Structural variations of the human genome and transcriptome induced by LINE-1 retrotransposons

Ashfaq Ali Mir

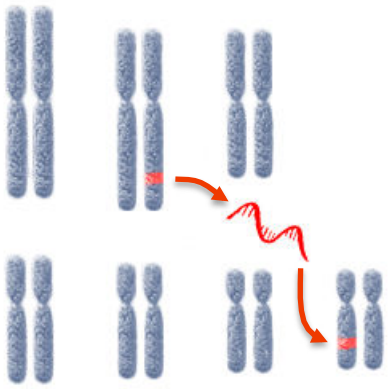
Thèse

En vue d'obtenir le grade de
Docteur de Université Nice-Sophia Antipolis
en Sciences de la Vie

Au titre de l'Ecole Doctorale n°85 Sciences de la Vie & Santé,
présentée et soutenue publiquement le 04 Décembre 2015 par

Ashfaq Ali MIR

Variations structurales du génome et du transcriptome humains induites par les rétrotransposons LINE-1



Jury composé de:
Docteur Pascal BARBRY, Président
Professeur Jean-Nicolas VOLFF, Rapporteur
Docteur Francois SABOT, Rapporteur
Docteur Gael CRISTOFARI, Directeur de thèse

50 ANS Université Nice Sophia Antipolis 1965-2015

For Irfan

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

Firstly, I would like to thank my supervisor Gael Cristofari for his guidance and support and Pascal Barbry, who has seen my evolution through all these years, and then Jean-Nicolas Volff and Francois Sabot for being part of my thesis jury.

I would also like to thank all the postdocs in my lab Jorge Vera-Otarola, Serdar Kasakyan, Javier Garcia-Pizarro, Natacha Broucqsault and Aurélien Doucet for all their help and guidance in my research.

Special thanks goes to Pilvi Nigumann, Claude Philippe, Nadira Lagha and my fellow past and present PhD students Monika Kuciak Stanishevska, Clement Monot, Sebastein Viollet, Ramona Galantonu and Tania Sultana for standing by my side through thick and thin.

I would also like to thank my colleagues Gianni Liti, Julien Cherfils, Alexandre Ottaviani, Sabrina Pisano, Marie-Josephe Giraud Panis, Valerie Renault, Mounir El Mai, Nadir Djerbi, Delphine Benarroch, Aaron Mendez-Bermudez, Benjamin Barre, Marina Shkreli, Margo Montandon, Lucile Yart, Rayane Ghandour, Karine Jamet, Anders Bergstorm, Maud Giroud, Floriane Tissot, Priyanka Sharma, Alexandre Leroy, Johnston Hereroa and Eloise Grasset for being part of my life for the last 4 years like a family.

Last but not the least, I would like to thank my parents and brothers for standing by my side and showering all their love on me.

It's never easy to adjust in a place far away from home, but I think I was lucky to meet some wonderful people who never let me feel alone. Thanks to all of you!

Ashfaq Ali Mir,
Nice, November 23, 2015.

LIST OF ABBREVIATIONS

ARMD	Alu recombination-mediated deletion
ASP	Antisense promoter
ATLAS	Amplification typing of L1 active subfamilies
CAGE	Cap Analysis of Gene Expression
cDNA	Complementary DNA
CNEE	Conserved non-exonic elements
DLEA	Direct L1 extension assay
DSB	DNA-double strand break
EMSA	Electrophoretic mobility shift assay
EN	Endonuclease
EST	Expressed Sequence Tag
HERV	Human endogenous retrovirus
HGR	Human genome reference
iCLIP	Individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation
INT	Integrase
IOSR	Independently-occurred shared retrofamilies
IRES	Internal ribosome entry site
L1RMD	L1 recombination-mediated deletion
LEAP	L1 Element Amplification Protocol
LINE	Long Interspersed Element
LSR	Lineage specific retrofamilies
LTR	Long-Terminal Repeat
MaLR	Mammalian apparent LTR-retrotransposon
MCC	Mutated in colorectal cancer
MIR	Mammalian-wide Interspersed Repeat
MITE	Miniature Inverted-repeat Transposable Element
NAHR	Non-allelic homologous recombination
NGS	Next-generation sequencing
NHEJ	Non-homologous end-joining
ORF	Open reading frame
PCR	Polymerase chain reaction
Pol III	RNA Polymerase III
PP	Processed pseudogene
PR	Protease
RH	RNase H
RNP	Ribonucleoprotein particle
RRM	RNA recognition motif
RT	Reverse transcriptase
SEBI	Starch-branching enzyme
SINE	Short Interspersed Element
SNP	Single Nucleotide Polymorphism
snRNA	Small nuclear RNA
TE	Transposable element

TIR	Terminal Inverted Repeat
TP	Target-primed
TPRT	Target-primed reverse transcription
tRNA	Transfer ribonucleic acid
TSD	Target Site Duplication
TSS	Transcription start site
UTR	Untranslated region
VNTR	Variable Number Tandem Repeat

RÉSUMÉ

Les rétrotransposons constituent presque la moitié de notre génome. Ce sont des éléments génétiques mobiles, également connus sous le nom de gènes sauteurs. Seule la sous-famille L1HS appartenant aux *Long Interspersed Elements* (LINEs) a gardé une capacité de mobilité autonome chez l'Homme moderne. Leur mobilisation dans la lignée germinale, mais aussi dans certains tissus somatiques, contribue à la diversité du génome humain ainsi qu'à certaines maladies comme le cancer. Ainsi, de nouvelles copies de L1s peuvent directement s'intégrer dans des séquences codantes ou régulatrices, et altérer leur fonction. Les séquences L1 contiennent elles-mêmes plusieurs éléments *cis*-régulateurs (promoteurs sens et antisens, signaux de polyadénylation, sites d'épissage cryptiques). Aussi, des insertions de L1 à proximité d'un gène ou dans des séquences introniques peuvent produire des altérations génétiques plus subtiles et dont l'impact est plus difficiles à prédire. Ce phénomène n'est pas limité aux nouvelles insertions. En effet, la dérépression de copies L1 préexistantes et héritées peut également altérer des gènes à proximité, notablement en générant des transcrits L1 chimériques. Cette situation se produit dans certains cancers et pourrait contribuer à la tumorigénicité. Afin d'explorer l'ensemble des altérations géniques induites par les éléments L1s, nous avons développé un logiciel dédié à l'analyse des données de séquençage d'ARN qui permet : (i) d'identifier des transcrits chimériques avec les L1s et les transcrits antisens produits par les L1s; et (ii) d'annoter ces transcrits chimériques en fonction des différents événements d'épissage alternatif subits, y compris ceux pouvant être dus à des éléments L1 récemment intégrés. Au cours de ce travail, il est apparu que la compréhension du lien entre polymorphisme des insertions et phénotype nécessite une vue complète des différentes copies L1HS présentes chez un individu donné. Afin de disposer d'un catalogue aussi complet que possible des polymorphismes d'insertions L1HS identifiés dans des échantillons humains sains ou pathologiques et publiés dans des journaux scientifiques, nous avons développé euL1db, la base de données des insertions de rétrotransposon L1HS chez l'Homme (disponible à l'adresse http://euL1db.unice.fr). Une particularité importante de cette base de données est de pouvoir extraire les insertions présentes dans un échantillon donné pour faciliter les corrélations entre présence ou absence d'insertion L1 et un phénotype spécifique ou une maladie. En conclusion, ce travail aidera à comprendre l'impact des insertions, notamment somatiques, sur l'expression des gènes, à l'échelle complète du génome. Il permettra aussi de mettre en lumière la façon dont l'ensemble des éléments LINE-1 présents chez un individu donné est régulé au niveau transcriptionnel et quels environnements cellulaire et génomique permettent leur expression.

ABSTRACT

Retrotransposons compose almost half of our genome. They are mobile genetics elements, also known as jumping genes. Only the L1HS subfamily of the Long Interspersed Elements (LINEs) has retained the ability to jump autonomously in modern humans. Their mobilization in the germline – but also in some somatic tissues – contributes to human genetic diversity and to diseases, such as cancer. L1 reactivation can be directly mutagenic by disrupting genes or regulatory sequences. In addition, L1 sequences themselves contain many regulatory cis-elements (sense and antisense promoters, polyadenylation signals, cryptic splicing sites). Thus, L1 insertions near a gene or within intronic sequences can also produce more subtle genic alterations. This phenomenon is not limited to tumor-specific L1 insertions: even the derepression of existing and inherited L1 copies in tumors can contribute to cancer progression by altering the expression of their neighboring genes, notably by generating L1 chimeric transcripts. To explore L1-mediated genic alterations in a genome-wide manner, we have developed a dedicated RNA-seq analysis software able: (i) to identify L1 chimeric transcripts and anti-sense L1 transcripts; and (ii) to annotate *de novo* assembled chimeric transcripts for different alternative splicing events caused by L1 elements, including newly integrated insertions. During the course of this work, it appeared that understanding the link between L1HS insertion polymorphisms and phenotype or disease requires a comprehensive view of the different L1HS copies present in a given individual or sample. To provide a comprehensive summary of L1HS insertion polymorphisms identified in healthy or pathological human samples and published in peer-reviewed journals, we developed euL1db, the European database of L1HS retrotransposon insertions in humans (available at http://euL1db.unice.fr). An important feature of euL1db is that insertions can be retrieved at a sample-by-sample level to facilitate correlations between the presence or absence of an L1 insertion with a specific phenotype or disease. This work will help understanding the overall impact of somatic insertions on gene expression, which has been poorly explored so far. It will also shed light on how the full set of LINE-1 elements present in a given individual are regulated at the transcriptional level, and which cellular or genomic environment are permissive for their expression.

LIST OF FIGURES AND TABLES

INTRODUCTION

L'origine de la biologie des transposons prend sa source aux débuts de la génétique moderne lorsque Mendel a publié ses travaux expérimentaux sur les plantes hybrides en 1865. En effet, la cause des mutations étudiées par Mendel et responsables du phénotype ridé des pois, a été depuis attribuée à l'insertion d'un élément transposable similaires aux éléments Ac/Ds du maïs identifiés plus tard par Barbara McClintock. Cette insertion conduit à interrompre le gène *SEBI* impliqué dans la biosynthèse de l'amidon (1). Barbara McClintock a été la première à découvrir les transposons à ADN dans les années 1940 en travaillant sur la cytogénétique du maïs.

La moitié du génome humain est constitué d'éléments transposables (ETs), dont 17% de rétrotransposons sans LTR de type LINE-1 (*long interspersed element-1*, ou L1), la famille la plus importante de rétroéléments à réplication autonome chez les Mammifères. Les ETs ont un impact significatif sur l'organisation et le fonctionnement des génomes de Mammifères, en particulier du fait de leur amplification continue au cours des dernières 170 millions d'années (2–4). La réplication de l'élément L1 se fait via une séquence d'ARN intermédiaire copiée en ADN au niveau du site d'intégration (5–7). Ce mécanisme de réplication génère souvent des copies défectives tronquées à leur extrémité 5'. Ces copies sont classées en famille contenant des centaines à des milliers d'éléments partageant les mêmes variants nucléotidiques, hérités d'un progéniteur commun (ou d'un groupe de progéniteurs proches). Chez l'homme moderne, seule une minuscule fraction des éléments L1 est capable de générer de nouvelles copies de façon autonome. Toutes les copies potentiellement actives appartiennent à la sous famille L1HS (HS signifie *human-specific*), un sous-groupe de la famille des L1. Les autres familles sont des fossiles moléculaires d'anciens événements de rétrotransposition et ne sont plus mobilisés. La machinerie de rétrotransposition du L1 est aussi capable de mobiliser en *trans* quelques familles de rétrotransposons non-autonomes faisant partie de la classe des SINEs (*short-interspersed elements*, comme les séquences Alu ou SVA) ou encore des ARNs cellulaires (U6, mRNA), ce qui conduit à la formation de pseudogènes processés.

Un élément L1 entier a une longueur de l'ordre de 6 kb et contient un promoteur interne, localisé dans sa région 5' non traduite et code deux protéines, ORF1p et ORF2p, les deux étant requises pour la rétrotransposition. ORF1p est une protéine de liaison à l'ARN (8) et ORF2p possède des activités endonucléase et reverse transcriptase (9, 10). Les protéines ORF1p et ORF2p s'associent avec l'ARNm du L1 pour former une particule ribonucléoprotéique considérée comme le noyau de la machinerie de rétrotransposition (11, 12). Une nouvelle copie est produite quand ORF2p coupe l'ADN génomique cible et allonge l'extrémité 3' ainsi formée en utilisant l'ARNm du L1 comme matrice, un processus appelé *target-primed reverse transcription* (TPRT) (5, 7, 10) et conduisant à une courte duplication du site cible (TSD, *target-site duplication*). Lorsque la rétrotransposition est abortive, les copies formées sont tronquées au niveau de leur extrémité 5' (13, 14). Certaines insertions L1 sont caractérisées à la fois par une troncation 5' et par une inversion 5', du fait

d'un double amorçage (15). Les insertions L1 peuvent aussi contenir des transductions 5' ou 3', qui correspondent aux séquences génomiques localisées directement en amont ou en aval de leurs copies progénitrices. Un tel évènement se produit suite à la rétrotransposition de transcrits L1 initiés par un promoteur en amont du L1 ou se terminant en aval du L1 en raison d'un faible signal de polyadénylation (14, 16, 17). Le mode de ciblage du L1 dans le génome, et une éventuelle préférence pour certaines régions, ne sont actuellement pas entièrement définis. Néanmoins, la spécificité de l'endonucléase envers sa séquence consensus (A/TTTT) et la possibilité du site ciblé à s'hybrider partiellement à la queue poly(A) de l'ARNm L1 contribuent à ce processus (10, 18, 19).

L'analyse détaillée des mécanismes mutationnels à l'échelle du génome indique qu'environ 20 à 30% des variations structurales sont causées par des rétrotransposons sans LTR (20–23). Les fréquences de rétrotransposition des Alu, L1 et SVA sont estimées à un événement toutes les 21, 212 et 916 naissances, respectivement. En moyenne, chaque génome humain contient 1000-2000 rétrotransposons sans LTR polymorphiques, dont 79-85% d'Alu, 12-17% de L1s et 3% de SVA (20–26).

