synthesis of numerous copies of the ORF1p per L1 RNA but only one or two copy of ORF2p (Wei et al. 2001; Gilbert et al. 2002). The inter-ORF spacer contains two in-frame stop codons. Efficient translation of the first cistron of L1 RNA ensures the high ORF1p/ORF2p ratio needed for RNP formation and retrotransposition (Taylor et al. 2013; Dmitriev et al. 2007). Supposedly, if each ORF1p trimer coats 50nt of the L1 RNA (Basame et al. 2006), 120 trimers would be needed to coat the 6 kb RNA whereas possibly just one dimer of ORF2p is needed requiring a ~100-fold excess of ORF1p compared to ORF2p. ORF1 is translated in an efficient cap dependent manner and accounts for the high

number of the RNA binding ORF1p (Dmitriev et al. 2007). According to this model, a translation initiation complex which includes the 30S ribosomal subunit, binds at or near the 5' end of the capped L1 RNA and scans the L1 RNA for the presence of the AUG translation start codon. Translation elongation begins by the joining of 60S ribosomal subunit once the start codon is found.

3.2.3. L1 ORF2p is translated by an unconventional termination/re-initiation mechanism

ORF2p is translated from the bicistronic L1 mRNA in an unconventional termination/re-initiation mechanism. Antisera reactive to native ORF1p or to an epitope-tagged version of ORF1p identified only ~40 kDa ORF1p, and no other accompanying proteins. This suggests that ORF2p translation is initiated separately and is not synthesized as a fusion protein of ORF1p (Leibold et al. 1990; McMillan and Singer 1993; Goodier et al. 2004; Kulpa and Moran 2005). Besides it was found that a stop codon between ORF1 and ORF2 is required for retrotransposition (Alisch et al. 2006). In vitro translation study led to the hypothesis that an internal ribosome entry sequence (IRES) in the L1 inter-ORF spacer is required for human ORF2p translation (McMillan and Singer 1993). However, deletion analysis of either the 3' end of ORF1 or the inter-ORF spacer of L1 vector did not L1 retrotransposition significantly in cell culture based assays (Alisch et al. 2006). Hence, ORF2p translation is not dependent on either ORF1p 3'end nor on the inter-ORF spacer. Rather, ORF2p translation was found to initiate from the first in-frame AUG codon of ORF2 although replacing the AUG codon with any other coding triplets did not hamper ORF2 translation and L1 retrotransposition implying an AUG-independent translation of ORF2p (Alisch et al. 2006; Dmitriev et al. 2007; Li et al. 2001). A stop codon between ORF1 and ORF2 was required for L1 retrotransposition, which means that the two ORF proteins need to be translated separately. However, introducing a premature termination codon in ORF1 or a thermostable hairpin in the inter-ORF spacer to block ribosome scanning reduced ORF2p translation and L1 retrotransposition (20 to 50 fold). Together these results suggest that ORF2p translation occurs by an unconventional termination/re-initiation mechanism where a translating ribosome from the upstream ORF is needed to scan through the spacer

and find a cis-acting sequence in the 5'end of ORF2 which would position the ribosome at or near the ORF2 AUG initiation codon (McMillan and Singer 1993; Alisch et al. 2006; Dmitriev et al. 2007). This explains the reduction in the transposition efficiency by a premature stop codon in ORF1 or by the hairpin block in the spacer.

3.2.4. ORF1p and ORF2p assemble in a ribonucleoprotein particle with the L1 RNA in cytoplasmic foci

L1 RNA associates with its encoded proteins, several ORF1p homotrimers and at least 2 ORF2p (ORF2p dimer) if L1 follows the same model as R2 (Christensen and Eickbush 2005) to form a ribonucleoprotein particle (L1 RNP) in the cytoplasm (Kulpa and Moran 2006; 2005; Martin 1991) (Figure 3-2C). It is suggested that ORF1p polymerizes at the site of translation, which facilitates their binding to their own RNA to form RNP complex, a phenomenon known as cis preference (see section 2.1.2) (Callahan et al. 2012; Furano 2000). Because ORF2p is present in very low quantities, it was difficult to physically detect and characterize how ORF2p is associated with the L1 RNA. However, lately it was possible to detect both of the proteins associated with L1 RNA using an epitope/RNA tagging strategy (Doucet et al. 2010). L1 RNA and proteins were found to accumulate in cytoplasmic foci, which often colocalize with stress granules (Doucet et al. 2010; Goodier et al. 2007). It has been mentioned previously in section 2.2 that L1 proteins can mobilize other kinds of cellular RNAs although it is not well understood how the RNP is formed in trans. ORF2p was found to preferentially associate with 3' poly(A) tracts in L1 and Alu RNAs (Doucet et al. 2015b). Replacing the polyA signal of L1 at the 3'end with a stabilizing triple helix derived from MALAT1 non-coding RNA blocks L1 retrotransposition although transcription and translation were not hampered (Doucet et al. 2015b). Addition of a polyA at the 3' end of the chimeric transcript restored L1's ability to retrotranspose (Doucet et al. 2015b).

3.2.5. L1 ribonucleoprotein particles enter the nucleus by an unknown mechanism

To integrate new L1 copies in the genome, L1 RNP must enter into the nucleus (Figure 3-2D). L1 RNP with a number of ORF1p trimers, where each ORF1p is 40KDa and with two 150KDa ORF2p, possibly do not diffuse into the nucleus passively. Other options for nuclear

entry is either energy dependent active transport through the NPC or to reach chromosomes during cell division when the nuclear membrane is disrupted (Görlich and Kutay 1999). From cell culture based retrotransposition assay in growth arrested cells, using an L1 construct with mouse phosphoglycerate kinase-1 promoter, Kubo et al. suggested that L1 retrotransposition can occur in non-dividing cells (Kubo et al. 2006). Another study used a codon-optimized hyperactive mouse L1 with tetracycline inducible promoter and showed that retrotranspositon is slightly more efficient in dividing cells than the non-dividing ones (Xie et al. 2013). Inducing L1 expression for the same amount of time, they found 2.6-fold higher retrotransposition in synchronized cells undergoing two mitoses than those undergoing one mitosis (Xie et al. 2013). These two studies are in agreement and suggest that L1 retrotransposition can occur independently of mitotic nuclear envelope breakdown. But the mechanism of nuclear import of the L1 RNP has yet to be revealed.

3.3. L1 DNA synthesis occurs at the genomic target site

3.3.1. L1 predominantly integrates in genomic sites cleaved by the ORF2p Endonuclease

Most L1 integration takes place via the classical endonuclease dependent target-primed reverse transcription (TPRT) (Luan et al. 1993; Cost et al. 2002). The most detailed model on TPRT derives from studies of the R2 element in Bombyx mori and Drosophila melanogaster. Although L1 and R2 share many similarities, they have two major difference. First, unlike L1 which has two ORFs, R2 encodes only one ORF displaying both EN (Xiong and Eickbush 1988; Luan et al. 1993) and RT activities. Second, R2 EN contains an RLE domain while L1 EN contains an APE domain . In an alternative pathway, L1 can integrate at pre-existing DNA lesions, and does not require any endonuclease cleavage. This is called the endonuclease-independent (ENi) retrotransposition or non-classical L1 insertion (NCLI) (Morrish et al. 2002; Sen et al. 2007).

3.3.1.A. L1 endonuclease recognizes and nicks at degenerate 5'-TTTT/A-3' sequence motif

L1 EN domain resembles the metal dependent Apurinic/apyrimidinic endonuclease domain (APE) (Figure 3-1). APE is a component of the basic excision repair pathway containing 3' exonuclease, 3' phosphatase, and an RNase H activities originating from a single active site (Barzilay and Hickson 1995). The L1 EN nicks defined consensus sequences at the genomic DNA target (5'-TTTT/A-3'; slash indicates the scissile phosphate), which liberates a 5' phosphate and 3' hydroxyl group (Feng et al. 1996; Jurka 1997; Morrish et al. 2002; Cost and Boeke 1998) (Figure 3-2E). The liberated 3' hydroxyl group is used as a primer by the ORF2p RT activity to initiate reverse transcription of the L1 RNA. Thus EN-mediated nicking of the target site is coupled to reverse transcription of the L1 RNA template. Variations of this consensus motif are often observed, although a number of pyrimidine before the scissile bond followed by purines are almost always observed (5'-(Y)n/(R)n-3') (variation of nicked sites are archived in (Hancks and Kazazian 2016)). Initial crystallographic studies proposed that L1 EN recognizes an extra helical "flipped" adenine residue located 3' of the scissile bond to mediate cleavage (Weichenrieder et al. 2004). Bendability of the pyrimidine/purine dinucleotide is known to facilitate the integration of DNA transposons and retroviruses (Pruss et al. 1994a; 1994b; Serrao et al. 2015; Maertens et al. 2010). Hallmarks of endonuclease-dependent integration include insertion at a consensus L1 endonuclease recognition motif, a target-site duplication flanking the insertion and ranging from 4 to 20bp in length, and always the polyA tail of varying length. Both endonuclease-dependent and independent integration share the occurrence of genomic rearrangements (see section 4.1), such as, 5' truncations, internal rearrangements, inversions and transductions.

3.3.1.B. Endonuclease-independent retrotransposition represent and alternative mobilization pathway

The non-classical mechanism for L1 integration is independent of EN-mediated cleavage (ENi retrotransposition pathway). ORF2p RT can start reverse transcription from the free 3' OH of pre-existing DNA lesions or of dysfunctional telomeres (Morrish et al. 2002; Sen et

al. 2007). Evidence-0 of ENi retrotransposition were only found in cells defective for p53 and non-homologous end-joining (NHEJ) DNA repair pathways but the reason for this dependency is unclear (Morrish et al. 2002; Coufal et al. 2011). ENi retrotransposition has been proposed to act as an ancestral mechanism of RNA-mediated DNA repair associated with non-LTR retrotransposons. This repair mechanism may have been used by the genome before the non-LTR retrotransposons acquired the endonuclease domain and is also reminiscent of telomerase-mediated telomere extension (Garcia-Perez et al. 2007). Eni retrotransposition events are characterized by the absence of TSD and frequent 3'truncations (and thus no polyA). Target site deletions are also frequently found. Interestingly, existing L1 copies with hallmarks ENi L1 insertions were found to be comparatively slightly enriched in gene rich regions as compared to EN dependent events (Sen et al. 2007). It has been suggested that repairing genomic lesions in gene rich regions may provide with selective advantage to ENi insertions compared to the classical L1 insertions (Sen et al. 2007).

3.3.2. L1 first strand cDNA synthesis is directly initiated at the endonuclease cleavage site

Reverse transcription of the L1 RNA starts following the recognition and nicking of one of the two strands of the target DNA. L1 RT shares sequence similarity to the RT domains encoded by telomerase, group II introns, and other classes of retroelements (Xiong and Eickbush 1990; Malik et al. 1999). Despite the similarity in homologs, non-LTR reverse transcriptases function very differently from the LTR-retroelement reverse transcriptases. The latter reverse transcribe their RNA templates in the cytoplasm within viral like particles using host tRNAs to prime reverse transcription. Upon completion of reverse transcription, dsDNA associated with the integration machinery is transported to the nucleus where integration can take place. In contrast, L1 RT works on its RNA template at the site of integration, using the genomic 3'hydroxyl liberated by EN cleavage to prime reverse transcription (Luan et al. 1993). L1 RT displays both RNA-dependent and DNA-dependent polymerase activities (Piskareva et al. 2003). Reverse transcription of the L1 RNA starts at the polyA tail of L1 RNA (Kulpa and Moran 2006; Doucet et al. 2015b; Monot et al. 2013) (Figure 3-2F).

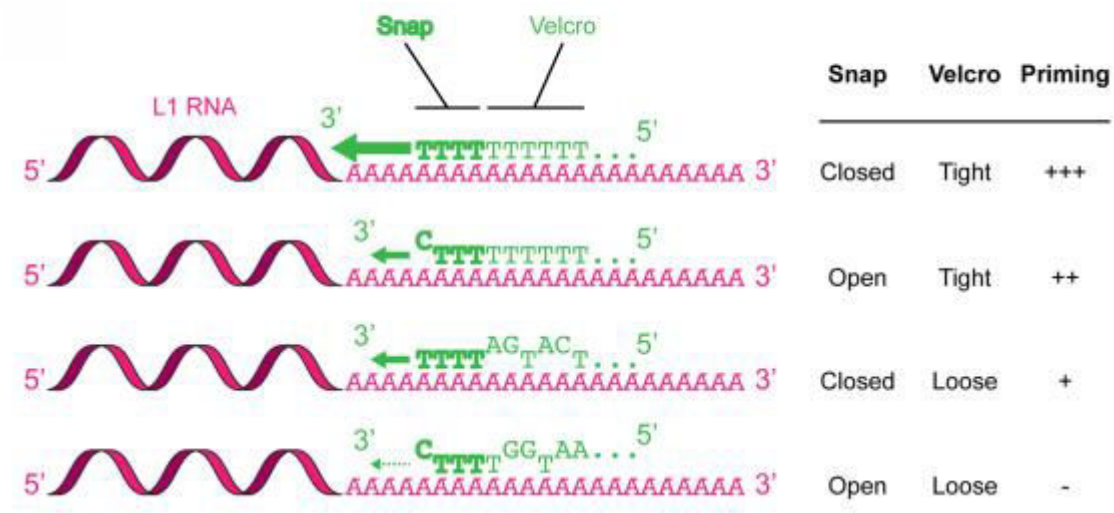

Figure 3-3. The snap-velcro model.

Reverse transcription priming requires base-pairing between the L1 RNA (pink) polyA tail and the target-site DNA (green). The snap (bold green) corresponds to the last 4 nucleotides at the 3' of the target DNA. The velcro (light green) contains the 6 bases upstream of the snap. The snap is considered as closed if 4 nucleotides are T. The velcro is tightly fastened if it is densed with T. The snap-velcro status predicts the efficiency of L1 reverse transcription priming in vitro (green arrow). Efficiency of priming is denoted with '+'. Reverse transcription is most efficient when snap is closed and Velcro is fastened. From (Viollet et al. 2014).

3.3.3. Integration site flanks contribute to the priming efficiency of reverse transcriptase

A major difference between R2 and L1 TPRT mechanism is that R2 does not require any complementarity between the target DNA and the R2 RNA while complementarity between the target DNA and the L1 RNA polyA tail promotes efficient priming (Monot et al. 2013; Luan and Eickbush 1996; 1995). Besides the presence of the recognition motif, base composition of their L1 flanking sequences and their chromatin status contribute to the efficient priming of reverse transcription (Cost et al. 2001; Monot et al. 2013). Lately, a model named snap-velcro has been proposed to illustrate the correlation between target DNA-L1 RNA complementarity and priming efficiency (Figure 3-3). According to this model, polyT tract (Velcro) downstream of the EN cleavage site can compensate for mismatches close to the priming site (snap). Requirement of polyA annealing to target site might stabilize the L1 reverse transcription complex to promote initiation of reverse transcription

(Monot et al. 2013). It is somehow striking that both the EN and RT have coevolved a preference for T-rich tracts. Consequently, local target site sequence preference is determined by the specificities of both enzymatic activities (Repanas et al. 2007; Monot et al. 2013).

3.3.4. The second strand cDNA synthesis starts from a nick on the second strand typically within 4 to 20bp from the first strand nick

For R2 element, cleavage of the second strand DNA takes place following initiation of first-strand cDNA synthesis and is mediated by R2 endonuclease activity (Christensen and Eickbush 2005). The genomic rearrangements of the L1 integration sites, namely 5'-inversions and target-site deletions, suggested that similar to the R2 TPRT mechanism, in L1 TPRT, the second strand cleavage occurs following initiation of first-strand cDNA synthesis (Hancks and Kazazian 2016). The second strand cleavage sites do not show any sequence preferences unlike the first strand (Jurka 1997; Cost et al. 2002). However, the second strand cleavage position may be influenced by the distance from the first strand cleavage position as target site duplications generally range from 4-20 bp (Lander et al. 2001; Hancks and Kazazian 2016; Gilbert et al. 2005). The length and the sequence of the TSDs created during L1 retrotransposition is determined by the distance between the first and the second nick. While the activity responsible for the second strand nick is not confirmed, for L1 it is assumed that L1 EN or an additional nuclease activity, which has been observed in in vitro L1 RNPs might be involved (Cost et al. 2002; Kopera et al. 2011). In vitro, ORF2p shows DNA-dependent DNA polymerase activity, which could participate to second strand cDNA synthesis but other cellular DNA polymerase activities cannot be excluded (Piskareva and Schmatchenko 2006). How insertion is resolved (ligation) is unknown, presumably achieved by cellular activities.

L1 DNA synthesis occurs at the genomic target site

4. L1 contributes to genome evolution and may cause disease

L1 Integration results in a variety of rearrangement of the genomic DNA at the target site. Most rearrangements have no immediate effect, if they are distant from genes and regulatory elements. However, sometimes, such rearrangements may have positive or negative impact on the genome. Accumulations of L1-mediated genomic alterations diversify the genome, and contributes to its genome evolution. On the other side, a particular retrotransposition event may result in a genetic disorder due to the disruption of DNA sequences necessary for cellular functions. The consequences of L1 presence in our genome will be detailed in this section.

4.1. L1-mediated genomic rearrangements shape genome architecture

Approximately 0.3% of all human mutations are attributable to L1-mediated *de novo* retrotransposition events (Callinan and Batzer 2006). Although frequency might appear limited, L1 insertions can have much more consequences them point mutation. L1 causes target site alterations in a range of ways. The extent of the effect due to DNA rearrangements depends on the genomic features around the integration sites. Some may have no visible effect; others may result in genetic disorders. Here, target site alteration will be discussed in two section, first, how L1 causes local genomic instability, and second how L1 affect human transcriptome.

4.1.1. L1-mediated genomic rearrangements can destabilize our genome

4.1.1.A. Non-homologous recombinations between L1 copies cause deletions and inversions of genomic segments

Both the abundance and the activity of non-LTR retrotransposons have affected human genome evolution (reviewed in (Cordaux and Batzer 2009)). Regardless the inability of most L1 copies in the genome to replicate, their high density impacts the genome through a variety of rearrangements caused by ectopic recombination between non-allelic homologous copies (Figure 4-1A). Such recombination events between two L1 elements

may result in deletions (Burwinkel and Kilimann 1998) and inversions (Lee et al. 2008) of intervening genomic sequences. Recombination-mediated deletions are generated via homologous recombination of two retrotransposon sequences in the same orientation on the same chromosome, while crossing over between two retrotransposon sequences inverted relative to each other may result in an inversion (Cordaux and Batzer 2009; Deininger and Batzer 1999). Since the divergence of human and chimpanzee genomes, L1 and Alu mediated recombinations caused one fifth of the total inversions (Lee et al. 2008). L1 recombination-mediated segmental duplication was observed in the mouse genome (Janoušek et al. 2013) but no such evidence has yet been found in humans. It has been proposed that recombination between Alu elements might represent an important mechanism for the origin and expansion of segmental duplications in the human genome (Bailey et al. 2003). In general, recombination-mediated rearrangements are more frequent for Alu elements compared to L1 due to their very high density. More than seventy cases of Alu recombination-mediated deletions responsible for various cancers and genetic disorders have been reported (Deininger and Batzer 1999; Callinan and Batzer 2006) while only three such cases are known for L1 (Han et al. 2008). Compared to Alu, L1 recombination-mediated deletions are larger and are seen more frequently in gene poor regions, which suggest that L1 mediated long deleterious deletions are prone to negative selection in human (Song 2007). Together, such deletions have removed nearly 1 Mb of genomic sequence from the human genome over the past few million years (Han et al. 2008; Cordaux 2008).

4.1.1.B. L1 destabilize genome by DNA double-strand breaks (DSBs) and target site deletion.

L1 mediates genomic instability by EN-mediated DNA breaks across the genome and integration-mediated deletions at integration sites (Figure 4-1C and Figure 4-1D). It has been found in cell culture-based assay that the number of EN-mediated double strand breaks (DSBs) in the genome are more frequent than actually used for L1-mediated insertions (Gasior et al. 2006). DSBs are highly mutagenic and prone to recombination and recombination-mediated deletions. Retrotransposition-independent DSB have been found on L1 body itself since it contains sequence motif recognized by L1. The increased

expression of L1 endonuclease during neural differentiation induces dsDNA breaks preferentially at L1 loci associated with deletion of proximal genomic regions (Erwin et al. 2016). Gamma-H2AX foci accumulates at the sites of DSBs, which is associated with an abnormal cell cycle progression through a G2/M accumulation and induction of apoptosis. Such cases have been evidenced in cancerous (Belgnaoui et al. 2006) and aging cells (Erwin et al. 2014). The second process leading to deletions comes directly from L1-mediated insertions. Target site deletions of variable sizes originates from the variable position of second strand cleavage and subsequent processing of double strand breaks by a 5'-3' exonuclease activity of unknown origin (Gilbert et al. 2002). In cell culture-based assay, this phenomenon can lead to deletions of a few base pairs to as long as 71kb (Gilbert et al. 2002).

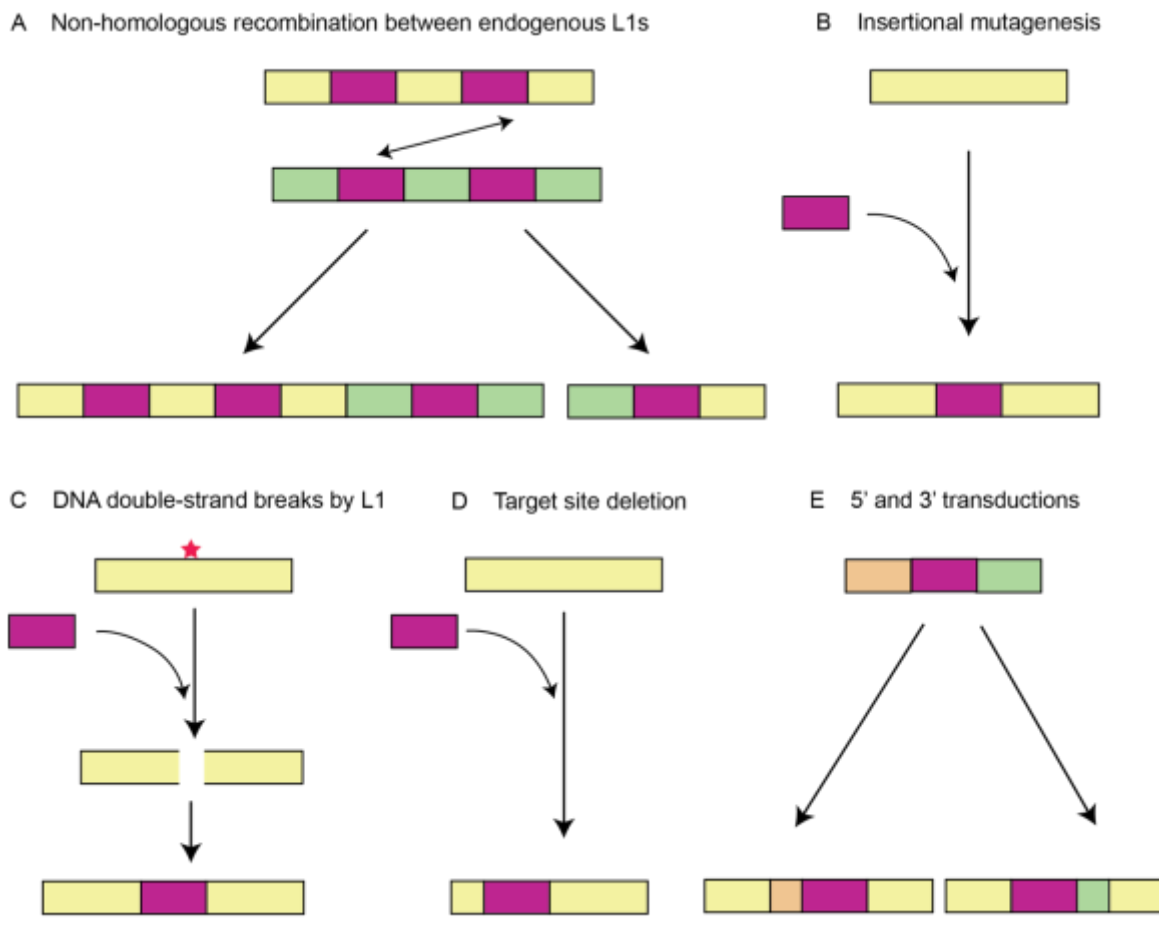

Figure 4-1. Impacts of L1 on human genome structure. Legend continued on next page.

(A) Ectopic recombination (double arrowhead) between non-allelic homologous retrotransposons may result in genomic rearrangements, such as deletions (left) or duplications (right) of intervening genomic sequences. (B) Typical insertion of a LINE-1 (L1), Alu or SVA retrotransposon (red box) at a new genomic site (dark grey). If the new genomic site is a genic region, the retrotransposon may cause insertional mutagenesis. (C) The protein product (green oval) of an L1 element may create DNA double-strand breaks (broken dark grey area). Alternatively, an existing double-strand break may be repaired by non-classical endonuclease-independent insertion of a retrotransposon. (D) The insertion of a retrotransposon is sometimes associated with the concomitant deletion of a target genomic sequence (light grey box). (E) During the duplication of a retrotransposon, the downstream 3′ flanking sequence or the upstream 5′ flanking sequence (dark grey boxes) may also be duplicated (known as 3′ or 5′ transduction, respectively). This results in the retrotransposition of the 3′ flanking sequence (left) or the 5′ flanking sequence (right) along with the retrotransposon.

4.1.2. L1 contributes to variations of the human transcriptome and proteome

4.1.2.A. Transcriptomic variation originates from the composition of L1 and its flanking genomic sequences

L1-mediated variations of the transcriptome may take place in a number of ways, collectively decided by the composition of L1 body and its flanking genomic sequences. For example, the most immediate phenotypic impact is visible when the transcriptome is affected by insertions in coding or regulatory sequences (Kazazian et al. 1988). L1 insertions in genic regions in antisense orientation can cause gene breakage producing two smaller transcripts: the first one contains the upstream exon and terminates in the major polyadenylation site of L1, the second one is transcribed from the L1 antisense promoter and includes the downstream exons (Wheelan et al. 2005). Variations of sequences composition greatly influences its mutagenic effect. For example, regulatory sequences within the L1 body or in 3′ L1 transduced sequences has the potential to elevate or repress expression of upstream and downstream genes. The L1 antisense promoter in 5′UTR may drive the transcription of the 5′ flanking genomic sequence giving rise to ectopic non-coding RNAs (Criscione et al. 2016; Speek 2001). Alternative transcription initiation by the L1 antisense promoter has also been evidenced in different studies and can alter tissue-specific gene expression, which increases the transcriptional flexibility of several human genes (Mätlik et al. 2006). Thus, antisense transcripts can lead to the production of non-

coding RNA (ncRNA) or chimeric transcripts. Such antisense RNAs could reduce mRNA levels through the formation of double-stranded RNA (dsRNA), triggering protein kinase R (PKR) degradation pathway (Heinicke et al. 2009), or leading to siRNA that induces silencing via the RNA-induced silencing complex, RISC (Yang and Kazazian 2006).

Another type of regulatory elements in L1 sequences are the canonical and non-canonical internal polyadenylation signals in both sense and antisense orientations (Perepelitsa-Belancio and Deininger 2003). These signals minimize full length L1mRNA transcripts accumulation. However, their presence in the body of L1 copies integrated in genes can lead to alternative mRNA transcripts or premature termination, thus affecting mRNA splicing and stability (Han et al. 2004).

Besides the direct influence of L1 sequence, epigenetic changes of L1 elements may also influence the expression of surrounding sequences through changes in their chromatin status. Hypomethylation of the L1 promoter is known to activate alternate transcripts leading to pathological conditions (Wolff et al. 2010).

4.1.2.B. L1 can mediate genetic innovation

New genes are continuously generated over evolutionary time. Re-arrangements between pre-existing genic structures is the major source of genetic innovation (Long et al. 2003). L1 contribute to the rise of new genes by three known mechanisms: formation of pseudogene (discussed in section 2.3), 5' and 3' transduction (see section 3.1.1 and 3.1.3), and exonization (discussed below).

- Gene duplication

L1-mediated retrotransposition of cellular mRNAs and small RNAs gives rise to a copy of the encoding. The new gene, which is called a retropseudogene, lacks regulatory sequences required for its expression. It can nevertheless be expressed if it acquires regulatory sequences or from regulatory sequences nearby the site of integration. An example has been described in section 2.3. L1 mediated retrotransposition events are responsible for emerging new genes in primates (Babushok et al. 2007; Sayah et al. 2004; Kaessmann et al.

2009). It has been estimated that altogether at least one novel gene has emerged every million years in the human lineage over the past ~65 Myr (Marques et al. 2005) .

- Transduction

The flanks of a progenitor L1 sequence may contain exons or regulatory sequences. During L1 transcription, this flanking non-L1 sequences may also be transcribed due to an upstream promoter or to the weak transcription termination signal of L1 (termed 5' and 3' transductions respectively, see section 3.1.1 and section 3.1.3) (Figure 4-1E). When such extended L1 transcripts are used by the retrotransposition machinery, the flanking genic or regulatory sequences can be copied to new genomic locations, thereby giving rise to new gene isoforms by exon or regulatory sequence shuffling (Moran et al. 1999; Tubio et al. 2014) or creating new genes by integration of regulatory sequences. L1-mediated transductions took place during human genome evolution and that it may account for 0.6–1% of total human DNA (Lander et al. 2001; Goodier et al. 2000; Pickeral et al. 2000). A recent analysis of SVA retrotransposons, which are mobilized by L1, has demonstrated the evolutionary significance of retrotransposon-mediated transductions by showing that this process is responsible for the creation of the acyl-malonyl condensing enzyme 1 (AMAC1) gene family, which has four members in the human genome (Xing et al. 2006). The ancestral AMAC1L3 gene copy at the source locus consisted of two exons separated by an intron. By contrast, the three transduced copies of AMAC1L3 (AMAC1, AMAC1L1 and AMAC1L2) were intronless as a result of the splicing of the intron during the retrotransposition process.

- Exonization

Exonization is the creation of a new exon from intronic sequences. L1 mediated insertions in introns may exonize part of the intron by transcribing it from one of the L1 promoters giving rise to new transcripts (Wheelan et al. 2005). Besides, both L1 and alu contains a lot of cryptic donor and splice sites. L1 contains numerous functional splice donor and acceptor sites in both sense and antisense (AS) orientations though most of them are weak (Belancio et al. 2006). A typical Alu sequence contains 9 GT dinucleotides and 14 AG dinucleotides that represent the same number of cryptic donor and acceptor splice sites,

respectively (Sorek et al. 2002; Lev-Maor et al. 2003). L1-mediated insertions of functional splice sites in intronic sequences may disrupt normal gene expression or forms alternative mRNA transcripts (Belancio et al. 2006; Sorek 2007). L1-mediated indirect exonization by Alu elements are more frequent than L1-mediated exonization and occurred consistently during primate evolution (Krull et al. 2005).

4.2. L1-mediated genomic rearrangements occasionally result in disease

4.2.1. Genetic diseases

Genomic rearrangements caused by L1 may affect the transcriptome and the proteome by a variety of mechanisms (see section 4.1), which occasionally leads to novel genetic diseases. 124 L1-mediated insertions have been reported to cause genetic diseases, such as cases of cystic fibrosis, muscular dystrophy, hemophilia, autoimmune diseases, and neurofibromatosis (Hancks and Kazazian 2016). Among the disease-causing insertions, 29 are L1 insertions, 77 are L1-mediated Alu retrotransposition, 13 are L1-mediated SVA retrotransposition, and 1 is an L1-mediated retrotransposition of CYBB gene (reviewed in (Hancks and Kazazian 2016; 2012)). The first report of L1-mediated disease came in 1987 from the Kazazian lab, demonstrating that L1s are still actively replicating in human somatic cells. Most of the 124 disease-causing insertions reported to date inactivate gene function through insertional mutagenesis or aberrant splicing (Hancks and Kazazian 2012; Chen et al. 2005; Belancio et al. 2008a; Kagawa et al. 2015).

4.2.2. Somatic L1 retrotransposition contribute to cancer genome mutagenesis load and can act as drivers of tumorigenesis.

Half of all human epithelial cancers have been found to re-express the L1 machinery (Rodic et al. 2014). Genome-wide sequencing studies have detected extensive somatic insertions in various epithelial carcinomas including colon, pancreas, esophagus, uterus, head and neck, liver, lung, gastrointestinal tract, ovary and prostate (Ewing et al. 2015; Helman et al. 2014; Tubio et al. 2014; Makohon-Moore et al. 2015; Solyom et al. 2012; Shukla et al. 2013; mechanisms (see section 4.1Iskow et al. 2010). While clear driver L1 insertions into –or

nearby–genes, which inactivate tumor suppressor genes or activate oncogenes, provide selective advantage and promote tumor growth (Helman et al. 2014; Shukla et al. 2013; Miki et al. 1992; Scott et al. 2016; Doucet-O'Hare et al. 2015), others have no defined impact, and might be passenger events. They might also contribute to tumor genome plasticity by shuffling genomic features through transductions of flanking genomic sequences or by pseudogene formation {Tubio:2014gm, Cooke:2014ib}. Besides cancerous and metastatic tissues, the observation of somatic L1 insertions in precancerous lesions and sometimes in the adjacent normal tissue, but not in blood DNA, is consistent with direct involvement of L1 in the early stages of tumorigenesis (Ewing et al. 2015).

4.3. Different cellular pathways counteract L1-mediated mutagenesis

L1 can influence the genome in a number of ways (see section 4) and can have harmful consequences. As predicted by the Red Queen's evolutionary hypothesis (Van Valen 1973), range of defense mechanisms have continuously evolved to protect the genome against such deleterious events. Control of L1 retrotransposition takes place at both transcriptional and post-transcriptional levels through the participation of a number of nuclear and cytoplasmic host factors (Figure 4-2).

4.3.1. Epigenetic silencing

L1 expression is silenced through CpG DNA methylation and histone modifications (Castro-Diaz et al. 2014; Jacobs et al. 2014; Bestor and Bourc'his 2004). DNA methylation restricts binding of transcription factors to promoters, and also attracts methyl-CpG-binding proteins (MBDs), which is associated with histone deacetylases (HDACs) and other heterochromatin proteins to remodel chromatin. Thus heterochromatinization of the surrounding region can limit L1 transcription (Castro-Diaz et al. 2014; Jacobs et al. 2014). Alterations of DNA methylations are recognized as an important feature of tumorigenesis L1 hypomethylation is associated with different stages of tumorigenesis (Suter et al. 2004; Schulz et al. 2002).