Les éléments L1 peuvent affecter notre génome de plusieurs façons. Premièrement, une insertion au niveau d'un exon peut modifier la séquence codante du gène affecté. D'autre part, la transduction d'une séquence flanquante en 3' d'un L1 peut contenir un exon, ou changer l'expression des gènes environnants en copiant des séquences régulatrices. Il a été estimé qu'environ 1% de l'ADN génomique humain a été transduit par L1, une proportion comparable à celle des exons dans le génome. Ceci souligne le rôle de L1 dans le brassage de l'ADN génomique et ainsi la plasticité du génome (27). Enfin, l'insertion du L1 dans un intron peut altérer significativement la structure de ce gène, en modifiant le processus d'épissage par rétention d'intron, par exonisation d'un fragment de L1 ou d'intron, ou par saut d'exon. Les transcrits altérés par des ETs ont souvent une expression spécifique de chaque tissu ou type cellulaire, apportant un niveau supplémentaire de régulation du transcriptome (28) .

La majorité des gènes humains subissent des phénomènes d'épissage alternatif (29). L'étude de la séquence des L1s révèle de nombreux sites donneurs et accepteur d'épissage potentiels. Certains de ces sites sont effectivement utilisés et conduisent à l'accumulation d'une large gamme de transcrits alternatifs de taille différente, réduisant l'accumulation d'ARN L1 complet et fonctionnel (30). D'autre part, l'étude des ESTs (*expressed-sequenced tags*) a montré que ces sites d'épissage internes aux éléments L1 peuvent être utilisés pendant la maturation des transcripts dans lesquels ces derniers sont insérés. Ce mécanisme contribue ainsi à la plasticité de notre génome et de notre transcriptome. L'introduction de nouveaux sites d'épissage par les rétrotransposons peut se traduire par une sévère perturbation des gènes de même qu'une création de nouveaux gènes codants ou non-codants (31–35).

La transcription de l'élément L1 par l'ARN Polymérase II est également interrompue par de nombreux signaux de polyadénylation présents tout le long de la séquence

du L1 (36). Certains de ces sites semblent même être plus efficaces que le signal de polyadenylation relativement faible présent à l'extrémité 3' de l'élément (37) . Ces signaux peuvent également impacter la terminaison de la transcription des gènes dans lesquels les L1s sont intégrés, en procurant des sites alternatifs de polyadénylation (38). Une polyadénylation prématurée peut ainsi aboutir à des transcrits, voire à de nouveaux isoformes protéiques tronqués à leur extrémité C-terminale.

Enfin, les L1s contiennent un promoteur antisens (ASP) dans leur extrémité 5' non-traduite. Cet ASP initie la transcription alternative de différents gènes comme *c-MET*, codant un récepteur tyrosine kinase dont l'activité peut causer la tumorigénecité dans différents types cancéreux (39–42).

Chez la plupart des Eucaryotes, dont l'Homme, les ETs jouent un rôle important dans l'expansion du répertoire des sites de fixation de facteur de transcription, et donc dans l'évolution des réseaux de régulation génique. Les ETs peuvent fournir des sites de liaison de facteurs de transcription prêts à utiliser, qu'ils apportent à leur site d'intégration (43–46). Ainsi, la transcription des gènes à proximité de ces ETs devient régulée par ces facteurs apportant une nouvelle forme de régulation (47–49). Les ETs, en dispersant et en combinant ces éléments régulateurs, ont largement contribué au développement de nouveaux réseaux de gènes chez les Eucaryotes (47).

Les L1s peuvent également générer et intégrer des rétrocopies d'ARNm cellulaires, produisant des pseudogènes "processés" dépourvues de certaines caractéristiques de leurs gènes parentaux, telles qu'introns ou promoteurs (50–53). Une partie des pseudogènes processés recrute parfois des séquences régulatrices en amont et peuvent devenir fonctionnels (54, 55), pour donner des rétrogènes. Environ 120 rétrogènes ont ainsi été répertoriés dans notre génome (50). Les rétrogènes font ainsi partie de la boîte à outil évolutive qui a conduit à la diversité transcirptionnelle (55–58).

La conservation évolutive de certaines copies d'ETs est susceptible de refléter des processus de domestication (47, 59–62). Environ 50 gènes codant des protéines humaines ont émergé par ce mécanisme et sont impliqués dans une grande variété de processus, parmi lesquels la régulation transcriptionelle, la prolifération et le cycle cellulaires, ou encore l'apoptose. Ils sont aussi à l'origine de longs ARNs non-codants (*long noncoding RNAs,* ou lncRNAs) (63). Le rôle moléculaire de ces derniers est encore mal connu, mais certains sont impliqués dans le remodelage de la chromatine et la régulation transcriptionelle (64).

Des transcrits L1 ou des transcrits chimériques contenant des séquences L1 ont été détectés dans différents types de cancer chez l'homme (tels que les cancers du testicule, de la vessie, du foie, du poumon, du sein ou du colon), aussi bien que dans différentes lignées cellulaires (65). L'hypométhylation du L1 peut être corrélée avec l'instabilité génomique dans différents cancers, comme dans le cas du cancer de poumon (66) ou bien avec des altérations transcriptionnelles, en particulier du fait de l'activité de ses promoteurs bidirectionnels (67, 68). Plusieurs études ont ainsi

montré l'implication des L1s dans la régulation épigénétique du développement embryonnaire et dans la tumorigenèse (69).

En conclusion, les éléments transposables, et plus particulièrement les L1s, sont une source importante de variation génétique qui a considérablement contribué à remodeler le transcriptome humain, à travers une grande variété de mécanismes (70).

BACKGROUND

1. Transposable elements have shaped the human genome

1.1. What are transposable elements (TEs)?

1.1.1. TEs are dispersed and repetitive genetic elements

Transposable elements - also known as “jumping genes” - are DNA sequences, capable of moving from one location to another within the genome. With rare exceptions, such as *Plasmodium falciparum*, "jumping genes" are present in all eukaryotic genomes (71).

Historically, the origin of transposon biology can be traced back from the beginning of genetics when Mendel published his experimental work on plant hybrids in 1865. Indeed, it was later shown that wrinkled (rr) seeds lack an isoform of the starch-branching enzyme (*SEBI*) present in round (RR or Rr) seeds. This is caused by a 0.8 kb insertion in the *SEBI* gene in (rr) lines, similar to the *Ac*/*Ds* family of transposable elements discovered later in maize (1).

Barbara McClintock first discovered DNA transposons in the 1940s. While working on maize cytogenetics, she observed spontaneous breakage and fusion of chromosome arms, which repeated over somatic and germinal cell divisions, at the same chromosomal position. She next identified two dominant and interacting genetic loci – *Dissociator* (*Ds*) and *Activator* (*Ac*), and in early 1948, she made the surprising discovery that both of them could change position on the chromosomes. McClintock observed that frequent chromosome breaks at the *Ds* locus on chromosome 9 appeared in an *Ac*- dependent manner. This was the first described case of interaction between mobile genetic elements later named non-autonomous and autonomous transposons. She also showed that mobilization of the *Ds* locus was correlated with the expression of the *C* gene (for color) and resulted in variegation of the kernel color. Based on these discoveries, McClintock proposed that *Ac* and *Ds* were ‘controlling elements’ that regulated the expression of other genes (72). Subsequently, other mobile elements were identified in different organisms: plants, bacteria, insects, mammals and also in humans (73).

1.1.2. TEs belong to different classes

Finnegan proposed the first classification of transposable elements in 1989 (74). He proposed two main categories: Class I transposons or retrotransposons, which use an RNA intermediate, and Class II transposons or DNA transposons, which use a DNA intermediate. These two classes of transposons are divided into sub-classes according to their structures and enzymatic properties (71, 75, 76). Most classes and subclasses comprise autonomous and non-autonomous elements. A general overview of transposable element classification and of the diversity of their structure is presented in Figure 1.

DNA transposons mobilize by cut-and-paste mechanisms in which the transposon is excised from one location and reintegrated elsewhere (2, 77). DNA transposons consist of a transposase gene, essential for their mobility, flanked by two Terminal Inverted Repeats (TIRs) (Figure 1). The transposase recognizes and cleaves TIRs to precisely excise transposon DNA, and reinsert it at a new genomic location. Upon insertion, the target site sequence is duplicated, resulting in Target Site Duplications (TSDs), a specific hallmark of each DNA transposon family. Generally, DNA transposons move through a non-replicative mechanism with the exception of *Helitron* and *Maverick* transposons (subclass-II), which do not generate double-strand DNA breaks during their mobilization but instead use a strand invasion mechanism (78). DNA transposons are classified into families depending on their sequence, TIRs or size. The known families in subclass-I are *Tc1*/*Mariner*, *PIF*/*Harbinger*, *hAT*, *Mutator*, *Merlin*, *Transib*, *P*, *PiggyBack,* and *CACTA*. The current families in subclass-II are *Helitron* and *Maverick.* As mentioned earlier, some families lack transposase-coding potential and are thus presumably dependent on autonomous DNA transposons for their mobilization. For example, *Miniature Inverted-repeat Transposable Elements* (MITEs) are short (80-500 bp) and abundant DNA transposon-like elements present in many eukaryotes, particularly plant species (79, 80), and occasionally in bacteria (81, 82). They are flanked by TSDs and have TIRs. DNA transposons have been extensively used as a functional genomics tools or transgenesis (83, 84).

Retrotransposons (class I) mobilize through the reverse transcription of an RNA intermediate, and the subsequent or concomitant integration into the genome. Thus retrotransposons are always replicative and their mobilization leads to an increase in copy number. Retrotransposons can be subdivided into two main groups: those containing Long-Terminal Repeats (LTR) and those that do not.

LTR-retrotransposons are very close to retroviruses since their structure and replication cycle share many characteristics. They are flanked by two LTRs, and contain PBS and PPT sequences, all required to achieve the synthesis of (-) and (+) strands during reverse transcription. LTRs have promoter and enhancer activities, and also contain functional polyadenylation signal, allowing retrotransposon RNA expression. The *GAG* gene encodes the structural protein of the viral capsid or virus-like, and the *POL* gene codes for a polyprotein with aspartic protease (PR), reverse transcriptase (RT), RNase H (RH) and integrase (INT) activities. Some elements, such as *Gypsy*, also encode an envelope gene (*ENV*) allowing an extracellular infectious phase.

Non-LTR-retrotransposons (also called target-primed (TP) retrotransposons), as implied by their name, do not contain LTRs and instead take on the likeness of an integrated mRNA. Non-LTR-retrotransposons are generally divided into two major groups: autonomous LINEs (Long INterspersed Elements) and non-autonomous SINEs (Short INterspersed Elements). This classification is based on the potential to code the replicative protein machinery necessary for “copy and paste” retrotransposition. LINEs can be further subdivided based on their RT domain into the R2, RTE, L1, I and Jockey clades. SINEs are often rearranged derivatives of

non-coding RNAs (tRNA, 7SL, 5S), which hijack the LINE machinery for their replication. They can be subdivided based on their RNA of origin.

Probably the ancestor of current retroelements was a retrotransposable element with both gag-like and pol-like genes (85). Further, comparison of RT sequences and mechanisms of mobility indicate that non-LTR-retrotransposons may have an evolutionary connection to group II introns (86, 87). Some studies also suggest an evolutionary link between non-LTR retroelements and the catalytic subunit of telomerase, based on the association of diverse non-LTR-retrotransposons with telomere-like functions in Drosophila, rotifers, stramenopiles, fungi, and plants (88, 89). Finally, modern retroviruses have emerged by the acquisition of an ENV gene by an LTR-retrotransposon (90).

1.1.3. TEs are abundant in eukaryotic genomes

Due to their mobility and their invasive nature, TEs can contribute to a significant portion of genomes. For example, they form at least 45% of the human genome (3), 37.5% of the mouse genome (91), 2.7% in the fugu fish, *Takifugu rubripes* (92), but nearly 85% of the genome of maize, *Zea mays* (93–95), and 41% of the dog genome (96). The proportion of transposable elements in plant genomes varies considerably (from 10%-85%).

Although polyploidy is common in plants, variability in genome size is also largely a consequence of mobile element expansion (97, 98).

The nature of the families, which have expended in distinct genomes is also highly variable. For example, LTR-retrotransposons are the most abundant transposable elements in plants. The corn genome is composed of 85% of transposable elements, including 75% of LTR-retrotransposons in which more than 300 families are represented (93). Many retrotransposons in LTRs families are relatively young (less than 4 million years), suggesting recent or contemporary mobilization (99). DNA transposons are also active in many plants, including the non-autonomous MITEs (100, 101). *Ac* and *Ds* transposons described in the historic preamble are also examples of active elements in a contemporary way.

Inversely, the genomes of *C. Elegans* and *D. melanogaster* are relatively compact and contain less TEs than many other organisms, representing 12% and 15% of their genome, respectively. The *C. Elegans* genome has mostly DNA transposons, some of which are still active (102), whereas the *D. melanogaster* genome contains a wide variety of active transposable elements, including both DNA transposons, such as *P* element, and retrotransposons from many distinct families (103, 104). *P* elements has invaded the wild population of *D. melanogaster* after the isolation of laboratory strains in the early 20th century (105), indicating a recent phenomenon.

Mammalian genomes are generally more consistent in size and widely invaded by TEs. With the exception of bats, LINEs are still widely active, including the L1 element. In brown bat *Myotis lucifugus*, DNA transposons are still largely active (106, 107).

Therefore, genome size is mainly due to the proportion of transposable elements, which results both from their rates of replication and of elimination, although other factors may also be involved, such as duplication mechanisms, polyploidy, and loss or gain of introns.

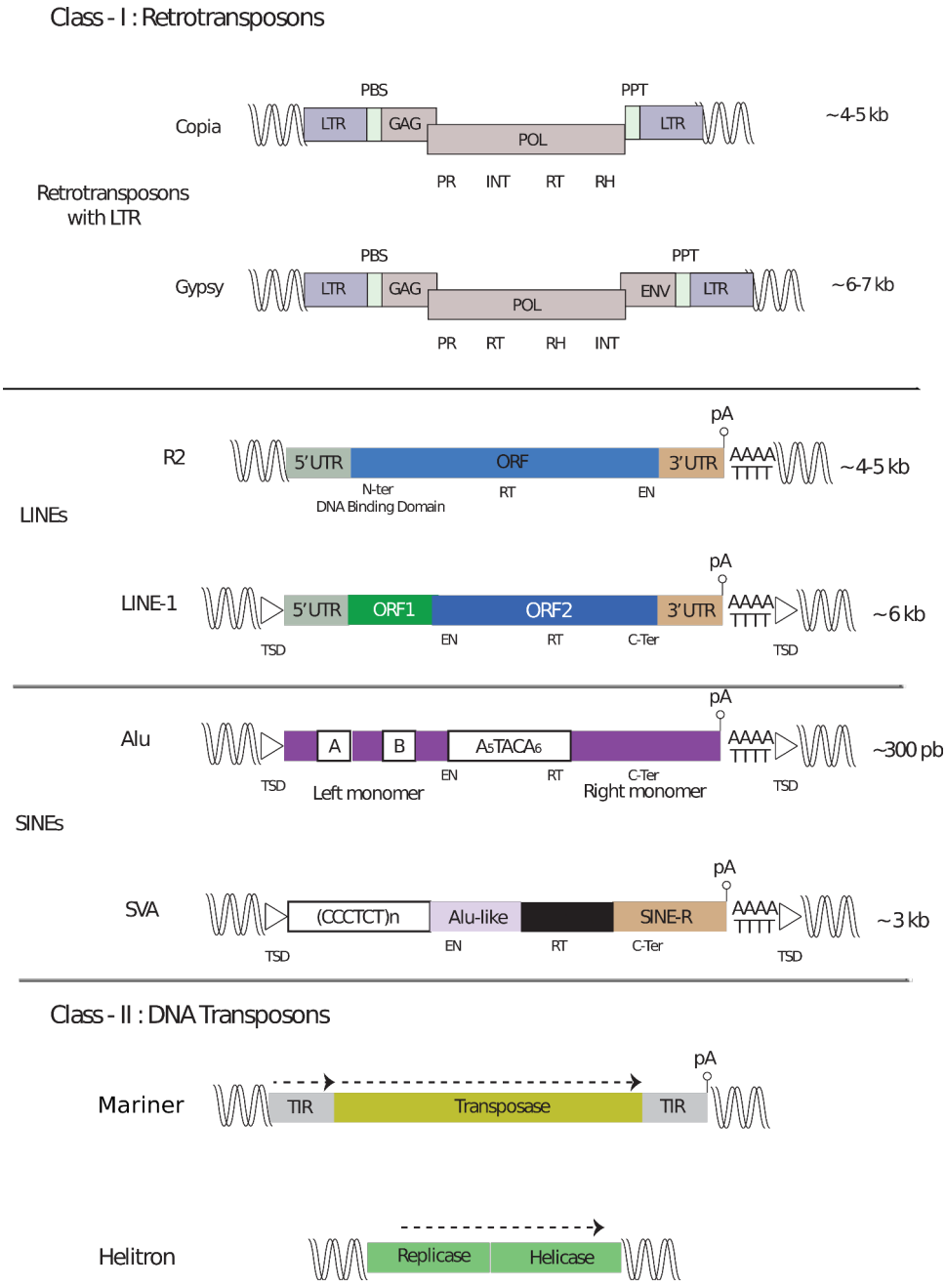


Figure 1: Schematic representation of selected elements in the main TE classes present in Eukaryotes. LTR-retrotransposons are represented by the *Gypsy* and *Copia* elements of *Drosophila melanogaster*. The LINEs are represented by the R2 element of *Bombyx mori* and human L1. The SINEs are represented by *Alu* and *SVA* elements. Alu consists of two monomers separated by a region rich in A. They have a bipartite promoter for the DNA polymerase III (A and B). The SVA element consists of a hexamer repeat (CCCTCT), followed by a region resembling (Alu-like), a minisatellite (Variable Number Tandem Repeat, VNTR) region and a SINE-R. DNA transposons are divided into two subclasses: The classical one encodes a transposase, flanked by inverted repeat sequence (TIR). The Helitron encodes recombinase (Rec) type "rolling circle" and DNA helicase (Hel).

1.2. Half or more of the human genome is composed of TEs

1.2.1. What are the TE classes present in the human genome, and which are the mobilized ones?

The composition of the human genome is depicted in Figure 2. TEs occupy nearly 45% of the genome (3). DNA transposons constitute only 3% and retrotransposons represent 42% of our DNA. The LINEs are the most abundant family representing 22% of the genome, from which L1 alone represents 17%. The SINEs are also present in abundance, Alu sequences representing 10% of the genome. Unlike plant genomes, LTR-containing elements (LTR-retrotransposons and human endogenous retroviruses, HERV) are less present (8% of the genome).

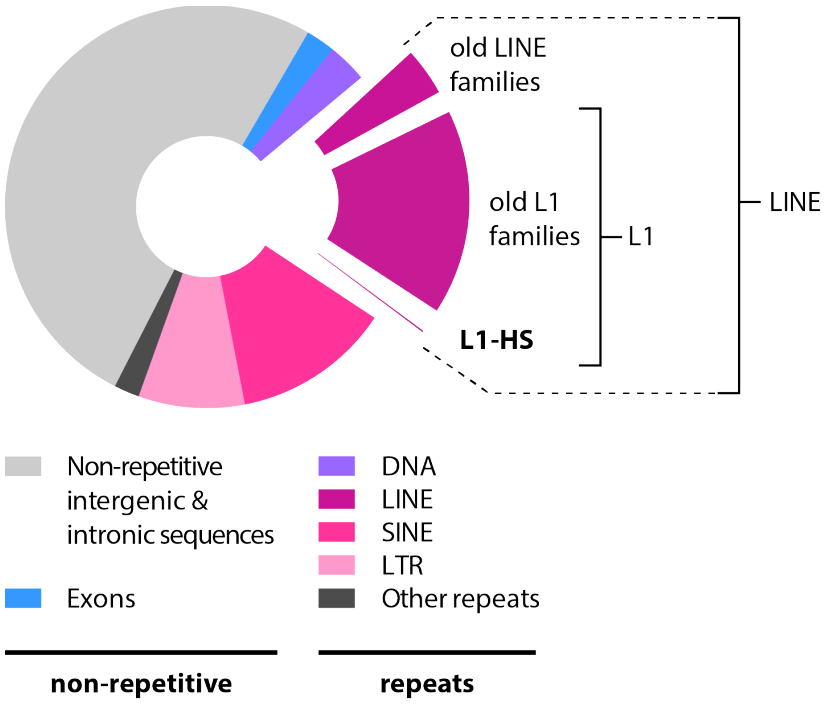


Figure 2: Proportion of repetitive elements in human reference genome. L1 forms ~17% of our genome (figure zoom out). L1HS represents ~3.3 Mb of the human reference genome (~0.1%). However, each individual also has additional non-reference L1HS copies, which contribute to our genetic diversity.

In humans, there are two major types of LTR-containing retroelements: human endogenous retroviruses (HERV) and mammalian apparent LTR-retrotransposons (MaLR) specific to mammals. Our genome contains ~ 200 000 HERV copies in size ranging from 6 to 11-kb, and encodes typical retroviral proteins such as a protease, reverse transcriptase, integrase, Gag structural protein and an Envelope protein (108). Human endogenous retroviruses (HERVs) are derived from ancient viral infections of germ cells, in which the viral DNA became permanently integrated within its host genome and as such is vertically transmitted to the next generation as any Mendelian trait (109). The MaLR elements are shorter (between 1.5 and 3 kb)

and have an open reading frame (ORF) with no clear homology with other known protein. However, this ORF is generally interrupted by multiple mutations, insertions, deletions, and truncations. At present, LTR-containing retroelements are incapable of replication, due to major deletions or nonsense mutations. However, the youngest HERV family, HERV-K, has been active after the divergence of humans and chimpanzees and some human individuals carry polymorphic copies of this virus (108). In addition, non-infectious HERV-K particle are produced in human embryonic cells (110).

Among the LINEs, the L1 clade has remained active in most mammals for ~100 million years and generated almost 17% of the human genome (111, 112). The first publication describing ~6.4 kb long LINE family derived sequence was published by J. Adams (73). They targeted the *beta-globin* gene in humans with various DNA probes and it was observed by Southern blotting that one of them binds to the DNA fragments of different sizes, suggesting the presence of a repeated sequence. The use of this probe in a library of human DNA confirmed that this sequence was at different locations in the genome. Kazazian published the first observation that L1 could still be active and create new insertions in the contemporary human genome (113). The first molecular clone of a competent retrotransposition element was isolated and studied by Dombroski (114). Only a tiny fraction of all L1 sequences is still able to autonomously generate new copies in modern humans. All the potentially active copies belong to the L1HS subfamily. Other families are molecular fossils of ancient retrotransposition events and are not mobilized anymore. A full-length human L1 is ~6.0 kb in length, contains an internal promoter located in the 5'-untranslated region (UTR) and two non-overlapping open-reading frames (*ORF1* and *ORF2*), separated by a short inter-ORF spacer. Both ORFs are required for retrotransposition. *ORF1* and *ORF2* encode a 40 kDa RNA-binding protein (ORF1p) and a 150 kDa protein with endonuclease (EN) and reverse transcriptase (RT) activities (ORF2p), respectively (115)(9, 10). The structural features of a full length L1 are shown in Figure 3. Shortly, a new L1 copy is produced when ORF2p nicks the genomic DNA and extends this newly formed 3' end using the L1 mRNA as a template, a process known as target-primed reverse transcription (TPRT) (5, 10, 14). Short duplications at the target site (TSD, target-site duplication) are formed as a result of this process. Abortive retrotransposition often leads to 5' truncated L1 copies (13, 14). Some L1 insertions exhibit both a 5' truncation and a 5' inversion, due to twin-priming (116). Finally, L1 insertions can also contain 5'- or 3'-transductions. L1 target site preference is currently not fully defined, but both the endonuclease consensus sequence and the ability of the target site to partially anneal to the L1 mRNA poly(A) tail contribute to this process (10, 18, 19). Each aspect of this process will be developed in the following sections.

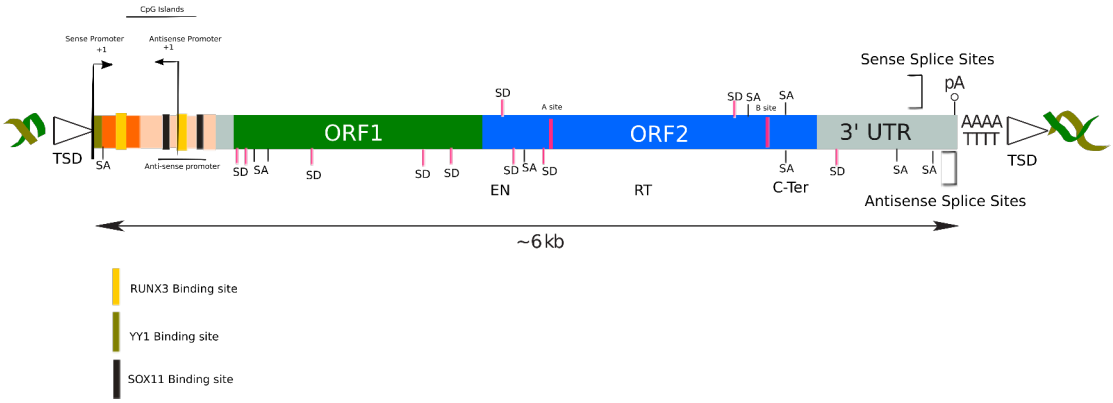


Figure 3: Structure of the L1 element. A prototype L1 element is approximately 6kb in length and is surrounded by target-site duplications (TSD). The 5' UTR region is shown in multiple colors depicting the location of transcription factor different binding sites, CpG Islands and bi-directional promoters. *ORF1* and *ORF2* are represented as green and blue boxes, respectively. The endonuclease (EN), reverse transcriptase (RT), C-terminal (C-ter) domains are shown below *ORF2*. The 3'UTR ends with a polyadenylation site and is followed by a poly(A) tail. Cryptic splice acceptor and donor sites are shown as short black and red ticks, respectively.

The L1 retrotransposon machinery is also able to mobilize *in trans* a restricted number of non-autonomous retrotransposons families belonging to the SINE class (Alu, SVA, see Figure 1). SINEs are very heterogeneous in sequence. Their lengths range from 100 bp to several kb (117–119). MIR (Mammalian-wide Interspersed Repeat) is an ancient family of tRNA-derived SINEs (120, 121) found in all mammals, which shows its ancient origin (122), with no evidence of recent retrotransposition activity. Alu sequences are primate-specific SINEs consisting of a duplicated region derived from the 7SL RNA (123, 124). *Alu* elements are the most abundant human retrotransposons (by number of copies), represented by ~ 1.2 million copies per haploid genome (3). They rose ~65 million years ago and radiated into nearly 30 Alu subfamilies. Only a small subset of *Alu* elements is thought to be currently retrotransposition competent in humans (125, 126). The active *Alu* elements within our genome derive from the Young (Y) subfamily and include Ya5, Ya5a2, Ya8, Yb8, Yb9, Yc1, and Yc2. Alu elements are 300 bp long, composed of two arms separated by an A-rich tract, and variable in polyA tail length (Figure 1). Alu are flanked by short direct repeats that are a remnant of the retrotransposition process. They harbor a bipartite RNA polymerase III promoter. *Alu* elements are non-coding elements and thus their mobilization depends on L1 replicative machinery. L1-encoded ORF2p is essential for *Alu* retrotransposition, whereas L1 ORF1p only enhances this process (127). Although, the criteria required for *Alu* activity are still not fully elucidated, the promoter integrity as well as the length and homogeneity of the polyA tail have been suggested as principal factors determining the retrotransposition capability of these elements (126, 128).

Finally, SVA elements form a composite SINE family. SVAs were originally named "SINEs-R", with the "R" indicating a sequence of retroviral origin. SVA consists of a hexamer repeat (CCCTCT), an Alu-like sequence, a GC-rich Variable Number Tandem Repeat (VNTR), a Short Interspersed Nuclear Elements (SINE) and a poly A-tail (129) (Figure 1). The flanking hexamer is also a VNTR (130). SVA elements

represent only 0.13% of the genome, with ~2 700 copies. Thus it constitutes the youngest retroelement in the human genome and is hominid-specific. Their replication mechanism is slightly different from that of Alu elements: it is likely transcribed by RNA Polymerase II, and requires both L1 ORF1p and ORF2p for its mobilization (131). SVA elements can vary in length from ~1000–4000 bp with 63% of SVA element insertions in the human genome being full-length, containing all five domains (129, 130). SVA elements are divided into subtypes (A-F) based on the SINE region and recently a 7th subtype has been identified to contain a 5′ transduction of the sequence from *MAST2* gene referred to as CpG-SVA, MAST2 SVA or SVA F1 element (132, 133).