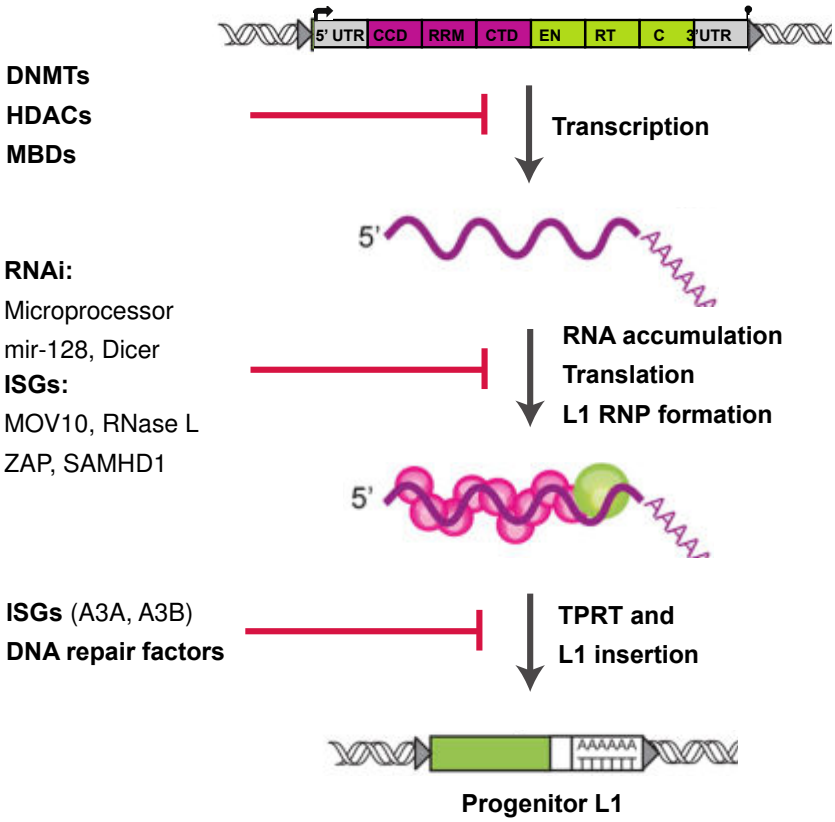

Figure 4-2. Cellular regulators limit L1 retrotransposition at different level.

DNMTs, DNA methyl transferases; HDAC, Histone deacetylases; MBD, methyl-CpG-binding domain proteins, mir, micro-RNA, RNAi, RNA interference; ISG, interferon-stimulated genes; MOV10, Moloney leukemia virus 10; SAMHD1, SAM domain and HD domain 1; ZAP, zinc-finger antiviral protein; RNaseL, ribonuclease L; A3A, APOBEC3A; A3B, APOBEC3B.

4.3.2. Post-transcriptional silencing

The abortive reverse transcription of L1 during the process of TPRT often produce 5' truncated progenitors, inactivating progenitor L1 copies to retrotranspose. Also, L1 contains multiple polyadenylation sites which limit the full length transcription of L1 (Perepelitsa-Belancio and Deininger 2003). Cryptic splice sites in L1 RNA transcripts induce a complex pattern of splicing that may remove portions of the ORFs or the 5'UTR (Belancio et al. 2006; 2008b).

Sequence specific post-transcriptional silencing of L1 is mediated by small RNAs. RNA-induced silencing through RNA interference has been suggested to reduce L1 retrotransposition in cultured cells (Soifer et al. 2005; Yang and Kazazian 2006). Piwi

proteins and Piwi-interacting RNAs (piRNA) silence L1 during genome reprogramming in the embryonic male germ line (De Fazio et al. 2011; Marchetto et al. 2013). Lately, Hamdorf et al. uncovered a new mechanism in which microRNA, miR-128 restrict L1 mobilization and L1-associated mutations in cancer cells, cancer-initiating cells and iPS cells by binding directly to L1 RNA {Hamdorf:2015ex}. Post-transcriptional regulation of L1 also involves interferon response pathways (reviewed in (Pizarro and Cristofari 2016)). A number of interferon-stimulated genes (ISGs), including APOBEC3, MOV10, BST-2, ISG20, MAVS, MX2, RNase L, SAMHD1, TREX1, and ZAP restrict L1 retrotransposition, indicating that ISGs are key players of the type I interferon anti-retroelement response (reviewed in (Ariumi 2016; Goodier 2016; Pizarro and Cristofari 2016))

5. Many transposable elements preferentially insert in specific genomic regions

The planet of mobile elements is highly diverse. Many of them show preferences for certain features or locations in the host genome. This is directly linked to their evolutionary strategies and has strong consequences for their use as biotechnological tools. The manuscript below, in preparation, reviews the experimental approaches and bottlenecks to study TE target site preferences and what we learnt from them. It also covers retroviruses since they are mechanistically and phylogenetically related to LTR-retrotransposons and endogenous retroviruses.

5.1. Integration site selection by retroviruses and transposable elements in eukaryotes

Integration site selection by retroviruses and transposable elements in eukaryotes

Tania Sultana[1, #], Alessia Zamborlini[2,3, #], Gael Cristofari[1,4, *], Pascale Lesage[2, *]

[1]IRCAN, INSERM U1081, Centre National de la Recherche Scientifique UMR 7284, Université Côte d'Azur, 06107 Nice Cedex 2, France

[2]Université Paris Diderot, Sorbonne Paris Cité, INSERM U944, CNRS UMR 7212, Institut Universitaire d'Hématologie, Hôpital St. Louis, 75010 Paris, France.

[3]Laboratoire PVM, Conservatoire nationale des arts et métiers (Cnam), 75010 Paris, France

[4]FHU OncoAge, Université Côte d'Azur, 06107 Nice, France

[#] Equal contribution

*Correspondence to PL or GC. e-mail: pascale.lesage@inserm.fr or gael.cristofari@unice.fr

Abstract | Transposable elements and retroviruses shape the genome of most organisms, can be pathogenic and are widely used as gene-delivery and functional genomics tools. Exploring whether these genetic elements have a target-site preference and how this is achieved at the molecular level is critical to our understanding of genome evolution, somatic genome plasticity in cancer and aging, host-parasite interactions and for many genome engineering applications. High-throughput profiling of integration sites by next-generation sequencing techniques, combined with large-scale genomic data mining, and cellular or biochemical approaches has revealed that insertions are most often non-

random. Rather, the DNA sequence and chromatin contexts, cellular proteins, and the 3D organization of the nucleus cooperate in guiding integration in eukaryotic genomes, leading to a remarkable diversity of insertion distribution and evolutionary strategies.

Introduction

Transposable elements (TEs) are present and active in nearly all organisms, including humans, and are widely used as genomic and gene-therapy tools. They comprise DNA transposons, LTR and non-LTR retrotransposons. Because retroviruses are mechanistically and evolutionary related to LTR-retrotransposons, from which they are sometimes undistinguishable, we will include them under the term of TEs for the purpose of this review. TEs have remarkably contributed to shape genome structure and function, with impact on the physiology and diseases of most - if not all - living organisms. Although generally less frequent than point mutations, TE insertions can have much more radical outcomes. Indeed, TEs carry and transplant multiple cis-regulatory sequences and therefore can considerably remodel gene structure and rewire gene networks in a very short evolutionary time frame. Such a phenomenon is illustrated by the discovery that the regulation of interferon-response genes, forming an essential antiviral pathway in vertebrates, has been rewired multiple times by endogenous retroviruses providing transcriptional enhancer functions [1].

The genomic distribution of a given TE or retrovirus results from a two-step process: first, site-specific (or not) integration directing the initial allocation of the insertions, and second, selective pressures leading to the loss of harmful events and perpetuation of insertions that benefit to the host. Somatic insertions might be subjected to additional selective mechanisms, such as cellular expansion (tumorigenesis) or elimination (immune system) [2]. Yet, TEs and retroviruses must continue replicating to avoid extinction. A TE remains active providing that: (i) it is full-length and does not contain any mutation that would hamper its replicative machinery; (ii) some of its integrated copies can be transcribed in a timely manner [3-5]; and (iii) its cellular environment is permissive. This duality is driving the coexistence of TEs with their host, and their co-evolutionary strategies.

TE integration site selection can define the spectrum of genetic outcomes resulting from its insertion and its potential pathogenicity, has profound consequences on the ability of individual copies to undergo additional rounds of replication, and underlies a variety of evolutionary strategies. Therefore, addressing where TEs integrate in the genome and

whether this process is random, is fundamental to understand the intricate relationship between TEs and their hosts. Here, we will review the molecular and cellular determinants guiding the integration of TEs in eukaryotic genomes, and how deep-sequencing techniques have radically changed our ability to address this question. We will limit our survey to TEs for which genome-wide *de novo* integration profiles or insights into integration molecular mechanisms are accessible and will highlight how a restricted number of related mechanisms can lead to a remarkable diversity of insertion distributions and evolutionary strategies, which can be exploited for functional genomics or gene therapy purposes. Many aspects of the genetic and epigenetic impact of TE insertions have been reviewed elsewhere (see for example [6-15]) and will not be covered in this review.

Genomic distribution vs integration site preference

Distribution in genomes is non-random. The distribution of TEs in eukaryotic genomes at the steady-state is non-random. Not only TEs accumulate in specific regions of genomes, but they also show species- and TE-specific patterns. These biases were observed in the early days of molecular genetics. Ty1 and Ty3, the prototype LTR-retrotransposons in yeast, were originally identified as responsible for frequent restriction fragment-length polymorphisms associated with tDNA 16-18. Similarly, R1 and R2 non-LTR retrotransposons, first identified in insects, are exclusively found at fixed, but distinct positions, within rDNA units 19,20. A hint that retroviruses might also have integration site preferences came from the use of Murine Leukemia Virus (MLV)-based vectors in the first clinical trial aiming to cure severe genetic immunodeficiencies. Indeed, few patients developed leukemia due to recurrent insertions of the retroviral vector near the promoter of the LMO2 proto-oncogene 21. Thus, some elements seem to be enriched in repetitive genomic regions, such as tDNA or rDNA, where their insertion is less likely to be detrimental. Inversely, other elements can give rise to recurrent mutagenic insertions and/or oncogenic clonal expansion. Although these examples suggest that some elements might preferentially integrate in specific genomic regions, they also indicate the importance of post-integration selective processes in the chromosomal distribution of TEs

and retroviruses observed at the steady-state. This will be discussed further in the light of TE-host coevolutionary strategies in the last section.

Study of de novo insertions is crucial. To investigate the mechanism of integration preference, it is critical to limit the effects of post-integration selective phenomena, and therefore to locate novel insertions as early as possible after integration, i.e. *de novo* insertions. Such studies generally require the availability of mobilization-competent or infectious molecular clones that can be used to experimentally induce transposition or infection. Performing such experiments in a short term and *ex vivo* limits potential biases linked to mutagenic effects on cell growth and avoid selection by the immune system. These effects are evidenced upon comparison of the genomic distribution of *de novo vs* fixed insertions, or recent *vs* older insertions [22]. A striking example comes from HERV-K an inactive endogenous retrovirus unable to produce infectious particles in modern humans. Resurrection of an infectious clone by recombining several defective copies [23,24] allowed the comparison of *de novo* and fixed insertion patterns [25]. *De novo* insertions were slightly enriched in transcriptional units, in gene-rich regions, and near active histone marks. In contrast, fixed elements are depleted from transcription units. Consistent with a progressive counter-selection of HERV-K genic insertions, the youngest endogenous copies show an intermediate distribution. In addition, fixed elements have more frequently an antisense orientation relative to genes, while *de novo* insertions hit genes indistinctly in sense or antisense orientation [25]. The importance of studying de novo insertions is also illustrated by the distinct distribution of LINE-1 (L1) and Alu sequences in the human genome. Fixed endogenous L1 and Alu insertions are enriched in opposite DNA isochores: L1 elements show a bias for AT-rich regions whereas Alu sequences are enriched in GC-rich regions 26. Of note, Alu elements are non-coding and are mobilized in trans by the L1 retrotransposition machinery, which could underlie a common site integration preference. Consistently, and in contrast to fixed copies, experimentally induced de novo Alu insertions are detected in the same AT-rich isochore as L1 elements 27.

Mapping integration sites

Technical bottlenecks. Mapping a large number of *de novo* TE insertion sites in a cellular population is technically challenging. Endogenous and novel copies are virtually undistinguishable. In addition, while endogenous copies might be present in all cells of the population, each new insertion is only present in one or few cells. Finally, TE mobilization is intrinsically infrequent, and thus most cells do not contain new insertions. To distinguish *de novo* integration events from preexisting endogenous copies and possibly to select cells containing new genomic insertions, transposition can be induced from a genetically marked TE copy. A popular marker for retrotransposition is in antisense orientation relative to the retroelement transcription and is interrupted by an intron, a setting allowing its expression only upon reverse transcription and integration [28-31]. The necessity to express a genetic marker upon integration may favor selection of insertions occurring in euchromatic regions. Nevertheless, comparison of unselected *vs* selected population shows only minimal or no difference [32-34]. The mapping of retroviral integration sites faces similar experimental difficulties, but is somehow facilitated by the absence of related endogenous copies and by the possibility to use infectious particles, with controlled multiplicity of infection. In the case of DNA transposons, mobilization from a chromosomal locus is prone to local hopping. Thus, addressing their target site preference genome-wide necessitates the use of a plasmid-borne donor element [35]. Early recovery methods were at low scale and labor consuming, often relying on the isolation of clones carrying a single event of mobilization [27,36-39]. Although they provided useful information on the mechanisms of mobilization, the number of recovered insertions was rarely sufficient to reveal insertion site preferences unless very pronounced. Deep-sequencing technologies combined with the availability of high-quality genome assembly and functional annotations have rapidly revolutionized the field.

Principle of insertion profiling by deep-sequencing. Several methods to map insertion sites have been described, but they all follow the same general outline (Figure 1). Permissive cells are infected by a retroviral vector or transfected with a plasmid-borne TE whose expression is driven by a constitutive/inducible promoter, and containing a genetic marker,

which can be either a simple oligonucleotide tag added to the construct or a selectable cassette. If required, an optional selection step allows enrichment of cells containing new insertions. Junctions between virus or TE and the host genome are amplified by PCR-based methods (inverse PCR or iPCR; ligation-mediated PCR or LM-PCR; linear amplification–mediated PCR or LAM-PCR) [40-42], which start by ligating adapter sequences to either enzymatically or mechanically fragmented DNA, or an initial linear extension amplicon product. Nested PCR, which includes several rounds of amplification, is often necessary to amplify low abundance junctions in a complex DNA population. Sequencing adapters are added during the PCR steps or ligated afterword. Finally, deep-sequencing is achieved through common sequencing technologies (Roche, Illumina or Ion Torrent). Independent experiments can be sequenced in the same run, if libraries are barcoded and multiplexed [43]. Other PCR- or capture-based next-generation sequencing methods have been used to identify germline or somatic polymorphic natural TE insertions [5,44-49]. Finally, whole-genome sequencing is currently not worth considering, even for small genomes, given the sequencing depth required to identify rare events in a heterogeneous population.

Toward quantitative measures of integration frequency. The small proportion of DNA molecules carrying a given insertion, the high number of PCR cycles required to prepare sequencing libraries and the stochastic nature of PCR in these conditions contribute to a large proportion of PCR-generated duplicate reads by deep-sequencing with some insertions being over amplified. As a consequence, the number of reads obtained for a given junction cannot be directly translated into a frequency of integration events. This phenomenon is exacerbated by post-integration cellular divisions and by the continuous mobilization of TEs during cell growth, with both early integration events present in many cells, and later events present in much less cells. To identify recurrent insertions one solution is to generate multiple independent libraries from independent experiments. Alternatively, Levin and colleagues designed a clever strategy called serial number tagging to address the integration preference of the LTR-retrotransposon Tf1 in *S. pombe*. This method relies on TE mobilization from a library of donor plasmids, each containing a random sequence tag, which can be used to discriminate *bona fide* independent integration events from molecular or clonal expansion [50]. This approach has been

successfully applied to quantify alterations in integration distribution following genetic disruption of host factors involved in Tf1 integration [51].

Bioinformatics analysis and functional annotations. The computational analysis of sequencing data comprises a step where the genomic sequences flanking the insertions are extracted from the reads, aligned to the reference genome of interest, and used to call precise integration sites. Following this initial analysis, motif search can be performed to identify local sequence preference at or nearby insertion sites, and the degree of association between insertions and a wide range of genomic features can be assessed. Classically, these features include the position relative to genes, GC-content, chromatin domains, DNase-sensitive sites, ChIP-seq peaks, nucleosome positioning or any other relevant available dataset. To this end, public data repositories of large scale functional genomic experiments, such as the Saccharomyces Genome Database (SGD) or ENCODE, are invaluable resources [52,53]. Consequently, our understanding of integration site selection is far more advanced for organisms for which these datasets are available, such as model organisms or humans, than for any other species. Empirical comparison of experimental insertions with multiple sets of randomly distributed control insertions and receiver operating characteristic (ROC) curves are often used to statistically test the probability that such associations arise by chance. For this purpose, computer simulated insertions, matching the experimental design, are generated [25,39,54-56]. Generalized linear models (GLM) represent a tool of choice to evaluate the respective contribution of the various genomic features to insertion site selection [22,57]. However, assaying integration site distributions upon genetic manipulation of the cellular host or of the mobilization machinery, is required to validate these association studies [58-60].

Preferred genomic integration sites. Deep-sequencing, as well as more classical approaches have revealed a remarkable diversity of integration site preferences among TEs and retroviruses, from very specific nucleotide sequences, to broad chromatin domains or chromosomal regions. Known molecular determinants guiding these preferences are discussed in the following sections.

Intrinsic specificities of mobilization machineries

DNA transposons, non-LTR retrotransposons and LTR-containing retroelements possess distinct mobilization machineries, with unique enzymatic properties, which play a key role in determining the preference for given DNA sequences and/or chromatin structures.

Local preference of non-LTR retrotransposon endonucleases. Non-LTR retrotransposons replicate by a mechanism known as target-primed reverse transcription (TPRT). This process is initiated by a nick in the target DNA, followed by the local reverse transcription of the retrotransposon RNA, using the resulting 3' hydroxyl group as a primer (Figure 2). The endonuclease activity (EN) mediating the initial cleavage is encoded by the retrotransposon itself and can belong to distinct enzymatic classes, either a restriction enzyme-like EN (RLE), related to type IIs restriction enzymes, or an apurinic/apyrimidinic EN (APE) [61,62]. A major difference between the two classes of ENs is that the APE domain directly contributes to the sequence preference of the target site, while one (or several) independent DNA-binding domain(s), outside of a non-specific EN domain, mediates the recognition of the target DNA by RLE-encoding elements [63-65]. Most of the RLE elements, such as R2, SART1 and TRAS1 in insects, insert in specific targets both from a sequence and location point-of-view (see below). In contrast, only a small subset of the APE-encoding elements is strictly site-specific, with the majority integrating into genomes in a much-dispersed manner [66]. Crystallographic studies and point mutagenesis have revealed that a variable β-hairpin loop protruding from the DNA-binding surface of APE-like ENs contacts the DNA minor groove adjacent to the scissile bond and participates to sequence recognition at the cleavage site [67-71].

Annealing of non-LTR retrotransposon RNA to target sites. During TPRT, target sites are substrates for both endonucleolytic cleavage and reverse transcription, since these processes are coordinated. With few exceptions, such as the RLE element R2, this implies base-pairing between the target site DNA and the retrotransposon RNA, which limits the possible target sites, beyond EN consensus sequence. For example, L1 reverse transcription priming is favored by annealing the L1 RNA poly(rA) tail to the T-rich tracts at the target site [72,73]. Consequently, only T-rich sites are efficiently used during TPRT by L1. Regarding

R1Bm, a site-specific element inserting in *rDNA*, the 3' end of its RNA contains a sequence matching the target site, which has been incorporated by transcriptional readthrough of the progenitor locus [74]. APE domain swapping between distinct elements is sometimes sufficient to redirect integration toward different target sequences *in vivo* [75-77]. However, in other situations, the engineered EN is capable of cleaving altered consensus sequences *in vitro*, but not to mediate integration in such sites *in vivo*, consistent with a role of the target site in priming reverse transcription after cleavage [71].

Short sequences targeted by integrases and transposases. DNA transposases with RNase H-like catalytic nuclease domain form the most common group of transposases and are closely related to the catalytic core domain of retroviral and LTR-retrotransposon integrases (INs) [78]. Both DNA transposases and INs cleave the bound target DNA through a phosphodiester transesterification reaction to integrate double-stranded DNA [79-81]. Several DNA transposons (Sleeping Beauty, Piggyback, MITEs) and LTR-retroelements also exhibit a very short nucleotide signature, often containing or limited to a flexible pyrimidine (Y)/purine (R) dinucleotide at the center of the integration site. Central flexibility facilitates the deformation of the target DNA required to position the scissile phosphodiester bond within the active site of the enzyme as evidenced in the crystal structures of the strand-transfer complexes of Prototype Foamy Virus (PFV) and Mos1 DNA transposon [82,83]. Interactions between residues belonging to the transposase or IN and the phosphate backbone of nucleotides flanking the flexible dinucleotide provide the molecular basis for target sequence selection [82-85]. In most cases, sequences recognized by INs or transposases are very short and highly frequent in the genome, and their contribution to the overall genomic distribution of their respective TEs is only limited. An exception is represented by the IN of *Drosophila* endogenous retrovirus ZAM, which specifically recognizes a CGCGCG consensus sequence [86,87].

Palindromic target site and enzyme multimerization. Beyond the central nucleotides surrounding the scissile bond, the alignment of a large number of integration sites sometimes reveals TE- or virus-specific weakly conserved palindromic sequences that extend on each side of the insertion [50,88-93]. This pattern reflects the multimerization of INs and transposases within the synaptic complex [94-97]. Of note, multimeric complexes do not

always translate into a palindromic target site. R2 acts as a dimer, but the two protomers are not functionally equivalent in the TPRT process, and only one subunit seems to directly contact the target site [63].

DNA bending. Many TEs or retroviruses favor integration in target sites where the central nucleotides are bent, widening one of the DNA grooves and allowing the catalytic residues to contact the scissile bond [71,82,98]. As a consequence, pre-bent and distorted DNA, particularly in the context of nucleosome wrapping, is a good substrate for many integration complexes both *in vitro* and *in vivo* [32,33,40,89,95,99-105]. In the case of retroviruses such as Human Immunodeficiency Virus (HIV) and MLV the rotational orientation of nucleosomal-associated DNA also influences the selectivity of integration, which is characterized by an enrichment of insertions in the widened DNA major groove facing out the nucleosome structure [99,103,105]. This is not the case for PFV, which is insensitive to the deformation of the target DNA [105].

Conclusive remarks. Overall, except for RLE-containing and some APE-containing site-specific retrotransposons, the intrinsic biochemical properties of the mobilization machineries are not sufficient to explain the genomic distribution of *de novo* integration events. Although local DNA sequence or structure, such as DNA bending, might favor integration, reaching these favorable sites in the context of a complex chromatin and nuclear architecture involves additional mechanisms.

Chromatin and nuclear context

Cellular chromatin represents the natural substrate of TE insertion and depending on its structure it can affect the efficiency and/or selectivity of integration at a local level. For retroelements, the integration complex is assembled in the cytoplasm and needs to enter the nucleus to access the target genome. Whether crossing the nuclear envelope is achieved during mitosis or through the nuclear pore complex (NPC) in interphase cells, can also impact the chromosomal territories accessible to integration.

The nuclear entry route may shape integration site selection. Many retroelements, including L1, yeast LTR-retrotransposons and HIV are able to transpose/integrate in cells that do non-divide or that undergo a close mitosis. For these TEs, the NPC is the only passageway to enter the nucleus where integration occurs. Several components of the integration complex of LTR-retroelements, including the INs of Ty1, Tf1 and HIV, have karyophilic properties and contribute to nuclear import [106-111].

Analysis of integration profile of chimeric HIV harboring MLV sequences demonstrated the existence of a link between the nuclear entry pathway and integration site selection. These studies confirmed that IN drives the integration preferences, while the structural Gag protein defines the ability of HIV to access the nucleus in non-dividing cells. Importantly they also revealed that Gag contributes to integration profile. Consistently, depletion of host proteins implicated in HIV integration complex nuclear import and interacting with the Gag-derived capsid (CA) protein, including Nup153, Nup358 and CPSF6, alters integration patterns [112-114]. This phenotype was reproduced using HIV harboring CA mutations that impair binding with the abovementioned cofactors [112]. The topology of the host genome, particularly the organization of chromatin in the vicinity of the NPCs, also influences integration site selection of HIV. Once in the nucleus, HIV integration complexes preferentially localize in euchromatin areas close to the nuclear envelope [115-117] and, acting on the closest targets, direct integration preferentially in actively transcribed genes near the NPCs, while disfavoring transcriptionally repressed heterochromatin of lamin-associated domains (LADs) or transcriptionally active regions located in the center of the nucleus [118]. Another player in HIV integration site selection is the nucleoporin Tpr nucleoporin which contributes to maintain the chromatin architecture underneath the nuclear envelope, recruits transcribed genes near the NPCs [119,120] and interacts with LEDGF/p75, the major co-factor of HIV insertion in actively transcribed genes (see below) [121]. In *S. pombe*, Tf1 retrotransposition requires that Gag interacts with Nup124, the homolog of human Nup153, and enters the nucleus [122]. Systematic screens for non-essential yeast genes involved in Ty1 and Ty3 retrotransposition have also identified NPC components [123,124] and Ty3 nuclear entry is initiated by virus-like particle docking on NPC

proteins. These observations suggest that the connection between nuclear import and integration might be conserved for these elements.

Chromatin accessibility. DNA-bound proteins have a significant impact on the integration process of many TEs by blocking or facilitating the access of the integration complex to the target DNA. However, open chromatin is not always the favored substrate, each TE displaying preferences for specific chromatin features. Both Tf1 and Ty5 target DNase sensitive sites, which represent nucleosome-free region [51,60,125]. This preference is particularly twisted for Ty5, which integrates in heterochromatin environments, but still selects nucleosome-free sites at a local level [54]. Consistent with this model, the distribution of Ty5 insertions correlates with the integration pattern of the housefly DNA transposon Hermes, which identifies open chromatin into the yeast genome [126]. Other retroelements such as Ty1 and several retroviruses integrate preferentially into nucleosome-bound DNA both *in vitro* and *in vivo* [32,33,95,104,127,128]. Ty1 preference for nucleosome is characterized by a 70-bp periodic integration profile upstream of *tDNA* indicating two major sites of integration per nucleosome near the H2A-H2B interface [32,33,129]. Interestingly MLV and PFV target stable and dense chromatin, while HIV and Avian Sarcoma virus (ASV) have a bias for regions of low nucleosome occupancy [104,130]. This preferential integration into nucleosomes within a defined chromatin conformation may be driven by the structural constrains of the retroviral integration complex. Evidences indicate that chromatin-remodeling factors also contribute to integration site selection both for Ty1 [131] and HIV [127]. A loose association between DNA and the histone octamer could allow the degree of flexibility required to fit the target DNA in the active site of the integration complex as discussed above.

Integration profiles also correlate with specific histone marks. Studies on the integration preferences of several TEs indicate that this correlation is in large part due to cellular cofactors that guide the integration complex to the site of insertion and interact with specific modified histones (see below and [132]).

A generalized tethering model

The mechanisms detailed above influence target site selection locally and in many cases are not sufficient to explain the non-random distribution of TEs insertions across the host genome. A major determinant of integration site selection is the tethering of the integration machinery at the site of integration by element-specific cellular DNA- or chromatin-binding proteins. This mechanism, popularized by Fred Bushman, is known as the "tethering model" [133].

General overview and experimental criteria. During the integration process of LTR-retrotransposons and retroviruses, IN catalyzes the processing of the viral DNA ends and their joining to target DNA. As an inherent component of the PIC, IN is ideally positioned to contribute to integration site selectivity. Based on the study of yeast LTR-retrotransposons, the original tethering model proposed that IN interacts with a cellular protein that binds to the site of integration. Since this model was proposed, element-specific tethering factors have been identified for many retroelements, including retroviruses and some non-LTR retrotransposons. Beside interacting with the integration complexes, two properties should be considered when evaluating the potential role of a host protein in the tethering of TEs: first, whether TE integration site preference parallels the distribution of its candidate tethering factor [42,59,134-137]; and second, whether artificial genomic relocation of the candidate tethering factor, or of its integrase-binding domain (IBD), is sufficient to relocate TE insertions in the same genomic regions [138-141]. Abrogating the interaction between a TE and its predominant tethering factor can lead to insertion profiles with partially conserved, more random or completely new distribution. This hard-to-predict outcome results from the redundancy between multiple tethering factors, and from additional multilayers regulations, as discussed below [54,58,59,142-145].

Tethering through an interaction with IN. Historically, studies on Ty3 and Ty5 were pioneer in establishing that INs interaction with tethering factors was at the basis of integration targeting. *In vivo*, Ty5 integration into heterochromatin requires an interaction between a hexapeptide of Ty5 IN (named TD for targeting domain) and the Sir4 heterochromatin protein [146,147] (Figure 3). A single amino acid change in this motif abolishes Ty5 integration

preference. Regarding Ty3, *in vitro* approaches revealed that an interaction between IN and the transcription initiation factor TFIIIB, and especially the subunits Brf1 and TBP, targets Ty3 within one or two nucleotides of the transcription initiation sites of RNA polymerase III-transcribed genes [148,149] (Figure 3). *In vivo,* interaction between IN and the Tfc1 subunit of TFIIIC also influences the orientation of Ty3 insertions with respect to the targeted gene [150]. Since then, tethering factors that interact with IN and dictate integration targeting have been characterized for other retroelements. Ty1 preferentially integrates into a 1-kb window upstream of Pol III transcribed-genes [32,33,151] (Figure 3). Recent studies described interactions between Ty1 IN and different subunits of Pol III, including AC40, C31 and C53 [59,152]. Ty1 integration preference for Pol III genes is virtually abolished in a AC40/IN loss-of-interaction mutant indicating that AC40 acts predominantly in Ty1 targeting [59]. In *S. pombe*, Sap1, an essential DNA-binding protein involved in replication fork arrest, interacts with Tf1 IN and contributes to the efficiency and the selectivity of Tf1 integration on the fork side and at nucleosome-free regions observed at promoters of RNA polymerase II-transcribed genes [51,60] (Figure 3). The IN of plant chromoviruses, which are related to Ty3/Gypsy LTR-retrotransposons, harbors a chromodomain that recognizes histone marks characteristic of heterochromatin and directs integration [153]. In the case of HIV, IN interacts with the ubiquitously expressed transcriptional coactivator LEDGF/p75 via its C-terminal IN-binding domain [154] (Figure 3). The interaction between these two proteins has been extensively studied because it is a potential target for antiviral therapy (reviewed in [155]). LEDGF/p75 stimulates and directs HIV integration into the body of active and highly spliced genes located within gene-dense regions of chromosomes [40,58,142,145,156-158]. Consistently, LEDGF/p75 interacts with splicing factors [145,159] and contains a conserved N-terminal PWWP domain, which binds to H3K36me3, a histone mark typical of active transcription [160]. Regarding gamma-retroviruses, typified by MLV, they rely on the interaction of IN with the bromodomain and extra-terminal (BET) proteins to integrate near transcription start sites and CpG islands, features associated with promoters and enhancers [93,136,137,144,161-166] (Figure 3). This targeting is mediated by two N-terminal bromodomains present in BET proteins, which specifically recognize acetylated H3 and H4 histone tails that are enriched at Pol II promoters and enhancers [132].

Tethering through Gag or ORF1p. Beside the essential role of IN in integration site selection for many LTR-retroelements, Gag or Gag-derived proteins might also play a role in this process. Mutations in HIV Gag-derived CA protein impair the interaction with nuclear import cofactors and alter its integration pattern, as discussed above. Unlike HIV, which can infect non-dividing cells, other retroviruses including MLV and PFV require mitosis to access the host cell genome. PFV Gag harbors a C-terminal chromatin-binding sequence (CBS), which interacts with the H2A/H2B core histones [167] and facilitates integration complexes tethering on mitotic chomosomes (Figure 3). Mutation in this region impairs integration, but whether the integration profile is affected has not been established yet. Similarly, MLV Gag-cleavage product $p12^{Gag}$ harbors a chromatin-binding domain able to attract integration complexes to mitotic chromosomes [168-170]. However, genetic manipulation of $p12^{Gag}$ chromatin binding domains does not significantly change MLV integration profile, suggesting that this viral protein does not play a major role.

Evidence for tethering of non-LTR retrotransposons is scarce, but all known cases so far involve an RNA-binding protein, called ORF1p (or sometimes Gag by analogy with LTR-retroelements), encoded by the element itself, and belonging to the retrotransposition complex. A striking example concerns the collaborative targeting to telomeres of HeT-A, TART, and TAHRE elements in *Drosophila* [171-174]. HeT-A ORF1p forms spherical structures at chromosome ends (the so-called 'Het dots'), which are necessary for TPRT [173,174]. TART and TAHRE ORF1p proteins rely on HeT-A ORF1p to access telomeres [173]. The *ver* protein, which is essential for telomere protection is required for HeT-A ORF1p recruitment to telomeres and may therefore act as a tethering factor [174]. Similarly, tethering to telomeres for the non-LTR retrotransposon SART1 is mediated by its ORF1p protein, in *Bombyx mori*, although it is unknown whether a host factor is involved in this process [175]. In *Dictyostelium discoideum*, the ORF1p protein of the non-LTR retrotransposon TRE5-A interacts with TFIIIB *in vivo* and *in vitro* and contributes to TRE-5A integration targeting ~50 bp upstream of *tDNA* genes [176,177]. Whether ORF1p-mediated tethering to their target sites can be generalized to most non-LTR retrotransposons remains unknown.

Tethering factor redundancy. In many cases, impairing the interaction between IN and the tethering factor reveals secondary integration biases indicating that additional proteins

and/or genomic features influence target site selection. For instance, genome-wide analysis of Ty1 *de novo* insertion events in the absence of AC40/IN interaction shows redistribution towards subtelomeric regions [59]. In the absence of LEDGF/p75, the related HRP2 protein that contains both a PWWP and an IBD, can direct HIV integration to active genes. Even upon concomitant depletion of LEDGF/p75 and HRP2 viral integration into genes remains significantly more frequent than expected randomly [178,179]. This residual targeting into intron-rich genes and gene-dense regions has been attributed to an interaction between HIV CA and splicing factor CPSF6 [114,180]. Based on these data, a two-steps model has been proposed according to which the initial interaction between CA and CPSF6 guides the PIC to euchromatin. Next, IN interacts with chromatin-bound LEDGF/p75 that directs integration preference within genes [114].

Beside predominant tethering factors, additional cofactors that also bind to DNA or chromatin, have been identified. These cofactors may target integration to a subset of specific loci or have specific functions at the site of integration. This is the case for the transcription activator Atf1, which binds Tf1 IN and directs integration to the promoter of *fbp1* [181,182] and for the separase Esp1, which interacts with Ty1 IN and may be required for cohesin removal at some Ty1 targeted loci [183].

In conclusion, the currently identified tethering factors are linked to a myriad of biological processes acting on chromatin (*e.g.*, transcription, splicing, replication, heterochromatin structure), and are major drivers of TE insertion site preference. Since the initial description of the tethering model, studies on TE integration targeting have not only confirmed its relevance but have also indicated that this model is not restricted to LTR-retroelements, is modular (different TE components can be involved, not only INs) and redundant (several distinct host factors can contribute to target a given TE to the same genomic territories or to complementary regions).

TE integration, cell adaptation and genome evolution

Integration site and reactivation potential. TEs must continuously continue to replicate to avoid extinction. However, active elements may lose their mobility over time by acquiring

mutations. Hence, new insertions must take over and continue their replication cycle to escape extinction. TE mobility is regulated in the genome at multiple levels to control mutagenesis. Yet a small fraction of the integrated elements may succeed to escape transcriptional regulation [5,184,185] and in general, even a smaller fraction of the transcribed elements is transposition competent and able to bypass a number of host defense checkpoints against their activity [186-188].