1.2.2. The human genome has fingerprints of primate TE evolutionary history

Each transposable element family and subfamilies have gone through distinct periods of transcriptional activity during which they have spread over the genome. This has been usually followed by insertions, deletions and rearrangements and then inactivation periods and formation of new subfamilies (134). Vertical persistence of non-LTR-retrotransposons on an evolutionary scale in both mammals and primates sets them apart from the other TEs in mammals (3, 135). Based on diagnostic nucleotides substitutions and indels, L1, Alu and SVA can be subdivided into subfamilies. Diagnostic sequence mutations, which define subfamilies, have been shown to accumulate hierarchically apart from age factor (125, 136).

Whereas L1, Alu and SVA have continued their amplification from million of years ago, other non-LTR-retrotransposons which comprise almost ~6% of the human genome represent molecular fossils which is a proof for long relationship between transposable elements and the human genome (3). For long term evolution retrotransposons have adopted attenuation of mobilization strategy (137, 138). Non-LTR-retrotransposons are thought to follow a «master gene» model of amplification. Thus, these so called source elements are responsible for the formation of all other subfamily members (136).

L1 (L1) retrotransposons are the most abundant family of autonomously replicating retroelements in mammals. Their continuous amplification over the last ~170 million years (Myr) has had a significant impact on the organization and function of mammalian genomes (2–4). L1 retrotransposition often generates defective copies that are truncated at their 5′ end. Resultant copies are classified into families of hundreds to thousands of elements based on the shared nucleotide differences they inherit from their common progenitor(s). Most L1 copies accumulate mutations at the neutral rate (139–142). Thus, older families are more divergent than younger. In humans phylogenetic studies have shown that, over the long-term, a single L1 lineage amplified over the last 25 Myr (143, 144). Families of closely related variants can occasionally coexist for short periods of time (142, 145) until one family dominates and prevails in the replicative process. Competition between L1 families, most probably for a limiting host factor, could account for this pattern of evolution (145, 146). L1 families have been frequently recruiting novel 5′ UTRs in the Primates. Similar patterns of evolution have been observed in mouse, where L1

families acquired novel 5′ UTRs at least twice in the past 5-6 Myr (147, 148). The lack of homology between primates, mouse, rat, and rabbit 5′ UTRs also suggests that the acquisition of novel 5′ UTRs in mammals is a fundamental feature of L1 evolution (147–154). The 5′ end of *ORF1* (from nucleotide 12 to 396) underwent an episode of positive selection that occurred during the evolution of families L1PA8-L1PA3 (155). In contrast, this region has remained amazingly conserved during the evolution of older (L1PA16 to L1PA8), with the exception of family L1PA13B) and younger (L1PA2 and L1PA1) families. It suggests that the strength or nature of the selective pressure that has driven the rapid evolution of this region has changed over time. It was recently proposed that positive selection in *ORF1* could reflect an adaptation of L1 to its hosts (144, 156).

The rate of L1 amplification has slowly decreased in the Primate lineage over the last 25 Myr (3). Correlations between evolutionary radiations and bursts of amplification (157) suggest that history of populations, especially the occurrence of population bottlenecks (158), can possibly affect the dynamics of L1 amplification. Positive or negative interactions of a host factor with L1 replicative machinery is also thought to be responsible for the episodic nature of L1 amplification (141, 144, 148, 156). After analysis, it has been found that L1 families show considerable variation in their copy numbers, which suggests large differences in their replicative success in the absence of known specific elimination process. The most intense period of L1 activity concerns families L1PA8 to L1PA3 and lasted from ~40 Myr to ~12 Myr. The amplification of these very successful families is also indirectly responsible for the amplification of the bulk of AluY elements and of many processed pseudogenes (125, 159).

The L1 subfamilies that are specific only for humans, L1HS-PreTa and L1HS-Ta (human specific, transcribed, subset a) emerged ~4 Myr, somewhat after divergence among humans and chimpanzees (~6 Myr). The PreTa subfamily is evolutionarily older and thus is believed to predate the amplification of the Ta subfamily in the human lineage (142, 160). The Ta subfamily has subsequently differentiated into two major subsets, Ta0 and Ta1, each of which spawned additional subsets. All of them harbor a distinctive trinucleotide sequence (ACA) in their 3’ UTR (at position 5930-5932), which is a diagnostic sequence for the L1HS elements (142). The L1HS-Ta1 accounts currently for a replicative dominant subfamily in the human genome. They have a distinctive T at nucleotide 5536 and G at position 5539. Out of 459 L1HS-Ta elements in the reference human genome, 192 belong to the Ta1 and 137 to the Ta0 subsets, respectively. The remaining 130 elements are either truncated or rearranged in the diagnostic region or represent the intermediates between the two subsets (160).

1.3. Ancient TE copies have been selected through human evolution and contribute to genomic or physiological functions

1.3.1. TEs sequences can be under positive selection

There are many evidences that TEs are significant players in the evolution of genomes (4, 100, 101, 161–165). Evolutionary conservation of TEs is likely to reflect

the molecular domestication of the respective elements (47, 59–62). Except for some kind of negative selective pressure, inserted TEs can become fixed in the genome of a species and serve as a source for novel genetic loci. In other cases, accumulated mutations have caused neofunctionalization of inserted TEs. This process is referred to as exaptation (or molecular domestication or co-option). Positive selective pressure for maintenance of co-opted TEs reflects a beneficial function performed by the novel gene product. The process of TE exaptation has contributed significantly to the human genome. Over 10,000 TE-derived genomic regions have been subject to strong purifying selection (166) and ~50 protein-coding genes have arisen via this mechanism (62). Domesticated genes have been found to be involved in a variety of cellular processes, including transcriptional regulation, proliferation, cell cycle progression, and apoptosis. A wider survey of conserved non-exonic elements (CNEEs) in 29 mammalian genomes, has revealed almost 280,000 putative regulatory elements originating from TEs (167).

1.3.2. TEs contribute to transcriptional networks through the dispersion of regulatory elements

Cis-regulatory sequences and their evolution is believed to change the transcriptional output and have an impact on speciation (168). Alternate gene promoters are presumed to contribute in this regard (169). According to a study, ~18% of human genes, are having alternative promoters (170). LTR seems to be acting as a gene promoter and is often one of the alternative promoters. Interestingly it does not alter the coding sequence and thus regulates nearby human genes (170–173). At a genome-wide level, the Faulkner laboratory has observed that retrotransposons, which are located next to the 5' of protein-coding loci, are frequently functioning as alternative promoters (or express noncoding RNAs) (174).

TEs can also provide new transcription factor binding sites to promoters or to create novel enhancers, without affecting transcription start sites (43–46). Indeed, TEs have played an important role in expanding the repertoire of protein binding sites in mammalian genomes. A large part of transcription factor binding sites, such as (ESR1, TP53, POU5F1, SOX2, CCTV, and CTCF) are embedded in distinctive families of transposable elements or relics of these elements (47, 175–177). In fact, transposable elements have facilitated species-specific binding sites. Finally, binding motifs within repeats seem to be under selection (47). Gene transcription near transposable elements is regulated by these factors, bringing a new form of regulation (47–49). TEs, their rearrangement and replication of these regulatory elements, have largely contributed to the development of new gene networks in eukaryotes (47). Thus, repeat elements bound by transcription factors act as critical "control elements" in eukaryotic genomes (178–182). Changes in the regulatory elements can possibly have important phenotypic effects across species (183–187) and also within populations. Examples include various human diseases, such as Alzheimer (188), obesity (189), and cancer (190). Below we describe, a few selected examples.

MIR elements, an ancient SINE family, can donate transcription-factor binding sites (191, 192), enhancers (43, 193, 194), microRNAs (195, 196) and *cis* natural

antisense transcripts (197) to the human genome. The association of MIRs with tissue-specific expression, along with their propensity to be exapted as regulatory sequences, suggests possibility of a role in providing numerous tissue-specific regulatory sequences across the human genome (198).

Another example of how TEs link genes within a network can be observed in embryonic stem cells. LTR-derived transcripts contribute to the complexity of the stem cell nuclear transcriptome. They were found to be associated with enhancer regions. Thus most probably involved in the maintenance of pluripotency (199). This is consistent with the recent findings showing that a transcriptional network, controlled by ERV LTRs, act as a switch to determine if embryonic stem cells can stay in pluripotent or transient phase of totipotency (49). This is controlled by epigenetic modifications of LTRs. ERVs are transcriptionally repressed in the pluripotent state by histone H3K9 trimethylation. Histone methyltransferase activity is recruited to ERVs by Kap1 (200). Embryonic stem cells, which are deficient for Kap1, can switch more easily to the totipotent state, indicating that relaxation of ERV repression could drive network activation (49). This shows the critical role ERVs are playing in host cell fate decisions by activating transcriptional networks.

A last striking case is related to the evolution of pregnancy in mammals, including humans. The differentiation of endometrial stromal cells during the decidual reaction, which precedes embryo implantation, is triggered by hormone progesterone (201). This phenomenon relies on a hormone-dependent transcriptional network under the control of a subfamily of hAT-Charlie DNA transposon, the *MER20* elements, which provides binding sites for transcription factors acting downstream of progesterone-responsive signaling molecules (202).

Therefore, TE can be coopted for the evolution of regulatory networks and of complex physiological processes in humans.

1.3.3. Exaptation of TE sequences has led to mammalian- or human-specific proteins

As mentioned previously (§ 1.3.1), TEs have also contributed coding sequences. One such prominent example are the mammalian-wide interspersed repeat elements (MIRs), an ancient family of tRNA-derived SINEs, whose retrotransposition history traces back to 130 million years ago, even before the mammalian radiation. MIRs have persisted and probably helped in evolving mammalian-specific or even hominoid-specific functions since the exaptation process can occur anytime after retrotransposition. Consistently, Krull *et al.* found that 107 out of 126 MIR-derived proteins identified in mammalian databases are also detected in humans (203). Interestingly, one of them, *CHRNA1,* which encodes an acetylcholine receptor, is specific to the great Apes.

Although the exact contribution of TEs to the proteome has been discussed, some authors have suggested that thousands of proteins contain sequences resulting from TE exonization in vertebrate genomes including humans (204)(205, 206).

1.3.4. Retrogenes are functionalized copies of retrotransposed mRNAs

Retrogene can be defined as an intact retrocopy of a gene showing evidence of transcription. Retrocopies are generated by the L1 machinery (see §2.2.4). Retrotransposition can provide raw material for generating new genes (54, 207). Most retrocopies are only processed pseudogenes and lack their parental gene features, such as introns or promoter (50, 51). However, some of them recruit upstream regulatory elements and can become functional (54, 55), thus turning into retrogenes.

Retrogenes have been identified in many genomes, and are particularly abundant in mammals (50, 51, 58, 208). Retrofamilies are shared between different species. The reason for this could be the homology between L1s among species, which drives their formation, leading to enzymatic activities with similar specificities. Therefore, the general pattern of retrotransposition dynamics could be similar among mammals. Consistently, retrogene formation of ribosome-related genes is particularly enriched in mammals, as shown by comparing LSRs (lineage specific retrofamilies), IOSRs (independently-occurred shared retrofamilies), and non-IOSRs retrogenes. Almost 28% of the IOSRs have ribosome-related gene families, in contrast to only 2.6% for the non-IOSRs retrofamilies (209). In humans, almost 120 cases of retroposed sequences have been found to evolve into *bona fide* genes (50).

The impact of the retrogenes can be important. Recently, for example, the oncogenic role of *NanogP8,* a human tumor-specific retrogene homolog of *Nanog*, was investigated in transgenic mice. High levels of *NanogP8* expression disrupts normal developmental programs and thus inhibit tumor development by depleting stem cells (210). Another example illustrates the ability of retrocopies to reshuffle functional domains. The *PIPSL* retrogene, which undergoes rapid adaptative evolution, is specific to the hominoid lineage and results from the fusion of phosphatidylinositol-4-phosphate 5-kinase (PIP5K1A) and 26S proteasome subunit (S5a/PSMD4) (211). Retrocopy-mediated domain shuffling provides extraordinary diverse functions to the proteins involved thus playing a role in phenotypic evolution.

Therefore, transposable elements have played a crucial role in the formation of new genes and diversification of gene functions in genomes.

2. L1 elements are the only autonomous TEs in the human genome and are endogenous mutagens

2.1. The L1 machinery is a ribonucleoprotein particle (RNP)

2.1.1. How is L1 RNA synthesized?

Polyadenylated L1 mRNA was first isolated from a human teratocarcinoma cell line (NTera2D1) (212). The majority of these RNAs corresponded to full length L1 transcripts from various loci, with ORFs interrupted by premature stop codons (213). Initial experiments suggested that L1 was transcribed by RNA Polymerase III (214), however L1 sequence is extremely AT-rich and has numerous Pol III termination signals (TTTT), excluding such a possibility.

The 5' UTR region L1 is a sequence of about 900 bp. It contains both sense and antisense internal promoters. Using chimeric constructs containing the L1 5' UTR upstream of reporter genes, such as chloramphenicol acyltransferase (CAT) or β-galactosidase (β-gal), it was shown that it contains an internal and TATA-less RNA Polymerase II promoter (215–217). Deletion analyses has further shown that the first 150 nucleotides form its core and, more broadly, the first 670 nucleotides contribute to transcriptional activation (216, 217) (Figure **4**). Although L1 transcription is initiated primarily from this internal promoter at the first nucleotide of the 5' UTR region, transcription may also occasionally start upstream of the element from a promoter located in the genomic 5' flanking sequence (218, 219).

The antisense promoter (ASP) resides between nucleotides 400 and 600 (Figure **4**) and drives transcription opposite to the L1 sense promoter and ORFs (39). Further characterization of the ASP identified two initiation sites around positions 378-431 and 480-497. The nucleotide sequence downstream of these sites increase the activity of the ASP (220). Other less-frequent transcription initiation sites have been identified within the 5' UTR indicating that different transcripts could be formed by the same DNA sequence (221). The activity of the ASP can impact the transcription of nearby regions (see §0).