Transcriptional regulation is the primary step to limit the activation of TEs. For instance, transposition rate of a Ty1 element is correlated with the relative abundance of its transcripts [29,189] Often only few elements are responsible for the bulk of transcripts [3,5,185]. A recent study [5] demonstrated the heterogeneity of L1 transcriptional activity by measuring the expression of several hundreds of full-length human L1 elements in a panel of 12 commonly used primary and transformed cells of different tissue origin, a phenomenon governed by locus- and cell-type-specific determinants [5,190]. Unlike L1 or Ty1 elements, R2 elements, which integrate preferentially in the 28S *rDNA,* are exposed to a relaxed transcriptional control, since only a small fraction of *rDNA* genes needs to be transcribed. The proportion of R2 transcription varies by strain and species type and a minor fraction of R2 elements remain active as most of the transcripts are truncated [4]. In many organisms, epigenetic regulatory mechanisms have been developed to silence TEs upon their integration, and several TEs target heterochromatin domains. However, targeting integration to these regions is generally detrimental for the element expression and leads to progressive extinction by accumulation of mutations, as suggested by the abundance of remnants TE sequences in heterochromatin and the absence of functional Ty5 copies in *S. cerevisiae*. This repression can be transiently relieved. In *S. cerevisiae*, pheromone exposure removes Ty5 transcriptional silencing at telomeres, thus allowing the element to propagate [191]. Pheromones are expressed in haploid cells to allow conjugation between cells of opposite mating type, a process, which induces chromosomal restructuring. Therefore, activation of Ty5 in response to pheromones may be a way for the cell to create gene diversity in response to stress. Pheromone activation has also been described for Ty3 which is located upstream of Pol III and repressed at this position [192].

Like other LTR and non-LTR elements, activation of integrated retroviruses (provirus) is specific to the environment of integration loci. Reactivation of retroviruses deserves special attention as they can remain active and continue to infect cells, or maintain a latent reservoir which can be reactivated later by epigenetic modifications leading to persistent infection. Provirus genes are expressed in only a small percentage of integrated cells [184]. Cells with defective HIV proviruses have a survival advantage allowing them to expand clonally. In parallel, immune surveillance tends to eliminate infected cells expressing provirus transcripts and viral proteins, selecting the latent reservoir of future infections [121,193,194]. Low viral expression associates with HIV integration in gene deserts, centromeric heterochromatin, and very highly expressed cellular genes [195,196]. Cases have been found where cells with HIV insertions in specific genes are strongly positively selected by promoting the survival and expansion of the infected cells, eventually resulting in malignancies [193,197]. Besides the integration site, post-integration epigenetic modifications of the provirus and of the surrounding locus also correlate with the establishment and maintenance of a viral latent state. Particularly, deacetylation [196,198-202] and methylation [203-208] of histones located at the viral long terminal repeat (LTR), and DNA CpG methylation of the HIV-1 promoter [209] contribute to transcriptional repression of proviruses. Finally, nucleosome positioning in the 5' LTR mediated by the BAF complex [210,211], and nucleosome remodeling by LEDGF/p75, along with Iws1 and Spt6, also contribute to post-integration HIV silencing by inducing repressive chromatin [212].

Coevolution of host-element impacts target site preference. TE integration is a major threat for the maintenance of genome integrity and host survival, especially in compact genomes. On the other hand, death of the host ultimately affects TEs. Evolutionary strategies have emerged from both sides to allow propagation of TEs while minimizing the genetic damages to the host and reached an equilibrium. This is particularly important in organisms such as *S. cerevisiae* and *dictyostelids*, which genomes contain 70% of protein coding genes, and only 3% and 10% of TEs, respectively. In these organisms, TEs insert either in heterochromatin, *i.e.* subtelomeres or centromeres, or close to/in *tDNA* or *rDNA* [134,151,213,214], *i.e.* multicopy genes and thus individually non-essential. Strikingly, *tDNA*

targeting has been independently developed several times during dictyostelid and *S. cerevisiae* evolution for both non-LTR and/or LTR-retroelements [124,213,214].

Regions of heterochromatin are also preferential targets in larger genomes such as those of plants and insects. Because protein-coding genes represent a limited fraction in these genomes, targeting heterochromatin may not have been primary selected to limit insertional mutagenesis, but could have rather evolved to fulfill structural and regulatory functions. For example, in *Drosophila* and some insects, integration of non LTR-retrotransposons at telomeres is essential to maintain telomere length homeostasis and to protect chromosome ends in the absence of telomerase [215]. The proximity of chromosome ends and centromeres to the nuclear periphery in many organisms may also facilitate TEs targeting to these regions just after nuclear entry [118,121]. Likewise, the proximity of transposons or transposons remnants sequences within *tDNA* has been described in Drosophila and *C. elegans*. In these organisms, piRNA clusters represent TEs cemeteries where the elements are repressed, accumulate mutations and become transposition incompetent. Transcription of these remnants allows the production of piRNAs, which repress the transcription of *de novo* insertions. Recently, a link between tRNA processing defect occurring at the site of Pol III transcription and nearby piRNA clusters activation was discovered, indicating that *tDNAs* may create a chromatin environment facilitating piRNAs transcription [216]. The proximity of *tDNA* genes and TEs sequences observed in yeast, amoeba and drosophila, which are phylogenetically distant, also suggests that Pol III transcription may create a conserved and favorable environment for TE insertions.

Many TEs integration sites are accessible to DNase, even in heterochromatin domains, and correspond to intergenic regions, for example, introns and upstream regulatory sequences, to ensure the minimum loss of function to the host system. This is demonstrated by the insertion preferences upstream of RNA Pol III transcribed genes, as mentioned above, in nucleosome-free regions (Tf1 and Tf2 from *S. pombe, Ty5* from *S. cerevisiae*), heterochromatins (Ty5, skipper-1 from *D. discoideum*), telomeres (HeTA, TART and TAHRE from *D. melanogaster*, SART1 and TRAS1 from *B. mori*), introns (HIV) and regulatory regions upstream of genes (MLV). Alteration of nucleosome positioning by mutated chromatin remodeling factors was shown to be associated with altered periodicity of Ty1 integration

[131]. Integrations in- or upstream of the regulatory sequences may cause transcriptional deregulation of the downstream genes and likewise, expression of the integrated elements can be deregulated by the downstream genes [217,218].

***Retargeting upon stress*.** Barbara McClintock predicted that TEs provide the cell with a prewired mechanism to reorganize the genome in response to environmental challenge [219]. Consistent with this hypothesis, the transcription of TEs can be activated by stress as shown in yeast [220,221] and plants [222,223]. This activation can induce new insertion events, impact the regulation of adjacent genes and have important outcomes on the fitness of the organism. For example, in melon female, stress conditions induce sex reversion by derepression of a TE adjacent to a female specific gene allowing male flower development and seed production [224]. In *S. pombe,* Tf1 preferentially integrates in promoters that are induced by environmental stresses [125] and the expression of these genes is enhanced upon Tf1 integration [225,226].

However, to date, the only direct piece of evidence supporting the notion that integration preferences could be modified by stress comes from studies in *S. cerevisiae*. For Ty1 and Ty5, targeting and integration per se itself can be genetically separated. When interaction with the tethering factor is abolished, integration is maintained at relatively wild-type levels but new integration patterns can be observed [59,146]. In the case of Ty5, nutrient starvation abolishes the phosphorylation of a serine residue of IN, which is required for the interaction with Sir4. In the absence of Sir4/IN interaction, Ty5 continues to integrate in nucleosome-free regions and as such avoids coding-regions, thereby limiting insertional mutagenesis. Nevertheless, this new targeting may favor nucleosome-depleted promoter regions, and consequently alter the regulation of adjacent genes, which could have important evolutionary outcomes. Similarly, when the AC40/Ty1 IN interaction is abolished in the presence of an AC40 loss-of-interaction mutant, Ty1 insertions are redistributed towards subtelomeric regions. These regions contain non-essential fast-evolving gene families generally needed to respond to environmental changes [227] and integration at subtelomeres can also shape chromosome ends structure through recombination between ectopic copies [228]. Therefore, targeting Ty1 integration to subtelomeres could further protect the yeast genome from Ty1 mobility, while potentially promoting evolutionary

adaptation and gene innovation in response to stress. However, physiological or environmental conditions that would naturally disrupt the AC40/IN interaction have not been discovered yet.

Design of vectors for gene delivery or functional genomic studies

The ability to stably integrate their genetic material into the genome of the host cell is a key property of retroviruses and TEs that has been central to their use as biotechnological tools. Their applications range from genetic modification of cell lines and animal transgenesis to gene therapy for the treatment of human genetic diseases. Retrovirus-derived vectors ensure efficient gene transfer and long-lasting expression and are commonly used to investigate the function of a gene of interest or in functional screens with cDNA libraries [229]. Vectors derived from MLV, a mouse tumor virus, have also been central to the development of gene therapy over decades and were instrumental to prove the feasibility of this approach in pioneer studies for the treatment of inherited immunodeficiencies in children [230,231]. However, the development of vector-related leukemia in a subset of patients, due to dysregulated expression of a proto-oncogene following integration of the therapeutic gene in its proximity [21], pinpointed the mutagenic potential of MLV-based vectors. Further analyses of the pattern of vector integration in patients from gene therapy trials revealed preferential insertion near or within transcription units [232-236]. Presence of a dominant cellular clone carrying the integrated vector, without malignant development, has been also reported in a patient suffering from beta-thalassemia and treated with a HIV-derived vector [237]. Altogether these observations fueled research aiming to uncover the molecular basis of integration site selection and to design strategies to manipulate integration targeting in genomic safe harbors or non-coding repeats. Foamy viruses, which are considerably less prone to integrate near or within genes than MLV and HIV, respectively, would potentially make safer vectors for gene therapy (reviewed in [238,239]). Of note, preferential targeting of genes or promoters such as MLV can be an advantage for functional genomic screens relying on insertional mutagenesis.

Besides retroviruses, the resurrected DNA transposon Sleeping Beauty has proven useful to identify new genes implicated in oncogenesis in mice (reviewed in [240]) and has attracted much attention for gene therapy applications. Sleeping Beauty presents several interesting properties including its fairly random integration profile across the genome, the low post-integration silencing and the possibility to physically separate the inverted repeat-flanked transgene and the transposase-coding sequence, opening the possibility to restrict the expression of the latter in a selected tissue (reviewed in [241]). However, this system is not risk-free and targeted integration in safe genomic location is ultimately required for therapeutic application.

Numerous studies confirmed that the integration profile of retroviruses can be modified by expression of fusion proteins composed of the integrase binding domain and a DNA-binding domain recognizing the desired chromatin sites [242]. However, this strategy seems difficult to apply for human gene therapy. Alternatively, IN could be manipulated in order to abolish the binding to the endogenous tethering factor(s) and take on a new interaction with a protein bound to a desired site. However, this strategy might negatively affect the efficiency of gene transfer. In the case of Sleeping Beauty targeted transposition could be achieved without manipulating the transposase, by co-delivery a chimeric protein obtained by fusion of a DNA-binding domain and a peptide spanning the N-terminus of the transposase, which interacts both with the transposase subunits and the IR flanking the transgene [243-245].

Figure legends

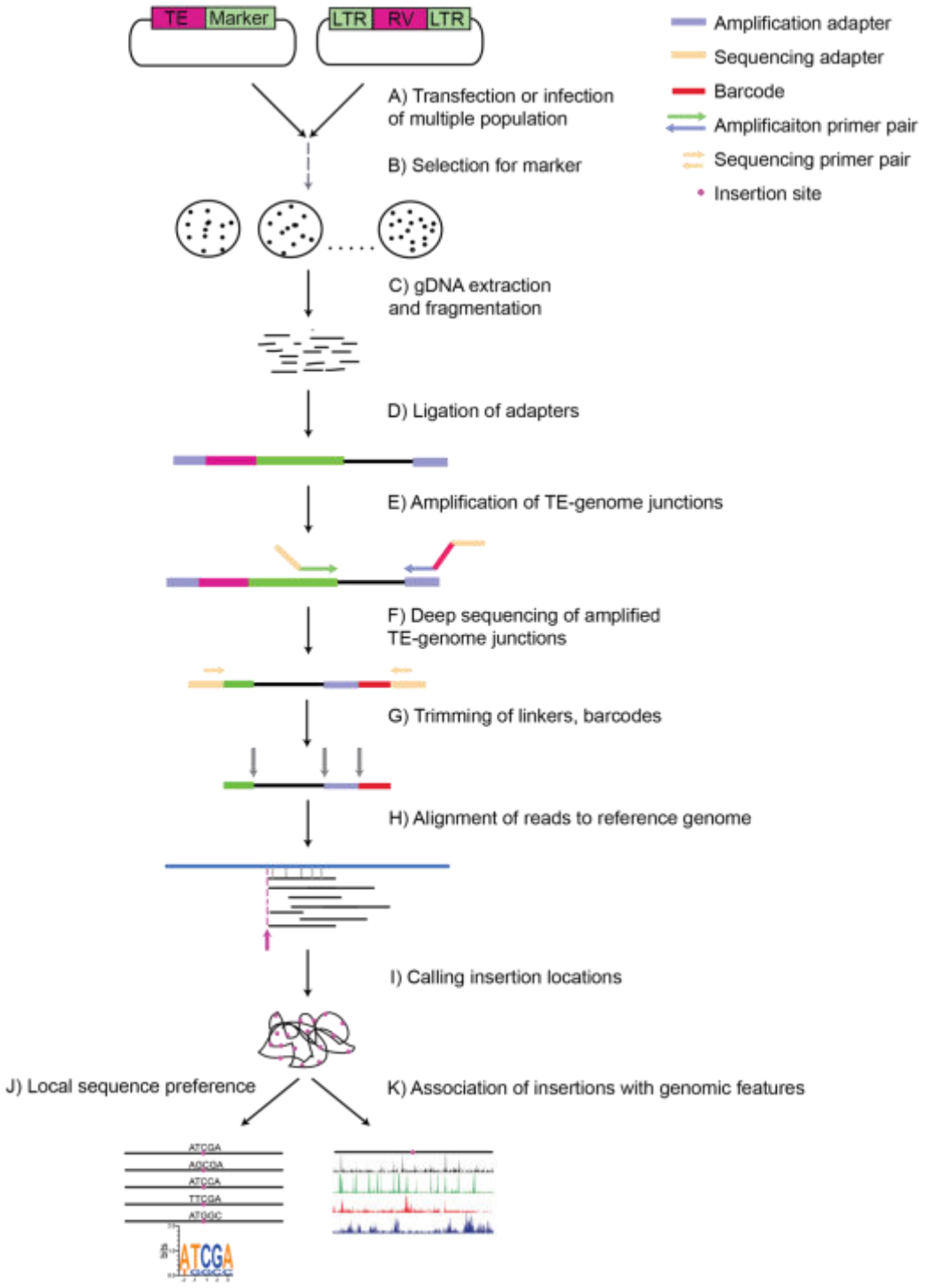

Figure 3. General experimental outline of insertion site mapping approaches. Figure legend continued on next page.

(A) Permissive cells are infected by a retroviral vector or transfected with a plasmid-borne TE containing a genetic marker. This process is repeated to obtain independent cellular populations. Alternatively, unique molecular identifiers (see main text) can be included in the construct used for transfection or infection. (B) An optional selection step (presented here with a dashed arrow) allows enrichment of cells containing new insertions. (C) Genomic DNA is extracted from each population and fragmented enzymatically or mechanically. (D) Adapters (violet boxes) are ligated to the fragmented DNA pieces. (E) Junctions between virus or TE and the host genome are amplified by PCR-based methods (ligation–mediated PCR is depicted here). Primers contain barcodes (red box) to multiplex sequencing of different populations and sequencing adapters (orange box). A second round of PCR (nested PCR) can be included but is not shown here. (F) PCR-enriched junctions are deep sequenced using next-generation sequencing technologies. (G) Non-genomic sequences, such as barcodes, linkers (and sometimes vector-originated sequences) are trimmed from the sequencing reads. (H-I) Alignment of trimmed reads to the reference genome of interest (blue line) is used to call precise integration sites (dashed line and arrow in pink indicate insertion site). (J-K) Flanking genomic sequences of insertion sites are examined to identify motif or local sequence preference (J) at or nearby insertion sites, and to quantify the degree of association between insertions and a wide range of genomic features (K).

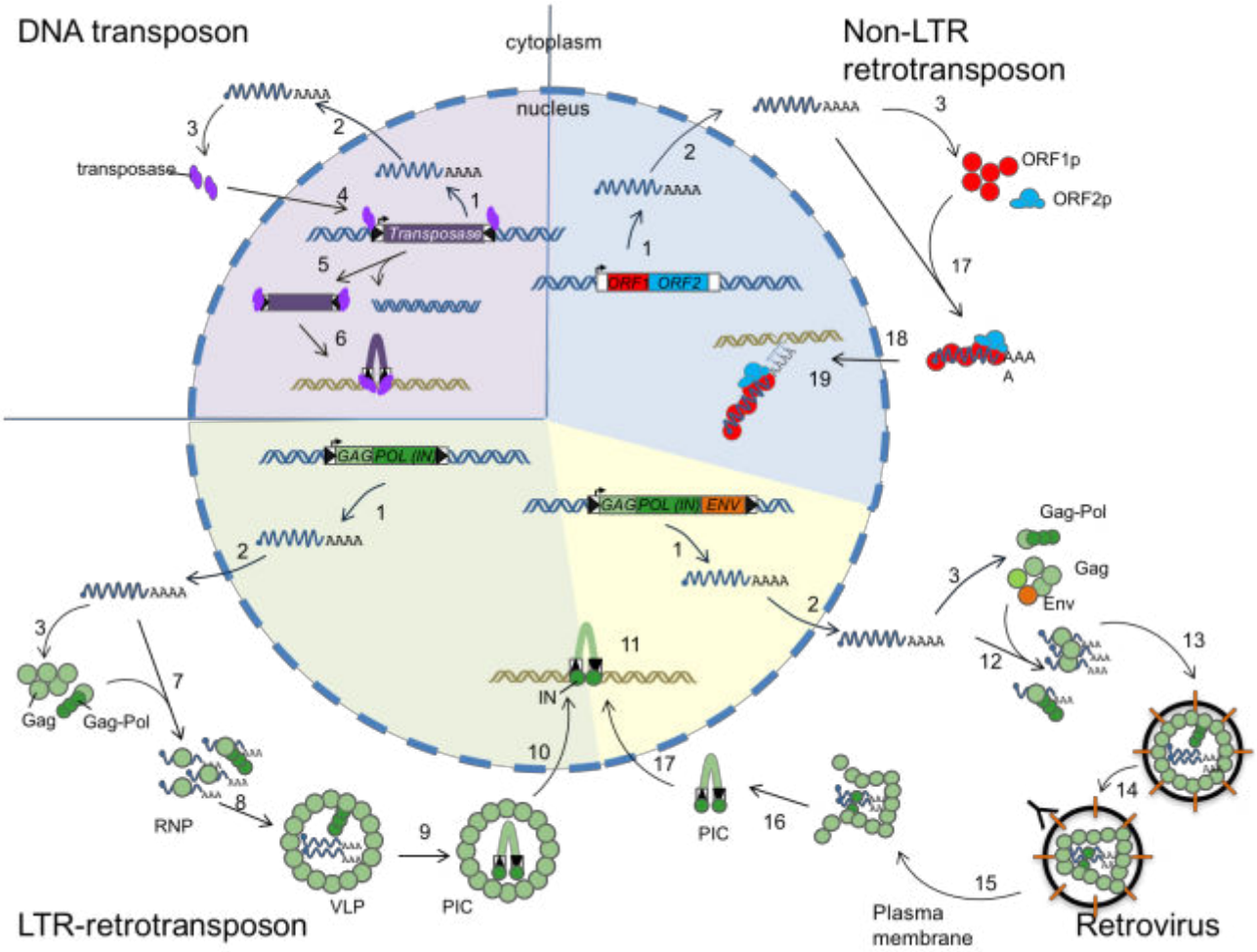

Figure 4. Broad overview of transposable element and retrovirus structure and replication cycle diversity.

In the case of all TE, following gene expression (1), mRNAs (wavy blue line) are maturated (blue dot, cap) and transported in the cytoplasm (2) where translation occurs (3). DNA transposons encode a transposase (purple circles) that once translated is imported in the nucleus (4) where it binds the inverted repeats sequences (black arrow within white rectangle) flanking the transposon, and resulting in its excision from the progenitor locus (blue DNA) (5). The transposase also catalizes the insertion of the transposon into a new genomic locus (grey DNA) (6). In contrast to DNA transposons, the mobilisation of retrotransposons and retroviruses requires the reverse transcription of a RNA intermediate. LTR-transposons harbor long terminal repeats (LTR, black arrows within white rectangles) which contain regulatory sequences including the Pol II promoter and the transcription start site (right-angled arrow) and flank the *GAG* and *POL* genes. Once expressed, Gag (light green circles) and Gag-Pol polyproteins (light and dark green circles) associate with the genomic mRNA to form ribonucleoprotein particles (RNP, aka retrosomes) (7) which give rise to virus-like particles (VLP) (8). During the following

maturation steps, Pol-derived protease (PR) cleaves the precursor proteins into mature Gag, integrase (IN) and reverse transcriptase (RT) proteins. Within the VLP, the genomic RNA is reverse transcribed by RT into a complementary double-stranded DNA molecule (cDNA) (9). The cDNA bound to IN forms the pre-integration complex (PIC) which is imported into the nucleus (10) where integration occurs (11). Retroviruses have a genomic organization and replication strategy similar to that of LTR-retroelements. The major difference lies in the presence of an envelope-coding gene (*ENV*), which allows cell-to-cell horizontal transmission through extracellular infectious particles. Viral proteins and RNA genome associate in the cytoplasm to form new viruses (12) which exit the producer cells by budding at the plasma membrane (13). Similar to LTR-retrotransposons, retroviruses undergo a maturation step consisting in the formation of a typical capsid shell that encloses the viral genome (14). This step is required to generate fully infectious viruses that are able to infect other target cells. Entry requires an interaction between the Env proteins at the surface of the virus and cellular proteins exposed at the plasma membrane (Y shapes) leading to fusion between the viral and cellular membranes (15). The capside enters the cytoplam where it undergoes disassembly in a step termed uncoating. Simultaneously the viral genome is reverse transcribed (16). The resulting PIC is imported in the nucleus where the viral DNA integrates into the target cell genome (11). Autonomous non-LTR retrotransposons, such as L1, encode a protein harboring endonuclease (EN) and reverse transcriptase (RT) activities (ORF2p in L1). Some elements also express a protein with RNA binding and nucleic acid chaperone activities (ORF1p in L1). Proteins and RNA genome associate in the cytoplam to form RNP (17) which are next imported into the nucleus (18) where integration occurs by Target-Primed Reverse Transcription (TPRT) (19). In this process, the target genome is nicked by the endonuclease (EN) and reverse transcription is directly initiated at the nick. In some elements, annealing of the retrotransposon RNA to the target site is required for efficient RT priming. The scheme depicts L1 structure and replication cycle. In addition to potential direct or inverted repeats, such as LTR or IR, most TEs are flanked by short target site duplications (TSD), formed during the integration process, and which are not shown here for the sake of simplicity.

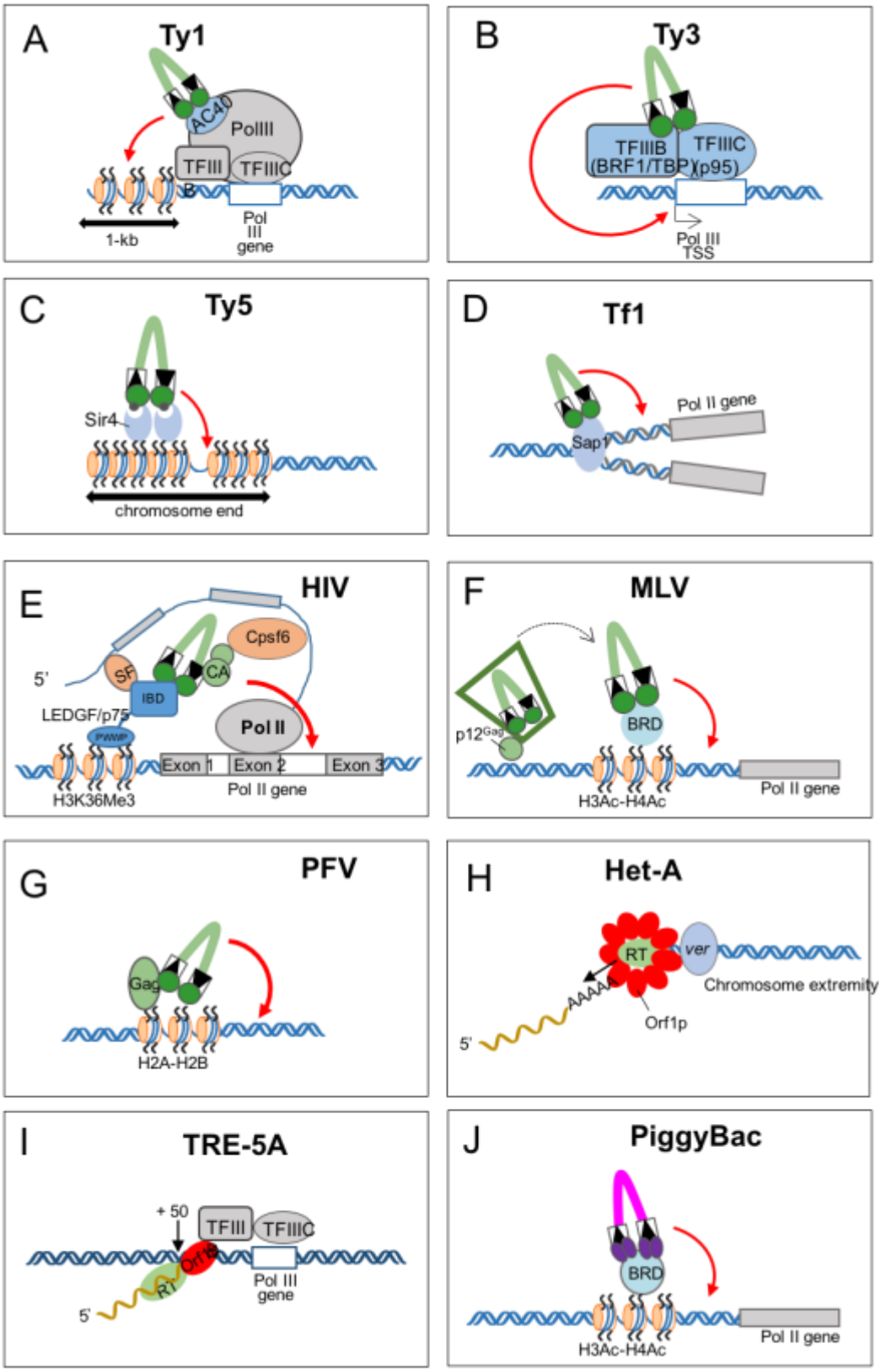

Figure 5. Extended tethering scheme. Figure legend continued on next page.

(A) Ty1 integration in nucleosomes within a 1-kb window upstream of Pol III-transcribed genes is driven by an interaction between IN and the AC40 subunit of Pol III. (B) Ty3 integration at Pol III transcription start site depends on an interaction between IN and RNAP III transcription initiation complexes composed of TFIIIB and TFIIIC. In vitro, TFIIIB subunits TBP and Brf1 are sufficient to target integration. (C) Ty5 integration into nucleosome-free regions at subtelomeres and HM loci requires an interaction between Ty5 IN, phosphorylated on serine 1095, and the Sir4 heterochromatin protein. (D) Tf1 integrates into promoters of RNA polymerase (Pol) II-transcribed genes which are nucleosome-free region upstream of the transcription start site. This targeting depends on an interaction between IN and the replication fork barrier factor Sap1. Integration occurs on the side of the fork arrest. (E) The HIV capsid protein (CA) interacts with both HIV preintegration complexes and the alternative polyadenylation complex Cpsf6 to direct HIV to transcriptionally active chromatin, where the IN-Ledgf/p75 interaction drives integration into gene bodies enriched in H3K36me3 modified histone. Ledgf/p75 interaction with splicing factors favors integration into highly spliced genes. (F) The MLV p12 protein encoded by Gag tethers the pre-integration complex (PIC) to condensed chromatin during mitosis, allowing PIC segregation to daughter cell nucleus. Release of p12 from MLV PIC allows IN to interact with BET proteins, which recognizes hyperacetylated histones H3 and H4 present at active promoters. (G) Chromatin-tethering of FVs genome is mediated by viral Gag protein which interacts with viral and cellular DNA. FVs Gag interaction with IN has not been demonstrated yet. (H) HeT-A Orf1p localizes in the nucleus and forms spherical structures that encapsulate HeT-A RNA and ORF2 (RT). The *Ver* protein, which is essential for telomere protection in *Drosophila* is required for HeT-A sphere formation. Currently, evidence that *Ver* contributes directly to HeT-A recruitment on telomeres is missing. (I) The TRE5-A preintegration complex consists of ORF1 and/or ORF2 (RT) proteins and TRE5-A RNA. ORF1 interacts with all TFIIIB subunits allowing TRE5-A integration in the +50 bp position relative to Pol III transcription start site. (J) Chromatin-tethering of PiggyBac transposon is thought to be mediated by a somewhat similar fashion to MLV via interaction of IN with BET proteins.

1. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351, 1083-1087 (2016).

2. Goodier, J. L. Retrotransposition in tumors and brains. Mob. DNA 5, 11 (2014).

3. Morillon, A., Bénard, L., Springer, M. & Lesage, P. Differential effects of chromatin and Gcn4 on the 50-fold range of expression among individual yeast Ty1 retrotransposons. Mol. Cell. Biol. 22, 2078-2088 (2002).

4. Ye, J. & Eickbush, T. H. Chromatin structure and transcription of the R1- and R2-inserted rRNA genes of Drosophila melanogaster. Mol. Cell. Biol. 26, 8781-8790 (2006).

5. Philippe, C., Vargas-Landin, D. B., Doucet, A. J., van Essen, D., et al. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. eLife 5, 166 (2016).

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

7. Weigel, D. & Colot, V. Epialleles in plant evolution. *Genome Biol.* **13,** 249 (2012).

8. Mirouze, M. & Vitte, C. Transposable elements, a treasure trove to decipher epigenetic variation: insights from Arabidopsis and crop epigenomes. *J Exp Bot* **65,** 2801-2812 (2014).

9. Bennetzen, J. L. & Wang, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol* **65,** 505-530 (2014).

10. Barron, M. G., Fiston-Lavier, A. -S., Petrov, D. A. & González, J. Population genomics of transposable elements in Drosophila. *Annu. Rev. Genet.* **48,** 561-581 (2014).

11. Kapusta, A. & Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends Genet.* **30,** 439-452 (2014).

12. Macia, A., Blanco-Jimenez, E. & García-Pérez, J. L. Retrotransposons in pluripotent cells: Impact and new roles in cellular plasticity. *Biochim Biophys Acta* **1849,** 417-426 (2015).

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

14. Schlesinger, S. & Goff, S. P. Retroviral transcriptional regulation and embryonic stem cells: war and peace. *Mol. Cell. Biol.* **35,** 770-777 (2015).

15. Göke, J. & Ng, H. H. CTRL+INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. *EMBO Rep* **17,** 1131-1144 (2016).

16. Cameron, J. R., Loh, E. Y. & Davis, R. W. Evidence for transposition of dispersed repetitive DNA families in yeast. *Cell* **16,** 739-751 (1979).

17. Sandmeyer, S. B. & Olson, M. V. Insertion of a repetitive element at the same position in the 5'-flanking regions of two dissimilar yeast tRNA genes. *Proc. Natl. Acad. Sci. U. S. A.* **79,** 7674-7678 (1982).

18. Brodeur, G. M., Sandmeyer, S. B. & Olson, M. V. Consistent association between sigma elements and tRNA genes in yeast. *Proc. Natl. Acad. Sci. U. S. A.* **80,** 3292-3296 (1983).

19. Burke, W. D., Calalang, C. C. & Eickbush, T. H. The site-specific ribosomal insertion element type II of Bombyx mori (R2Bm) contains the coding sequence for a reverse transcriptase-like enzyme. *Mol. Cell. Biol.* **7,** 2221-2230 (1987).

20. Xiong, Y. & Eickbush, T. H. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. *Mol. Cell. Biol.* **8,** 114-123 (1988).

21. Hacein-Bey-Abina, S., Von Kalle, C., Schmidt, M., McCormack, M. P., *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302,** 415-419 (2003).

22. Campos-Sánchez, R., Kapusta, A., Feschotte, C., Chiaromonte, F. & Makova, K. D. Genomic landscape of human, bat, and ex vivo DNA transposon integrations. *Mol Biol Evol* **31,** 1816-1832 (2014).

23. Dewannieux, M., Harper, F., Richaud, A., Letzelter, C., *et al.* Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* **16,** 1548-1556 (2006).

24. Lee, Y. N. & Bieniasz, P. D. Reconstitution of an infectious human endogenous retrovirus. *PLoS Pathogens* **3,** e10 (2007).

25. Brady, T., Lee, Y. N., Ronen, K., Malani, N., *et al.* Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23,** 633-642 (2009).

26. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860-921 (2001).

27. Wagstaff, B. J., Hedges, D. J., Derbes, R. S., Campos Sanchez, R., *et al.* Rescuing Alu: recovery of new inserts shows LINE-1 preserves Alu activity through A-tail expansion. *PLoS Genet* **8,** e1002842 (2012).

28. Heidmann, T., Heidmann, O. & Nicolas, J. F. An indicator gene to demonstrate intracellular transposition of defective retroviruses. *Proc Natl Acad Sci U S A* **85,** 2219-2223 (1988).

29. Curcio, M. J. & Garfinkel, D. J. Single-step selection for Ty1 element retrotransposition. *Proc Natl Acad Sci U S A* **88,** 936-940 (1991).

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

31. Rangwala, S. H. & Kazazian, H. H. J. The L1 retrotransposition assay: a retrospective and toolkit. *Methods* **49,** 219-226 (2009).

32. Baller, J. A., Gao, J., Stamenova, R., Curcio, M. J. & Voytas, D. F. A nucleosomal surface defines an integration hotspot for the Saccharomyces cerevisiae Ty1 retrotransposon. *Genome Res.* **22,** 704-713 (2012).

33. Mularoni, L., Zhou, Y., Bowen, T., Gangadharan, S., *et al.* Retrotransposon Ty1 integration targets specifically positioned asymmetric nucleosomal DNA segments in tRNA hotspots. *Genome Res.* **22,** 693-703 (2012).

34. de Jong, J., Wessels, L. F. A., van Lohuizen, M., de Ridder, J. & Akhtar, W. Applications of DNA integrating elements: Facing the bias bully. *Mob Genet Elements* **4,** 1-6 (2014).

35. Li, X., Ewis, H., Hice, R. H., Malani, N., *et al.* A resurrected mammalian hAT transposable element and a closely related insect element are highly active in human cell culture. *Proc Natl Acad Sci U S A* **110,** E478-E487 (2013).

36. Ji, H., Moore, D. P., Blomberg, M. A., Braiterman, L. T., *et al.* Hotspots for unselected Ty1 transposition events on yeast chromosome III are near tRNA genes and LTR sequences. *Cell* **73,** 1007-1018 (1993).