Several transcription factors binding sites were identified in L1 5' UTR (217, 218, 220, 222, 223). A binding site for the transcription factor YY1 has been located between nucleotides +13 to +21 (217, 218, 222). Although this site does not seem essential for the transcription and expression of L1, it is essential for the accuracy of initiation at nucleotide +1 (218). Two binding sites for the SRY family of transcription factors (SOX11) were identified at nucleotides 472 and 572 and this factor modulates L1 transcription levels (223). More recently, it has been shown that RUNX3 binds to the 5' UTR region from nucleotide 83-101 (220) and modulates the transcription and retrotransposition of L1. Finally, the 5' UTR region also contains a CpG island, which can be highly methylated (224). L1 promoter activities are repressed by the methyl-CpG-binding protein-2 (MeCP2) and DNA methylation (225, 226).

At the other extremity, L1 contains a polyadenylation signal, which is moderately effective. As a consequence, L1 transcripts frequently extend into the 3' flanking genomic sequence (16, 17, 27, 227). The initiation and termination of transcription of L1 are thus influenced by the genomic context where the element is inserted. RNAs initiated or completed in the flanking region can ultimately produce 5' or 3' transductions when used as a template during reverse transcription.

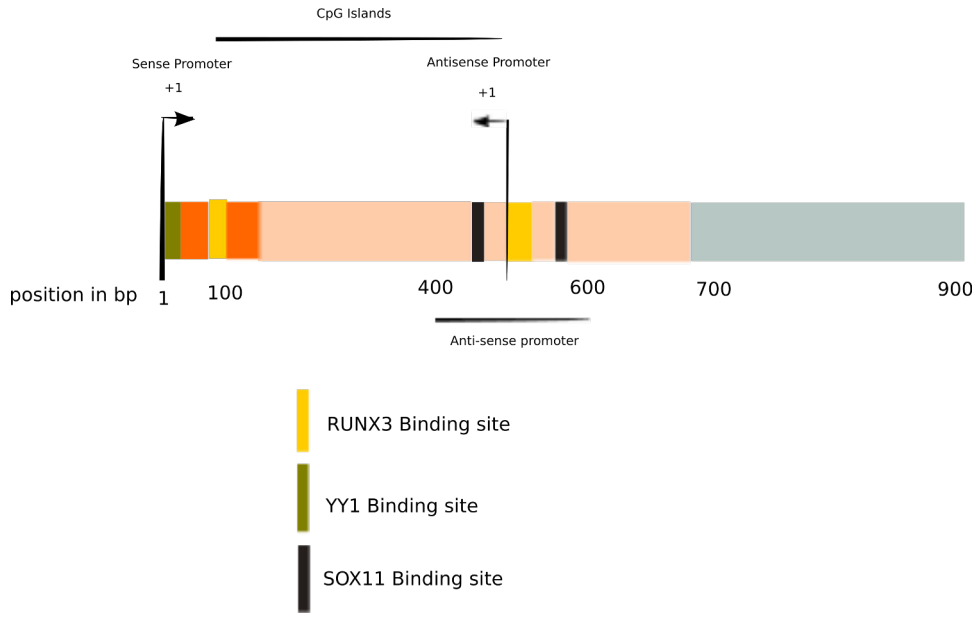


Figure 4: Structure of the L1 5' UTR region and its internal promoters. The L1 5' UTR region is a sequence of ~900 bp. Arrows indicate transcription start sites for sense and antisense promoters. The main regions responsible for promoter activities are shown in bright colors. The region required for the antisense promoter activity is indicated by a brace. Binding sites for the various transcription factors involved in L1 transcription are colored as indicated in the legend. A bar above the promoter marks a CpG island region.

Several cryptic splice sites and polyadenylation signals are also dispersed within L1 sequence and can lead to alternative or truncated transcripts not competent for retrotransposition (30, 32, 36, 228)(251). In cells, the L1 element is expressed from multiple loci (69, 174, 227). The flexibility of initiation and termination of transcription, coupled with alternative splicing, can therefore, explain the heterogeneity and diversity of transcripts observed in various cell types (228). Additionally, Cap Analysis Gene Expression (CAGE) approaches have also highlighted the possibility that a significant number of truncated L1 fragments can also generate transcripts from their 3' region (174). This phenomenon could be related to the presence of *Sox/LEF* sites in the inner region of L1, especially in *ORF2* (230).

Finally, there is little information about the export of full length (unspliced) L1 RNA from the nucleus to the cytoplasm. Unspliced or partially spliced RNAs are retained in the nucleus by commitment factors (231). It has been suggested that L1 mRNA might contain *cis*-acting elements required for its export from the nucleus to the cytoplasm (232, 233). Indeed, some intronless mRNAs expressed from transfected complementary DNA (cDNAs) are not exported efficiently, and several viruses have evolved *cis*-acting elements to facilitate nuclear export of unspliced RNA (234, 235).

However, the existence of *cis*-acting factors that affect L1 mRNA nuclear export is still a speculation that awaits experimental validation.

2.1.2. L1 encodes two functional proteins, ORF1p and ORF2p

The L1 RNA is bicistronic, encoding two non-overlapping open reading frames, *ORF1* and *ORF2,* separated by a 63-base spacer. Their protein products (ORF1p and ORF2p) bind the L1 RNA to form a ribonucleoprotein (RNP) complex that is presumed to be a critical retrotransposition intermediate. ORF2p is expressed at a significantly lower level than ORF1p. This difference likely results from the mechanism of ORF2p translation, a low-frequency ribosome reinitiation mechanism (236).

The first intact *ORF1* coding sequence was found by sequence analysis of a mouse L1 element called L1Md-A2 (237). Subsequently human ORF1p has been detected in human teratocarcinoma cell lines (238, 239). ORF1p, also known as p40, is a basic RNA-binding protein of 40k Da, able to formed a ribonucleoprotein particle (RNP) complex with the L1 RNA (115, 240, 241), a property necessary for retrotransposition (242).

ORF1p protein contains three domains: a coiled-coil domain with a leucine zipper motif, a non-canonical RNA recognition motif (RRM) (243) and a C-terminal domain (CTD) (244). In 3D, ORF1p folds into a trimeric and asymetric dumbbell structure (245) (Figure 5). The coiled-coil domain forms a supercoiled helix allowing trimerization (238, 240, 246), the RRM has a globular shape and is located at right angles to the coiled-coil domain. CTD and RRM domains are located one above the other and cooperate to bind nucleic acids (247). Interestingly, the coiled-coil domain has been submitted to positive selection suggesting that it is linked to evolutionary adaptation or extinction of human L1 lineages, and likely reflects the ability of ORF1p to attract or avoid interactions with other factors (144).
Experiments using murine ORF1p (mORF1p), which is very close to human ORF1p, showed that it can bind RNA of at least 38 nucleotides (nt), with no apparent sequence-specificity (246), except a slight preference for the sense transcript of the L1 relative to an antisense transcript (248). Human ORF1p (hORF1p) stably binds poly(rA) RNA oligonucleotides of 27 nt. It can also bind DNA, but with a clear preference for oligo(dT) sequences compared to oligo(dA) (247). Several mutants reducing the capacity of the ORF1p to bind RNA also reduce or abolish retrotransposition. ORF1p also binds a variety of different cellular RNAs *in vivo* as shown by PAR-CLIP (249). Finally, it has been recently demonstrated that L1 activity requires phosphorylation of ORF1p protein at S/T residues in the context of four conserved proline-directed protein kinase (PDPK) target sites (250).

Figure 5: Structure of the human L1 ORF1p trimer. The figure above shows the trimeric form of ORF1p. Each monomer is represented by a different color tint. (PDB accession Number: 2YKO). Made using CBSN PDB protein workshop from (247). (Source PDB)

ORF1p has nucleic acid chaperone properties similar to retroviral nucleocapsid proteins. This class of factors facilitates rearrangements of nucleic acid structures to their thermodynamically most stable form (251–253). Chaperoning activity was first associated with murine ORF1p purified from baculovirus-infected insect cells. It was found to greatly enhance annealing, strand exchange, and duplex melting of short DNA oligonucleotides *in vitro*. These properties were sequence-independent and occurred at an equimolar concentration of protein and DNA (254). The nucleic acid chaperone activity of both human and murine ORF1p is required for retrotransposition. A single-point mutation that abrogates chaperone activity (R297K) without affecting RNA- or single-stranded-DNA binding affinity, or RNP formation also diminishes or abolishes L1 retrotransposition (12, 255, 256)(242). The precise role of this activity in L1 replication is unknown, but it was hypothesized that it may be required during reverse transcription, to allow or to stabilize the formation of RNA-DNA duplexes during first and/or second strand DNA synthesis.

L1 ORF2p is a 150 kDa protein with two known enzymatic activities that can be assigned to specific domains (9, 10). The N-terminal part of the protein contains an endonuclease domain (EN), the sequence and structure of which are very similar to apurinic/apyrimidinic endonucleases. The central part of the protein is a reverse transcriptase domain (RT), which allows the synthesis of an L1 cDNA from the L1 mRNA. ORF2p also includes a C-terminal cysteine-rich domain of unknown function with a predicted zinc finger. ORF2p is 40 times less expressed than ORF1p (257), presumably due to its non-canonical mechanism of translation (see §2.2.1). ORF2p

is notoriously difficult to express in human cells or in a heterologous host. Therefore, this protein has been only poorly studied from a biochemical perspective.

The endonuclease activity of ORF2p was first identified in 1996. Recombinant ORF2p was expressed in and purified from bacteria, and its crystal structure was obtained in 2004 (258). L1 EN belongs to an enzyme family of metal-dependent phosphohydrolases that cleave variable phosphoester substrates (259, 260). Purified L1 EN protein (L1 ENp) can nick supercoiled plasmids *in vitro* (10) and hence is believed to cleave the L1 target site, initiating the insertion process and generating an extremity for reverse transcription priming. EN targets a consensus sequence 5'-AA/TTTT-3' but various variants are tolerated (13, 261–263). Accordingly, these *in vitro* cleavage sites are very similar to those found at Alu and L1 retrotransposon insertions *in vivo* (121, 122, 263). Point mutations in EN catalytic site destroy its activity, and abolish L1 retrotransposition in most cell types, demonstrating the importance of the endonuclease in this process.

Another essential property of ORF2p is its reverse transcriptase (RT) activity. RTs are RNA- and DNA-dependent DNA polymerases, able to generate complementary DNA (cDNA) from an RNA template by a process termed reverse transcription. L1 RT activity was first detected in macromolecular complexes purified from the teratocarcinoma cell line NTera2-D1 (264). After cloning the first active human L1 (114), the RT activity of ORF2p was demonstrated by domain swapping with a well-characterized yeast LTR-retrotransposon, for which genetic tools were uniquely available at the time (9). This was later confirmed by adapting a genetic system originally developed by T. Heidmann for retroviruses, showing that L1 replication in mammalian cells is mediated by an RNA intermediate and a reverse transcription step, which absolutely requires the conserved catalytic residues of L1 RT (37). It is sensitive to several reverse transcriptase inhibitors, such as AZT or d4T (265–267). Finally, recombinant ORF2p purified from insect cells was able to recapitulate several aspects of the retrotransposition reaction *in vitro*, although with very low efficiency (7). L1 RT seems to be very processive compared to the other viral reverse transcriptase (75, 268). In addition, ORF2p, in complex with the ORF1p and its RNA, is capable of extending a primer containing one or more terminal mismatches (11, 269).

The carboxy-terminal region has been characterized recently and was shown to bind single-stranded RNA but not double-stranded DNA by electrophoretic mobility shift assay (EMSA) *in vitro* (268). Although zinc finger motifs can be involved in nucleic acid binding, cysteine mutations do not affect the ability of this domain to bind RNA *in vitro*.

Figure 6: Structure of the endonuclease domain of ORF2p. The protein chain is colored from the N-terminal to the C-terminal using a rainbow color gradient. Made with CBSN PDB Protein Workshop using data from (258). (PDB accession Number 1VYB, Source PDB).

2.1.3. L1-encoded proteins assemble with the L1 RNA to form an RNP

Early crosslinking experiments in human teratocarcinoma cells indicated that ORF1p binds directly to the L1 RNA *in vivo* to form sedimentable RNP complexes (240). Using genetically and biochemically tagged L1 elements, it was later shown that ORF1p, ORF2p and the L1 RNA form a stable RNP complex and that L1 proteins preferentially bind *in cis* on their encoding RNA (11, 12, 52, 53). The so-called L1 RNP is considered as a major functional intermediate in the retrotransposition process.

2.2. L1 retrotransposition can occur through multiple mechanisms

2.2.1. Overview of L1 replication cycle

The replication cycle of the L1 element (Figure 7) consists in 3 major steps:

- L1 transcription;
- L1 proteins translation and assembly of a functional L1 RNP;
- L1 reverse transcription and integration.

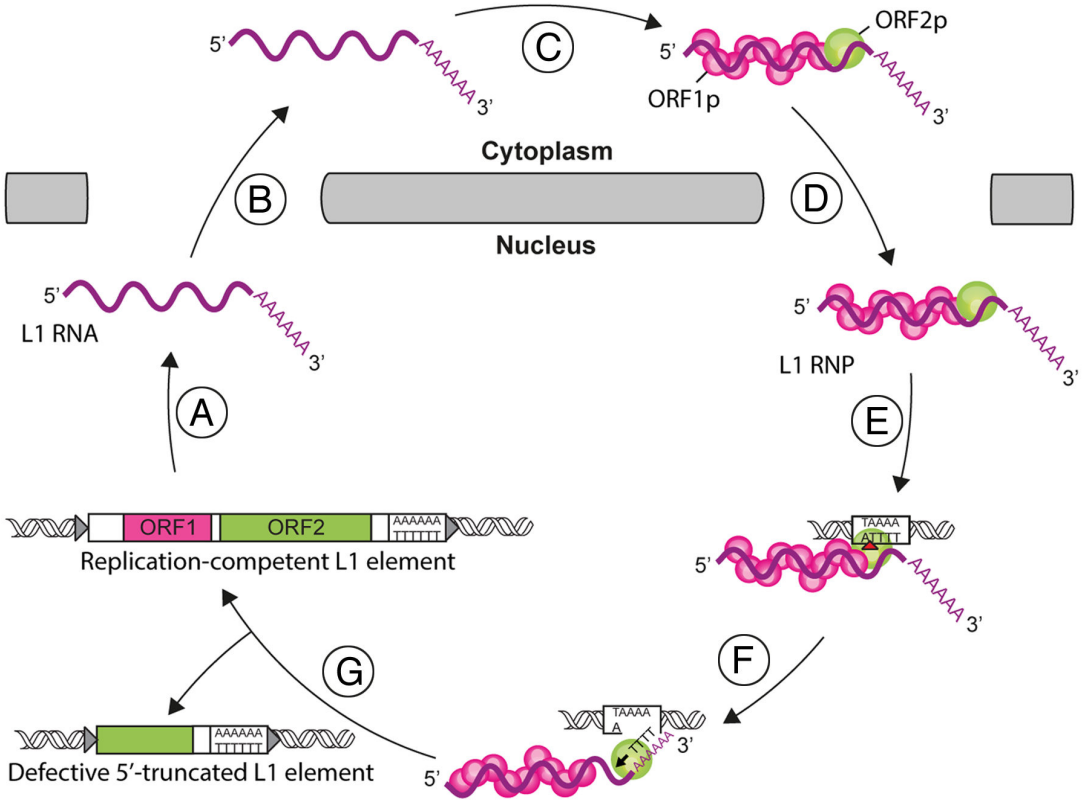


Figure 7: L1 life cycle. L1 replication starts with the transcription of a bicistronic mRNA (A). L1 RNA is then exported to the cytoplasm (B). Next, ORF1p and ORF2p proteins are translated and bind to the L1 RNA forming L1 ribonucleoprotein particles (RNP) (C). L1 RNP is then imported into the nucleus (D). Integration and reverse transcription process occur at the genomic target site. First, the L1 endonuclease (EN) activity nicks the target DNA (red arrowhead, E). Then, the L1 reverse transcriptase (RT) initiates the reverse transcription of L1 RNA through annealing between the target site and the poly(A) tail of the L1 RNA (black arrowhead, F). The mechanisms involved in the final steps of this process and the resolution of the integration are unresolved yet (G). Partial reverse transcription leads to the 5'-truncated L1 copies. Source: (19).

L1 transcription has been explained earlier in chapter 2.1.1. Once transcribed, L1 mRNA is transported to the cytoplasm, where the host ribosomal machinery is subsequently used to synthesize L1 proteins. In Eukaryotes, there are two major mechanisms of translation initiation: cap-dependent scanning and internal ribosome entry sites (IRES). The first one is the main mechanism for the majority of cellular mRNAs, whereas many viruses and some cellular mRNAs that are translated under particular conditions use the latter. IRES are functionally defined by their ability to promote independent translation of the second cistron in a bicistronic RNA (270).

Canonical cap-dependent translation follows a scanning model, which postulates that the 40S ribosome subunit binds to the m7G cap at the 5‘ end of the transcript, followed by linear scanning until the first AUG in the appropriate Kozak initiation context (271). Insertion of a stable secondary structure hairpin in the 5’ UTR of L1 greatly decreases the expression of ORF1p (272), suggesting that the initiation of translation of the ORF1p takes place according to this model. *ORF2* is located downstream of *ORF1*. Thus it raises the question of the mechanism of its translation. In principle, ORF2p could be translated from the long bicistronic L1 transcript, but also from a sub-genomic L1 transcript. Indeed, as mentioned previously, L1 is capable of generating different types of transcripts, by alternative splicing and/or premature polyadenylation events (30, 32, 228). Some of these spliced forms can also lead to the synthesis of functional ORF2p sufficient to mobilize SINEs (228). However, replication of L1 based on ORF1p or ORF2p expression from distinct constructs (trans-complementation) is inefficient (11, 52, 53). Indeed, L1 proteins have a cis-preference for their own RNA reinforcing the idea that ORF2p is translated from a bicistronic RNA. Early studies have suggested the presence of an IRES (Internal Ribosome Entry Site) in the inter-ORF region for synthesizing ORF2p (272, 273). IRES are RNA structures that allow assembly of the ribosome independently of the cap and thus enable an internal translation initiation (274, 275).

A study by Alisch helped to better understand the characteristics of the translation of the second ORF of L1 (236). First, deleting the inter-ORF sequence does not drastically reduce L1 retrotransposition. Second, the addition of a premature stop codon in *ORF1* prevents L1 retrotransposition and mobilization of Alu (which relies on ORF2p expression only). Third, the distance between the stop codon of *ORF1* and the start of *ORF2* is crucial for enabling ORF2p translation. Finally, mutating ORF2p start codon from AUG (methionine) to CCC (Proline), or UAA (stop codon) has no significant effect on the mobilization of the L1, suggesting that initiation is AUG independent. Altogether, these observations go against an IRES-mediated mechanism, and rather support a model by which ORF2p translation would be led by an unconventional mechanism of termination-reinitiation.

Various cellular factors have been identified over time that could be required for the translation of L1 proteins such as Nucleolin, which promotes the translation of ORF2p (276). Different members of the poly(A)-binding protein family (PABP) found to interact with L1 RNPs were also strong candidates (257). Indeed, these proteins are known to be necessary for the stabilization of RNA but also for translation. Among them, PABPC1 binds to mRNA within the cytoplasm and interacts with eIF4E to enable mRNAs to adopt a circularized structure necessary for the initiation of translation. PABPC1 positively regulate L1 retrotransposition, as shown in the knockdown of PABPC1 (277). However, the translation of L1 proteins is only very slightly affected by PABPC1, suggesting that it could be involved in stages downstream of translation such as the assembly and stability of L1 RNP or reverse transcription itself.

Post-translational modifications or protein processing of L1 ORF1p and ORF2p proteins are currently unknown. Since proteins larger than approximately 60 kDa are

too large to enter the nucleus by passive diffusion through the nuclear pore, the access of L1 RNPs to genomic DNA should either occur by energy-dependent, active transport through a nuclear pore, or by entry during nuclear membrane breakdown during cell division (278). Against the second possibility, L1 is able to retrotranspose in non-dividing cells (279).

Next, the integration of new copies of the L1 element can take place using two distinct molecular mechanisms, involving different biochemical properties of the L1 ribonucleoprotein complexes. The first is called target-primed reverse transcription (TPRT), requires the endonuclease activity of ORF2p, and is the preferred integration route. The second is endonuclease-independent and utilizes pre-existing DNA lesions.

2.2.2. Target-primed reverse transcription (TPRT) is a major pathway of L1 insertion

Non-LTR-retrotransposons insert into eukaryotic genomes by target-primed reverse transcription (TPRT), a process by which cleaved DNA targets are used to prime reverse transcription using retrotransposon RNA as a template. This mechanism of insertion possibly originates from mobile group II introns found in bacteria. The TPRT model was established through the study of the R2 non-LTR-retrotransposon in *Bombyx mori*. This element, consisting of a single open reading frame encoding a protein with site-specific endonuclease and reverse transcriptase activities, specifically fits in the ribosomal DNA (encoding the 28S RNA) (5, 280). In *in-vitro* assays, recombinant R2 protein is able to nick DNA, but only perform double-stranded DNA cleavage in presence of RNA. In the case of R2 element, the last 250 nucleotides of the 3' UTR are necessary to enable reverse transcription to initiate (6). The R2 protein has two DNA binding domains at the N-terminal and C-terminal, which may respectively link sequences downstream and upstream of the cleavage site in a dimeric complex (281).

The current model of R2 retrotransposition includes the following steps: (i) the endonuclease of the upstream monomer cleaves the first (bottom) DNA strand, (ii) the reverse transcriptase of the upstream monomer uses the free 3′ OH from the newly created nick to initiate target-primed reverse transcription (TPRT) using the R2 RNA as the template, (iii) the downstream monomer cleaves the second (top) DNA strand, and (iv) the second DNA strand is synthesized. It is not known if R2 or cellular DNA polymerases are responsible for the fourth step, however, the R2 reverse transcriptase is capable of displacing RNA from nucleic acid templates and the second subunit is likely to be in the correct orientation to perform second strand synthesis (281–283). The basic steps of this TPRT reaction appear to be part of the integration reaction of other non-LTR-retrotransposons (37, 284) as well as in the integration of SINEs (Alu) and processed pseudogenes (52, 285). TPRT is also thought to be involved in retrohoming of group II introns (286).

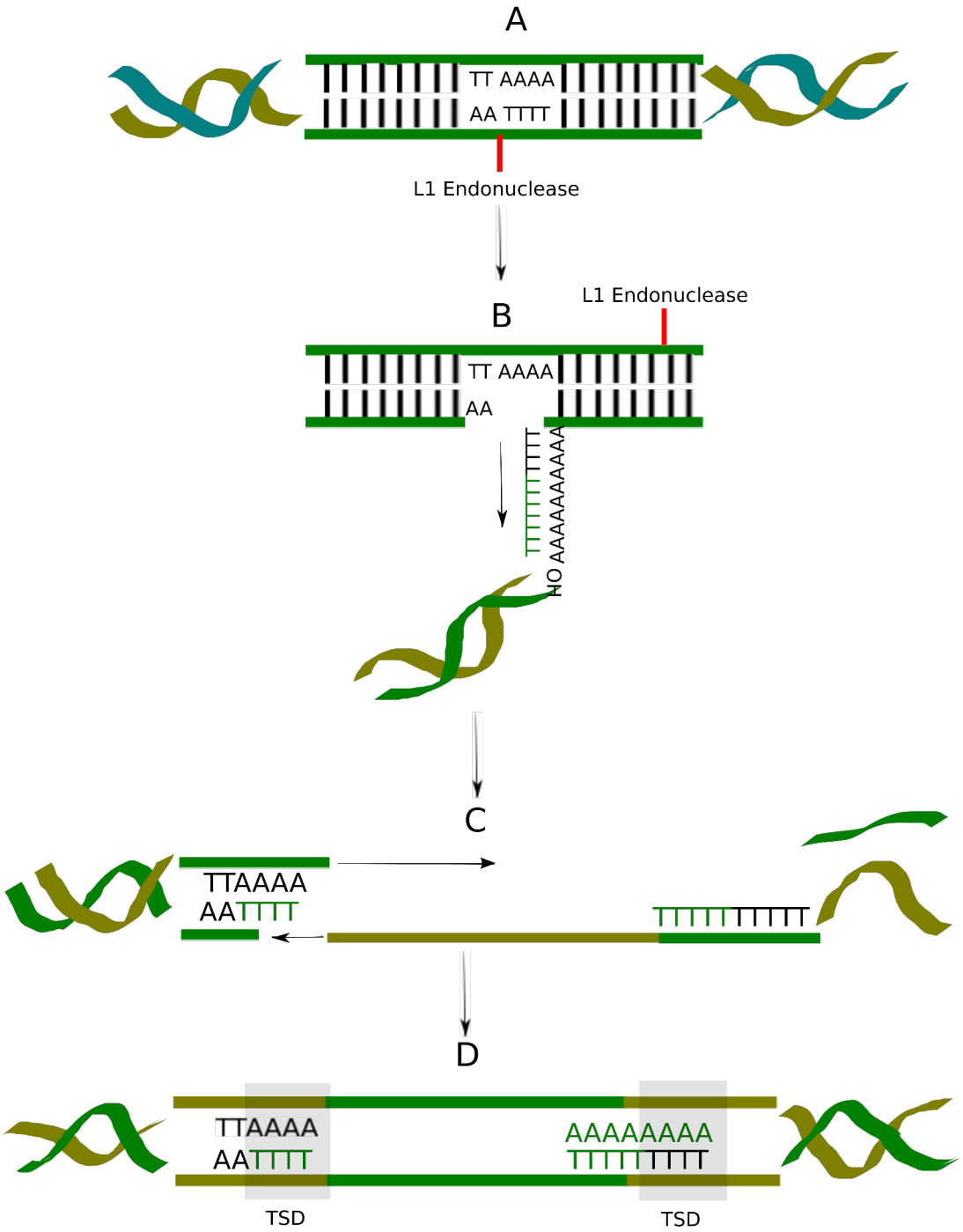


Figure 8: Reverse transcription at the integration site (TPRT). (A) L1 endonuclease generates a single DNA strand cleavage at its target sequence. (B) The reverse transcription is initiated using the free 3' OH end as primer. (C) A second cleavage at the other DNA strand is produced. (D) The second L1 DNA strand is synthesized and the DNA of L1 is ligated to the chromosomal DNA by unknown mechanisms. This process leads to a new insertion with the integration site duplication (TSD). The size of the TSD is the distance between the two cleavage sites and is generally between 4 and 20 nt.

L1 belongs to a different non-LTR-retrotransposon clade and encodes an additional protein (ORF1p), as compared to R2. Therefore, the question of a possible common insertion mechanism for all non-LTR-retrotransposons arises. Early *in vitro* studies using ORF2p from purified L1 showed that ORF2p was capable of synthesizing a

cDNA of the L1 RNA at the target site of the endonuclease (7). However, this experimental approach does not take into account the presence of ORF1p nor the specificity of the native L1 RNPs (assembled in cis). The vast majority of insertions obtained in cell culture are 5' truncated. Only 5% produce a new full-length L1 element (262). They are often padded with duplication of the target sequence (TSD) of variable size. They also contain a variable length of repeating 'A' which corresponds to the reverse transcription of the poly(A) tail. Most of the L1 insertions occur into sequences related to the L1 EN consensus sequence (degenerate 5′-TTTT/A-3′ sites) and frequently preceded by imperfect T-tracts. Nonetheless, less frequently the cut may take place between C/A, G/A or A/A. Assuming that reverse transcription is initiated by matching the poly(A) tail at the insertion site, this suggests that L1 RT can tolerate terminal mismatches, which was confirmed *in vitro* by the LEAP technique (L1 Element Amplification Protocol) (11).

One of the unresolved questions related to L1 reverse transcription priming was whether or to which degree the 3' end of the nicked genomic DNA needs to be accessible and to base-pair with the poly(A) tail of the L1 RNA. Although the consensus sequence released upon L1 EN cleavage (5′-TTTT-3) could in principle anneal to the poly(A) tail of the L1 RNA, it is extremely short for maintaining a stable interaction and the actual sequences cleaved by the L1 EN can significantly differ from the consensus sequence. Monot *et al.* addressed this question by quantifying the efficiency of extension of a vast collection of primers by direct L1 extension assay (DLEA), and found that efficiency of reverse transcription initiation is influenced by the last 10 nucleotides of the target DNA.

Inserts containing an entire L1 element are usually padded with duplication at the site of insertion (262). Sometimes they also contain additional non-templated guanosine at their 5', which could result from the reverse transcription of the cap. The truncated elements can also be associated with deletions of the target site (13, 14) or with an inverted 5' L1 fragment (262). The latter events result from a phenomenon called twin-priming (116) (see Figure 9). This is a variant of the canonical TPRT process, wherein the second strand of the target DNA is cleaved prior to the end of reverse transcription and primes a second reverse transcription reaction from an internal region of the L1 RNA. These two parallel reverse transcripts will then result in two inverted L1 fragments flanked by TSDs.