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

38. Symer, D. E., Connelly, C., Szak, S. T., Caputo, E. M., *et al.* Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* **110,** 327-338 (2002).

39. Mitchell, R. S., Beitzel, B. F., Schroder, A. R. W., Shinn, P., *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2,** E234 (2004).

40. Wang, G. P., Ciuffi, A., Leipzig, J., Berry, C. C. & Bushman, F. D. HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. *Genome Res* **17,** 1186-1194 (2007).

41. Schmidt, M., Schwarzwaelder, K., Bartholomae, C., Zaoui, K., *et al.* High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat Methods* **4,** 1051-1057 (2007).

42. Qi, X., Daily, K., Nguyen, K., Wang, H., *et al.* Retrotransposon profiling of RNA polymerase III initiation sites. *Genome Res.* **22,** 681-692 (2012).

43. Wang, G. P., Garrigue, A., Ciuffi, A., Ronen, K., *et al.* DNA bar coding and pyrosequencing to analyze adverse events in therapeutic gene transfer. *Nucleic Acids Res* **36,** e49 (2008).

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

45. Iskow, R. C., McCabe, M. T., Mills, R. E., Torene, S., *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141,** 1253-1261 (2010).

46. Witherspoon, D. J., Xing, J., Zhang, Y., Watkins, W. S., *et al.* Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* **11,** 410 (2010).

47. Shukla, R., Upton, K. R., Muñoz-Lopez, M., Gerhardt, D. J., *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153,** 101-111 (2013).

48. Rodić, N., Steranka, J. P., Makohon-Moore, A., Moyer, A., *et al.* Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat Med* **21,** 1060-1064 (2015).

49. Streva, V. A., Jordan, V. E., Linker, S., Hedges, D. J., *et al.* Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements between individuals. *BMC Genomics* **16,** 220 (2015).

50. Chatterjee, A. G., Esnault, C., Guo, Y., Hung, S., *et al.* Serial number tagging reveals a prominent sequence preference of retrotransposon integration. *Nucleic Acids Res.* **42,** 8449-8460 (2014).

51. Hickey, A., Esnault, C., Majumdar, A., Chatterjee, A. G., *et al.* Single-Nucleotide-Specific Targeting of the Tf1 Retrotransposon Promoted by the DNA-Binding Protein Sap1 of Schizosaccharomyces pombe. *Genetics* **201,** 905-924 (2015).

52. Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40,** D700-D705 (2012).

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

54. Baller, J. A., Gao, J. & Voytas, D. F. Access to DNA establishes a secondary target site bias for the yeast retrotransposon Ty5. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 20351-20356 (2011).

55. Ciuffi, A., Ronen, K., Brady, T., Malani, N., *et al.* Methods for integration site distribution analyses in animal cell genomes. *Methods* **47,** 261-268 (2009).

56. Gogol-Döring, A., Ammar, I., Gupta, S., Bunse, M., *et al.* Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4(+) T Cells. *Mol. Ther.* **24,** 592-606 (2016).

57. Fungtammasan, A., Walsh, E., Chiaromonte, F., Eckert, K. A. & Makova, K. D. A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome? *Genome Res* **22,** 993-1005 (2012).

58. Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., *et al.* A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11,** 1287-1289 (2005).

59. Bridier-Nahmias, A., Tchalikian-Cosson, A., Baller, J. A., Menouni, R., *et al.* Retrotransposons. An RNA polymerase III subunit determines sites of retrotransposon integration. *Science* **348,** 585-588 (2015).

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

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

62. Yang, J., Malik, H. S. & Eickbush, T. H. 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 (1999).

63. Christensen, S. M., Bibillo, A. & Eickbush, T. H. Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* **33,** 6461-6468 (2005).

64. Thompson, B. K. & Christensen, S. M. Independently derived targeting of 28S rDNA by A- and D-clade R2 retrotransposons: Plasticity of integration mechanism. *Mob Genet Elements* **1,** 29-37 (2011).

65. Shivram, H., Cawley, D. & Christensen, S. M. Targeting novel sites: The N-terminal DNA binding domain of non-LTR retrotransposons is an adaptable module that is implicated in changing site specificities. *Mob Genet Elements* **1,** 169-178 (2011).

66. Fujiwara, H. Site-specific non-LTR retrotransposons. *Microbiol Spectr* **3,** MDNA3-0001--2014 (2015).

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

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

69. Maita, N., Anzai, T., Aoyagi, H., Mizuno, H. & Fujiwara, H. Crystal structure of the endonuclease domain encoded by the telomere-specific long interspersed nuclear element, TRAS1. *J Biol Chem* **279,** 41067-41076 (2004).

70. Maita, N., Aoyagi, H., Osanai, M., Shirakawa, M. & Fujiwara, H. Characterization of the sequence specificity of the R1Bm endonuclease domain by structural and biochemical studies. *Nucleic Acids Res* **35,** 3918-3927 (2007).

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

72. Monot, C., Kuciak, M., Viollet, S., Mir, A. A., *et al.* The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. *PLoS Genet.* **9,** e1003499 (2013).

73. Viollet, S., Monot, C. & Cristofari, G. L1 retrotransposition: The snap-velcro model and its consequences. *Mob Genet Elements* **4,** e28907 (2014).

74. Anzai, T., Osanai, M., Hamada, M. & Fujiwara, H. Functional roles of 3'-terminal structures of template RNA during in vivo retrotransposition of non-LTR retrotransposon, R1Bm. *Nucleic Acids Res* **33,** 1993-2002 (2005).

75. Takahashi, H. & Fujiwara, H. Transplantation of target site specificity by swapping the endonuclease domains of two LINEs. *EMBO J.* **21,** 408-417 (2002).

76. Yoshitake, K. & Fujiwara, H. Creation of a novel telomere-cutting endonuclease based on the EN domain of telomere-specific non-long terminal repeat retrotransposon, TRAS1. *Mob. DNA* **1,** 13 (2010).

77. Osanai-Futahashi, M. & Fujiwara, H. Coevolution of telomeric repeats and telomeric repeat-specific non-LTR retrotransposons in insects. *Mol. Biol. Evol.* **28,** 2983-2986 (2011).

78. Dyda, F., Chandler, M. & Hickman, A. B. The emerging diversity of transpososome architectures. *Q. Rev. Biophys.* **45,** 493-521 (2012).

79. Engelman, A. & Craigie, R. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67,** 1211-1221 (1991).

80. Mizuuchi, K. & Adzuma, K. Inversion of the phosphate chirality at the target site of Mu DNA strand transfer: evidence for a one-step transesterification mechanism. *Cell* **66,** 129-140 (1991).

81. Kennedy, A. K. Single active site catalysis of the successive phosphoryl transfer steps by DNA transposases: insights from phosphorothioate stereoselectivity. *Cell* **101,** 295-305 (2000).

82. Maertens, G. N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468,** 326-329 (2010).

83. Morris, E. R., Grey, H., McKenzie, G., Jones, A. C. & Richardson, J. M. A bend, flip and trap mechanism for transposon integration. *Elife* **5,** (2016).

84. Serrao, E., Krishnan, L., Shun, M. -C., Li, X., *et al.* Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. *Nucleic Acids Res.* **42,** 5164-5176 (2014).

85. Aiyer, S., Rossi, P., Malani, N., Schneider, W. M., *et al.* Structural and sequencing analysis of local target DNA recognition by MLV integrase. *Nucleic Acids Res.* **43,** 5647-5663 (2015).

86. Leblanc, P., Dastugue, B. & Vaury, C. The integration machinery of ZAM, a retroelement from Drosophila melanogaster, acts as a sequence-specific endonuclease. *J Virol* **73,** 7061-7064 (1999).

87. Faye, B., Arnaud, F., Peyretaillade, E., Brasset, E., *et al.* Functional characteristics of a highly specific integrase encoded by an LTR-retrotransposon. *PLoS One* **3,** e3185 (2008).

88. Liao, G. C., Rehm, E. J. & Rubin, G. M. Insertion site preferences of the P transposable element in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U.S.A.* **97,** 3347-3351 (2000).

89. Vigdal, T. J., Kaufman, C. D., Izsvák, Z., Voytas, D. F. & Ivics, Z. Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. *J Mol Biol* **323,** 441-452 (2002).

90. Holman, A. G. & Coffin, J. M. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc. Natl. Acad. Sci. U.S.A.* **102,** 6103-6107 (2005).

91. Linheiro, R. S. & Bergman, C. M. Testing the palindromic target site model for DNA transposon insertion using the Drosophila melanogaster P-element. *Nucleic Acids Res.* **36,** 6199-6208 (2008).

92. Linheiro, R. S. & Bergman, C. M. Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLoS ONE* **7,** e30008 (2012).

93. Serrao, E., Ballandras-Colas, A., Cherepanov, P., Maertens, G. N. & Engelman, A. N. Key determinants of target DNA recognition by retroviral intasomes. *Retrovirology* **12,** 39 (2015).

94. Nesmelova, I. V. & Hackett, P. B. DDE transposases: Structural similarity and diversity. *Adv Drug Deliv Rev* **62,** 1187-1195 (2010).

95. Maskell, D. P., Renault, L., Serrao, E., Lesbats, P., *et al.* Structural basis for retroviral integration into nucleosomes. *Nature* **523,** 366-369 (2015).

96. Yin, Z., Shi, K., Banerjee, S., Pandey, K. K., *et al.* Crystal structure of the Rous sarcoma virus intasome. *Nature* **530,** 362-366 (2016).

97. Ballandras-Colas, A., Brown, M., Cook, N. J., Dewdney, T. G., *et al.* Cryo-EM reveals a novel octameric integrase structure for betaretroviral intasome function. *Nature* **530,** 358-361 (2016).

98. Voigt, F., Wiedemann, L., Zuliani, C., Querques, I., *et al.* Sleeping Beauty transposase structure allows rational design of hyperactive variants for genetic engineering. *Nat Commun* **7,** 11126 (2016).

99. Pryciak, P. M. & Varmus, H. E. Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection. *Cell* **69,** 769-780 (1992).

100. Pryciak, P. M., Müller, H. P. & Varmus, H. E. Simian virus 40 minichromosomes as targets for retroviral integration in vivo. *Proc Natl Acad Sci U S A* **89,** 9237-9241 (1992).

101. Muller, H. P. & Varmus, H. E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J.* **13,** 4704-4714 (1994).

102. Pruss, D., Reeves, R., Bushman, F. D. & Wolffe, A. P. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J Biol Chem* **269,** 25031-25041 (1994).

103. Pruss, D., Bushman, F. D. & Wolffe, A. P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc Natl Acad Sci U S A* **91,** 5913-5917 (1994).

104. Benleulmi, M. S., Matysiak, J., Henriquez, D. R., Vaillant, C., *et al.* Intasome architecture and chromatin density modulate retroviral integration into nucleosome. *Retrovirology* **12,** 13 (2015).

105. Pasi, M., Mornico, D., Volant, S., Juchet, A., *et al.* DNA minicircles clarify the specific role of DNA structure on retroviral integration. *Nucleic Acids Res* **44,** 7830-7847 (2016).

106. Kenna, M. A., Brachmann, C. B., Devine, S. E. & Boeke, J. D. Invading the yeast nucleus: a nuclear localization signal at the C terminus of Ty1 integrase is required for transposition in vivo. *Mol. Cell. Biol.* **18,** 1115-1124 (1998).

107. Moore, S. P., Rinckel, L. A. & Garfinkel, D. J. A Ty1 integrase nuclear localization signal required for retrotransposition. *Mol. Cell. Biol.* **18,** 1105-1114 (1998).

108. Dang, V. D. & Levin, H. L. Nuclear import of the retrotransposon Tf1 is governed by a nuclear localization signal that possesses a unique requirement for the FXFG nuclear pore factor Nup124p. *Mol. Cell. Biol.* **20,** 7798-7812 (2000).

109. Katz, R. A., Greger, J. G., Darby, K., Boimel, P., *et al.* Transduction of interphase cells by avian sarcoma virus. *J. Virol.* **76,** 5422-5434 (2002).

110. McLane, L. M., Pulliam, K. F., Devine, S. E. & Corbett, A. H. The Ty1 integrase protein can exploit the classical nuclear protein import machinery for entry into the nucleus. *Nucleic Acids Res.* **36,** 4317-4326 (2008).

111. Levin, A., Loyter, A. & Bukrinsky, M. Strategies to inhibit viral protein nuclear import: HIV-1 as a target. *Biochim. Biophys. Acta* **1813,** 1646-1653 (2011).

112. Schaller, T., Ocwieja, K. E., Rasaiyaah, J., Price, A. J., *et al.* HIV-1 capsid-cyclophilin interactions determine nuclear import pathway, integration targeting and replication efficiency. *PLoS Pathog* **7,** e1002439 (2011).

113. Koh, Y., Wu, X., Ferris, A. L., Matreyek, K. A., *et al.* Differential effects of human immunodeficiency virus type 1 capsid and cellular factors nucleoporin 153 and LEDGF/p75 on the efficiency and specificity of viral DNA integration. *J Virol* **87,** 648-658 (2013).

114. Sowd, G. A., Serrao, E., Wang, H., Wang, W., *et al.* A critical role for alternative polyadenylation factor CPSF6 in targeting HIV-1 integration to transcriptionally active chromatin. *Proc. Natl. Acad. Sci. U. S. A.* **113,** E1054-E1063 (2016).

115. Albanese, A., Arosio, D., Terreni, M. & Cereseto, A. HIV-1 pre-integration complexes selectively target decondensed chromatin in the nuclear periphery. *PLoS One* **3,** e2413 (2008).

116. Di Primio, C., Quercioli, V., Allouch, A., Gijsbers, R., *et al.* Single-cell imaging of HIV-1 provirus (SCIP). *Proc Natl Acad Sci U S A* **110,** 5636-5641 (2013).

117. Burdick, R. C., Hu, W. S. & Pathak, V. K. Nuclear import of APOBEC3F-labeled HIV-1 preintegration complexes. *Proc Natl Acad Sci U S A* **110,** E4780-E4789 (2013).

118. Marini, B., Kertesz-Farkas, A., Ali, H., Lucic, B., *et al.* Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521,** 227-231 (2015).

119. Krull, M., Brosius, J. & Schmitz, J. Alu-SINE exonization: en route to protein-coding function. *Mol. Biol. Evol.* **22,** 1702-1711 (2005).

120. Vaquerizas, J. M., Suyama, R., Kind, J., Miura, K., *et al.* Nuclear pore proteins nup153 and megator define transcriptionally active regions in the Drosophila genome. *PLoS Genet.* **6,** e1000846 (2010).

121. Lelek, M., Casartelli, N., Pellin, D., Rizzi, E., *et al.* Chromatin organization at the nuclear pore favours HIV replication. *Nat. Commun.* **6,** 6483 (2015).

122. Varadarajan, P., Mahalingam, S., Liu, P., Ng, S. B., *et al.* The functionally conserved nucleoporins Nup124p from fission yeast and the human Nup153 mediate nuclear import and activity of the Tf1 retrotransposon and HIV-1 Vpr. *Mol Biol Cell* **16,** 1823-1838 (2005).

123. Irwin, B., Aye, M., Baldi, P., Beliakova-Bethell, N., *et al.* Retroviruses and yeast retrotransposons use overlapping sets of host genes. *Genome Res.* **15,** 641-654 (2005).

124. Curcio, M. J., Lutz, S. & Lesage, P. The Ty1 LTR-Retrotransposon of Budding Yeast, Saccharomyces cerevisiae. *Microbiol Spectr* **3,** MDNA3-0053--2014 (2015).

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

126. Gangadharan, S., Mularoni, L., Fain-Thornton, J., Wheelan, S. J. & Craig, N. L. DNA transposon Hermes inserts into DNA in nucleosome-free regions in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 21966-21972 (2010).

127. Lesbats, P., Botbol, Y., Chevereau, G., Vaillant, C., *et al.* Functional coupling between HIV-1 integrase and the SWI/SNF chromatin remodeling complex for efficient in vitro integration into stable nucleosomes. *PLoS Pathogens* **7,** e1001280 (2011).

128. Naughtin, M., Haftek-Terreau, Z., Xavier, J., Meyer, S., *et al.* DNA Physical Properties and Nucleosome Positions Are Major Determinants of HIV-1 Integrase Selectivity. *PLoS One* **10,** e0129427 (2015).

129. Bachman, N., Eby, Y. & Boeke, J. D. Local definition of Ty1 target preference by long terminal repeats and clustered tRNA genes. *Genome Res* **14,** 1232-1247 (2004).

130. Taganov, K. D., Cuesta, I., Daniel, R., Cirillo, L. A., *et al.* Integrase-specific enhancement and suppression of retroviral DNA integration by compacted chromatin structure in vitro. *J Virol* **78,** 5848-5855 (2004).

131. Gelbart, M. E., Bachman, N., Delrow, J., Boeke, J. D. & Tsukiyama, T. Genome-wide identification of Isw2 chromatin-remodeling targets by localization of a catalytically inactive mutant. *Genes Dev.* **19,** 942-954 (2005).

132. Kvaratskhelia, M., Sharma, A., Larue, R. C., Serrao, E. & Engelman, A. Molecular mechanisms of retroviral integration site selection. *Nucleic Acids Res* **42,** 10209-10225 (2014).

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

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

135. De Rijck, J., Bartholomeeusen, K., Ceulemans, H., Debyser, Z. & Gijsbers, R. High-resolution profiling of the LEDGF/p75 chromatin interaction in the ENCODE region. *Nucleic Acids Res.* **38,** 6135-6147 (2010).

136. De Rijck, J., de Kogel, C., Demeulemeester, J., Vets, S., *et al.* The BET family of proteins targets moloney murine leukemia virus integration near transcription start sites. *Cell Rep.* **5,** 886-894 (2013).

137. Sharma, A., Larue, R. C., Plumb, M. R., Malani, N., *et al.* BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 12036-12041 (2013).

138. Zhu, Y., Dai, J., Fuerst, P. G. & Voytas, D. F. Controlling integration specificity of a yeast retrotransposon. *Proc Natl Acad Sci U S A* **100,** 5891-5895 (2003).

139. Ferris, A. L., Wu, X., Hughes, C. M., Stewart, C., *et al.* Lens epithelium-derived growth factor fusion proteins redirect HIV-1 DNA integration. *Proc Natl Acad Sci U S A* **107,** 3135-3140 (2010).

140. Gijsbers, R., Ronen, K., Vets, S., Malani, N., *et al.* LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Mol Ther* **18,** 552-560 (2010).

141. Silvers, R. M., Smith, J. A., Schowalter, M., Litwin, S., *et al.* Modification of integration site preferences of an HIV-1-based vector by expression of a novel synthetic protein. *Hum Gene Ther* **21,** 337-349 (2010).

142. Shun, M. -C., Raghavendra, N. K., Vandegraaff, N., Daigle, J. E., *et al.* LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* **21,** 1767-1778 (2007).

143. Marshall, H. M., Ronen, K., Berry, C., Llano, M., *et al.* Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* **2,** e1340 (2007).

144. Gupta, S. S., Maetzig, T., Maertens, G. N., Sharif, A., *et al.* Bromo- and extraterminal domain chromatin regulators serve as cofactors for murine leukemia virus integration. *J Virol* **87,** 12721-12736 (2013).

145. Singh, P. K., Plumb, M. R., Ferris, A. L., Iben, J. R., *et al.* LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev.* **29,** 2287-2297 (2015).

146. Gai, X. & Voytas, D. F. A single amino acid change in the yeast retrotransposon Ty5 abolishes targeting to silent chromatin. *Mol. Cell* **1,** 1051-1055 (1998).

147. Xie, W., Gai, X., Zhu, Y., Zappulla, D. C., *et al.* Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Mol. Cell. Biol.* **21,** 6606-6614 (2001).

148. Kirchner, J., Connolly, C. M. & Sandmeyer, S. B. Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. *Science* **267,** 1488-1491 (1995).

149. Qi, X. & Sandmeyer, S. In vitro targeting of strand transfer by the Ty3 retroelement integrase. *J. Biol. Chem.* **287,** 18589-18595 (2012).

150. Aye, M., Dildine, S. L., Claypool, J. A., Jourdain, S. & Sandmeyer, S. B. A truncation mutant of the 95-kilodalton subunit of transcription factor IIIC reveals asymmetry in Ty3 integration. *Mol Cell Biol* **21,** 7839-7851 (2001).

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

152. Cheung, S., Ma, L., Chan, P. H. W., Hu, H. -L., *et al.* Ty1 Integrase Interacts with RNA Polymerase III-specific Subcomplexes to Promote Insertion of Ty1 Elements Upstream of Polymerase (Pol) III-transcribed Genes. *J. Biol. Chem.* **291,** 6396-6411 (2016).

153. Gao, X., Hou, Y., Ebina, H., Levin, H. L. & Voytas, D. F. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* **18,** 359-369 (2008).

154. Cherepanov, P., Maertens, G., Proost, P., Devreese, B., *et al.* HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* **278,** 372-381 (2003).

155. Engelman, A. & Cherepanov, P. The lentiviral integrase binding protein LEDGF/p75 and HIV-1 replication. *PLoS Pathog.* **4,** e1000046 (2008).

156. Emiliani, S., Mousnier, A., Busschots, K., Maroun, M., *et al.* Integrase mutants defective for interaction with LEDGF/p75 are impaired in chromosome tethering and HIV-1 replication. *J. Biol. Chem.* **280,** 25517-25523 (2005).

157. De Rijck, J., Vandekerckhove, L., Gijsbers, R., Hombrouck, A., *et al.* Overexpression of the lens epithelium-derived growth factor/p75 integrase binding domain inhibits human immunodeficiency virus replication. *J Virol* **80,** 11498-11509 (2006).

158. Llano, M., Vanegas, M., Hutchins, N., Thompson, D., *et al.* Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75. *J Mol Biol* **360,** 760-773 (2006).

159. Morchikh, M., Naughtin, M., Di Nunzio, F., Xavier, J., *et al.* TOX4 and NOVA1 proteins are partners of the LEDGF PWWP domain and affect HIV-1 replication. *PLoS ONE* **8,** e81217 (2013).

160. Eidahl, J. O., Crowe, B. L., North, J. A., McKee, C. J., *et al.* Structural basis for high-affinity binding of LEDGF PWWP to mononucleosomes. *Nucleic Acids Res.* **41,** 3924-3936 (2013).

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

162. Larue, R. C., Plumb, M. R., Crowe, B. L., Shkriabai, N., *et al.* Bimodal high-affinity association of Brd4 with murine leukemia virus integrase and mononucleosomes. *Nucleic Acids Res.* **42,** 4868-4881 (2014).

163. Aiyer, S., Swapna, G. V., Malani, N., Aramini, J. M., *et al.* Altering murine leukemia virus integration through disruption of the integrase and BET protein family interaction. *Nucleic Acids Res* **42,** 5917-5928 (2014).

164. El Ashkar, S., De Rijck, J., Demeulemeester, J., Vets, S., *et al.* BET-independent MLV-based Vectors Target Away From Promoters and Regulatory Elements. *Mol Ther Nucleic Acids* **3,** e179 (2014).

165. De Ravin, S. S., Su, L., Theobald, N., Choi, U., *et al.* Enhancers are major targets for murine leukemia virus vector integration. *J Virol* **88,** 4504-4513 (2014).

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

167. Tobaly-Tapiero, J., Bittoun, P., Lehmann-Che, J., Delelis, O., *et al.* Chromatin tethering of incoming foamy virus by the structural Gag protein. *Traffic* **9,** 1717-1727 (2008).

168. Elis, E., Ehrlich, M., Prizan-Ravid, A., Laham-Karam, N. & Bacharach, E. p12 tethers the murine leukemia virus pre-integration complex to mitotic chromosomes. *PLoS Pathog* **8,** e1003103 (2012).

169. Wight, D. J., Boucherit, V. C., Nader, M., Allen, D. J., *et al.* The gammaretroviral p12 protein has multiple domains that function during the early stages of replication. *Retrovirology* **9,** 83 (2012).

170. Schneider, W. M., Brzezinski, J. D., Aiyer, S., Malani, N., *et al.* Viral DNA tethering domains complement replication-defective mutations in the p12 protein of MuLV Gag. *Proc Natl Acad Sci U S A* **110,** 9487-9492 (2013).

171. Rashkova, S., Karam, S. E., Kellum, R. & Pardue, M. L. Gag proteins of the two Drosophila telomeric retrotransposons are targeted to chromosome ends. *J Cell Biol* **159,** 397-402 (2002).

172. Rashkova, S., Karam, S. E. & Pardue, M. L. Element-specific localization of Drosophila retrotransposon Gag proteins occurs in both nucleus and cytoplasm. *Proc Natl Acad Sci U S A* **99,** 3621-3626 (2002).

173. Fuller, A. M., Cook, E. G. & Kelley, K. J. Gag proteins of Drosophila telomeric retrotransposons: collaborative targeting to chromosome ends. *Genetics* **184,** 629-636 (2010).

174. Zhang, L., Beaucher, M., Cheng, Y. & Rong, Y. S. Coordination of transposon expression with DNA replication in the targeting of telomeric retrotransposons in Drosophila. *EMBO J.* **33,** 1148-1158 (2014).

175. Matsumoto, T., Takahashi, H. & Fujiwara, H. Targeted nuclear import of open reading frame 1 protein is required for in vivo retrotransposition of a telomere-specific non-long terminal repeat retrotransposon, SART1. *Mol. Cell. Biol.* **24,** 105-122 (2004).

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

177. Siol, O., Boutliliss, M., Chung, T., Glöckner, G., *et al.* Role of RNA polymerase III transcription factors in the selection of integration sites by the dictyostelium non-long terminal repeat retrotransposon TRE5-A. *Mol Cell Biol* **26,** 8242-8251 (2006).

178. Schrijvers, R., Vets, S., De Rijck, J., Malani, N., *et al.* HRP-2 determines HIV-1 integration site selection in LEDGF/p75 depleted cells. *Retrovirology* **9,** 84 (2012).

179. Wang, H., Jurado, K. A., Wu, X., Shun, M. C., *et al.* HRP2 determines the efficiency and specificity of HIV-1 integration in LEDGF/p75 knockout cells but does not contribute to the antiviral activity of a potent LEDGF/p75-binding site integrase inhibitor. *Nucleic Acids Res* **40,** 11518-11530 (2012).

180. Chin, C. R., Perreira, J. M., Savidis, G., Portmann, J. M., *et al.* Direct Visualization of HIV-1 Replication Intermediates Shows that Capsid and CPSF6 Modulate HIV-1 Intra-nuclear Invasion and Integration. *Cell Rep* **13,** 1717-1731 (2015).

181. Leem, Y. -E., Ripmaster, T. L., Kelly, F. D., Ebina, H., *et al.* Retrotransposon Tf1 is targeted to Pol II promoters by transcription activators. *Mol. Cell* **30,** 98-107 (2008).

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

183. Ho, K. L., Ma, L., Cheung, S., Manhas, S., *et al.* A role for the budding yeast separase, Esp1, in Ty1 element retrotransposition. *PLoS Genet.* **11,** e1005109 (2015).

184. Mok, H. -P. & Lever, A. M. Chromatin, gene silencing and HIV latency. *Genome Biol.* **8,** 228 (2007).

185. Scott, E. C., Gardner, E. J., Masood, A., Chuang, N. T., *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* (2016).

186. Brouha, B., Schustak, J., Badge, R. M., Lutz-Prigge, S., *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A* **100,** 5280-5285 (2003).

187. Beck, C. R., Collier, P., Macfarlane, C., Malig, M., *et al.* LINE-1 Retrotransposition Activity in Human Genomes. *Cell* **141,** 1159-1170 (2010).

188. Tubio, J. M., Li, Y., Ju, Y. S., Martincorena, I., *et al.* Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345,** 1251343 (2014).

189. Bryk, M., Banerjee, M., Murphy, M., Knudsen, K. E., *et al.* Transcriptional silencing of Ty1 elements in the RDN1 locus of yeast. *Genes Dev.* **11,** 255-269 (1997).

190. Lavie, L., Maldener, E., Brouha, B., Meese, E. U. & Mayer, J. The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res.* **14,** 2253-2260 (2004).

191. Ke, N., Irwin, P. A. & Voytas, D. F. The pheromone response pathway activates transcription of Ty5 retrotransposons located within silent chromatin of Saccharomyces cerevisiae. *EMBO J* **16,** 6272-6280 (1997).

192. Kinsey, P. T. & Sandmeyer, S. B. Ty3 transposes in mating populations of yeast: a novel transposition assay for Ty3. *Genetics* **139,** 81-94 (1995).

193. Maldarelli, F., Wu, X., Su, L., Simonetti, F. R., *et al.* Specific HIV integration sites are linked to clonal expansion and persistence of infected cells. *Science* **345,** 179-183 (2014).

194. Cohn, L. B., Silva, I. T., Oliveira, T. Y., Rosales, R. A., *et al.* HIV-1 integration landscape during latent and active infection. *Cell* **160,** 420-432 (2015).

195. Lewinski, M. K., Bisgrove, D., Shinn, P., Chen, H., *et al.* Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J. Virol.* **79,** 6610-6619 (2005).

196. Sherrill-Mix, S., Lewinski, M. K., Famiglietti, M., Bosque, A., *et al.* HIV latency and integration site placement in five cell-based models. *Retrovirology* **10,** 90 (2013).

197. Wagner, T. A., McLaughlin, S., Garg, K., Cheung, C. Y. K., *et al.* HIV latency. Proliferation of cells with HIV integrated into cancer genes contributes to persistent infection. *Science* **345,** 570-573 (2014).

198. Coull, J. J., Romerio, F., Sun, J. M., Volker, J. L., *et al.* The human factors YY1 and LSF repress the human immunodeficiency virus type 1 long terminal repeat via recruitment of histone deacetylase 1. *J. Virol.* **74,** 6790-6799 (2000).

199. Ylisastigui, L., Lehrman, G. & Bosch, R. J. Coaxing HIV-1 from resting CD4 T cells: histone deacetylase inhibition allows latent viral expression. *AIDS* **18,** 1101-1108 (2004).

200. Williams, S. A., Chen, L. -F., Kwon, H., Ruiz-Jarabo, C. M. & Greene, W. C. NF-kappaB p50 promotes HIV latency through HDAC recruitment and repression of transcriptional initiation. *EMBO J.* **25,** 139-149 (2006).

201. Imai, K. & Okamoto, T. Transcriptional repression of human immunodeficiency virus type 1 by AP-4. *J. Biol. Chem.* **281,** 12495-12505 (2006).

202. Tyagi, M. CBF-1 promotes transcriptional silencing during the establishment of HIV-1 latency. *EMBO J.* **26,** 4985-4995 (2007).

203. Marban, C., Suzanne, S., Dequiedt, F., de Walque, S., *et al.* Recruitment of chromatin-modifying enzymes by CTIP2 promotes HIV-1 transcriptional silencing. *EMBO J.* **26,** 412-423 (2007).

204. du Chéné, I., Basyuk, E., Lin, Y. -L., Triboulet, R., *et al.* Suv39H1 and HP1gamma are responsible for chromatin-mediated HIV-1 transcriptional silencing and post-integration latency. *EMBO J.* **26,** 424-435 (2007).

205. Pearson, R., Kim, Y. K., Hokello, J., Lassen, K. & Tyagi, M. Epigenetic silencing of human immunodeficiency virus (HIV) transcription by formation of restrictive chromatin structures at the viral long terminal repeat drives the progressive entry of HIV into latency. *J. Virol.* **82,** 12291-12303 (2008).

206. Tyagi, M. & Pearson, R. J. Establishment of HIV latency in primary CD4+ cells is due to epigenetic transcriptional silencing and P-TEFb restriction. *J. Virol.* **84,** 6425-6437 (2010).

207. Imai, K., Togami, H. & Okamoto, T. Involvement of histone H3 lysine 9 (H3K9) methyltransferase G9a in the maintenance of HIV-1 latency and its reactivation by BIX01294. *J Biol Chem* **285,** 16538-16545 (2010).

208. Friedman, J., Cho, W. K., Chu, C. K., Keedy, K. S., *et al.* Epigenetic silencing of HIV-1 by the histone H3 lysine 27 methyltransferase enhancer of Zeste 2. *J Virol* **85,** 9078-9089 (2011).

209. Koiwa, T., Hamano-Usami, A., Ishida, T., Okayama, A., *et al.* 5'-long terminal repeat-selective CpG methylation of latent human T-cell leukemia virus type 1 provirus in vitro and in vivo. *J. Virol.* **76,** 9389-9397 (2002).

210. Taniguchi, Y., Nosaka, K., Yasunaga, J., Maeda, M., *et al.* Silencing of human T-cell leukemia virus type I gene transcription by epigenetic mechanisms. *Retrovirology* **2,** 64 (2005).

211. Rafati, H., Moshkin, Y., Mahmoudi, T., Parra, M. & Hakre, S. New transcription regulatory mechanisms of latent HIV LTR. *Retrovirology* **9,** O3 (2012).

212. Gérard, A., Ségéral, E., Naughtin, M., Abdouni, A., *et al.* The integrase cofactor LEDGF/p75 associates with Iws1 and Spt6 for postintegration silencing of HIV-1 gene expression in latently infected cells. *Cell Host Microbe* **17,** 107-117 (2015).

213. Sandmeyer, S., Patterson, K. & Bilanchone, V. Ty3, a Position-specific Retrotransposon in Budding Yeast. *Microbiol Spectr* **3,** MDNA3-0057-2014 (2015).

214. Spaller, T., Kling, E., Glöckner, G., Hillmann, F. & Winckler, T. Convergent evolution of tRNA gene targeting preferences in compact genomes. *Mob DNA* **7,** 17 (2016).

215. Pardue, M. L. & DeBaryshe, P. G. Retrotransposons that maintain chromosome ends. *Proc Natl Acad Sci U S A* **108,** 20317-20324 (2011).

216. Molla-Herman, A., Vallés, A. M., Ganem-Elbaz, C., Antoniewski, C. & Huynh, J. R. tRNA processing defects induce replication stress and Chk2-dependent disruption of piRNA transcription. *EMBO J* **34,** 3009-3027 (2015).

217. Kinsey, P. T. & Sandmeyer, S. B. Adjacent pol II and pol III promoters: transcription of the yeast retrotransposon Ty3 and a target tRNA gene. *Nucleic Acids Res* **19,** 1317-1324 (1991).

218. Bolton, E. C. & Boeke, J. D. Transcriptional interactions between yeast tRNA genes, flanking genes and Ty elements: a genomic point of view. *Genome Res.* **13,** 254-263 (2003).

219. McClintock, B. The significance of responses of the genome to challenge. *Science* **226,** 792-801 (1984).

220. Morillon, A., Springer, M. & Lesage, P. Activation of the Kss1 invasive-filamentous growth pathway induces Ty1 transcription and retrotransposition in Saccharomyces cerevisiae. *Mol Cell Biol* **20,** 5766-5776 (2000).

221. Todeschini, A. -L., Morillon, A., Springer, M. & Lesage, P. Severe adenine starvation activates Ty1 transcription and retrotransposition in Saccharomyces cerevisiae. *Mol. Cell. Biol.* **25,** 7459-7472 (2005).

222. Capy, P., Gasperi, G., Biémont, C. & Bazin, C. Stress and transposable elements: co-evolution or useful parasites? *Heredity (Edinb)* **85 (Pt 2),** 101-106 (2000).

223. Grandbastien, M. A., Audeon, C., Bonnivard, E., Casacuberta, J. M., *et al.* Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. *Cytogenet Genome Res* **110,** 229-241 (2005).

224. Martin, A., Troadec, C., Boualem, A., Rajab, M., *et al.* A transposon-induced epigenetic change leads to sex determination in melon. *Nature* **461,** 1135-1138 (2009).

225. Sehgal, A., Lee, C. -Y. S. & Espenshade, P. J. SREBP controls oxygen-dependent mobilization of retrotransposons in fission yeast. *PLoS Genet.* **3,** e131 (2007).

226. Feng, G., Leem, Y. -E. & Levin, H. L. Transposon integration enhances expression of stress response genes. *Nucleic Acids Res.* **41,** 775-789 (2013).

227. Brown, C. A., Murray, A. W. & Verstrepen, K. J. Rapid expansion and functional divergence of subtelomeric gene families in yeasts. *Curr Biol* **20,** 895-903 (2010).

228. Zou, S., Kim, J. M. & Voytas, D. F. The Saccharomyces retrotransposon Ty5 influences the organization of chromosome ends. *Nucleic Acids Res* **24,** 4825-4831 (1996).

229. Kitamura, T., Onishi, M., Kinoshita, S., Shibuya, A., *et al.* Efficient screening of retroviral cDNA expression libraries. *Proc Natl Acad Sci U S A* **92,** 9146-9150 (1995).

230. Hacein-Bey-Abina, S., Le Deist, F., Carlier, F., Bouneaud, C., *et al.* Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N Engl J Med* **346,** 1185-1193 (2002).

231. Aiuti, A., Slavin, S., Aker, M., Ficara, F., *et al.* Correction of ADA-SCID by stem cell gene therapy combined with nonmyeloablative conditioning. *Science* **296,** 2410-2413 (2002).

232. Cattoglio, C., Facchini, G., Sartori, D., Antonelli, A., *et al.* Hot spots of retroviral integration in human CD34+ hematopoietic cells. *Blood* **110,** 1770-1778 (2007).

233. Deichmann, A., Hacein-Bey-Abina, S., Schmidt, M., Garrigue, A., *et al.* Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. *J Clin Invest* **117,** 2225-2232 (2007).

234. Schwarzwaelder, K., Howe, S. J., Schmidt, M., Brugman, M. H., *et al.* Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo. *J Clin Invest* **117,** 2241-2249 (2007).

235. Deichmann, A., Brugman, M. H., Bartholomae, C. C., Schwarzwaelder, K., *et al.* Insertion sites in engrafted cells cluster within a limited repertoire of genomic areas after gammaretroviral vector gene therapy. *Mol Ther* **19,** 2031-2039 (2011).

236. Biasco, L., Baricordi, C. & Aiuti, A. Retroviral integrations in gene therapy trials. *Mol Ther* **20,** 709-716 (2012).

237. Cavazzana-Calvo, M., Hacein-Bey, S., de Saint Basile, G., Gross, F., *et al.* Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288,** 669-672 (2000).

238. Trobridge, G. D. Foamy virus vectors for gene transfer. *Expert Opin Biol Ther* **9,** 1427-1436 (2009).

239. Erlwein, O. & McClure, M. O. Progress and prospects: foamy virus vectors enter a new age. *Gene Ther* **17,** 1423-1429 (2010).

240. Copeland, N. G. & Jenkins, N. A. Harnessing transposons for cancer gene discovery. *Nat Rev Cancer* **10,** 696-706 (2010).

241. Narayanavari, S. A., Chilkunda, S. S., Ivics, Z. & Izsvák, Z. Sleeping Beauty transposition: from biology to applications. *Crit Rev Biochem Mol Biol* 1-27 (2016).

242. Craigie, R. Targeting HIV-1 DNA integration by swapping tethers. *Proc Natl Acad Sci U S A* **107,** 2735-2736 (2010).

243. Ivics, Z., Katzer, A., Stüwe, E. E., Fiedler, D., *et al.* Targeted Sleeping Beauty transposition in human cells. *Mol Ther* **15,** 1137-1144 (2007).

244. Voigt, K., Gogol-Döring, A., Miskey, C., Chen, W., *et al.* Retargeting sleeping beauty transposon insertions by engineered zinc finger DNA-binding domains. *Mol Ther* **20,** 1852-1862 (2012).

245. Ammar, I., Gogol-Döring, A., Miskey, C., Chen, W., *et al.* Retargeting transposon insertions by the adeno-associated virus Rep protein. *Nucleic Acids Res* **40,** 6693-6712 (2012).

6. Problem statement, scope, and approach of the study

6.1. Problem statement

6.1.1. Recurrent and independent L1 integration events in restricted genomic regions support the non-randomness of L1-mediated retrotransposition

The targeting of L1 in its host genome site is influenced by the L1 endonuclease, whose preferred site is a degenerate consensus recognition motif (5' TTTT/A 3') in a local A+T rich context (discussed in section 3.3.1.A) (Monot et al. 2013; Feng et al. 1996; Cost and Boeke 1998; Jurka 1997). However, given the abundance of such favorable sites in the genome, one would expect a relatively dispersed genomic distribution. Moreover this is in contrast with observations that specific chromosomal regions seem to be particularly susceptible to the L1 machinery and behave as hotspots of L1-insertions (Gasior et al. 2007; Wimmer et al. 2011; Amariglio et al. 1991). To date, two large genomic regions and six genomic positions (three different nucleotides in NF1 gene: c.1642 in intron 14, c.2835 in exon 21, and c.4319 in exon 33; one nucleotide in BTK gene: 12 bp before the end of exon 9; one nucleotide in codon 1526 of APC gene; and one nucleotide in codon 96 in exon V of F9 gene) have been reported to be subjected to recurrent L1 retrotransposition. Analysis of ˜100 novel L1 insertions in HeLa cells in a cell culture-based assay revealed a cluster of four novel L1 insertions into the 470 kb region with c-myc locus at the center. The c-myc locus in human is flanked by the POU5F1P1 pseudogene (~300 kb upstream) and the PVRT oncogene downstream. Retrotransposition events occurred into a known breakpoint region within 3 kb of the last coding exon of c-myc, into c-myc regulatory region, into the nearby oncogenic PVRT locus and into the POU5F1P1 pseudogene (Gasior et al. 2007) (Figure 6-1). In a breast carcinoma patient, rearrangement of the locus caused by L1 integration in the second intron of c-myc gene has been found (Morse et al. 1988). Another L1-mediated integration hotspot has been observed in neurofibromatosis type I patients. Altogether six different L1-mediated Alu insertions have been found to be clustered in a relatively small 1.5-kb region (NF1 exons 21–23) within the 280-kb NF1 gene (Wimmer et

al. 2011; Wallace et al. 1991). Furthermore, three different specific integration sites, one of them located in this cluster region, were each used twice.

Given the size of human genome, it is unlikely that two independent insertions will take place at the same nucleotide by chance (Boissinot et al. 2000). Such independent L1-mediated insertions in the same nucleotide have also been observed in BTK, APC, and F9 genes besides the NF1 gene. Retrotransposition of an SVA and an AluY sequence at exactly the same nucleotide with typical hallmarks of a retrotransposon insertion including target site duplication and a long poly A tail, have been found within the coding region of BTK, the gene responsible for X-linked agammaglobulinemia (Conley et al. 2005). A somatic L1 integration and a germline Alu integration have been reported at exactly the same location in the APC gene in two individuals (Miki et al. 1992; Halling et al. 1999). Another example of this phenomenon is provided by integration of two Alu elements from two different Alu family in the F9 gene causing severe hemophilia B (Vidaud et al. 1993; Wulff et al. 2000).

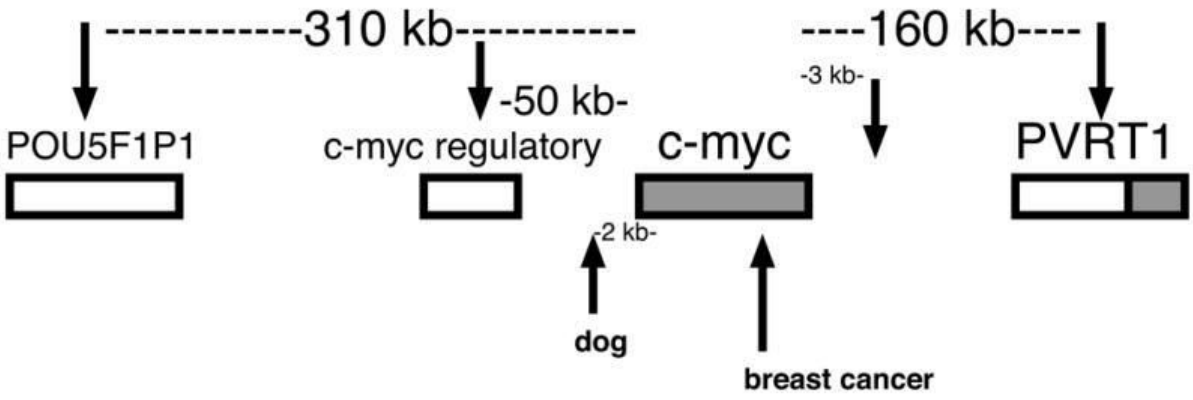

Figure 6-1.Insertions near the c-myc locus.

A schematic of the c-myc locus with 5' flanking pseudogene POU5F1P1 and 3' flanking PVRT gene is presented. The locations of 4 *de novo* L1 insertions are marked with arrows above the genes pointing down. The locations into c-myc of a somatic L1 insertion/rearrangement from a breast cancer and the site of a canine L1 insertion shown with arrows pointing up. From (Gasior et al. 2007).

6.1.2. *De novo* L1 integration sites have not been investigated in a large scale and genome wide

The extent of genomic rearrangements and the consequences of the rearrangements on host phenotype largely depend on the environment of the site where retrotransposition takes place (discussed in section 4.1 and section 4.2). In addition, the ability of an integrated retrotransposon to maintain its activity, also partly depends on its genomic environment of the retrotransposition site (discussed in the review article in section 5.1). Thereby, understanding the genomic context of the sites prone to L1 insertions will provide valuable information on the evolution of our genome and on the etiology of diseases caused by L1. The occurrences of multiple independent retrotransposon events at exactly the same nucleotide, or clusters of independent retrotransposon events in a relatively small genomic region supports the notion of a non-random phenomenon and that certain genomic sites are more vulnerable to L1-mediated retrotransposition. Moreover, little is known about the genomic context of the sites vulnerable to L1 insertions. It is likely that beside the site selectivity of the L1 EN, additional factors contribute to L1 integration site selection. A better understanding of the genomic environment that makes a site vulnerable will shed light on the mechanisms of L1-mediated insertional mutagenesis. However, apart from dispersed studies and occasional observations in disease, there had been no large scale, genome-wide, and unbiased investigation *on de novo* L1 retrotransposition sites to explore a possible preference for particular genomic location.

6.2. Goal of the study

Considering the observations described above, we wanted to test whether new L1 insertions occur randomly in the genome or not, and if unidentified features at the target site might favor L1 integration.

6.3. Experimental approach chosen

6.3.1. Study of *de novo* insertions

Analysis of the genomic environment of the L1 retrotransposition sites must be done on the flanking sequences of novel L1 retrotransposition sites as existing endogenous L1 copies have been subjected to evolutionary selective pressure due to the host-L1 interaction (discussed in section 1.3.1). Host-L1 interactions over an evolutionary time results in biased, non-random distribution of endogenous L1s, analysis of which will not provide any information in the initial site-specificity of L1 integration. Upon selection deleterious events are eliminated and harmless or profiting ones can be maintained. The distribution of younger human-specific L1 distributions is different from the distribution of the older primate specific L1 distribution (Ovchinnikov et al. 2001). This difference most likely originates from the selective pressure. Similarly, alhough human L1 and Alu elements are both mobilized by the L1 retrotransposition machinery and exploit the same endonuclease target sequence recognition bias to integrate into AT-rich sites in the host genome (Feng et al. 1996; Gasior et al. 2006; Monot et al. 2013), fixed L1 and Alu insertions show contrast in genomic distribution. L1 elements accumulate in AT rich regions whereas genomic fixed Alu elements are enriched in GC-rich regions when compared to the average genome composition (Soriano et al. 1983; Korenberg and Rykowski 1988; Moyzis et al. 1989; Boyle et al. 1990; Baker and Kass 1994). This could be explained by the contrasting characteristics of L1 and Alu of L1 and Alu sequences on genes, also in AT-rich regions and how they can be tolerated in the genome. These findings suggest that element-specific differential selective pressure is operating on L1 mediated retrotransposition events and emphasize the necessity of investigating *de novo* L1 retrotransposition sites.

6.3.2. Genomic flanks of *de novo* L1 insertions generated in cell culture have been analyzed using bioinformatic and statistical approaches

To verify our hypothesis, first, we induced novel L1 insertions in the genome. Next, we rescued the integration sites using sequencing approaches. Finally, novel L1 integration sites were examined for their overlapping or proximity with a number of genomic features.

To generate novel L1 insertions, we transfected HeLa S3 cells with an L1 containing plasmid. In this context, L1 is expressed from its native promoter, complete its life cycle and integrate in the HeLa S3 genome. The L1 element contains a retrotransposition reporter gene which provides neomycin-resistance to the host cell only upon transcription, splicing, reverse transcription, and integration. Upon integration in the genome, expression of this cassette provides resistance to a neomycin derivative. The, sequence of this cassette was also exploited to discriminate novel L1 insertions from the existing numerous L1 sequences in the genome. Integration sites are enriched by suppression PCR and sequenced by Ion Torrent sequencing. Sequencing reads are mapped in the reference genome, to locate the novel L1 insertions at nucleotide resolution. To map integration sites, we adapted an in house technique to locate *de novo* L1 insertions, originally developed in the lab, named ATLAS-seq, to locate endogenous L1 copies. Using bioinformatic and statistical tools, the proximity of the integration sites towards a large number of genomic features

6.3.3. Limitations

We used HeLa S3cells in our study, which is a transformed cell line obtained from cervical cancer tissue and adapted to grow in suspension. No other transformed cell lines, which have been well studied by the ENCODE project, were found permissive to L1 mobility to obtain a sufficient number of new insertions for our study. Thus, we could not investigate if L1 preferred integration sites vary across cell types and might be due to the availability of tissue-specific factors.

Influence of the genomic context on integration site selection by human L1 retrotransposons

Influence of the genomic context on integration site selection by human L1 retrotransposons

Sultana Tania[1], Siol Oliver[2], van Essen Dominic[1], Philippe Claude[1], Nigumann Pilvi[1], Saccani Simona[1], Gilbert Nicolas[3], Cristofari Gaël[1, 4] *

[1] IRCAN, INSERM U1081, Centre National de la Recherche Scientifique UMR 7284, University of Nice-Sophia Antipolis, 06107 Nice Cedex 2, France

[2] Institut de Génétique Humaine (IGH), Montpellier, France

[3] Institut Pasteur of Shanghai, INSERM, Xuhui District, Shanghai, China

[4] FHU OncoAge, University of Nice-Sophia Antipolis, 06107 Nice, France

* Corresponding author. E-mail: gael.cristofari@unice.fr

Keywords: non-LTR retrotransposons, LINE-1, retrotransposition, integration site preference, chromatin state, ENCODE.

Retrotransposons are mobile genetic elements that employ an RNA intermediate and a reverse transcription step for their replication. Long INterspersed Elements-1 (LINE-1 or L1) form the only autonomously active retrotransposon family in humans. Although most copies are defective due to the accumulation of mutations, each individual genome contains an average of 100 retrotransposition-competent L1 copies, which contribute to the dynamics of contemporary human genomes. L1 integration sites in the host genome directly determine the genetic consequences of the integration and the fate of the integrated copy. Thus, where L1 integrates in the genome, and whether this process is random, is critical to our understanding of human genome evolution, somatic genome plasticity in cancer and aging, and host-parasite interactions. To characterize L1 insertion sites, rather than studying

endogenous L1 which have been subjected to evolutionary selective pressure, we induced *de novo* L1 retrotransposition by transfecting a plasmid-borne active L1 element into HeLa S3 cells. Then, we mapped *de novo* insertions in the human genome at nucleotide resolution by a dedicated deep-sequencing approach, named ATLAS-seq. Finally, *de novo* insertions were examined for their proximity towards a large number of genomic features. We found that L1 preferentially integrates in the lowly-expressed and weak enhancer chromatin segments. We also detected several hotspots of recurrent L1 integration. Our results indicate that the distribution of *de novo* L1 insertions is non-random both at local and regional scales, and pave the way to identify potential cellular factors involved in the targeting of L1 insertions.

Introduction

Transposable elements are present in almost all species and significantly contribute to shape host genome structure and function. Long INterspersed Elements (LINE-1 or L1) are the only autonomously-active class of transposable element in humans. L1s belong to the non-LTR retrotransposon class and replicate in the genome by an RNA-mediated "copy-and-paste" mechanism. Our genome has approximately 500,000 copies of L1 occupying 17% of the genome, although most of these copies are functionally inactive due to the accumulation of mutations (Lander et al. 2001). However, only ~100 copies are estimated to be still retrotransposition-competent (RC-L1), some being polymorphic among individuals (Beck et al. 2010; Brouha et al. 2003), all of them belonging to the youngest human-specific L1 subfamily, L1HS (Khan et al. 2006). Finally, expression of a particular L1 copy in somatic cells is dependent on locus-, and cell-type specific determinants (Philippe et al. 2016). Thus the subset of active L1 copies vary in populations, individuals, and cell types. L1 encoded proteins, ORF1p and ORF2p, preferentially bind their own mRNA to form a ribonucleoprotein particle (RNP), a phenomenon known as cis preference (Esnault et al. 2000; Wei et al. 2001; Kulpa and Moran 2006). L1 proteins can also occasionally act in trans to mobilize non-autonomous non-LTR retrotransposons (e.g., human Alu and SVA elements) (Raiz et al. 2012; Dewannieux et al. 2003; Hancks et al. 2012; 2011) and cellular mRNAs leading to formation of processed pseudogenes (retropseudogenes) (Wei et al. 2001; Esnault et al. 2000; Doucet et al. 2015a). L1 elements are active in germ cells and early embryo (Brouha et al. 2003), occasionally leading to genetic diseases (Hancks and Kazazian 2016; 2012), but also in some somatic tissues such as brain (Erwin et al. 2016; 2014) or epithelial tumors (Ewing et al. 2015; Helman et al. 2014; Tubio et al. 2014; Makohon-Moore et al. 2015; Solyom et al. 2012; Shukla et al. 2013; Iskow et al. 2010), where they participate to tumor genome instability.

ORF1p has both nucleic acid binding and chaperone activities (Kolosha and Martin 2003; Martin and Bushman 2001). When the RNP complex reaches in the target genomic site, the endonuclease (EN) activity of ORF2p nicks the genomic DNA target at loosely defined consensus sequences at, which liberates a 5' phosphate and 3' hydroxyl group (Feng et al. 1996; Jurka 1997; Morrish et al. 2002; Cost and Boeke 1998). The liberated 3' hydroxyl group

is used as a primer by ORF2p reverse transcriptase activity (RT) to synthesize the L1 cDNA (Luan et al. 1993; Kulpa and Moran 2006; Doucet et al. 2015b; Monot et al. 2013). Second strand DNA cleavage and second strand L1 DNA synthesis can be achieved in vitro by ORF2p but have not been confirmed in vivo so far (Cost et al. 2002; Kopera et al. 2011; Piskareva and Schmatchenko 2006). Many insertions are 5' truncated due to abortive reverse transcription (Myers et al. 2002). Altogether, this process, known as target-primed reverse transcription, usually leads to short target-site duplication (4-16 bp), but can also be coupled to additional genomic rearrangements, such as target site deletion (Gilbert et al. 2002), transduction (Lander et al. 2001; Goodier et al. 2000; Pickeral et al. 2000), and exonization of target sites (Gasior et al. 2006; Erwin et al. 2016; Gilbert et al. 2002; Xing et al. 2006; Sayah et al. 2004; Kaessmann et al. 2009; Moran et al. 1999). The genetic and epigenetic consequences of L1 insertions are collectively determined by the size of the L1 insertion and its location in the genome. An L1 element carries a number of cis regulatory sequences, for example, sense and antisense promoters (Swergold 1990; Speek 2001; Nigumann et al. 2002), a number of splice sites (Belancio et al. 2006), and transcription termination signal (Moran et al. 1999). Depending on the target site, introduction of these features can considerably remodel gene structure and gene networks in a very short evolutionary time frame (Speek 2001; Wheelan 2005; Han et al. 2004). Where L1 integrates in the host genome also dictates the fate of the integrated copy (i.e., whether it can be subsequently expressed and mobilized). Thus, understanding L1 target site selection process is critical to our understanding of genome evolution, somatic genome plasticity in cancer or aging, and for host-parasite interactions.

The targeting of L1 in the genome is influenced by L1 endonuclease and reverse transcriptase activities, which show a preference for a degenerate consensus motif (5' TTTT/A 3') in a local A+T rich context (Monot et al. 2013; Feng et al. 1996; Cost and Boeke 1998; Jurka 1997). Given the abundance of such sites in the genome, one would expect a relatively dispersed genomic distribution. However, some observations suggest that specific chromosomal regions could be particularly susceptible to the L1 machinery and behave as hotspots of L1 insertions. To date, two local genomic regions (NF1 and c-myc) (Gasior et al. 2007; Wimmer et al. 2011; Amariglio et al. 1991) and six nucleotide positions have been reported to be subjected to recurrent L1 retrotransposition (three in the NF1 gene; one in the BTK gene; one in the APC gene; and one

in the F9 gene) (Conley et al. 2005; Wimmer et al. 2011; Rohrer et al. 1999; Halling et al. 1999; Miki et al. 1992; Wulff et al. 2000; Vidaud et al. 1993). Given the size of the human genome, it is unlikely that two independent insertions will take place at the same nucleotide by chance (Entezam et al. 2004). The occurrences of multiple independent retrotransposition events at exactly the same nucleotide, or cluster of independent retrotransposition events in a relatively small genomic region supports the notion of non-random L1 retrotransposition in the human genome. It is likely that beside the weak site selectivity of the L1 EN, additional factors may contribute to L1 integration site selection. A better understanding of the genomic environment that makes a site vulnerable to L1 insertion will shed light on the mechanisms of L1-mediated insertional mutagenesis. Apart from some dispersed studies and occasional observations in diseases, the landscape of *de novo* L1 insertions has not been explored in a genome-wide and unbiased manner. Considering the observations described above, we wanted to test whether new L1 insertions occur randomly in the genome or not, and whether unidentified genomic features at the target site might favor L1 integration.

To identify a potential insertion site preference, we focused on *de novo* L1 retrotransposition events. Indeed, existing endogenous L1 copies have been subjected to various selective pressures. Over evolutionary times, host-L1 interactions resulted in biased, non-random distribution of endogenous L1 copies, the analysis of which will not provide conclusive information relative to the initial site-specificity of L1 integration. This phenomenon is evidenced by the distinct distribution of the younger human-specific L1 copies from the older primate-specific ones (Ovchinnikov et al. 2001). Similarly, endogenous Alu sequences and L1 elements exhibit distinct isochore distribution, while they are both mobilized by the L1 retrotransposition machinery and both exploit the same endonuclease recognition motif (Gasior et al. 2006; Monot et al. 2013; Soriano et al. 1983; Korenberg and Rykowski 1988; Moyzis et al. 1989; Boyle et al. 1990; Baker and Kass 1994), suggesting element-specific selective processes. Therefore, to characterize L1 insertions, we induced *de novo* L1 retrotransposition by transfecting a plasmid-borne active L1 element into HeLa S3 cells, and we mapped novel insertion sites by a dedicated deep-sequencing approach, named ATLAS-seq.

Results

De novo L1 insertions display hallmarks of L1-mediated retrotransposition.

To facilitate the genomic characterization of pre-integration sites *in silico*, we performed retrotransposition assays in cell lines also studied by the Encyclopedia of DNA Elements (ENCODE) consortium, such as we could benefit from a considerable amount of publicly available genomic data obtained in the same cell types. We screened six ENCODE cell lines from tier 1 and 2 for their ability to sustain high levels of retrotransposition (K562, GM12878, HeLa S3, MCF-7, HepG2, IMR90) using the assay described in Figure 1A. Among them, the HeLa S3 cell line was the most permissive to L1 retrotransposition (1-5% of transfected cells, as observed in other permissive cell lines) and was selected to obtain a large number of independent retrotransposition events (see details in 'Materials and Methods' section). We induced retrotransposition from a plasmid-borne active L1 element, which is expressed by its natural promoter, and contains a neomycin-resistance gene (NeoR) in its 3' untranslated region. This genetic marker allows us to discriminate new copies from endogenous ones and to select cells containing retrotransposition events. Of note, this retrotransposition cassette only becomes functional upon transcription, reverse transcription and integration (Figure 1A and (Freeman et al. 1994; Moran et al. 1996)). Then, we adapted ATLAS-seq, a deep-sequencing approach, originally developed to map endogenous L1 elements genome-wide (Philippe et al. 2016), to locate new L1 insertions. By applying ATLAS-seq, we mapped 1136 de novo L1 insertion sites from 24 independent populations (an example of sequencing reads mapping and insertion site calling is illustrated in Figure 1B), which were used in subsequent downstream analyses (Figure S 1). 45% of the de novo L1 insertions were recovered from a single non-redundant read spanning the NeoR-genome junction. Among them 43% were reproducibly found in duplicate libraries generated from the same pool of cells (Figure S 2A and Figure S 2B). When considering all insertions, independently of the number of reads supporting them, as much as 65% were reproducibly found in duplicate libraries (Figure S 2C), consistent with the idea that any given insertion is present in very few cells, possibly in a single cell. Validation of 67% of de novo L1 somatic insertions have been reported previously (Solyom et al. 2012). In an attempt to detect rare insertions possibly representing the late retrotransposition events, which are undetectable or irreproducible by conventional PCR

method, we used digital droplet PCR (ddPCR) (White et al. 2014). In ddPCR, input DNA along with the PCR reagents of each reaction mix is partitioned into approximately 20,000 droplets as a water-in-oil emulsion (Pekin et al. 2011). Some droplets contain no copies of the template target DNA while others contain one or more (Pekin et al. 2011). Thus, it is possible to amplify a single insertion event in the template DNA inside a droplet and quantify amplification by counting fluorescent droplets. In our assay, identification of amplified *de novo* L1 insertion junctions were achieved through fluorescence analysis of EvaGreen dye intercalated into the double stranded amplicons. Indeed, as much as 227 insertion sites were recovered from a single pool of cells, which is more than five-fold higher than the insertions recovered by conventional PCR from the same pool of cells (Figure S 2). *De novo* L1 integration sites display the known hallmarks of L1 retrotransposition such as a typical endonuclease consensus sequence (Figure 1C) and the presence of a polyA tail (Figure 1B and Figure S 4). The presence of target site duplications could not be tested since only the 3' junction is sequenced.

L1 inserts non-randomly in the human genome.

De novo L1 insertions are uniformly distributed among the chromosomes when normalized by chromosome length (linear regression test, R^2= 0.8071, p<0.0001) (Figure 2A). We did not observe any orientation bias of *de novo* L1 insertions, 51.14% of integrations were in sense orientation and 48.86% were in the antisense orientation (binomial test, p=0.4583) (Figure 2B). However, when scanning the genome by small 1Mb-windows for *de novo* L1 insertions, we detected hotspots containing as many as 6 insertions (Figure 2C). One particular hotspot located on chromosome 1p11.2 carried 6 insertions originating from five independent retrotransposition assays (Figure S 4). Among these six insertions, four are very close (i.e., less than 1 kb apart, two in each orientation, Figure S 4). The observed frequency of 1Mb genomic windows that contain 0 to 6 *de novo* L1 insertions significantly deviates from the expected frequency (chi square test, $p<2.2*10^{-16}$) (Figure 2D and Supplementary table 2). While examining larger genomic windows for hotspots, six 10Mb regions were found to contain more than 10 *de novo* L1 insertions (Figure S 3). This observation supports an additional layer of regulation influencing L1 integration *in vivo*, apart from the specificities of L1 EN and RT. Apart from these local hotspots, we examined the overall spacing of *de novo* L1 insertions. Distances between two adjacent insertions were computed for both *de novo* L1 insertions and

in silico generated random insertions. *De novo* L1 insertions are more spaced than expected (Kolmogorov-Smirnov test, $p<2.2\times10^{-16}$) (Figure 2E). However, ~3 % (30 out of 1,113) of *de novo* L1 insertions are located within 100 bp of another one, in contrast to 0.0143 % (16 out of 111,300) of random insertions. Thus, *de novo* insertions significantly deviate from the randomness and some genomic locations might be more attractive than others.

De novo L1 insertions are enriched in ENCODE annotated low-expression and weak enhancer chromatin segments.

To identify potential regulators of L1 integration, we examined the extent of association of L1 insertion sites with functional genomic features (see 'association analysis' in 'Materials and Methods' section), including gene bodies of different categories, promoters, enhancers, exons, CpG islands, transcription start sites, nuclear lamina binding sites, DNA hypersensitive sites, multiple histone marks, nucleosome sites, replication timing, repeat elements, transposon free regions, and chromatin segments. Through this approach, we found enrichment of *de novo* L1 insertions in low activity chromatin segments (Low), characterized as regions of low frequency of epigenetic signals, low level of transcription, and proximal to active elements (Hoffman et al. 2013; Ernst et al. 2011). De novo L1 insertions were also enriched in candidate weak enhancers (EnhW). In contrast, HeLa S3 endogenous L1 insertions (data obtained from (Philippe et al. 2016)) are enriched in quiescent chromatin segment (Quies), characterized as neutral chromatin regions with near-zero epigenetic and transcription signals (Figure 3A). Together, Low and Quies state comprise the majority of the genome. These observations are in agreement with the association between *de novo* L1 insertions and histone marks (data obtained by ENCODE/Broad) (Figure 3B). *De novo* L1 insertions are enriched in genomic regions containing histone marks associated with transcriptional activation (H3K4me1, H3K4me2, H3K27ac, and H3K36me3), in contrast to HeLa S3 endogenous L1 insertions displaying no association with these histone modifications (Figure 3B). As a control of our computational approach, we also analyzed publicly available *de novo* insertion datasets obtained for other classes of transposable elements or retroviruses, with known target site preference (LaFave et al. 2014; Gogol-Döring et al. 2016; Ikeda et al. 2007). As previously found, de novo HIV insertions are enriched in transcriptionally active units (Gen3', Gen5', Elon, and ElonW ENCODE chromatin states; H3K36me3, H3K79me2

histone marks) (Ikeda et al. 2007; Schroder et al. 2002; Singh et al. 2015); de novo MLV insertions are enriched in promoters and enhancers (Enh, EnhF, PromF, Tss, TssF ENCODE chromatin states) and depleted in quiescent chromatin regions (LaFave et al. 2014; Wu et al. 2003; Mitchell et al. 2004); sleeping beauty DNA transposon insertions are only slightly enriched in transcriptional units (Gogol-Döring et al. 2016; Liu et al. 2005). We also generated two additional *in silico* control datasets for this analysis, named 'Background' and 'Random'. 'Background' corresponds to random genomic sites with base composition matching *de novo* L1 insertions (see details in 'Materials and Methods' section), while 'Random' dataset represents completely random set of genomic coordinates from the reference genome. As expected, none of these two datasets showed association with any of the chromatin states or histone marks (Figure 3).

De novo L1 insertions are depleted in genes.

Only 35% of *de novo* L1 insertions occurred in RefSeq genes, although the latter represent 46% of the genome, indicating a moderate depletion (Fisher's exact test, $p<0.0001$) (Figure 4A). Endogenous HeLa S3 L1 insertions were further depleted in genic regions (27%), presumably due to post-integrative negative selection (Fisher's exact test, $p<0.0001$). Genic *de novo* L1 insertions are equally oriented relative to RefSeq genes (binomial test, $p=0.08$) (Figure 4B), in contrast to endogenous L1 insertions, which are enriched in the antisense orientation (Han et al. 2004). Since a majority of disease-causing L1 insertions are in the sense orientation relative to the disrupted gene (Chen et al. 2005), this suggests that sense insertions are more likely to be detrimental and counter-selected after integration. Since *de novo* L1 insertions are depleted in genic regions but enriched in low expression chromatin segment, we tested whether new genic insertions preferentially integrate in genes with a particular level of expression. To this end, we measured the overlap of *de novo* L1 genic insertions with genes with different expression levels (see 'Association analysis' in 'Materials and Methods' section), categorized as quantiles of ENCODE HeLa S3 RNA-seq expression data higher than 0 FPKM. We did not observe any significant association of *de novo* L1 insertions and endogenous insertions with any particular gene expression category, in contrast to HIV or MLV insertions, which are both strongly enriched in highly expressed genes (FPKM>13) (Figure 4C). Association of *de novo* L1 insertions with exon density were also examined following the same method.

No association of *de novo* L1 insertions with any particular level of exon density was observed. Both HIV and MLV insertions displayed enrichment in regions highly occupied with exons, categorized by more than 9.3% exon occupancy.

Discussion

L1 retrotransposition contributes to shape genome structure and function, and sometimes can be pathogenic (reviewed in (Hancks and Kazazian 2016)), the various consequences being determined by the nature of insertion sites. Little is known about L1 target site preference and most information originates from studies of either disease causing L1 insertions or endogenous L1 copies. For unbiased study of target site preference, it is critical to limit the effects of post-integration selective phenomena. Thus, to investigate if L1 displays preference in targeting genomic sites, we induced de novo L1 insertions ex vivo and analyzed their genomic distribution. Using high throughput sequencing of novel L1 integrated HeLa S3 genomes, we characterized a set of more than one thousand de novo L1 somatic insertion sites. Congruous with earlier observations (Gasior et al. 2007; Wimmer et al. 2011), we detected a number of regions containing 5 or 6 insertions in a genomic window of maximum 1Mb. These insertions were well supported by sequencing reads spanning the L1-genome junctions, by the presence of EN-specific cleavage sites and polyA tail. De *novo* L1 insertions were enriched in lowly expressed chromatin segments, in weak enhancer candidates and depleted in genic regions. We defined a 'Background' control dataset that represents genome wide potential sites for L1 insertions, matching with base composition of the *de novo* L1 inserted sites. Association of *de novo* L1 insertions did not parallel with background insertions in respect to the features analyzed (chromatin states, gene expression level, or exon density). This indicates that EN preference does not solely define L1 integration sites and additional factors might be involved to make certain genomic regions more permissive than others.

Distribution of endogenous L1 copies differed from the *de novo* L1 ones, presumably due to active selective pressure on endogenous copies (Lander et al. 2001). Studies on other classes of transposable elements comparing the de novo versus the fixed insertions, or the younger versus the older insertions also evidenced differences in insertion distribution (Brady et al. 2009; Ovchinnikov et al. 2001; Barr et al. 2005). We observed enrichment of endogenous L1

copies in dead or quiescent chromatin states, in contrast to the enrichment of *de novo* insertions in low expression chromatin. Stronger depletion in genic regions relative to *de novo* insertions was also found. We could not detect any association of endogenous L1 copies with histone marks, gene expression level, or exon density. It seems that endogenous L1 copies have been cleared from genomic regions associated with any important function. Distribution of transposable elements in genome arise both from their integration site preferences and from a variety of selective pressures. Indeed, our analysis on endogenous copies suggest that deleterious L1 insertions have been lost from the genome over an evolutionary period regardless the initial insertion distribution. Similarly, active selective force on L1 insertions in diseased cells also result in distinct distribution bias relative to the initial integration bias. Analysis of 2756 somatic L1 insertions from 290 tumors showed accumulation of insertions in heterochromatin, possibly because genic insertions were deleterious to the cancer clones and therefore subjected to negative selection (Tubio et al. 2014).