Chimeric L1 insertions or pseudogenes were also observed (262). A similar phenomenon was observed for R2 in *Bombyx mori* (282). In the case of R2, the RT is able to add additional nucleotides at the end of the synthesis of cDNA, which can serve as a primer for another RNA by template switching. This could also explain the formation of chimeric pseudogenes with L1 fragments (287).

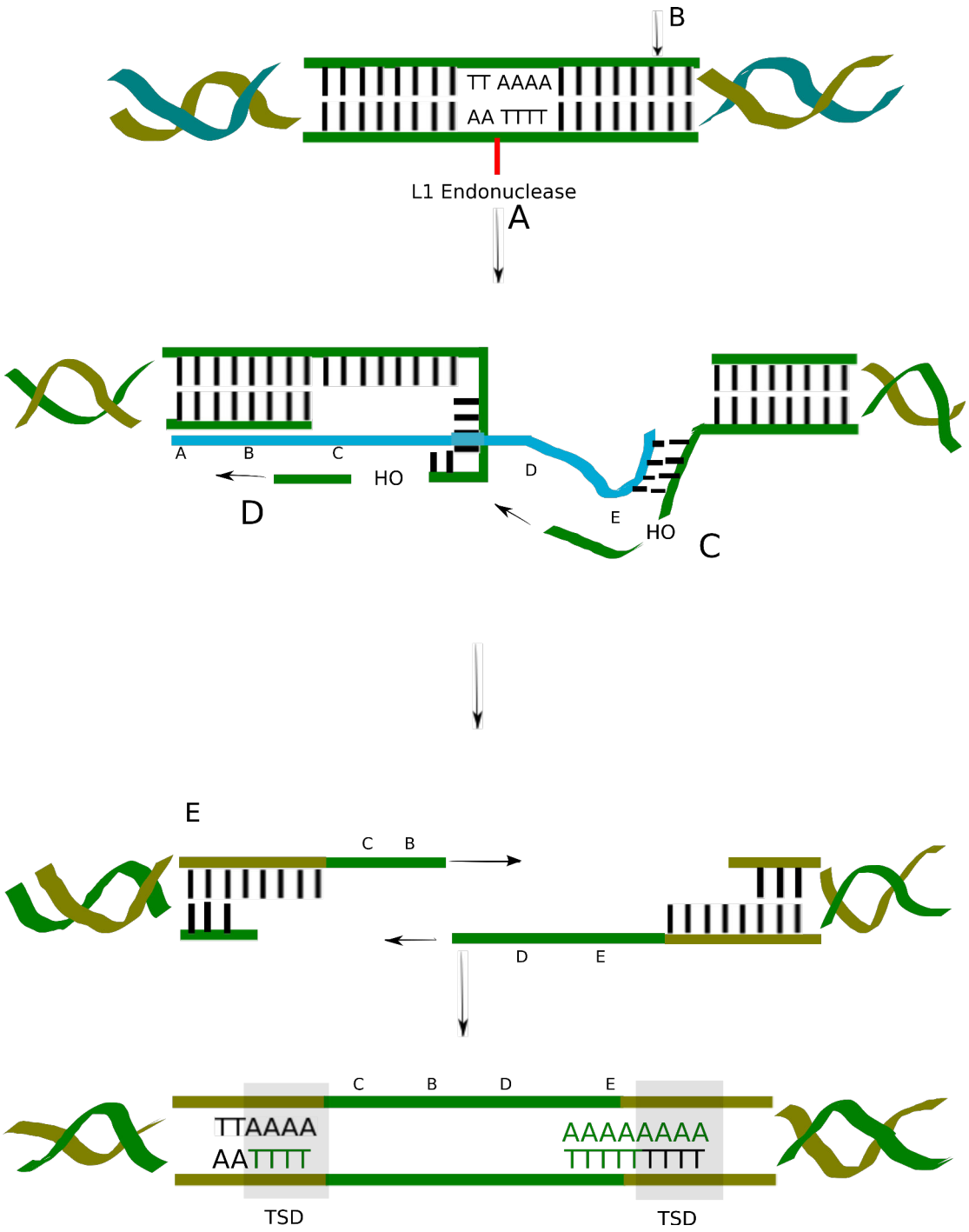


Figure 9: Twin-priming mechanism. (A) The L1 endonuclease generates a single DNA strand cleavage at its target sequence. (B) Second cleavage at the other DNA strand is produced. (C) Reverse transcription is initiated at the 3' end of the L1 RNA using the free 3'-OH end as a primer. (D) L1 RNA invades the second DNA strand and a second reverse transcription reaction is initiated internally. (E) The second L1 DNA strand is synthesized and the DNA of L1 is ligated to the chromosomal DNA by unknown mechanisms. This phenomenon is characterized by the insertion of a 5'-truncated element with a 5' inversion, bordered by TSD. Adapted from (116).

Many aspects of the TPRT process still remain unclear, such as the necessity to unwind DNA at the target site after cutting by the endonuclease. Similarly, the need for hybridization between the target genomic DNA and the L1 RNA has not been demonstrated. Finally, the steps between first strand cDNA synthesis and the

resolution of the integration are still very poorly understood. Analysis of the 5' L1 junctions with genomic DNA reveals pairings between the L1 and the target DNA at the insertion of a truncated element (288).

The majority of L1 copies have been inserted through a TPRT mechanism, however some inserts do not show the hallmarks of this process (TSD, poly(A), EN consensus sequence) suggesting that other mechanisms insertions are possible.

2.2.3. L1 can also insert through endonuclease-independent mechanisms

In an alternative to TPRT integration pathway, some L1 can initiate reverse transcription independently of their EN activity. This phenomenon was first observed in cells deficient in non-homologous end-joining (NHEJ), a DNA double-strand break repair pathway, and mutated L1s with point mutations in the EN domain (289). The characteristics of such events are: (i) the integration site does not correspond to the consensus sequence of the endonuclease; (ii) the insertion is not flanked by a duplication at the integration site, but rather often associated with deletions at the integration site; (iii) L1 sequence may be 3'-truncated and, therefore, presents no poly(dA) (289). Bioinformatics analysis identified 21 insertions in the reference human genome as endonuclease independent (290). Alu sequences can also be mobilized through this alternative pathway and act as DNA-patch to repair double-strand DNA breaks (290–292).

Finally, another study reported that EN-deficient L1 could integrate into dysfunctional telomeres, taking advantage of free 3'OH groups at the ends of chromosomes (293). Actually, polymerase chain reaction (PCR) amplification of genomic DNA of cells deficient for DNA-PKcs (an NHEJ factor) showed that 30% of all endonuclease-independent insertions occur at telomeres (293)(269), reinforcing the idea of a potential mechanistic and evolutionary link between telomerase reverse transcriptase and ORF2p (294, 295).

2.2.4. The L1 machinery can mobilize other RNA in trans

L1 encoded proteins can mobilize other cellular RNAs like SINEs Alu, SVA and also U6 snRNAs in *trans*. There mobilization has be shown to be L1 dependent (129, 285, 296). ORF2p is required for the mobilization of SINEs as shown by trans-complementation tests. However, *ORF1* does not seem to be necessary for Alu mobilization (285) (297), but might stimulate it when the expression of the tested Alu construct is driven by RNA polymerase II instead of RNA polymerase III (267). Both *ORF1* and *ORF2* are required for efficient retrotransposition of U6 snRNA though (287). ORF1p presence seems be to be required for SVA.10 unlike SVA.2. SVA.10 is longer and hence the difference could be because of transcript size (298). However, it should be noted that these trans-mobilization tests do not exclude the possibility that endogenous ORF1p is sufficient for this trans-mobilization.

L1 machinery also mobilize in trans cellular RNA which leads to the formation of pseudogenes (287, 299). This mobilization requires both ORF1p and ORF2p. Thus, overtaking of the L1 machinery by the host gene mRNA leads to host gene retrotransposition and results in processed pseudogene (PPs) formation or retrogenes creation. Processed pseudogenes are copies of mRNAs, which are

reverse transcribed into DNA and inserted into the genome using the enzymatic activities of active L1 elements. The human genome contains numerous copies of pseudogenes from coding or noncoding genes (300–303). Processed pseudogenes have following features: 1) their sequences are very similar to the transcribed portion of the parent gene; 2) they lack all or most introns, so they appear to be cDNA copies of processed mRNAs; 3) they have a poly(dA) tail attached to their 3' end; and 4) they are flanked by target site duplications (TSDs) of 5 to 20 nucleotides. Some processed pseudogenes are formed by template switching and are called chimeras. Processed pseudogenes differ from other pseudogenes, which arise by DNA duplication, contain introns and are located in close proximity to their active gene copies.

Among more than 14,000 pseudogenes present in the human genome (207), at least 10% are no longer 'pseudo'-genes and are active (207, 304). Processed pseudogenes are signs of mobilization by the endonuclease and reverse transcriptase activities of active L1 (L1) elements (13, 52). More than 2,075 human genes are represented by at least one PP in the genome, while some genes, such as *GAPDH*, ribosomal proteins, and actin β have 50 to 100 PPs (Pei 2012). Recently, Mandal found 48 novel PP insertion sites among 939 low pass genomes from the 1,000 genomes project (249). They also found first instances of somatic insertion of PPs; three PPs were predicted to occur in lung cancers that were absent from paired normal tissue. Other studies have demonstrated PP polymorphism in humans (305–307). The majority of PP insertions in cancer have TSDs of 5 to 20 base pairs, 74% were 5' truncated (a percentage similar to that of human-specific L1s), 20% had inversions at their 5' ends due to 'twin priming' (15), and long poly(dA) tracts. In lung adenocarcinoma, one insertion was observed to be associated with an 8 kb deletion of the promoter and exon 1 of a tumor suppressor gene, *MGA1*, leading to a functional knock out as determined by RNA-seq. De Boer *et al.* recently showed the potential for PP formation during early development in humans in a case of X-linked disorder (chronic granulomatous disease) (308). Overall, there is overwhelming evidence that PPs continue to insert in the germline and in somatic cells of human beings.

Finally, Doucet *et al.* identified distinct recruiting steps during the L1 retrotransposition cycle for the formation of snRNA-processed pseudogenes by analyzing genomic structures and retrotransposition signatures associated with small nuclear RNA (snRNA) sequences. They found that some of these recruiting steps take place in the nucleus, and established that snRNA amplification by template switching is common to many LINE families from several LINE clades. They suggest that U6 snRNA copies can serve as markers of L1 retrotransposition dynamics in mammalian genomes (309).

2.3. L1 retrotransposition is a source of structural variation and a mutagenic process

2.3.1. Multiple methods have been developed to track L1 retrotransposition in humans

Next generation sequencing technologies have been pivotal in mapping L1 insertions and exploring the extent of L1 insertion polymorphisms or somatic retrotransposition in humans. Therefore, we will start by giving an overview of these methods.

- **Introduction to sequencing technologies**

Sanger sequencing technology was introduced by Frederick Sanger, which is based on the chain termination method. Later on Walter Gilbert developed another method, which was based on chemical modification of DNA. First generation of sequencing technologies involved Sanger sequencing. In 1987, Applied Biosystems introduced capillary electrophoresis. Sanger capillary sequencing was the technology behind the completion of the human genome project in 2001. Later on, Roche 454, Illumina (previously Solexa) and SOLiD brought next-generation sequencing technologies on the market, followed by Ion Torrent and Pacific Biosciences. A major breakthrough of these approaches was to massively parallelize sequencing by performing single-molecule DNA amplification in partitioned populations. Each company developed unique approaches to achieve this, coupled to different sequencing methods, with variable outputs. This is summarized in Table 1. The development and commercialization of 454 and SOLiD systems are now discontinued, showing that this field is extremely quickly evolving.

	Sanger	Roche 454[a]	Illumina[b]	SOLiD[c]	PacBio[d]	Ion Torrent[e]
Partitioning	n/a	emulsion PCR	Cluster formation on flow cell	Amplification on flow chip[c]	Single-molecule	emulsion PCR
Sequence by	synthesis	synthesis	synthesis	ligation	synthesis	synthesis
Detection	Radiolabeled or fluo.	Indirect luciferase	Fluo.	Fluo.	Fluo.	pH
Throughput	n/a	700 Mb	1 Tb	120 Gb	16 Gb	2 Gb
Read length	1 kb	800 bp	2x125 bp	75 bp	20 kb	400 bp
Sequence format	sequence-space	sequence-space	sequence-space	color-space	sequence-space	sequence -space

Table 1: Summary of next-generation sequencing techniques. Throughputs and read lengths are indicative since they highly dependent on a particular model of machine, given a specific technology. Except for Illumina, the indicated read length is a median. Fluo, fluorecence. *a*, GS FLX+ System; *b*, HiSeq 2500 with high-output option (note that other machines with lower throughput can output 2x250 bp paired reads); *c*, SOLiD 5500 W system with wildfire technology and single-end fragments; *d*, PacBio RS II system (1 Gb/SMRT cell; up to 16 SMRTcells/run); *e*, Ion Torrent PGM with Ion 318 chip and 400 bp mode.

Two major methodological improvements have played an important role in expending the range of next generation sequencing (NGS) applications. First, paired-end sequencing has been developed to reduce mapping and assembly ambiguities due to short reads. It involves sequencing of both ends of DNA

fragments in a sequencing library and then aligning forward and reverse reads as pairs. Library with different fragment lengths can be generated to resolve variations at different scales. Paired-end reads can be aligned more accurately and used to detect larger indels or other forms of variations in contrast to single-end reads. it can also help in discriminating and removing PCR duplicates. Second, multiplexing which allows to pool many libraries together in a single run, has increased the sample throughput per run. Unique index sequences are added to DNA fragments during library preparation. This allows identifying and sorting each read before final data analysis. This has dramatically reduced the processing time and sequencing costs. Latest NGS platforms are highly scalable and available for every method and the scale of study.

Whole genome sequencing has been used to obtain full genomes of various plant species, livestock or disease-causing microbes. It is also useful to sequence wide range of human genomes to understand disease and variation marks across populations. A striking example was the sequencing of an E. coli bacterial strain in 2011, which caused a disease outbreak in Europe, allowing tracing its origin and understanding its increased virulence.