Targeted integration is known for many transposable elements across a number of eukaryotic genomes. Criteria known to be associated with guiding most classes of non-randomly distributed transposable elements are local DNA sequences (Liao et al. 2000; Serrao et al. 2015; Holman and Coffin 2005; Linheiro and Bergman 2008; 2012; Serrao et al. 2014; Aiyer et al. 2015; Maertens et al. 2010), chromatin contexts (Hickey et al. 2015; Baller et al. 2012; Maskell et al. 2015; Mularoni et al. 2012; Lesbats et al. 2011), cellular proteins(Qi et al. 2012; Sharma et al. 2013; Bridier-Nahmias et al. 2015; De Rijck et al. 2010; Ciuffi et al. 2005; Singh et al. 2015), and the 3D organization of the nucleus (Marini et al. 2015; Lelek et al. 2015). To verify the sensitivity and reproducibility of our statistical approaches, we analyzed publicly available de novo insertion data sets of two retroviruses (HIV, MLV) and DNA transposon (Sleeping Beauty) for which target site preference have been extensively studied. Our approaches successfully reproduced the preferred target sites of these elements. As expected, HIV and MLV were respectively enriched in transcription units and in cis regulatory sequence (Singh et al. 2015; Schroder et al. 2002; Sharma et al. 2013; LaFave et al. 2014). Sleeping beauty transposes in a comparatively random manner with slight enrichment in transcription units (Liu et al. 2005; Gogol-Döring et al. 2016), which was also detected in our analysis.

Our results might be influenced by two technical limitations. First, sufficient number of de novo insertions to saturate the genome could not be obtained because of very low frequency of L1 retrotransposition. Thus, we might have missed or underestimated genomic features involved in L1 targeting. Second, we used a neomycin resistance reporter gene to select cells with novel integrations as insertions in non-selected cells were extremely diluted to detect., However, we did not observe any detectable influence of the reporter cassette, for e.g., enrichment or depletion of *de novo* L1 insertions respectively in transcriptionally permissive or repressed chromatin states. In a study on piggyBac insertions De jong et al. showed that a sample size of 120 integrations was sufficient to distinguish the influence of reporter gene expression (de Jong et al. 2014). Using a reporter can reduce statistical power, thus size of the dataset is important to detect features weakly associated with insertions.

A major determinant of integration site selection by a transposable element is the tethering of the integration machinery at the site of integration by element-specific cellular DNA- or chromatin-binding proteins, termed as tethering model (Bushman 2003). Evidences of tethering model has been demonstrated for a number of LTR retrotransposons and retroviruses catalyzed by the interaction of mostly integrase (in some cases Gag or ORF1p) of retrotransposition machinery with element specific host cellular partners which guide the retrotransposition machinery to their DNA binding sites (Devine and Boeke 1996; Baller et al. 2012; Gai and Voytas 1998; Xie et al. 2001; Hickey et al. 2015; Mularoni et al. 2012; Bridier-Nahmias et al. 2015; Serrao et al. 2015; Sharma et al. 2013; Aiyer et al. 2015; LaFave et al. 2014). Although evidence for tethering of non-LTR retrotransposons are limited, all known cases so far involve ORF1p (or sometimes Gag by analogy with LTR-retroelements) {Rashkova 2002a; Rashkova 2002b; Fuller 2010; Zhang 2014}. Whether ORF1p-mediated tethering to their target sites can be generalized to most non-LTR retrotransposons remains unknown. Our data would require further analyses to verify the involvement of cellular partners in non-random targeting of L1: i) identification of motifs in the flanking DNA of the novel insertion sites which might correspond to direct or indirect chromatin reader binding site, ii) study interaction of L1 proteins with potential chromatin reader hits, iii) validation with *in vitro* or *ex vivo* approaches.

L1 is the only autonomously-active class of transposable element and responsible for almost all the retrotransposition activities in our genome. A number of host defense mechanisms restrict L1 retrotransposition in somatic cells, thus limiting L1-mediated insertional mutagenesis. However, L1 can occasionally bypass restriction and retrotranspose with various frequencies in germ cells, early embryos, brain cells, and in epithelial tumors. Understanding how targeted distribution is achieved at the molecular level is critical to our understanding of genome evolution, genome plasticity, and host-parasite interactions. We characterized *de novo* L1 insertions across the genome with a large number of features and found non-random distribution of L1. Future studies will be focused on understanding the mechanisms leading to this non-random distribution.

Materials and Methods

Cell Culture

ENCODE tier group 1 (K562, GM12878), tier group 2 (HeLa S3, MCF7, HepG2, IMR90) cell lines were used to verify their permissibility to plasmid borne L1 retrotransposition activity. Cell lines were maintained in a tissue culture incubator (37'C at a 5% CO2 level) in Dulbecco's modified Eagle medium (DMEM) containing 4.5 g/L D-Glucose, 110 mg/L Sodium Pyruvate, and supplemented with 10% FBS, 100 U/mL penicillin, and 100 mg/mL streptomycin. Growth medium was also supplemented with 862 mg/mL L-Alanyl-L-Glutamine (Glutamax), or 2mM Glutamine.

Plasmid constructs

pCEP4 backbone vector with the active human L1.3 clone and a neomycin resistant indicator cassette (NeoR) at the 3' end of the L1 (Figure 1 and (Moran et al. 1996)) was used to transfect cells. L1.3 was expressed by its natural promoter. NeoR allowed discrimination of de novo copies from endogenous L1 elements in our genome. As negative control of L1 retrotransposition activity, we used an identical plasmid with a point mutation in the reverse transcriptase domain, which completely blocks retrotransposition. As a transfection control, we used a phrGFP (Stratagene) plasmid.

Oligonucleotides

Oligonucleotides, described Supplementary table 1, were synthesized by Integrated DNA Technologies (IDT, Coralville, IA).

The L1 Retrotransposition Assay

The cultured cell retrotransposition assay was conducted as described previously (Moran et al. 1996). Briefly, 2×105 cells/well were plated in 6-well plates. The next day, cells were transfected with 1 μg of plasmid DNA and 3 μL of Lipofectamine 2000 (Life Technologies) diluted in 200 μL of Opti-MEM (Life Technologies). Medium was replaced with fresh medium after 5 hr. For retrotransposition assays in T75 and T175 flasks, 2×10^6 and 5×10^6 cells were plated, respectively. Two days post-transfection, medium was supplemented with G418 (Life Technologies) at 400 μg/mL to select for retrotransposition events. The media was changed daily. After 10 days of selection, surviving cells in one well per batch of retrotransposition assay was washed with 1X Phosphate-Buffered Saline (PBS), fixed, and stained with crystal violet to visualize colonies. If stained cells surviving the G418 selection were present in the well, cells were collected from other wells. gDNA was extracted using a QiaAmp DNA Blood mini kit (Qiagen). In parallel, HeLa S3 cells were plated in 6-well plates and transfected with 0.5 μg of the same plasmid and hrGFP (Stratagene). Three days post-transfection, cells were subjected to flow cytometry and the transfection efficiency was determined based on the number of GFP positive cells by FACS.

Library preparation and high throughput sequencing

Mechanical fragmentation, end-repair, A-tailing, and adapter ligation. Libraries were prepared as described previously (Philippe et al. 2016). Briefly, 1μg of genomic DNA was sonicated for 6-12 cycles (5s on/90s off) at 4˚C with a Bioruptor sonicator (Diagenode), to reach an average fragment size of 1200 bp. DNA ends were repaired using the End-It DNA End-Repair Kit (Epicentre, Madison, WI). A-tailing of the repaired blunt ends was performed with Klenow Fragment (3'-to-5' exo-, New Englands Biolabs, Ipswich, MA) following the manufacturer's protocol. Adapter and dummy oligonucleotides were ligated to the A-tailed DNA. Between each of the above steps, DNA was purified with Agencourt AMPure XP beads

(Beckman Coulter, Brea, CA) using a 0.8:1 ratio of beads to DNA solution (v/v) and DNA was quality-controlled by Bioanalyzer 2100 (DNA high sensitivity kit, Agilent Technologies, Santa Clara, CA).

Library preparation by suppression PCR. Junctions of novel L1 3' end and genome were selectively enriched by suppression PCR. To reduce PCR stochasticity, the ligated genomic DNA of each sample was amplified in 8 independent parallel reactions of 40 µL each, containing 20 ng of ligated genomic DNA under the following cycling conditions: 1 cycle at 95˚C for 4 min; followed by 30 cycles at 95˚C for 30 s, 68˚C for 30 s, and 72˚C for 1 min; and a final extension step at 72˚C for 7 min. Primers are described in Supplementary table 1. Each primer pair contains trP1 and oligoA fragments, to be used for subsequent Ion Torrent library quantification and Ion Torrent sequencing. PCR products from the 8 reactions corresponding to the same population were pooled.

Library preparation in emulsion. One of the libraries was amplified by digital droplet PCR (ddPCR) in parallel to the PCR method described above to verify if stochasticity in PCR amplification could be reduced and the complexity of the reactions could be preserved. Massive partitioning of template DNA into 20,000 droplets allow capturing of late and rare retrotransposition events and also minimize over-amplification of insertions. ddPCR amplification was done using QX200 ddPCR EvaGreen supermix from Bio-Rad under the following cycling conditions: 1 cycle at 95˚C for 5 min; followed by 30 cycles at 95˚C for 30 s, 64˚C for 1 min; followed by signal stabilization of 1 cycle at 4˚C for 5 min, 1 cycle at 90˚C for 5 min (Bio-Rad's C1000 touch thermal cycle). For each sample, 9 reactions were done. For control of amplification, droplets of one reaction were read with a QX200TM droplet reader and analyzed with the QuantaSoft software. Droplets from the remaining 8 reactions were pooled and amplified DNA was extracted from the droplets by the chloroform extraction method.

Size selection. Pooled amplicons from either PCR method were subjected to double size selection to retain amplicons ranging between 300 and 450 bp by two consecutive Agencourt AMPure XP bead purifications using beads-to-DNA ratios of 0.6:1 and 0.7:1, respectively. The supernatant of the first bead purification using beads:DNA ratio of 0.6:1 contains DNA

fragments larger than 300 bp. This supernatant is applied to a second selection step with a beads:DNA ratio of 0.7:1 (i.e. addition of 0.1X beads to the supernatant), where fragments smaller than 450 bp are bound to the beads and subsequently eluted. To eliminate any traces of primers, a last step of purification using beads to DNA ratio of 1:1 was performed.

Library quantification. Each library was quantified for copy number using a quantitative PCR based assay (library quantification kit for Ion Torrent, Kappa Biosystems, Wilmington, MA). Average amplicon length was quantified by Bioanalyzer 2100 (DNA high sensitivity kit, Agilent Technologies, Santa Clara, CA). Library concentration was deduced from amplicons' average length and copy number.

Ion Torrent PGM sequencing. For sequencing, three to five libraries were pooled in equimolar amounts (final concentration of 20 pM). Emulsion PCR and enrichment for positive Ion Sphere Particles (ISPs) was performed on the Ion OneTouch 2 and ES enrichment modules, respectively, using the Ion PGM Template OT2 400 Kit (Life Technologies), and sequenced on the Ion Torrent PGM, using the Ion PGM Sequencing 400 Kit and Ion 318 v2 Chips (Life Technologies), according to the manufacturer's protocols.

Integration site mapping

Ion Torrent sequencing reads were processed and mapped to the reference human genome (hg19) in order to locate de novo L1 insertion sites, using a modified ATLAS-seq pipeline (Philippe et al. 2016), summarized below (see Figure S 1).

FASTQ files were de-multiplexed according to the sample-specific barcode using cutadapt (https://github.com/marcelm/cutadapt). Reads from each barcoded library were then trimmed using cutadapt to remove barcodes, ATLAS-seq primers, and adapters. Trimmed reads were mapped to the hg19 human reference genome using the Burrows-Wheeler Aligner (BWA) program with the 'mem' algorithm allowing softclipping (Li and Durbin 2010). Mapped reads were filtered to remove secondary alignment and ambiguously mapped reads (MAPQ 20) using SAMtools (Li et al. 2009). Softclipped reads with a polyA or polyT at the junction were recovered and insertion sites were called based on softclipped position for each read.

PCR duplicate reads were removed with Picard tools (http://broadinstitute.github.io/picard, MarkDuplicates function), keeping only the longest representative read. Reads were considered redundant if they started from the same linker position, which corresponds to the initial genomic DNA break during sonication. Same insertion sites from independent pool of cells which were sequenced together were merged into clusters using BEDtools (Quinlan and Hall 2010).

Generation of controls datasets

In this study, we used two different control datasets, called 'Random' and 'Background'. We generated an *in silico* insertion dataset of random locations from the reference genome, which is called 'random' and which we compared with the *de novo* L1 insertions. Equal number of random locations as the number of *de novo* L1 insertions were randomly picked 100 times from the genome using the random function of the bedtools package to generate an exhaustive random control.

Background dataset was generated keeping in mind that association of certain genomic features with L1 insertions may actually originate from the affinity of L1 EN for 5'-TTTT/A-3' consensus sequence without having any true association with that feature itself. Base composition around the integration sites may cause bias for a genomic feature although the feature itself has no association with L1 insertion. To verify that, we have generated a base composition matched control. 10 nt up- and down-stream flanking DNA sequences of 1136 L1 insertions were extracted from the reference genome. For each of these 20 nt-flanking DNA sequences, 43 DNA sequences of matched base composition were extracted from the reference genome using homer 2 package (-dumpFasta option). The base composition matched fasta sequences were then mapped to the genome to locate the genomic coordinates of L1 preferred local sequences in our genome using bwa, samtools, and bedtools package.

Association analysis

A perl script was used to compare the frequency of overlap between the L1 integration dataset and a given genomic feature, with the frequency of overlap between a random set of

chromosomal coordinates and the same feature (Figure 2, Figure 4C and Figure 4D). In this script, the level of overlap between *de novo* L1 integration sites is compared with a variety of genomic features, and ranked according to the level of statistical significance of the overlap relative to that expected for a random distribution. Dataset containing the chromosomal coordinates of insertion sites and the datasets containing the chromosomal coordinates of genomic markers were randomized 1000 times and each randomized set was compared to find overlap. For each randomization, the positions of chromosomes are shuffled and then the same coordinates as in the subject and query datasets are compared to find overlap. The significance of each overlap is expressed as a Z-score, calculated as the number of standard deviations by which the observed similarity between datasets differs from the similarity level expected by chance.

Statistical tests

Chromosomal distribution of *de novo* L1 insertions was tested by linear regression analysis and multinomial test (Figure 2A). Strand distribution of *de novo* L1 insertions and *de novo* L1 genic insertions were tested by binomial test (Figure 2B, Figure 4B). Expected *de novo* L1 insertion distribution across the genomic bins was derived from binomial distribution, and chi square test was done to compare with the observed distribution (Figure 2D). Kolmogorov-Smirnov test of cumulative frequency was done to compare the distances between adjusted *de novo* L1 insertions to the distances of random insertion dataset (Figure 2E). Fisher's exact test was done to compare enrichment or depletion of *de novo* L1 insertions in genes (Figure 4A). Chi square test for given probabilities was done to test for deviation of *de novo* L1 insertion distribution in different gene expression category and exon density category from random insertions. Categorization of gene expression and exon density were done per quantile in *R* (Figure 4C and Figure 4D).

Figure Legends

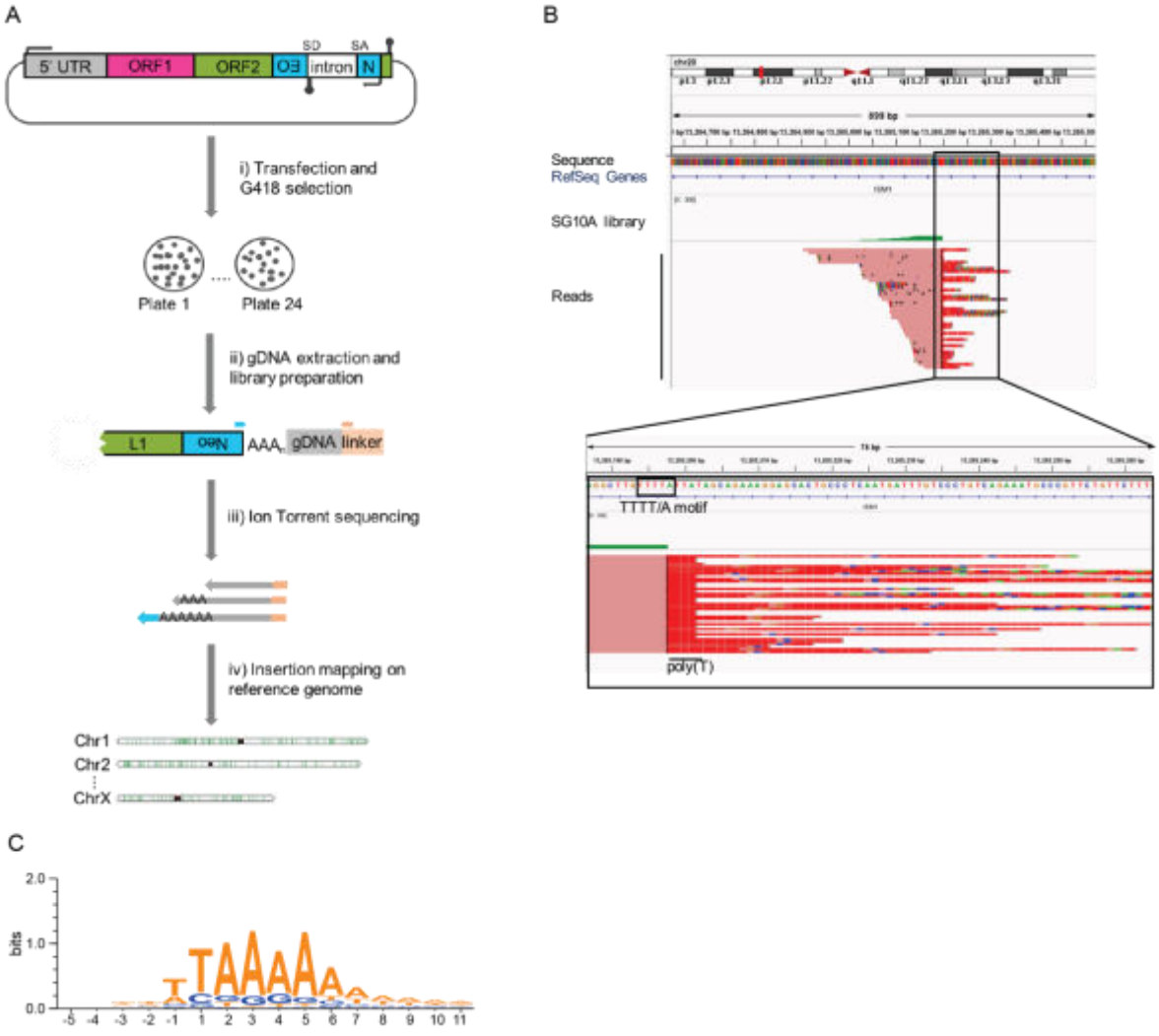

Figure 1. De novo L1 integration shows typical hallmarks of L1 retrotranposition.

(A) Overall experimental workflow. (i) *De novo* L1 retrotransposition was induced by transfecting transformed cells with an episomal active L1. Cells with new L1 insertions were selected by G418. This process was repeated to obtain 24 independent cellular populations. (ii) Genomic DNA (gDNA) was extracted from cells surviving the G418 selection and L1 3' junctions with host DNA were selectively amplified to prepare deep-sequencing libraries. (iii) Amplified junctions were then sequenced by Ion Torrent sequencing technology. (iv) Reads were mapped on the human reference genome hg19 to locate the chromosomal coordinates of integration sites using an adapted ATLAS-seq bioinformatic pipeline. (B) Integrative Genome Viewer screenshot of aligned ATLAS-seq reads on chromosome 20 supporting an antisense L1 insertion (bottom). Reads contain two parts, an aligned region corresponding to the flanking genomic sequence and an unaligned segment absent from the genome (softclipped) and corresponding to the inserted L1 copy. A polyT (or polyA) is found at the junction. The soft-clipped region of the reads is shown in color (base code: T, red; A, green; C, blue; G; orange). Integration site contains endonuclease recognition motif (5' TTTT/A 3'), and integrated L1 is followed by a polyA tail (polyT here since L1 is located on the reverse genomic

strand). Note that in long polyT homopolymeric sequences, indels are frequent and thus the L1 sequence next to the polyT is not phased in all reads. (C) Consensus sequence motif at *de novo* L1 integration sites corresponds to L1 endonuclease recognition sequence. Motif was generated with WebLogo 3.5.0.

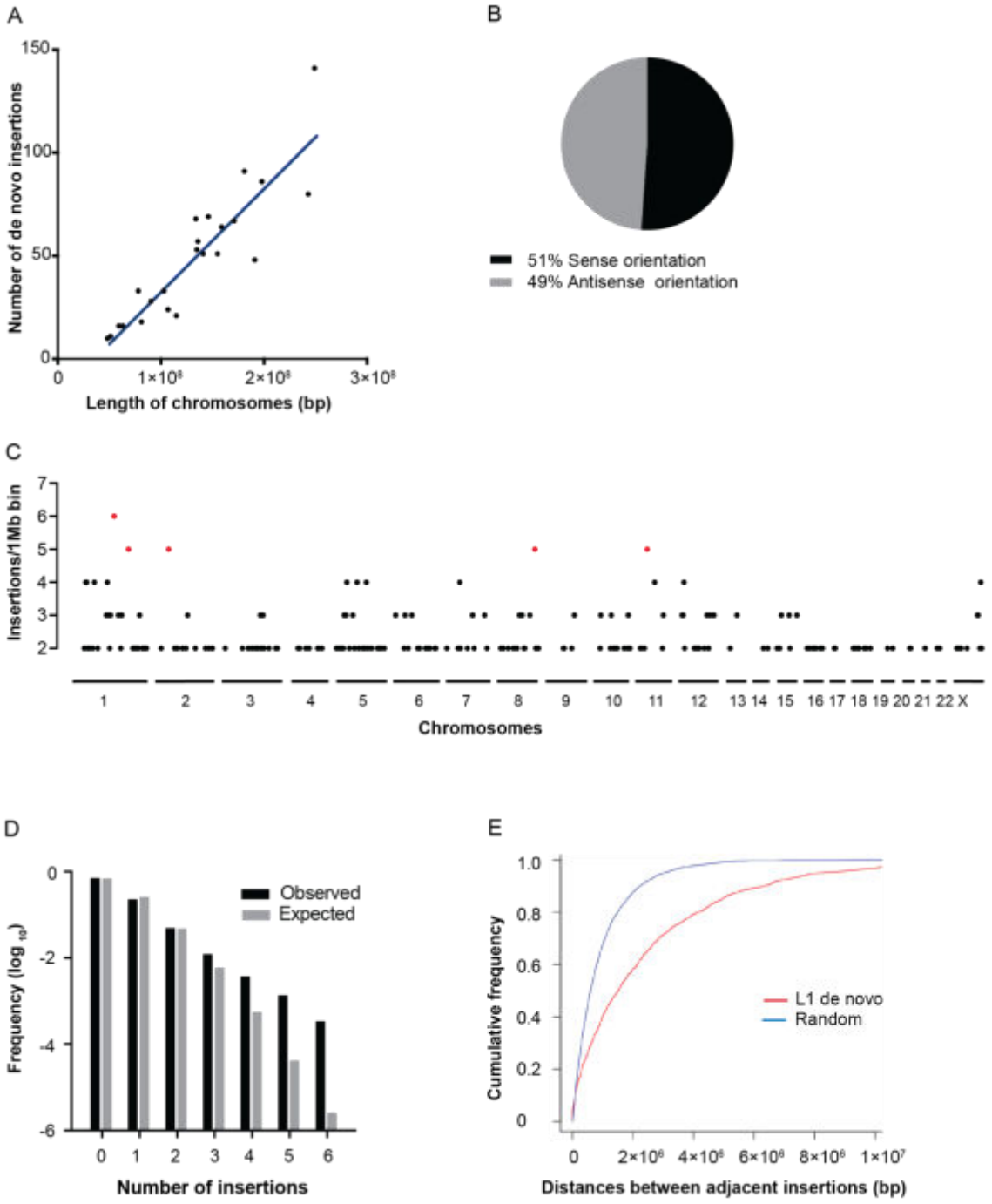

Figure 2. Distribution of de novo L1 integration is non-random.

(A) *De novo* L1 insertions are uniformly distributed across the chromosomes when normalized by the length of chromosomes (linear regression test, R^2= 0.8071, p<0.0001). (B) *De novo* L1 orientation. *De novo* L1 insertions are evenly distributed in both sense and antisense orientation (binomial test, p=0.4583). (C-D) L1 integration hotspots. (C) Dots show the number of integrated L1 per 1Mb genomic bin. Red dots, over-represented regions that contain a cluster of five or more *de novo* L1 insertions. (D) The expected and observed

frequency of 1Mb genomic window that contain 0 to 6 *de novo* L1 insertions. The observed frequency significantly deviates from the expected one (chi square test, $p<2.2*10^{-16}$). (E) Distances between adjacent *de novo* L1 insertions. Experimental L1-L1 distances were compared with *in silico* generated random datasets. *De novo* L1 insertions are less closely spaced than random insertions (two-sample Kolmogorov-Smirnov test, $p<2.2\times10^{-16}$).

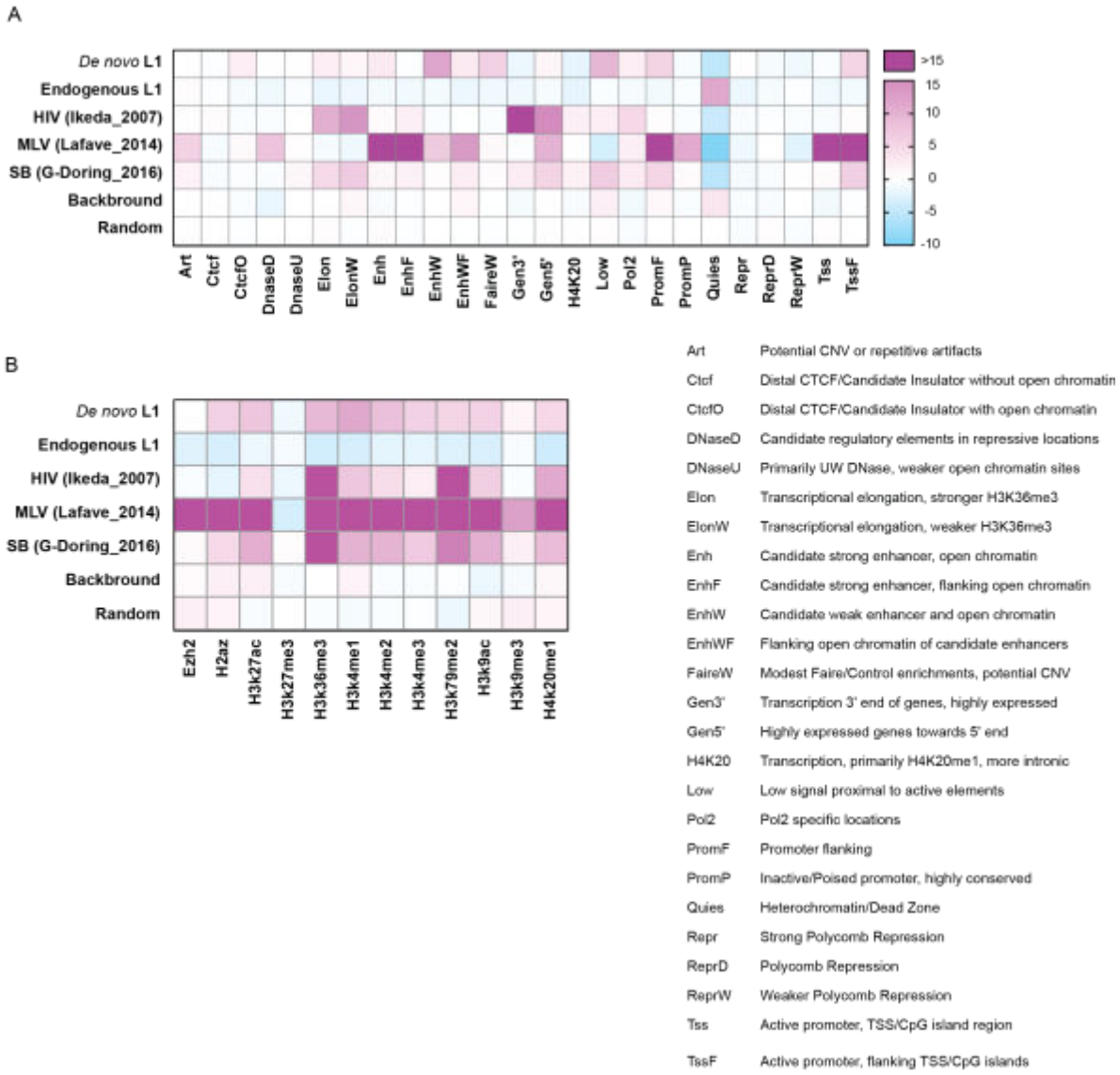

Figure 3. De novo L1 insertions are moderately enriched in chromatin states characteristics of low expression and weak enhancer activities.

(A) Association of *de novo* L1 insertions with chromatin states. *De novo* L1 insertions are moderately enriched in specific ENCODE chromatin states in HeLa S3 genome; chromatin states were defined by the ChromHMM chromatin state annotation algorithm. Heatmap displays z score for the overlap of each chromatin state. Z score is defined as the number of standard deviations by which the observed level of overlap between *de novo* insertions and each chromatin state differs from the expected one. Expected level of overlapping was deduced from 1000 randomized experimental datasets with the chromatin states. Double color gradient from blue to pink indicates depletion (observed overlapping is lower than the expected level) to enrichment (observed overlapping is higher than the expected level). A brief description of each chromatin state is presented below the heatmap. As control for the computational analysis, we analyzed in parallel publicly available *de novo* insertion

datasets previously obtained for other classes of transposable elements or retroviruses (LaFave et al. 2014; Gogol-Döring et al. 2016; Ikeda et al. 2007). Endogenous L1 correspond to existing L1 copies present in the reference human genome. As previously found, HIV is enriched in transcriptionally active units, MLV is enriched in promoters and enhancers, and sleeping beauty did not show much deviation from the expected level. 'Background' and 'Random' represent two *in silico*-generated integration data. The first corresponds to random genomic sites with base composition matching *de novo* L1 integration sites (see details in 'Materials and Methods' section), while the second is completely random. (B) Association of *de novo* L1 insertions with histone marks. Heatmap displays z score of observed overlap with various histone ChIP-seq peak obtained by ENCODE/Broad. Color scale and datasets are as in (A).

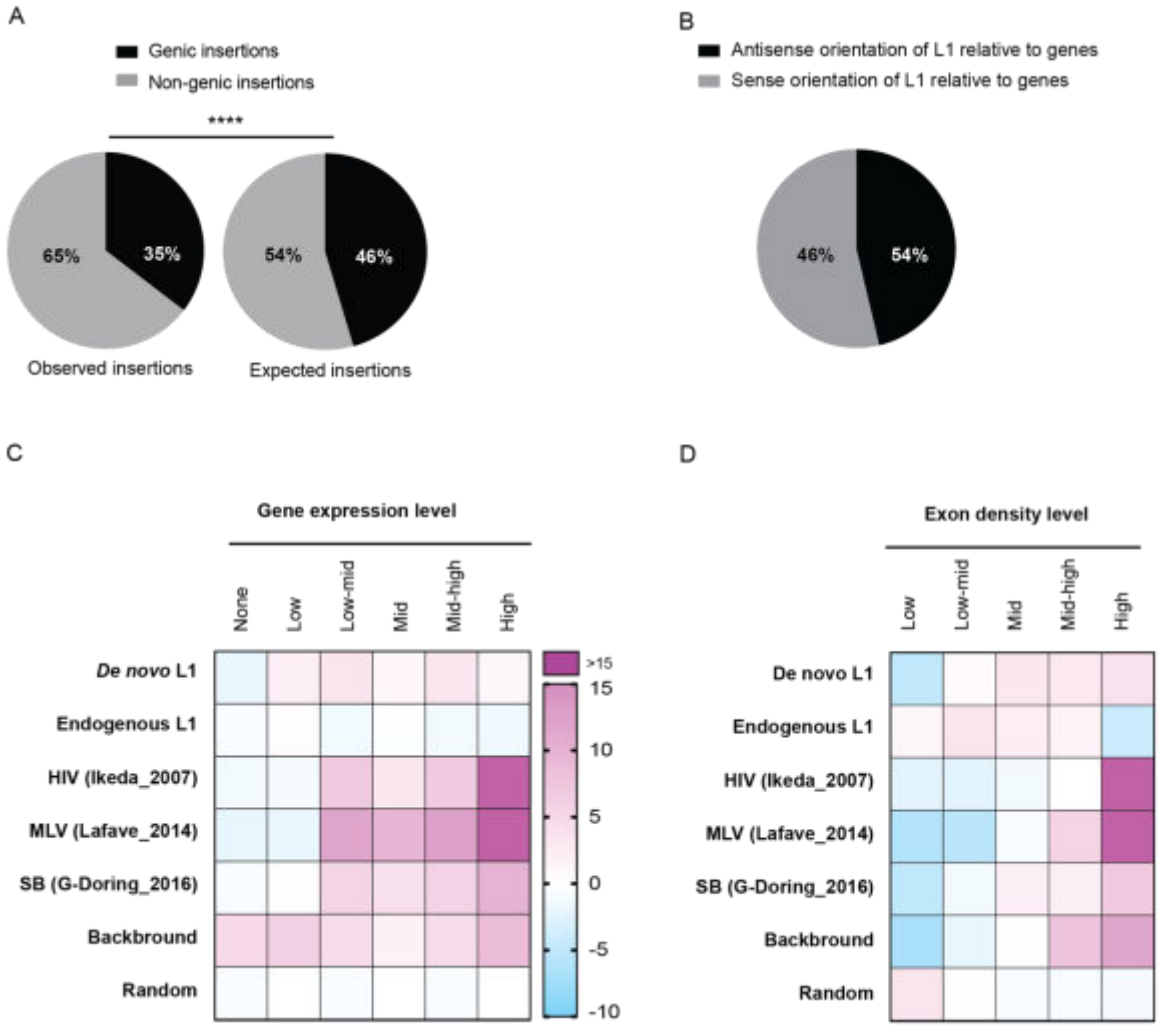

Figure 4. De novo L1 insertions are depleted in genic regions.

A) Distribution of *de novo* L1 insertions in genic regions. *De novo* L1 insertions are depleted in genic regions, thin bars represent confidence interval (Fisher's exact test, $p<0.0001$). (B) Orientation of *de novo* L1 insertions relative to genes. *De novo* L1 genic insertions are slightly enriched in sense orientation relative to the RefSeq genes (binomial test, one tailed $p=0.0444$). (C) Association of *de novo* L1 genic insertions with gene expression level. Heatmap displays z score of observed overlap of insertion datasets with various gene expression category. HIV and MLV insertions show strong association with high expression genes while *de novo* L1, endogenous L1, and sleeping beauty did not show association with gene expression level. Color scale and datasets are as in Figure 3A. D) Association of *de novo* L1 genic insertions with exon density. Heatmap displays z score of observed overlap of insertion datasets with regions of different exon density. HIV and MLV insertions show strong association with highly dense regions while *de novo* L1, endogenous L1, and sleeping beauty did not show association with exon density. Color scale and datasets are as in Figure 3A.

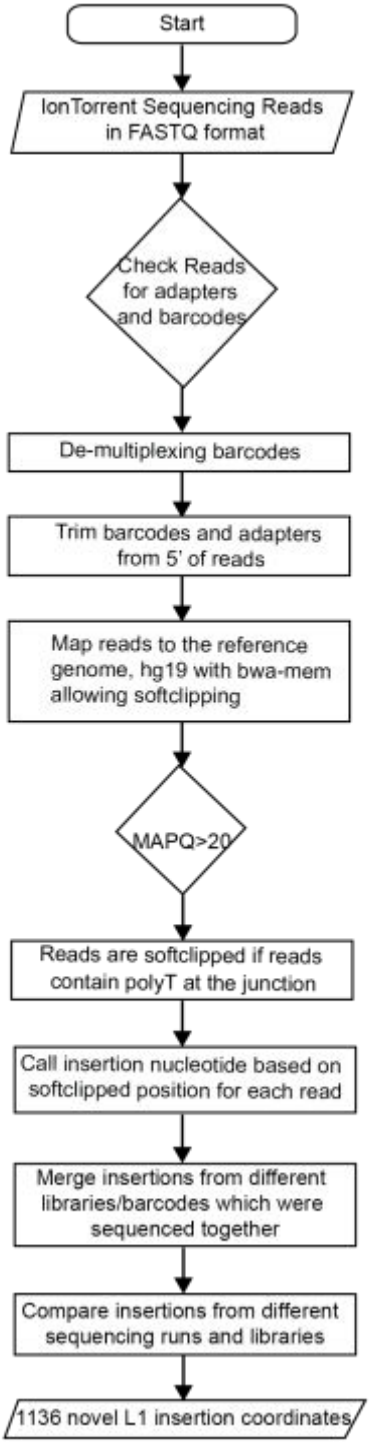

Figure S 1. ATLAS-seq integration site mapping workflow.

The details of each step is covered in the 'Materials and Methods' section. In brief, Ion Torrent sequencing reads were checked for the presence of linkers and barcodes, barcodes were de-multiplexed to obtain the reads originating from each sample, barcodes and linkers were removed from the reads, reads were then aligned with hg19 reference genome, good quality aligned reads with a soft-clipped non-aligned poly(T) sequences at their 3'end were used to

call integration sites. Insertion points were filtered and softclipped, integration junctions were obtained from the alignment of non-redundant reads based on the softclipped sequence position. Identical insertion points from different libraries sequenced together in the same run were merged (to exclude trace levels of barcode contamination).

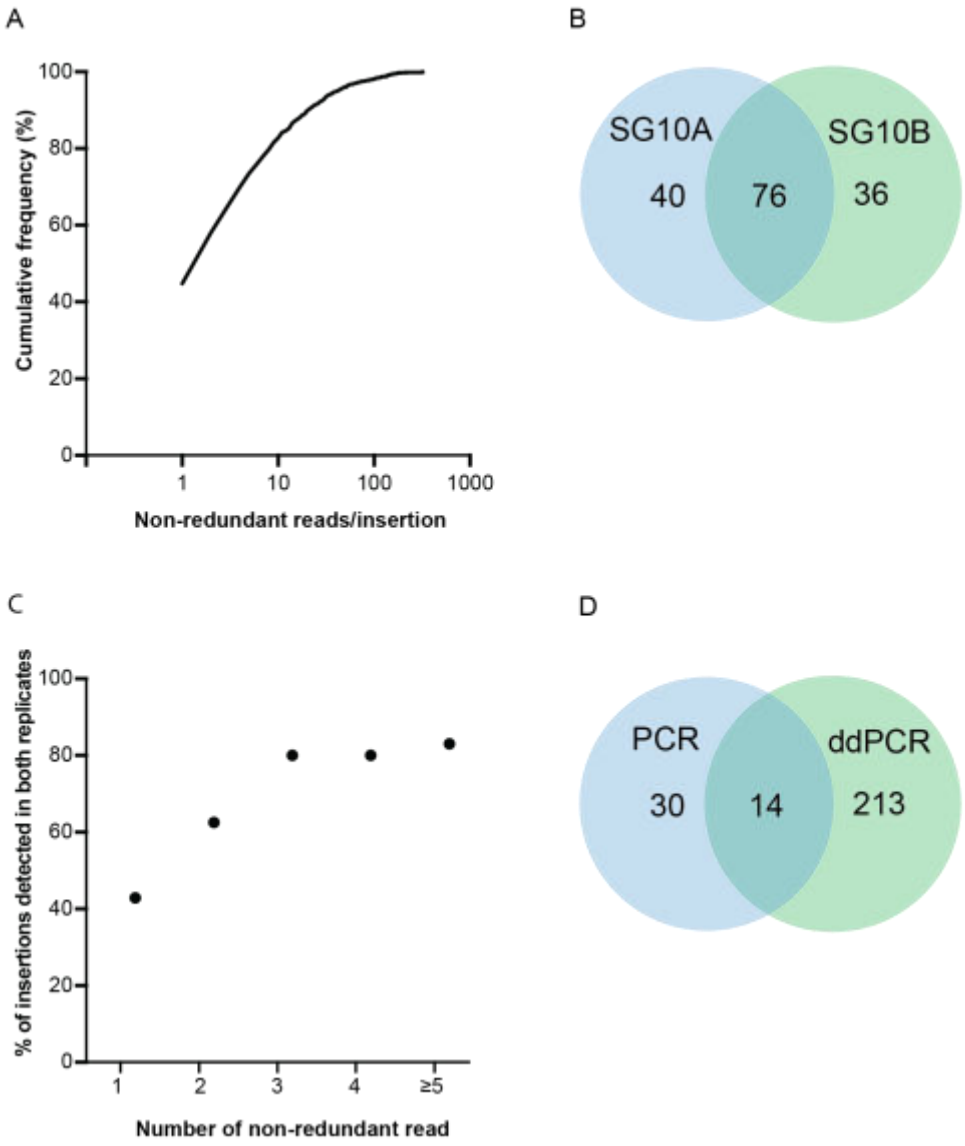

Figure S 2. Quality control of recovery of de novo L1 insertions.

(A) Recovery of insertion sites called from non-redundant reads. 45% of *de novo* L1 insertions were recovered from a single non-redundant read spanning the L1-genome junction, while 90% of the insertions were supported by less than 20 non-redundant reads. Note that given the average size of the reads and the necessity to span the junction, the maximum number of non-redundant reads is somehow limited (B-C) Reproducibility of *de novo* L1 insertions. (B) Approximately 65% of the *de novo* L1 insertions were reproducibly detected in a duplicate library obtained from the same sample. (C) 43% of insertions called from a single non-redundant read was reproducible. Recovery rate of insertions increased with increase in number of non-redundant read supporting an insertion. As high as 83% of insertions supported by more than 4 non-redundant reads was reproducible. (D) L1-genome junction enrichment by emulsion PCR enhanced recovery of insertions. Use of digital droplet PCR for selective amplification of *de novo* L1 integrated sites resulted in more than 5-fold increase in recovery of integration sites compared to the conventional PCR.

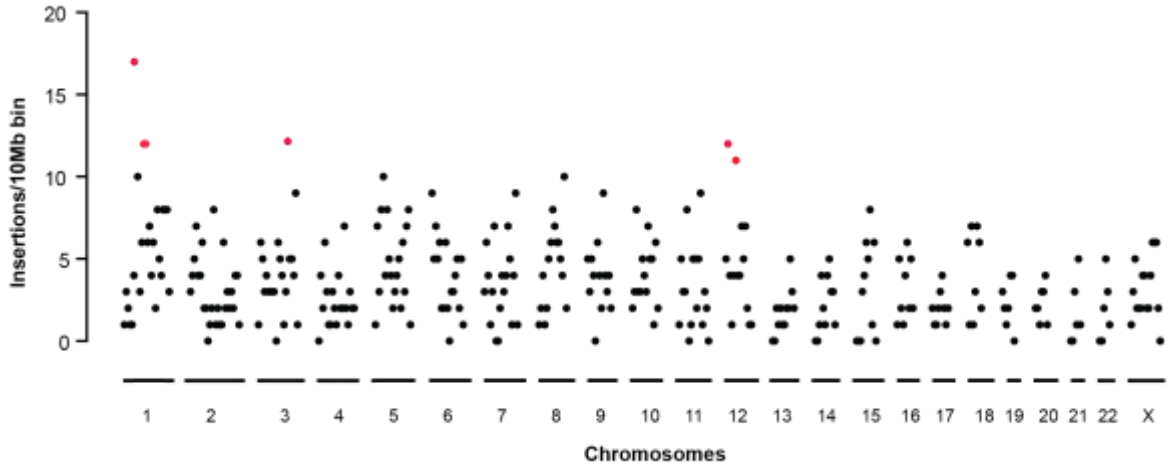

Figure S 3. Cluster of de novo L1 insertions in 10Mb genomic windows.

L1 integration hotspots. Dots show the number of integrated L1 per 10 Mb bins. Red dots, over-represented regions that contain a cluster of more than 10 *de novo* L1 insertions.

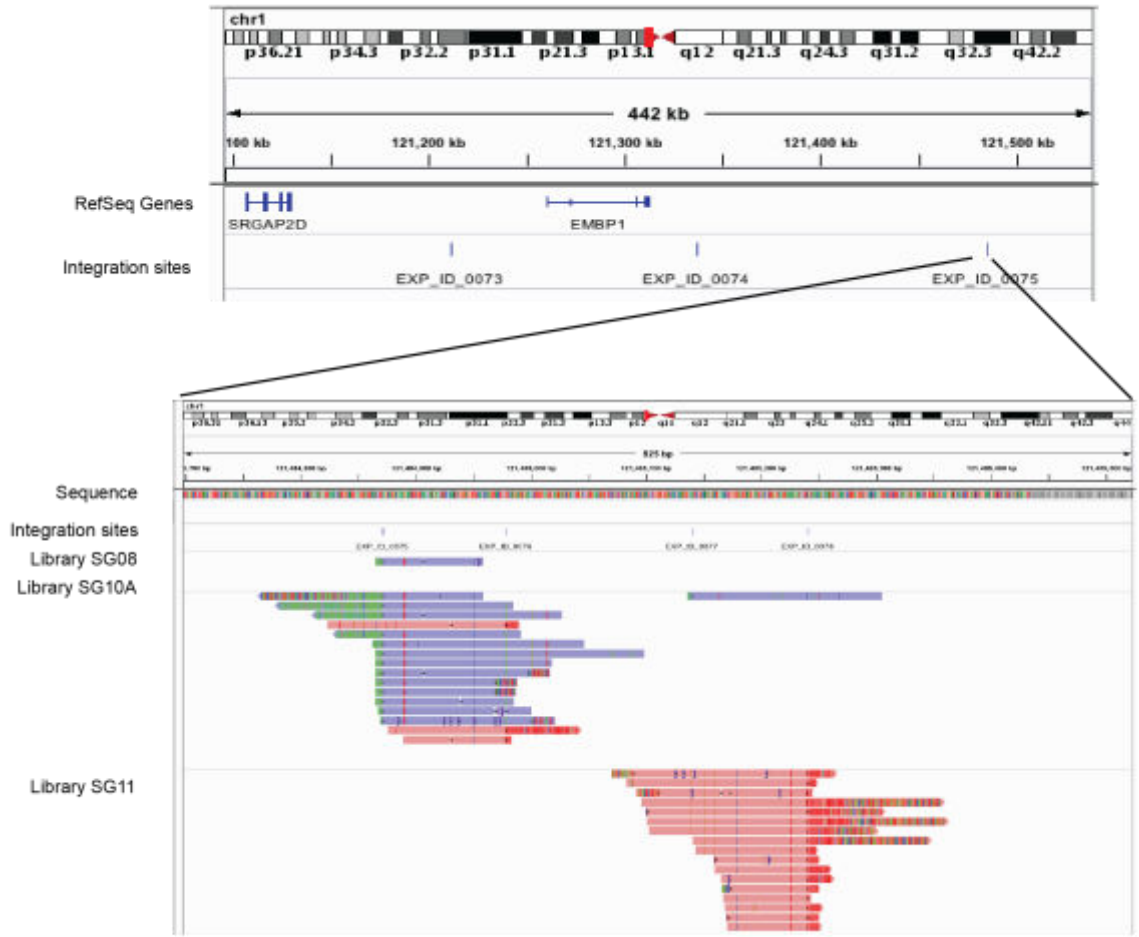

Figure S 4. Representation of a regional hotspot in chromosome 1 containing 6 independent de novo L1 retrotransposition events.

Integrative genome viewer screenshot of an L1 retrotransposition hotspot. Screenshot showing a 0.5Mb region in chromosome 1 with 6 independent integration events from five independent cell populations. A small 800bp region (zoomed view, bottom) contains 4 independent insertions from three cell populations (SG08, SG10, SG11). Two are sense insertions (supporting non-redundant reads shown in blue), and two antisense (supporting non-redundant reads shown in red). For details on five L1 retrotransposition hotspots recovered in this study, see Supplementary table 2.

Supplementary tables

Supplementary table 1. List of oligonucleotides used in this study.

Primer name	Sequence (5'to3')	Target	Comment
LOU1362	GCGCCCGGTTCTTTTTG	mneoI cassette on L1 3' terminus	
LOU1363	GCCTCGTCCTGAAGCTCATT	mneoI cassette on integraed L1 in the genome	
LOU365	GTGGCGGCCAGTATTCGTAGGAGGGCGCGTAGCATAGAACGT		ATLAS-seq linker
LOU366	CGTTCTATGCTACGC		Dummy for ATLAS-seq linker
LOU1078	CCTCTCTATGGGCAGTCGGTGATCGATACCGTAAGCCGAATTG	L1 3' terminus downstream to mneoI	Ion Torrent oligo-trP1/L1
LOU1109	CCATCTCATCCCTGCGTGTCTCCGACTCAG AAGAGGATTC GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- IonXpress_003 -linker)
LOU1111	CCATCTCATCCCTGCGTGTCTCCGACTCAG CAGAAGGAAC GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- IonXpress_005 -linker)
LOU1112	CCATCTCATCCCTGCGTGTCTCCGACTCAG CTGCAAGTTC GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- IonXpress_006 -linker)
LOU1113	CCATCTCATCCCTGCGTGTCTCCGACTCAG TTCGTGATTC GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- IonXpress_007 -linker)
LOU1364	CCATCTCATCCCTGCGTGTCTCCGACTCAG AGCACTGTAG GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_004 -linker)
LOU1365	CCATCTCATCCCTGCGTGTCTCCGACTCAG CGTGTCTCTA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_007 -linker)
LOU1366	CCATCTCATCCCTGCGTGTCTCCGACTCAG TCTCTATGCG GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_010 -linker)
LOU1367	CCATCTCATCCCTGCGTGTCTCCGACTCAG TGATACGTCT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_011 -linker)
LOU1368	CCATCTCATCCCTGCGTGTCTCCGACTCAG CATAGTAGTG GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_013 -linker)
LOU1369	CCATCTCATCCCTGCGTGTCTCCGACTCAG ATACGACGTA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_015 -linker)
LOU1370	CCATCTCATCCCTGCGTGTCTCCGACTCAG TCACGTACTA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_016 -linker)
LOU1371	CCATCTCATCCCTGCGTGTCTCCGACTCAG TACTCTCGTG GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_023 -linker)
LOU1372	CCATCTCATCCCTGCGTGTCTCCGACTCAG TCGTCGCTCG GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_025 -linker)
LOU1373	CCATCTCATCCCTGCGTGTCTCCGACTCAG ACATACGCGT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_026 -linker)
LOU1374	CCATCTCATCCCTGCGTGTCTCCGACTCAG ACTACTATGT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_028 -linker)
LOU1375	CCATCTCATCCCTGCGTGTCTCCGACTCAG AGACTATACT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_030 -linker)
LOU1376	CCATCTCATCCCTGCGTGTCTCCGACTCAG AGTACGCTAT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_032 -linker)
LOU1377	CCATCTCATCCCTGCGTGTCTCCGACTCAG ATAGAGTACT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_033 -linker)
LOU1378	CCATCTCATCCCTGCGTGTCTCCGACTCAG CAGTAGACGT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_035 -linker)
LOU1379	CCATCTCATCCCTGCGTGTCTCCGACTCAG TACAGATCGT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_039 -linker)
LOU1382	CCATCTCATCCCTGCGTGTCTCCGACTCAG TAGTGTAGAT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_041 -linker)
LOU1383	CCATCTCATCCCTGCGTGTCTCCGACTCAG TCGCACTAGT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_043 -linker)
LOU1384	CCATCTCATCCCTGCGTGTCTCCGACTCAG TCTATACTAT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_045 -linker)
LOU1385	CCATCTCATCCCTGCGTGTCTCCGACTCAG TGTGAGTAGT GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_047 -linker)

Primer name	Sequence (5'-to-3')	Target	Comment
LOU1386	CCATCTCATCCCTGCGTGTCTCCGACTCAG ACAGTATATA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_048 -linker)
LOU1387	CCATCTCATCCCTGCGTGTCTCCGACTCAG ACTAGCAGTA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_050 -linker)
LOU1388	CCATCTCATCCCTGCGTGTCTCCGACTCAG AGCTCACGTA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_051 -linker)
LOU1389	CCATCTCATCCCTGCGTGTCTCCGACTCAG AGTATACATA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_052 -linker)
LOU1390	CCATCTCATCCCTGCGTGTCTCCGACTCAG AGTCGAGAGA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_053 -linker)
LOU1391	CCATCTCATCCCTGCGTGTCTCCGACTCAG CGATCGTATA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_055 -linker)
LOU1392	CCATCTCATCCCTGCGTGTCTCCGACTCAG CGTACAGTCA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_058 -linker)
LOU1393	CCATCTCATCCCTGCGTGTCTCCGACTCAG CGTACTCAGA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_059 -linker)
LOU1394	CCATCTCATCCCTGCGTGTCTCCGACTCAG CTACGCTCTA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_060 -linker)
LOU1395	CCATCTCATCCCTGCGTGTCTCCGACTCAG CTATAGCGTA GTGGCGGCCAGTATTC	ATLAS-seq linker	Barcoded Ion Torrent fusion primer (A- MID_061 -linker)

Supplementary table 2. Position and orientation of de novo L1 insertions in hotspots.

Hotspots	Number of integration	Integration nucleotide	Integration name	Integration strand
chr1 (121010000-121535434)	6	121211805	EXP_ID_0073	-
		121337448	EXP_ID_0074	-
		121484871	EXP_ID_0075	+
		121484978	EXP_ID_0076	-
		121485140	EXP_ID_0077	+
		121485240	EXP_ID_0078	-
chr1 (189535434-190535434)	5	189961379	EXP_ID_0100	-
		190045426	EXP_ID_0101	-
		190075926	EXP_ID_0102	-
		190155201	EXP_ID_0103	+
		190232174	EXP_ID_0104	+
chr2 (33010000-34010000)	5	33091840	EXP_ID_0506	-
		33091880	EXP_ID_0507	-
		33092041	EXP_ID_0508	-
		33092082	EXP_ID_0509	-
		33092085	EXP_ID_0510	-
chr8 (127838887-128838887)	5	127897611	EXP_ID_1023	-
		128438899	EXP_ID_1024	+
		128655093	EXP_ID_1025	-
		128655744	EXP_ID_1026	-
		128655792	EXP_ID_1027	-
chr11 (47010000-48010000)	5	47030057	EXP_ID_0212	-
		47650145	EXP_ID_0213	-
		47869551	EXP_ID_0214	+
		47978246	EXP_ID_0215	-
		47978470	EXP_ID_0216	-

References

Aiyer S, Rossi P, Malani N, Schneider WM, Chandar A, Bushman FD, Montelione GT, Roth MJ. 2015. Structural and sequencing analysis of local target DNA recognition by MLV integrase. *Nucleic Acids Res* **43**: 5647–5663.

Amariglio EN, Hakim I, Brok-Simoni F, Grossman Z, Katzir N, Harmelin A, Ramot B, Rechavi G. 1991. Identity of rearranged LINE/c-MYC junction sequences specific for the canine transmissible venereal tumor. *Proc Natl Acad Sci USA* **88**: 8136–8139.

Baker RJ, Kass DH. 1994. Comparison of chromosomal distribution of a retroposon (LINE) and a retrovirus-like element mys in Peromyscus maniculatus and P. leucopus. *Chromosome Res* **2**: 185–189.

Baller JA, Gao J, Stamenova R, Curcio MJ, Voytas DF. 2012. A nucleosomal surface defines an integration hotspot for the Saccharomyces cerevisiae Ty1 retrotransposon. *Genome Res* **22**: 704–713.

Barr SD, Leipzig J, Shinn P, Ecker JR, Bushman FD. 2005. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J Virol* **79**: 12035–12044.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* **141**: 1159–1170.

Belancio VP, Hedges DJ, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res* **34**: 1512–1521.

Boyle AL, Ballard SG, Ward DC. 1990. Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence in situ hybridization. *Proc Natl Acad Sci USA* **87**: 7757–7761.

Brady T, Lee YN, Ronen K, Malani N, Berry CC, Bieniasz PD, Bushman FD. 2009. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev* **23**: 633–642.

Bridier-Nahmias A, Tchalikian-Cosson A, Baller JA, Menouni R, Fayol H, Flores A, Saïb A, Werner M, Voytas DF, Lesage P. 2015. Retrotransposons. An RNA polymerase III subunit determines sites of retrotransposon integration. *Science* **348**: 585–588.

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci USA* **100**: 5280–5285.

Bushman FD. 2003. Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell* **115**: 135–138.

Chen J-M, Stenson PD, Cooper DN, Férec C. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* **117**: 411–427.

Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P, Ecker JR, Bushman F. 2005. A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11**: 1287–1289.

Conley ME, Partain JD, Norland SM, Shurtleff SA, Kazazian HH. 2005. Two independent

retrotransposon insertions at the same site within the coding region of BTK. *Hum Mutat* **25**: 324–325.

Cost GJ, Boeke JD. 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 GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**: 5899–5910.

de Jong J, Wessels LFA, van Lohuizen M, de Ridder J, Akhtar W. 2014. Applications of DNA integrating elements: Facing the bias bully. *Mob Genet Elements* **4**: 1–6.

De Rijck J, Bartholomeeusen K, Ceulemans H, Debyser Z, Gijsbers R. 2010. High-resolution profiling of the LEDGF/p75 chromatin interaction in the ENCODE region. *Nucleic Acids Res* **38**: 6135–6147.

Devine SE, Boeke JD. 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, Heidmann T. 2003. LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* **35**: 41–48.

Doucet AJ, Droc G, Siol O, Audoux J, Gilbert N. 2015a. U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. *Mol Biol Evol* **32**: 1815–1832.

Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, Moran JV. 2015b. A 3' Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol Cell* **60**: 728–741.

Entezam A, Young L, Munson PJ, Furano AV. 2004. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res* **14**: 1221–1231.

Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M, et al. 2011. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**: 43–49.

Erwin JA, Marchetto MC, Gage FH. 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* **15**: 497–506.

Erwin JA, Paquola ACM, Singer T, Gallina I, Novotny M, Quayle C, Bedrosian TA, Alves FIA, Butcher CR, Herdy JR, et al. 2016. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci*.

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

Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim MS, Manda SS, 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.

Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905–916.

Freeman JD, Goodchild NL, Mager DL. 1994. A modified indicator gene for selection of retrotransposition events in mammalian cells. *BioTechniques* **17**: 46–48–9– 52.

Gai X, Voytas DF. 1998. A single amino acid change in the yeast retrotransposon Ty5 abolishes

targeting to silent chromatin. *Mol Cell* **1**: 1051–1055.

Gasior SL, Preston G, Hedges DJ, Gilbert N, Moran JV, Deininger PL. 2007. Characterization of pre-insertion loci of de novo L1 insertions. *Gene* **390**: 190–198.

Gasior SL, Wakeman TP, Xu B, Deininger PL. 2006. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol* **357**: 1383–1393.

Gilbert N, Lutz-Prigge S, Moran JV. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**: 315–325.

Gogol-Döring A, Ammar I, Gupta S, Bunse M, Miskey C, Chen W, Uckert W, Schulz TF, Izsvák Z, Ivics Z. 2016. Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4(+) T Cells. *Mol Ther* **24**: 592–606.

Goodier JL, Ostertag EM, Kazazian HH. 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* **9**: 653–657.

Halling KC, Lazzaro CR, Honchel R, Bufill JA, Powell SM, Arndt CA, Lindor NM. 1999. Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. *Hum Hered* **49**: 97–102.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**: 268–274.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH. 2011. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum Mol Genet* **20**: 3386–3400.

Hancks DC, Kazazian HH. 2012. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev* **22**: 191–203.

Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**: 9.

Hancks DC, Mandal PK, Cheung LE, Kazazian HH. 2012. The minimal active human SVA retrotransposon requires only the 5'-hexamer and Alu-like domains. *Mol Cell Biol* **32**: 4718–4726.

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

Hickey A, Esnault C, Majumdar A, Chatterjee AG, Iben JR, McQueen PG, Yang AX, Mizuguchi T, Grewal SIS, Levin HL. 2015. Single-Nucleotide-Specific Targeting of the Tf1 Retrotransposon Promoted by the DNA-Binding Protein Sap1 of Schizosaccharomyces pombe. *Genetics* **201**: 905–924.

Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, Giardine B, Ellenbogen PM, Bilmes JA, Birney E, et al. 2013. Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res* **41**: 827–841.

Holman AG, Coffin JM. 2005. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc Natl Acad Sci USA* **102**: 6103–6107.

Ikeda T, Shibata J, Yoshimura K, Koito A, Matsushita S. 2007. Recurrent HIV-1 integration at the BACH2 locus in resting CD4+ T cell populations during effective highly active antiretroviral therapy. *J Infect Dis* **195**: 716–725.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**: 1253–1261.

Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci USA* **94**: 1872–1877.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**: 19–31.

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

Kolosha VO, Martin SL. 2003. High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J Biol Chem* **278**: 8112–8117.

Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2011. Similarities between long interspersed element-1 (LINE-1) reverse transcriptase and telomerase. *Proc Natl Acad Sci U S A* **108**: 20345–20350.

Korenberg JR, Rykowski MC. 1988. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**: 391–400.

Kulpa DA, Moran JV. 2006. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat Struct Mol Biol* **13**: 655–660.

LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, Burgess SM. 2014. MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Res* **42**: 4257–4269.

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

Lelek M, Casartelli N, Pellin D, Rizzi E, Souque P, Severgnini M, Di Serio C, Fricke T, Diaz-Griffero F, Zimmer C, et al. 2015. Chromatin organization at the nuclear pore favours HIV replication. *Nat Commun* **6**: 6483.

Lesbats P, Botbol Y, Chevereau G, Vaillant C, Calmels C, Arneodo A, Andreola M-L, Lavigne M, Parissi V. 2011. Functional coupling between HIV-1 integrase and the SWI/SNF chromatin remodeling complex for efficient in vitro integration into stable nucleosomes. *PLoS Pathogens* **7**: e1001280.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Liao GC, Rehm EJ, Rubin GM. 2000. Insertion site preferences of the P transposable element in Drosophila melanogaster. *Proc Natl Acad Sci USA* **97**: 3347–3351.

Linheiro RS, Bergman CM. 2008. Testing the palindromic target site model for DNA transposon insertion using the Drosophila melanogaster P-element. *Nucleic Acids Res* **36**: 6199–6208.

Linheiro RS, Bergman CM. 2012. Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. *PLoS ONE* **7**: e30008.

Liu G, Geurts AM, Yae K, Srinivasan AR, Fahrenkrug SC, Largaespada DA, Takeda J, Horie K, Olson WK, Hackett PB. 2005. Target-site preferences of Sleeping Beauty transposons. *J Mol Biol* **346**: 161–173.

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

Maertens GN, Hare S, Cherepanov P. 2010. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**: 326–329.

Makohon-Moore A, Moyer A, Kohutek ZA, Huang CR, Ahn D, Barker NJ, Hruban RH, Iacobuzio-Donahue CA, Boeke JD, Burns KH. 2015. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat Med* **21**: 1060–1064.

Marini B, Kertesz-Farkas A, Ali H, Lucic B, Lisek K, Manganaro L, Pongor S, Luzzati R, Recchia A, Mavilio F, et al. 2015. Nuclear architecture dictates HIV-1 integration site selection. *Nature* **521**: 227–231.

Martin SL, Bushman FD. 2001. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* **21**: 467–475.

Maskell DP, Renault L, Serrao E, Lesbats P, Matadeen R, Hare S, Lindemann D, Engelman AN, Costa A, Cherepanov P. 2015. Structural basis for retroviral integration into nucleosomes. *Nature* **523**: 366–369.

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

Mitchell RS, Beitzel BF, Schroder ARW, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. ed. Michael Emerman. *PLoS Biol* **2**: E234.

Monot C, Kuciak M, Viollet S, Mir AA, Gabus C, Darlix J-L, Cristofari G. 2013. The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. *PLoS Genet* **9**: e1003499.

Moran JV, DeBerardinis RJ, Kazazian HH. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530–1534.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**: 917–927.

Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, Taccioli GE, Batzer MA, Moran JV. 2002. DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat*

Genet **31**: 159–165.

Moyzis RK, Torney DC, Meyne J, Buckingham JM, Wu JR, Burks C, Sirotkin KM, Goad WB. 1989. The distribution of interspersed repetitive DNA sequences in the human genome. *Genomics* **4**: 273–289.

Mularoni L, Zhou Y, Bowen T, Gangadharan S, Wheelan SJ, Boeke JD. 2012. Retrotransposon Ty1 integration targets specifically positioned asymmetric nucleosomal DNA segments in tRNA hotspots. *Genome Res* **22**: 693–703.

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71**: 312–326.

Nigumann P, Redik K, Mätlik K, Speek M. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**: 628–634.

Ovchinnikov I, Troxel AB, Swergold GD. 2001. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res* **11**: 2050–2058.

Pekin D, Skhiri Y, Baret J-C, Le Corre D, Mazutis L, Salem CB, Millot F, Harrak El A, Hutchison JB, Larson JW, et al. 2011. Quantitative and sensitive detection of rare mutations using droplet-based microfluidics. *Lab Chip* **11**: 2156–2166.

Philippe C, Vargas-Landin DB, Doucet AJ, van Essen D, Vera-Otarola J, Kuciak M, Corbin A, Nigumann P, Cristofari G. 2016. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *eLife* **5**: 166.

Pickeral OK, Makałowski W, Boguski MS, Boeke JD. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* **10**: 411–415.

Piskareva O, 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.

Qi X, Daily K, Nguyen K, Wang H, Mayhew D, Rigor P, Forouzan S, Johnston M, Mitra RD, Baldi P, et al. 2012. Retrotransposon profiling of RNA polymerase III initiation sites. *Genome Res* **22**: 681–692.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG. 2012. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res* **40**: 1666–1683.

Rohrer J, Minegishi Y, Richter D, Eguiguren J, Conley ME. 1999. Unusual mutations in Btk: an insertion, a duplication, an inversion, and four large deletions. *Clin Immunol* **90**: 28–37.

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

Schroder ARW, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**: 521–529.

Serrao E, Ballandras-Colas A, Cherepanov P, Maertens GN, Engelman AN. 2015. Key determinants of target DNA recognition by retroviral intasomes. *Retrovirology* **12**: 39.

Serrao E, Krishnan L, Shun M-C, Li X, Cherepanov P, Engelman A, Maertens GN. 2014. Integrase residues that determine nucleotide preferences at sites of HIV-1 integration: implications for the mechanism of target DNA binding. *Nucleic Acids Res* **42**: 5164–5176.

Sharma A, Larue RC, Plumb MR, Malani N, Male F, Slaughter A, Kessl JJ, Shkriabai N, Coward E, Aiyer SS, et al. 2013. BET proteins promote efficient murine leukemia virus integration at transcription start sites. *Proc Natl Acad Sci U S A* **110**: 12036–12041.

Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, Brennan PM, Baillie JK, Collino A, Ghisletti S, et al. 2013. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**: 101–111.

Singh PK, Plumb MR, Ferris AL, Iben JR, Wu X, Fadel HJ, Luke BT, Esnault C, Poeschla EM, Hughes SH, et al. 2015. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev* **29**: 2287–2297.

Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, et al. 2012. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* **22**: 2328–2338.

Soriano P, Meunier-Rotival M, Bernardi G. 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci USA* **80**: 1816–1820.

Speek M. 2001. Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes. *Mol Cell Biol* **21**: 1973–1985.

Swergold GD. 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* **10**: 6718–6729.

Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, 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**: 1251343–1251343.

Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, Meyer D, Goossens M, Lavergne JM. 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.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**: 1429–1439.

Wheelan SJ. 2005. Gene-breaking: A new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* **15**: 1073–1078.

White TB, McCoy AM, Streva VA, Fenrich J, Deininger PL. 2014. A droplet digital PCR detection method for rare L1 insertions in tumors. *Mob DNA* **5**: 30.

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

Wu X, Li Y, Crise B, Burgess SM. 2003. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**: 1749–1751.

Wulff K, Gazda H, Schröder W, Robicka-Milewska R, Herrmann FH. 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**: 299–299.

Xie W, Gai X, Zhu Y, Zappulla DC, Sternglanz R, Voytas DF. 2001. Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Mol Cell Biol* **21**: 6606–6614.

Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci USA* **103**: 17608–17613.

Conclusion

Targeting of L1 retrotransposons in human genome is poorly understood. In recent years, profiling of insertions using deep sequencing technology, along with large-scale genomic data mining, and cellular or biochemical approaches has led to interesting discoveries regarding targeted integration by many transposable elements and retroviruses. Our work led to the discovery of non-random integration site selection by human L1 retrotransposons. Future work should focus on exploring molecular mechanism of L1 integration site selection.

Literature cited

Adhya SL, Shapiro JA. 1969. The galactose operon of E. coli K-12. I. Structural and pleiotropic mutations of the operon. *Genetics* **62**: 231–247.

Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV. 2006. Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev* **20**: 210–224.

Amariglio EN, Hakim I, Brok-Simoni F, Grossman Z, Katzir N, Harmelin A, Ramot B, Rechavi G. 1991. Identity of rearranged LINE/c-MYC junction sequences specific for the canine transmissible venereal tumor. *Proc Natl Acad Sci USA* **88**: 8136–8139.

Ariumi Y. 2016. Guardian of the Human Genome: Host Defense Mechanisms against LINE-1 Retrotransposition. *Front Chem* **4**: 761–12.

Babushok DV, Ohshima K, Ostertag EM, Chen X, Wang Y, Okada N, Abrams CS, Kazazian HH. 2007. A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. *Genome Res* **17**: 1129–1138.

Bailey JA, Liu G, Eichler EE. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73**: 823–834.

Baker RJ, Kass DH. 1994. Comparison of chromosomal distribution of a retroposon (LINE) and a retrovirus-like element mys in Peromyscus maniculatus and P. leucopus. *Chromosome Res* **2**: 185–189.

Baller JA, Gao J, Voytas DF. 2011. Access to DNA establishes a secondary target site bias for the yeast retrotransposon Ty5. *Proc Natl Acad Sci U S A* **108**: 20351–20356.

Barr SD, Leipzig J, Shinn P, Ecker JR, Bushman FD. 2005. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J Virol* **79**: 12035–12044.

Barzilay G, Hickson ID. 1995. Structure and function of apurinic/apyrimidinic endonucleases. *Bioessays* **17**: 713–719.

Basame S, Wai-lun Li P, Howard G, Branciforte D, Keller D, Martin SL. 2006. Spatial assembly and RNA binding stoichiometry of a LINE-1 protein essential for retrotransposition. *J Mol Biol* **357**: 351–357.

Beauregard A, Curcio MJ, Belfort M. 2008. The take and give between retrotransposable elements and their hosts. *Annu Rev Genet* **42**: 587–617.

Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, Eichler EE, Badge RM, Moran JV. 2010. LINE-1 retrotransposition activity in human genomes. *Cell* **141**: 1159–1170.

Beck CR, Garcia-Perez JL, Badge RM, Moran JV. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genom Hum Genet* **12**: 187–215.

Becker KG, Swergold GD, Ozato K, Thayer RE. 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.

Belancio VP, Hedges DJ, Deininger P. 2006. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res* **34**: 1512–1521.

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

Belancio VP, Roy-Engel AM, Deininger P. 2008b. The impact of multiple splice sites in human L1 elements. *Gene* **411**: 38–45.

Belgnaoui SM, Gosden RG, Semmes OJ, Haoudi A. 2006. Human LINE-1 retrotransposon induces DNA damage and apoptosis in cancer cells. *Cancer Cell Int* **6**: 13.

Bellen HJ, Levis RW, He Y, Carlson JW, Evans-Holm M, Bae E, Kim J, Metaxakis A, Savakis C, Schulze KL, et al. 2011. The Drosophila gene disruption project: progress using transposons with distinctive site specificities. ed. N. Perrimon. *Genetics* **188**: 731–743.

Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE. 2008. Active Alu retrotransposons in the human genome. *Genome Res* **18**: 1875–1883.

Bestor TH, Bourc'his D. 2004. Transposon silencing and imprint establishment in mammalian germ cells. *Cold Spring Harb Symp Quant Biol* **69**: 381–387.

Biessmann H, Mason JM. 2003. Telomerase-independent mechanisms of telomere elongation. *Cell Mol Life Sci* **60**: 2325–2333.

Boeke JD. 1997. LINEs and Alus — the polyA connection. *Nat Genet* **16**: 6–7.

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

Bowerman B, Brown PO, Bishop JM, Varmus HE. 1989. A nucleoprotein complex mediates the integration of retroviral DNA. *Genes Dev* **3**: 469–478.

Boyle AL, Ballard SG, Ward DC. 1990. Differential distribution of long and short interspersed element sequences in the mouse genome: chromosome karyotyping by fluorescence in situ hybridization. *Proc Natl Acad Sci USA* **87**: 7757–7761.

Brady T, Lee YN, Ronen K, Malani N, Berry CC, Bieniasz PD, Bushman FD. 2009. Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev* **23**: 633–642.

Britten RJ, Davidson EH. 1969. Gene regulation for higher cells: a theory. *Science* **165**: 349–

357.

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

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci USA* **100**: 5280–5285.

Bukhari AI, Froshauer S. 1978. Insertion of a transposon for chloramphenicol resistance into bacteriophage Mu. *Gene* **3**: 303–314.

Burton FH, Loeb DD, Voliva CF, Martin SL, Edgell MH, Hutchison CA. 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, Kilimann MW. 1998. Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol* **277**: 513–517.

Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, Reforgiato-Recupero G, Martin C. 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *The Plant Cell* **24**: 1242–1255.

Cabot EL, Angeletti B, Usdin K, Furano AV. 1997. Rapid evolution of a young L1 (LINE-1) clade in recently speciated Rattus taxa. *J Mol Evol* **45**: 412–423.

Callahan KE, Hickman AB, Jones CE, Ghirlando R, Furano AV. 2012. Polymerization and nucleic acid-binding properties of human L1 ORF1 protein. *Nucleic Acids Res* **40**: 813–827.

Callinan PA, Batzer MA. 2006. Retrotransposable elements and human disease. *Genome Dyn* **1**: 104–115.

Casacuberta E, Pardue M-L. 2005. HeT-A and TART, two Drosophila retrotransposons with a bona fide role in chromosome structure for more than 60 million years. *Cytogenet Genome Res* **110**: 152–159.

Casavant NC, Hardies SC. 1994. Shared sequence variants of Mus spretus LINE-1 elements tracing dispersal to within the last 1 million years. *Genetics* **137**: 565–572.

Castro-Diaz N, Ecco G, Coluccio A, Kapopoulou A, Yazdanpanah B, Friedli M, Duc J, Jang SM, Turelli P, Trono D. 2014. Evolutionally dynamic L1 regulation in embryonic stem cells. *Genes Dev* **28**: 1397–1409.

Chen J-M, Stenson PD, Cooper DN, Férec C. 2005. A systematic analysis of LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease. *Hum Genet* **117**: 411–427.

Christensen SM, Eickbush TH. 2005. R2 target-primed reverse transcription: ordered cleavage

and polymerization steps by protein subunits asymmetrically bound to the target DNA. *Mol Cell Biol* **25**: 6617–6628.

Chu WM, Liu WM, Schmid CW. 1995. RNA polymerase III promoter and terminator elements affect Alu RNA expression. *Nucleic Acids Res* **23**: 1750–1757.

Ciuffi A. 2008. Mechanisms governing lentivirus integration site selection. *Curr Gene Ther* **8**: 419–429.

Claeys Bouuaert C, Chalmers RM. 2010. Gene therapy vectors: the prospects and potentials of the cut-and-paste transposons. *Genetica* **138**: 473–484.

Coffin JM, Hughes SH, Varmus HE, Coffin JM, Hughes SH, Varmus HE. 1997. The Interactions of Retroviruses and their Hosts.

Comeaux MS, Roy-Engel AM, Hedges DJ, Deininger PL. 2009. Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die? *Genome Res* **19**: 545–555.

Conley ME, Partain JD, Norland SM, Shurtleff SA, Kazazian HH. 2005. Two independent retrotransposon insertions at the same site within the coding region of BTK. *Hum Mutat* **25**: 324–325.

Cook PR, Jones CE, Furano AV. 2015. Phosphorylation of ORF1p is required for L1 retrotransposition. *Proc Natl Acad Sci U S A* **112**: 4298–4303.

Cordaux R. 2008. The human genome in the LINE of fire. *Proc Natl Acad Sci U S A* **105**: 19033–19034.

Cordaux R, Batzer MA. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet* **10**: 691–703.

Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A* **103**: 8101–8106.

Cost GJ, Boeke JD. 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 GJ, Feng Q, Jacquier A, Boeke JD. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J* **21**: 5899–5910.

Cost GJ, Golding A, Schlissel MS, Boeke JD. 2001. Target DNA chromatinization modulates nicking by L1 endonuclease. *Nucleic Acids Res* **29**: 573–577.

Coufal NG, Garcia-Perez JL, Peng GE, Marchetto MCN, Muotri AR, Mu Y, Carson CT, Macia A, Moran JV, Gage FH. 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.

Criscione SW, Theodosakis N, Micevic G, Cornish TC, Burns KH, Neretti N. 2016. Genome-wide characterization of human L1 antisense promoter-driven transcripts. *BMC Genomics* **17**: 740.

Curcio MJ, Derbyshire KM. 2003. The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol* **4**: 865–877.

Dai J, Xie W, Brady TL, Gao J, Voytas DF. 2007. Phosphorylation regulates integration of the yeast Ty5 retrotransposon into heterochromatin. *Mol Cell* **27**: 289–299.

Damert A, Raiz J, Horn AV, Löwer J, Wang H, Xing J, Batzer MA, Löwer R, Schumann GG. 2009. 5'-Transducing SVA retrotransposon groups spread efficiently throughout the human genome. *Genome Res* **19**: 1992–2008.

De Fazio S, Bartonicek N, Di Giacomo M, Abreu-Goodger C, Sankar A, Funaya C, Antony C, Moreira PN, Enright AJ, O'Carroll D. 2011. The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. *Nature* **480**: 259–263.

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

Denli AM, Narvaiza I, Kerman BE, Pena M, Benner C, Marchetto MCN, Diedrich JK, Aslanian A, Ma J, Moresco JJ, et al. 2015. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell* **163**: 583–593.

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

Dmitriev SE, Andreev DE, Terenin IM, Olovnikov IA, Prassolov VS, Merrick WC, Shatsky IN. 2007. Efficient translation initiation directed by the 900-nucleotide-long and GC-rich 5' untranslated region of the human retrotransposon LINE-1 mRNA is strictly cap dependent rather than internal ribosome entry site mediated. *Mol Cell Biol* **27**: 4685–4697.

Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH. 1991. Isolation of an active human transposable element. *Science* **254**: 1805–1808.

Doucet AJ, Droc G, Siol O, Audoux J, Gilbert N. 2015a. U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. *Mol Biol Evol* **32**: 1815–1832.

Doucet AJ, Hulme AE, Sahinovic E, Kulpa DA, Moldovan JB, Kopera HC, Athanikar JN, Hasnaoui M, Bucheton A, Moran JV, et al. 2010. Characterization of LINE-1 ribonucleoprotein particles. ed. G.S. Barsh. *PLoS Genet* **6**: e1001150.

Doucet AJ, Wilusz JE, Miyoshi T, Liu Y, Moran JV. 2015b. A 3’ Poly(A) Tract Is Required for LINE-1 Retrotransposition. *Mol Cell* **60**: 728–741.

Doucet-O'Hare TT, Rodić N, Sharma R, Darbari I, Abril G, Choi JA, Young Ahn J, Cheng Y, Anders RA, Burns KH, et al. 2015. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc Natl Acad Sci U S A* **112**: E4894–900.

Eickbush TH, Jamburuthugoda VK. 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* **134**: 221–234.

Eissenberg JC. 2001. Molecular biology of the chromo domain: an ancient chromatin module comes of age. *Gene* **275**: 19–29.

Erwin JA, Marchetto MC, Gage FH. 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat Rev Neurosci* **15**: 497–506.

Erwin JA, Paquola ACM, Singer T, Gallina I, Novotny M, Quayle C, Bedrosian TA, Alves FIA, Butcher CR, Herdy JR, et al. 2016. L1-associated genomic regions are deleted in somatic cells of the healthy human brain. *Nat Neurosci*.

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

Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim MS, Manda SS, 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.

Fanning TG, Singer MF. 1987. LINE-1: a mammalian transposable element. *Biochim Biophys Acta* **910**: 203–212.

Feng Q, Moran JV, Kazazian HH, Boeke JD. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**: 905–916.

Finnegan DJ. 1983. Retrovirus evolution: Retroviruses and transposable elements — which came first? *Nature* **302**: 105–106.

Fujiwara H. 2015. Site-specific non-LTR retrotransposons. *Microbiol Spectr* **3**: MDNA3–0001–2014.

Furano AV. 2000. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. *Prog Nucleic Acid Res Mol Biol* **64**: 255–294.

Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res* **18**: 359–369.

Garcia-Perez JL, Sekiguchi J, Moran JV. 2007. Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* **446**: 208–212.

Gasior SL, Preston G, Hedges DJ, Gilbert N, Moran JV, Deininger PL. 2007. Characterization of pre-insertion loci of de novo L1 insertions. *Gene* **390**: 190–198.

Gasior SL, Wakeman TP, Xu B, Deininger PL. 2006. The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol* **357**: 1383–1393.

Gilbert N, Lutz S, Moran JV. 2005. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* **25**: 7780–7795.

Gilbert N, Lutz-Prigge S, Moran JV. 2002. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**: 315–325.

Gogol-Döring A, Ammar I, Gupta S, Bunse M, Miskey C, Chen W, Uckert W, Schulz TF, Izsvák Z, Ivics Z. 2016. Genome-wide Profiling Reveals Remarkable Parallels Between Insertion Site Selection Properties of the MLV Retrovirus and the piggyBac Transposon in Primary Human CD4(+) T Cells. *Mol Ther* **24**: 592–606.

Goodier JL. 2016. Restricting retrotransposons: a review. *Mob DNA* **7**: 16.

Goodier JL, Ostertag EM, Engleka KA, Seleme MC, Kazazian HH. 2004. A potential role for the nucleolus in L1 retrotransposition. *Hum Mol Genet* **13**: 1041–1048.

Goodier JL, Ostertag EM, Kazazian HH. 2000. Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* **9**: 653–657.

Goodier JL, Vetter MR, Kazazian HH. 2007. LINE-1 ORF1 protein localizes in stress granules with other RNA-binding proteins, including components of RNA interference RNA-induced silencing complex. *Mol Cell Biol* **27**: 6469–6483.

Görlich D, Kutay U. 1999. Transport Between the Cell Nucleus and the Cytoplasm. *Annu Rev Cell Dev Biol* **15**: 607–660.

Grow EJ, Flynn RA, Chavez SL, Bayless NL, Wossidlo M, Wesche DJ, Martin L, Ware CB, Blish CA, Chang HY, et al. 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* **522**: 221–225.

Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, Lim A, Osborne CS, Pawliuk R, Morillon E, et al. 2003. LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**: 415–419.

Halling KC, Lazzaro CR, Honchel R, Bufill JA, Powell SM, Arndt CA, Lindor NM. 1999. Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. *Hum Hered* **49**: 97–102.

Han JS, Szak ST, Boeke JD. 2004. Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* **429**: 268–274.

Han K, Lee J, Meyer TJ, Remedios P, Goodwin L, Batzer MA. 2008. L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A* **105**: 19366–19371.

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

Hancks DC, Ewing AD, Chen JE, Tokunaga K, Kazazian HH. 2009. Exon-trapping mediated by the human retrotransposon SVA. *Genome Res* **19**: 1983–1991.

Hancks DC, Goodier JL, Mandal PK, Cheung LE, Kazazian HH. 2011. Retrotransposition of

marked SVA elements by human L1s in cultured cells. *Hum Mol Genet* **20**: 3386–3400.

Hancks DC, Kazazian HH. 2012. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev* **22**: 191–203.

Hancks DC, Kazazian HH. 2016. Roles for retrotransposon insertions in human disease. *Mob DNA* **7**: 9.

Hancks DC, Mandal PK, Cheung LE, Kazazian HH. 2012. The minimal active human SVA retrotransposon requires only the 5'-hexamer and Alu-like domains. *Mol Cell Biol* **32**: 4718–4726.

Harris CR, Dewan A, Zupnick A, Normart R, Gabriel A, Prives C, Levine AJ, Hoh J. 2009. p53 responsive elements in human retrotransposons. *Oncogene* **28**: 3857–3865.

Heinicke LA, Wong CJ, Lary J, Nallagatla SR, Diegelman-Parente A, Zheng X, Cole JL, Bevilacqua PC. 2009. RNA dimerization promotes PKR dimerization and activation. *J Mol Biol* **390**: 319–338.

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

Huang CRL, Schneider AM, Lu Y, Niranjan T, Shen P, Robinson MA, Steranka JP, Valle D, Civin CI, Wang T, et al. 2010. Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**: 1171–1182.

Hughes SH. 2015. Reverse Transcription of Retroviruses and LTR Retrotransposons. *Microbiol Spectr* **3**: MDNA3–0027–2014.

Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, Van Meir EG, Vertino PM, Devine SE. 2010. Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**: 1253–1261.

Jacob F, Perrin D, Sánchez C, Monod J, Edelstein S. 2005. *[The operon: a group of genes with expression coordinated by an operator. C.R.Acad. Sci. Paris 250 (1960) 1727-1729].*

Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**: 242–245.

Janoušek V, Karn RC, Laukaitis CM. 2013. The role of retrotransposons in gene family expansions: insights from the mouse Abp gene family. *BMC Evol Biol* **13**: 107.

Jurka J. 1997. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci USA* **94**: 1872–1877.

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

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* **10**: 19–31.

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.

Kajikawa M, Okada N. 2002. LINEs Mobilize SINEs in the Eel through a Shared 3' Sequence. *Cell* **111**: 433–444.

Kapitonov VV, Tempel S, Jurka J. 2009. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. *Gene* **448**: 207–213.

Kaulfers PM, Laufs R, Jahn G. 1978. Molecular properties of transmissible R factors of Haemophilus influenzae determing tetracycline resistance. *J Gen Microbiol* **105**: 243–252.

Kazazian HH. 2004. Mobile elements: drivers of genome evolution. *Science* **303**: 1626–1632.

Kazazian HH Jr., Wong C, Youssoufian H, Phillips DG, Antonarakis SE. 1988. Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**: 164–166.

Ke N, Gao X, Keeney JB, Boeke JD, Voytas DF. 1999. The yeast retrotransposon Ty5 uses the anticodon stem-loop of the initiator methionine tRNA as a primer for reverse transcription. *RNA* **5**: 929–938.

Khan H, Smit A, 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, Büttner R, Schmidt S, Coles M, Weichenrieder O. 2011. Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat Struct Mol Biol* **18**: 1006–1014.

Kolosha VO, Martin SL. 2003. High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J Biol Chem* **278**: 8112–8117.

Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. 2011. Similarities between long interspersed element-1 (LINE-1) reverse transcriptase and telomerase. *Proc Natl Acad Sci U S A* **108**: 20345–20350.

Korenberg JR, Rykowski MC. 1988. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**: 391–400.

Krull M, Brosius J, Schmitz J. 2005. Alu-SINE exonization: en route to protein-coding function. *Mol Biol Evol* **22**: 1702–1711.

Kubo S, Seleme MDC, Soifer HS, Perez JLG, Moran JV, Kazazian HH, Kasahara N. 2006. L1

retrotransposition in nondividing and primary human somatic cells. *Proc Natl Acad Sci USA* **103**: 8036–8041.

Kulpa DA, Moran JV. 2006. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat Struct Mol Biol* **13**: 655–660.

Kulpa DA, Moran JV. 2005. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet* **14**: 3237–3248.

Kushner PJ, Blair LC, Herskowitz I. 1979. Control of yeast cell types by mobile genes: a test. *Proc Natl Acad Sci USA* **76**: 5264–5268.

Kuwabara T, Hsieh J, Muotri A, Yeo G, Warashina M, Lie DC, Moore L, Nakashima K, Asashima M, Gage FH. 2009. Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. *Nat Neurosci* **12**: 1097–1105.

LaFave MC, Varshney GK, Gildea DE, Wolfsberg TG, Baxevanis AD, Burgess SM. 2014. MLV integration site selection is driven by strong enhancers and active promoters. *Nucleic Acids Res* **42**: 4257–4269.

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

Lavie L, Maldener E, Brouha B, Meese EU, Mayer J. 2004. The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res* **14**: 2253–2260.

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

Lee J, Han K, Meyer TJ, Kim H-S, Batzer MA. 2008. Chromosomal Inversions between Human and Chimpanzee Lineages Caused by Retrotransposons ed. J.C. Fay. *PLoS ONE* **3**: e4047–9.

Leibold DM, Swergold GD, Singer MF, Thayer RE, Dombroski BA, Fanning TG. 1990. Translation of LINE-1 DNA elements in vitro and in human cells. *Proc Natl Acad Sci USA* **87**: 6990–6994.

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

Levin HL, Moran JV. 2011. Dynamic interactions between transposable elements and their hosts. *Nat Rev Genet* **12**: 615–627.

Li X, Scaringe WA, Hill KA, Roberts S, Mengos A, Careri D, Pinto MT, Kasper CK, Sommer SS. 2001. Frequency of recent retrotransposition events in the human factor IX gene. *Hum Mutat* **17**: 511–519.

Lisch D. 2012. Regulation of transposable elements in maize. *Curr Opin Plant Biol* **15**: 511–516.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**: 865–875.

Luan DD, Eickbush TH. 1996. Downstream 28S gene sequences on the RNA template affect the choice of primer and the accuracy of initiation by the R2 reverse transcriptase. *Mol Cell Biol* **16**: 4726–4734.

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

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

Maertens GN, Hare S, Cherepanov P. 2010. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**: 326–329.

Makohon-Moore A, Moyer A, Kohutek ZA, Huang CR, Ahn D, Barker NJ, Hruban RH, Iacobuzio-Donahue CA, Boeke JD, Burns KH. 2015. Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat Med* **21**: 1060–1064.

Malfavon-Borja R, Wu LI, Emerman M, Malik HS. 2013. Birth, decay, and reconstruction of an ancient TRIMCyp gene fusion in primate genomes. *Proc Natl Acad Sci U S A* **110**: E583–92.

Malik HS, Burke WD, Eickbush TH. 1999. The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol* **16**: 793–805.

Malik HS, Henikoff S, Eickbush TH. 2000. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* **10**: 1307–1318.

Marchetto MCN, Narvaiza I, Denli AM, Benner C, Lazzarini TA, Nathanson JL, Paquola ACM, Desai KN, Herai RH, Weitzman MD, et al. 2013. Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503**: 525–529.

Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. ed. K. Wolfe. *PLoS Biol* **3**: e357.

Martin SL. 1991. Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. *Mol Cell Biol* **11**: 4804–4807. /pmc/articles/PMC361385/?report=abstract.

Martin SL, Branciforte D, Keller D, Bain DL. 2003. Trimeric structure for an essential protein in L1 retrotransposition. *Proc Natl Acad Sci USA* **100**: 13815–13820.

Martin SL, Bushman FD. 2001. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* **21**: 467–475.

Mathias SL, Scott AF. 1993. Promoter binding proteins of an active human L1 retrotransposon. *Biochem Biophys Res Commun* **191**: 625–632.

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

Mätlik K, Redik K, Speek M. 2006. L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* **2006**: 71753–71716.

McClintock B. 1950. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci USA* **36**: 344–355.

McMillan JP, Singer MF. 1993. Translation of the human LINE-1 element, L1Hs. *Proc Natl Acad Sci USA* **90**: 11533–11537.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res* **12**: 1483–1495.

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

Mills RE, Bennett EA, Iskow RC, Devine SE. 2007. Which transposable elements are active in the human genome? *Trends Genet* **23**: 183–191.

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

Mitchell RS, Beitzel BF, Schroder ARW, Shinn P, Chen H, Berry CC, Ecker JR, Bushman FD. 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. ed. Michael Emerman. *PLoS Biol* **2**: E234.

Monot C, Kuciak M, Viollet S, Mir AA, Gabus C, Darlix J-L, Cristofari G. 2013. The specificity and flexibility of l1 reverse transcription priming at imperfect T-tracts. *PLoS Genet* **9**: e1003499.

Moran JV, DeBerardinis RJ, Kazazian HH. 1999. Exon shuffling by L1 retrotransposition. *Science* **283**: 1530–1534.

Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH. 1996. High frequency retrotransposition in cultured mammalian cells. *Cell* **87**: 917–927.

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

Morse B, Rotherg PG, South VJ, Spandorfer JM, Astrin SM. 1988. Insertional mutagenesis of

the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature* **333**: 87–90.

Moyzis RK, Torney DC, Meyne J, Buckingham JM, Wu JR, Burks C, Sirotkin KM, Goad WB. 1989. The distribution of interspersed repetitive DNA sequences in the human genome. *Genomics* **4**: 273–289.

Myers JS, Vincent BJ, Udall H, Watkins WS, Morrish TA, Kilroy GE, Swergold GD, Henke J, Henke L, Moran JV, et al. 2002. A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* **71**: 312–326.

Narezkina A, Taganov KD, Litwin S, Stoyanova R, Hayashi J, Seeger C, Skalka AM, Katz RA. 2004. Genome-wide analyses of avian sarcoma virus integration sites. *J Virol* **78**: 11656–11663.

Naufer MN, Callahan KE, Cook PR, Perez-Gonzalez CE, Williams MC, Furano AV. 2016. L1 retrotransposition requires rapid ORF1p oligomerization, a novel coiled coil-dependent property conserved despite extensive remodeling. *Nucleic Acids Res* **44**: 281–293.

Nielsen ML, Hermansen TD, Aleksenko A. 2001. A family of DNA repeats in Aspergillus nidulans has assimilated degenerated retrotransposons. *Mol Genet Genomics* **265**: 883–887.

Nigumann P, Redik K, Mätlik K, Speek M. 2002. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**: 628–634.

Ohno S. 1972. So much "junk" DNA in our genome. *Brookhaven Symp Biol* **23**: 366–370.

Ovchinnikov I, Troxel AB, Swergold GD. 2001. Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res* **11**: 2050–2058.

Pardue M-L, DeBaryshe PG. 2000. Drosophila telomere transposons: genetically active elements in heterochromatin. *Genetica* **109**: 45–52.

Pardue M-L, DeBaryshe PG. 2003. Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu Rev Genet* **37**: 485–511.

Pardue M-L, Rashkova S, Casacuberta E, DeBaryshe PG, George JA, Traverse KL. 2005. Two retrotransposons maintain telomeres in Drosophila. *Chromosome Res* **13**: 443–453.

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

Pickeral OK, Makałowski W, Boguski MS, Boeke JD. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res* **10**: 411–415.

Piskareva O, Denmukhametova S, Schmatchenko V. 2003. Functional reverse transcriptase encoded by the human LINE-1 from baculovirus-infected insect cells. *Protein Expr Purif* **28**: 125–130.

Piskareva O, 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.

Pizarro JG, Cristofari G. 2016. Post-Transcriptional Control of LINE-1 Retrotransposition by Cellular Host Factors in Somatic Cells. *Front Cell Dev Biol* **4**: 14.

Pruss D, Bushman FD, Wolffe AP. 1994a. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc Natl Acad Sci USA* **91**: 5913–5917.

Pruss D, Reeves R, Bushman FD, Wolffe AP. 1994b. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J Biol Chem* **269**: 25031–25041.

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

Quentin Y. 1992b. Origin of the Alu family: a family of Alu-like monomers gave birth to the left and the right arms of the Alu elements. *Nucleic Acids Res* **20**: 3397–3401.

Raiz J, Damert A, Chira S, Held U, Klawitter S, Hamdorf M, Löwer J, Strätling WH, Löwer R, Schumann GG. 2012. The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res* **40**: 1666–1683.

Rashkova S, Karam SE, Kellum R, Pardue M-L. 2002a. Gag proteins of the two Drosophila telomeric retrotransposons are targeted to chromosome ends. *J Cell Biol* **159**: 397–402.

Rashkova S, Karam SE, Pardue M-L. 2002b. Element-specific localization of Drosophila retrotransposon Gag proteins occurs in both nucleus and cytoplasm. *Proc Natl Acad Sci USA* **99**: 3621–3626.

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

Robart AR, Zimmerly S. 2005. Group II intron retroelements: function and diversity. *Cytogenet Genome Res* **110**: 589–597.

Salem AH, Myers JS, Otieno AC, Watkins WS, Jorde LB, Batzer MA. 2003. LINE-1 preTa elements in the human genome. *J Mol Biol* **326**: 1127–1146.

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

Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, Gabriel A, Swergold GD, Kazazian HH. 1997. Many human L1 elements are capable of retrotransposition. *Nat Genet* **16**: 37–43.

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

Schroder ARW, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. 2002. HIV-1 integration in the

human genome favors active genes and local hotspots. *Cell* **110**: 521–529.

Schulz WA, Elo JP, Florl AR, Pennanen S, Santourlidis S, Engers R, Buchardt M, Seifert H-H, Visakorpi T. 2002. Genomewide DNA hypomethylation is associated with alterations on chromosome 8 in prostate carcinoma. *Genes Chromosomes Cancer* **35**: 58–65.

Scott AF, Schmeckpeper BJ, Abdelrazik M, Comey CT, O'Hara B, Rossiter JP, Cooley T, Heath P, Smith KD, Margolet L. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics* **1**: 113–125.

Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. 2016. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res* **26**: 745–755.

Sen SK, Huang CT, Han K, Batzer MA. 2007. Endonuclease-independent insertion provides an alternative pathway for L1 retrotransposition in the human genome. *Nucleic Acids Res* **35**: 3741–3751.

Serrao E, Ballandras-Colas A, Cherepanov P, Maertens GN, Engelman AN. 2015. Key determinants of target DNA recognition by retroviral intasomes. *Retrovirology* **12**: 39.

Shapiro JA, Adhya SL. 1969. The galactose operon of E. coli K-12. II. A deletion analysis of operon structure and polarity. *Genetics* **62**: 249–264.

Shukla R, Upton KR, Muñoz-Lopez M, Gerhardt DJ, Fisher ME, Nguyen T, Brennan PM, Baillie JK, Collino A, Ghisletti S, et al. 2013. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**: 101–111.

Singh PK, Plumb MR, Ferris AL, Iben JR, Wu X, Fadel HJ, Luke BT, Esnault C, Poeschla EM, Hughes SH, et al. 2015. LEDGF/p75 interacts with mRNA splicing factors and targets HIV-1 integration to highly spliced genes. *Genes Dev* **29**: 2287–2297.

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

Siol O, Boutliliss M, Chung T, Glöckner G, Dingermann T, Winckler T. 2006. Role of RNA polymerase III transcription factors in the selection of integration sites by the dictyostelium non-long terminal repeat retrotransposon TRE5-A. *Mol Cell Biol* **26**: 8242–8251.

Siol O, Spaller T, Schiefner J, Winckler T. 2011. Genetically tagged TRE5-A retrotransposons reveal high amplification rates and authentic target site preference in the Dictyostelium discoideum genome. *Nucleic Acids Res* **39**: 6608–6619.

Smit AF, Tóth G, Riggs AD, Jurka J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* **246**: 401–417.

Soifer HS, Zaragoza A, Peyvan M, Behlke MA, Rossi JJ. 2005. A potential role for RNA interference in controlling the activity of the human LINE-1 retrotransposon. *Nucleic Acids*

Res **33**: 846–856.

Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, Burns MB, Harris RS, Sigmon DF, Casella A, Erlanger B, et al. 2012. Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* **22**: 2328–2338.

Song M. 2007. Selection against LINE-1 retrotransposons results principally from their ability to mediate ectopic recombination. *Gene* **390**: 206–213.

Sorek R. 2007. The birth of new exons: mechanisms and evolutionary consequences. *RNA* **13**: 1603–1608.

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

Soriano P, Meunier-Rotival M, Bernardi G. 1983. The distribution of interspersed repeats is nonuniform and conserved in the mouse and human genomes. *Proc Natl Acad Sci USA* **80**: 1816–1820.

Speek M. 2001. Antisense Promoter of Human L1 Retrotransposon Drives Transcription of Adjacent Cellular Genes. *Mol Cell Biol* **21**: 1973–1985.

Stoye JP. 2012. Studies of endogenous retroviruses reveal a continuing evolutionary saga. *Nat Rev Microbiol* **10**: 395–406.

Suter CM, Martin DI, Ward RL. 2004. Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue. *Int J Colorectal Dis* **19**: 95–101.

Swergold GD. 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* **10**: 6718–6729.

Taylor MS, LaCava J, Mita P, Molloy KR, Huang CRL, Li D, Adney EM, Jiang H, Burns KH, Chait BT, et al. 2013. Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell* **155**: 1034–1048.

Tchénio T, Casella JF, Heidmann T. 2000. Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res* **28**: 411–415.

Temin HM. 1980. Origin of retroviruses from cellular moveable genetic elements. *Cell* **21**: 599–600.

Terzian C, Ferraz C, Demaille J, Bucheton A. 2000. Evolution of the Gypsy endogenous retrovirus in the Drosophila melanogaster subgroup. *Mol Biol Evol* **17**: 908–914.

Terzian C, Pélisson A, Bucheton A. 2001. Evolution and phylogeny of insect endogenous retroviruses. *BMC Evol Biol* **1**: 3.

Torrents D, Suyama M, Zdobnov E, Bork P. 2003. A genome-wide survey of human pseudogenes. *Genome Res* **13**: 2559–2567.

Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, Gundem G, Pipinikas CP, 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**: 1251343–1251343.

Ullu E, Murphy S, Melli M. 1982. Human 7SL RNA consists of a 140 nucleotide middle-repetitive sequence inserted in an alu sequence. *Cell* **29**: 195–202.

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

Ullu E, Weiner AM. 1984. Human genes and pseudogenes for the 7SL RNA component of signal recognition particle. *EMBO J* **3**: 3303–3310.

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

Van Valen L. 1973. *A new evolutionary law*. Evolutionary theory.

Vidaud D, Vidaud M, Bahnak BR, Siguret V, Gispert Sanchez S, Laurian Y, Meyer D, Goossens M, Lavergne JM. 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.

Viollet S, Monot C, Cristofari G. 2014. L1 retrotransposition: The snap-velcro model and its consequences. *Mob Genet Elements* **4**: e28907.

Voliva CF, Martin SL, Hutchison CA, Edgell MH. 1984. Dispersal process associated with the L1 family of interspersed repetitive DNA sequences. *J Mol Biol* **178**: 795–813.

Wallace MR, Andersen LB, Saulino AM, Gregory PE, Glover TW, Collins FS. 1991. A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353**: 864–866.

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

Wei SQ, Mizuuchi K, Craigie R. 1997. A large nucleoprotein assembly at the ends of the viral DNA mediates retroviral DNA integration. *EMBO J* **16**: 7511–7520.

Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, Boeke JD, Moran JV. 2001. Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* **21**: 1429–1439.

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

Weiner AM. 2002. SINEs and LINEs: the art of biting the hand that feeds you. *Curr Opin Cell Biol* **14**: 343–350.

Weiss RA. 2006. The discovery of endogenous retroviruses. *Retrovirology* **3**: 67.

Wheelan SJ, Aizawa Y, Han JS, Boeke JD. 2005. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* **15**: 1073–1078.

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

Winckler T, Dingermann T, Glöckner G. 2002. Dictyostelium mobile elements: strategies to amplify in a compact genome. *Cell Mol Life Sci* **59**: 2097–2111.

Wolff EM, Byun H-M, Han HF, Sharma S, Nichols PW, Siegmund KD, Yang AS, Jones PA, Liang G. 2010. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet* **6**: e1000917.

Wulff K, Gazda H, Schröder W, Robicka-Milewska R, Herrmann FH. 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**: 299–299.

Xie Y, Mátés L, Ivics Z, Izsvák Z, Martin SL, An W. 2013. Cell division promotes efficient retrotransposition in a stable L1 reporter cell line. *Mob DNA* **4**: 10.

Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. 2006. Emergence of primate genes by retrotransposon-mediated sequence transduction. *Proc Natl Acad Sci USA* **103**: 17608–17613.

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

Xiong Y, Eickbush TH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* **9**: 3353–3362.

Xiong Y, Eickbush TH. 1988. The site-specific ribosomal DNA insertion element R1Bm belongs to a class of non-long-terminal-repeat retrotransposons. *Mol Cell Biol* **8**: 114–123.

Yang J, Malik HS, Eickbush TH. 1999. Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements. *Proc Natl Acad Sci USA* **96**: 7847–7852.

Yang N, Kazazian HH. 2006. L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. *Nat Struct Mol Biol* **13**: 763–771.

Yang N, Zhang L, Zhang Y, Kazazian HH. 2003. An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res* **31**: 4929–4940.

Zhang Z, Harrison P, Gerstein M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* **12**: 1466–1482.