In order to zoom down to the coding part of our genome, exome sequencing has been used more often recently. It is a cheaper alternative and can be more effective for population's genetics, cancer or disease genetics studies, in which a large number of individuals or samples need to be analyzed. Recently, even more targeted sequencing has been extensively used to focus on areas of interest, thus enabling higher coverage than usually achieved for whole genome sequencing (500x – 1000x or even higher, instead of 20x-50x). This is required to detect and identify rare variants, such as somatic mutations in cancer samples. Two methods are currently used: target enrichment or amplicon generation methods. While target enrichment can capture around 20 kb to 62 Mb regions, amplicon sequencing can sequence 26-1536 targets at a time, which could span 150 bp to 1.5 kb per target. Applications are diverse like targeting specific pathways, phylogenic or taxonomic studies especially metagenomics samples.

De novo sequencing has been extensively used to sequence novel genomes for which no existing reference genome is available, to assemble it into contigs, and eventually into chromosomes. It often combines sequencing of long insert mate pairs along with short insert paired-end reads to get maximal coverage across the genome. This enables the resolution of repetitive regions of the genome and detection of a wide range of structural variation types to identify even more complex rearrangements. A huge research effort has been done to develop efficient and accurate assembling softwares, and several high-class genome and transcriptome assemblers are currently available.

More specifically, a number of L1 detection methods at the genome-wide level have been described. They been described. They all extensively use next-generation sequencing. Some are based on PCR- or based on PCR- or capture-based enrichment of retrotransposon junction sequences, followed by targeted followed by targeted resequencing. Others are computational approaches to identify L1 insertion

L1 insertion polymorphisms in whole genome or exome sequencing data, generally based on discordant read pairs.

Table 2 gives an overview of these techniques, and the following paragraphs provide their detailed description.

Method Name	Type	Starting material	Through-put	Approach	Reference
RC-seq	enrichment	genomic DNA	high	capture by hybridization & PE-sequencing	(310)(311)
Fosmid sequencing	library screening	fosmid	medium	southern-blot & Sanger sequencing	(312)
L1-seq	enrichment	genomic DNA	high	ligation-mediated PCR & SE-sequencing	(313)
Ewing PCR	enrichment	genomic DNA	high	hemi-specific PCR & SE-sequencing	(314)
ATLAS-seq	enrichment	genomic DNA	high	anchored PCR & SE-sequencing	(315) and unpublished
TranspoSeq	computational	WGS or WES (PE) data	high	discordant read pair identification	(316)
Ewing pipeline	computational	WGS (PE) data	high	discordant read pair identification	(317)
Tea	computational	WGS (PE) data	high	discordant read pair identification	(26)
TraFic	computational	WGS (PE) data	high	discordant read pair identification	(318)
Mobster	computational	WGS (PE) data	high	discordant read pairs and split read identification	(319)
Tangram	computational	WGS (PE) data	high	discordant read pair and split read identification	(320)
RetroSeq	computational	WGS (PE) data	high	discordant read pair identification	(321)

Table 2: A summary of L1 insertion detection methods. PE, paired-end; SE, single-end; WGS, whole genome sequence; WES, whole exome sequence.

- **RC-seq**

Retrotrotransposon capture sequencing (RC-seq, Figure 10) was first introduced by Baillie *et al*. in 2011 (310) and then further enhanced in 2013 (311). The initial method by Baillie used capture by hybridization followed by paired-end sequencing. Firstly, sheared genomic DNA is hybridized to custom tiling arrays probing full-length retrotransposons. Captured DNA fragments are eluted and analyzed with an Illumina sequencer, producing $\sim 2.5 \times 10^7$ paired-end reads per library that are subsequently aligned to the reference genome. Then the reads mapping as a pair to a single locus are indicative of known retrotransposon insertions present in the reference genome. Next, unpaired reads showing discordant behavior are indicative of novel retrotransposition events. Improvements included a multiplex and liquid-phase sequence capture step using refined probes and reduced insert size. This enabled high confidence assembly of overlapping paired-end reads and recognition

of integration sites at higher resolution. RC-seq was applied to show somatic retrotransposition in human brain and liver cancer (310, 311).

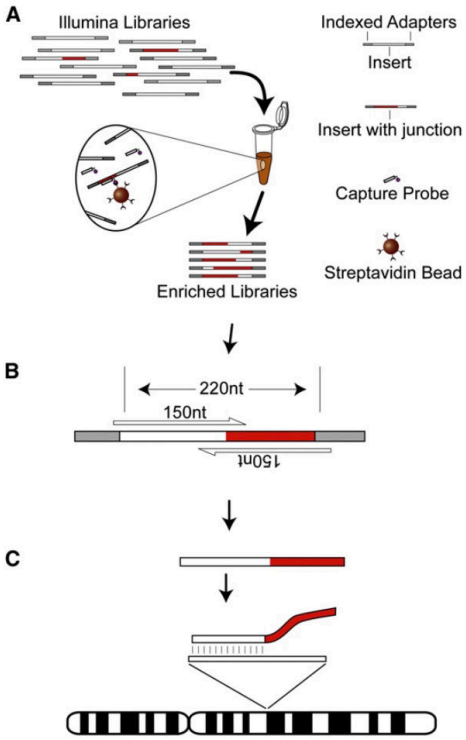


Figure 10: RC-Seq flow diagram (311)**.** (A) 5' and 3' ends of recently active and human-specific retrotransposons present in multiplexed Illumina libraries are hybridized to liquid-phase sequence capture probes. (B) Paired-end 150 bp sequencing of ~ 220 nt inserts enables «contig » assembly of each read pair into a single sequence. (C) Assembled reads with 3' or 5' side of active retrotransposon at one end are retained (shown in red). Opposite end is then aligned to the reference genome, indicating the location of known and novel insertions.

- **L1-seq**

Iskow introduced this technique, based on targeted amplification of retrotransposon junctions, to detect young human retrotransposon insertions, L1HS-Ta and Alu (313). The principle of this technique is depicted in Figure 11. By applying L1-seq, they showed that young and polymorphic insertions are abundant in human populations and that new somatic L1 insertions occur in human lung cancer genomes. Genome-wide analysis suggested that altered DNA methylation might be responsible for the high levels of L1 mobilization observed in these tumors. This data indicated that transposon-mediated mutagenesis is extensive in human genomes and is likely to have a major impact on human biology and diseases.

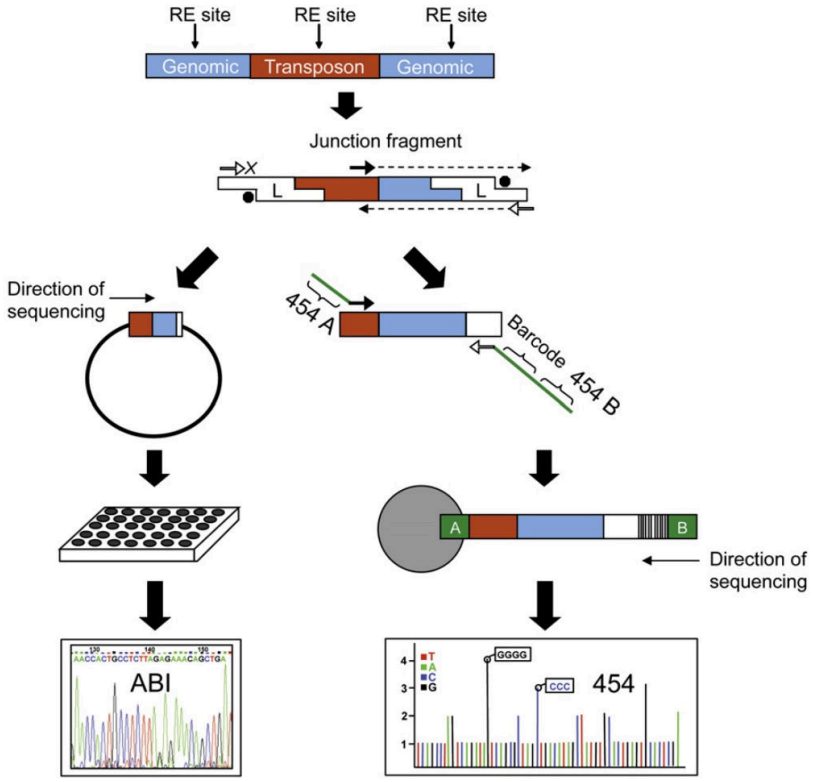


Figure 11: L1-seq flow diagram (313)**.** Human genomic DNA is digested with restriction endonucleases and ligated to a linker. To prevent random amplification of genomic DNA the linker is partially double-stranded with 3' amine group on the short strand. Amplification only occurs if there is an extension from transposon specific primer, thus completing double stranded linker primer to anneal to, and therefore, allowing the PCR reaction to proceed. Amplicons were either cloned and sequenced by Sanger sequencing (left side), or directly sequenced by 454 (right side). This was achieved after reamplification with a second set of nested containing A- and B-adaptor sequences for 454 sequencing and a sample-specific barcode of 8 bp. Samples were pooled in equal molar ratios for emulsion PCR with beads binding only the ''A'' end. Thus, sequencing occurs from the ''B'' end only, avoiding possible problems with sequencing through the poly(A) tail of L1. The same principle was used for Alu except that the 5' junctions were amplified and sequenced.

- **ATLAS and ATLAS-seq**

Richard Badge developed a technique, called ATLAS (amplification typing of L1 active subfamilies, Figure 12), to identify polymorphic L1 insertions (315). In its original form, this low-throughput technique was based on the specific amplification of L1HS 5' and 3' junctions from restriction-digested genomic DNA, followed by comparison of electrophoretic migration profiles. Polymorphic bands, corresponding to potential polymorphic insertions, were excised, PCR-amplified, cloned and sequenced by the Sanger method. The different steps are shown in Figure 12. Our laboratory has adapted this method to render it high-throughput (unpublished). This was achieved by: (i) replacing restriction-enzyme digestion by mechanical fragmentation; (ii) adding sequencing adaptors to the suppression PCR primers; (iii) performing relatively long read (400 bp) Ion Torrent sequencing of the amplified DNA.

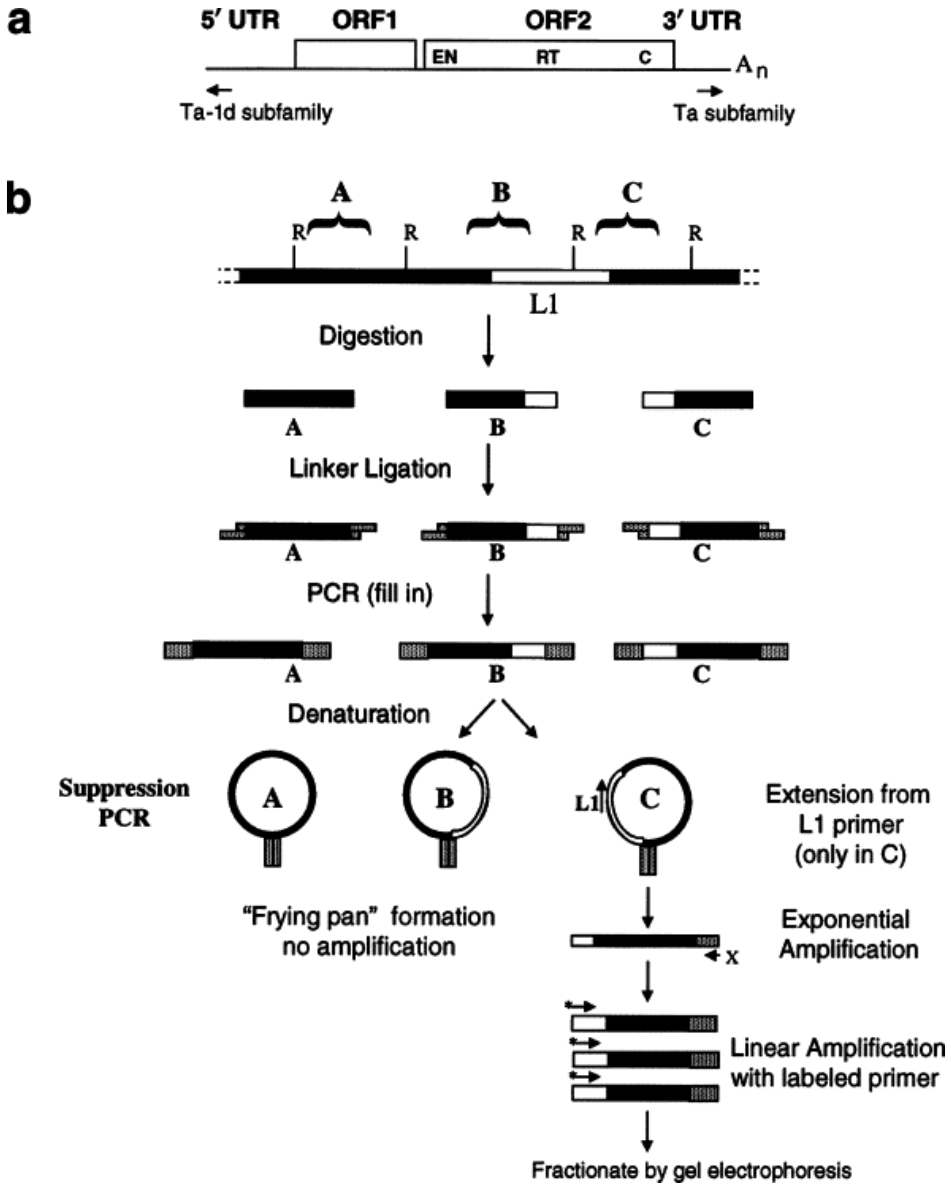


Figure 12: ATLAS method flow diagram (315)**.** Genomic DNA digestion is performed with selected restriction enzymes, which possess restriction sites within L1 sequences and are insensible to CpG methylation. The second step is linker ligation followed by suppression PCR (sPCR) using L1 and linker-specific primers. Then linear amplification of sPCR product with radiolabeled L1-specific primer resolved by polyacrylamide gel electrophoresis is performed.

- **Fosmid sequencing**

To identify full-length L1 elements not present in the reference human genome, a fosmid sequencing strategy (Figure 13) has been developed by the Moran laboratory (312). The extremities of ~40 kb DNA fragments cloned in fosmids were sequenced and their spacing was compared with the human genome reference (HGR). Clones with discordant lengths were further screened by southern-blot for the presence of an L1 element, followed by Sanger sequencing or ATLAS to identify the precise junctions.

This approach identified 68 full-length L1s that are differentially present among individuals but absent from the reference genome sequence (312). The majority

these L1s were highly active in a cultured cell retrotransposition assay. Genotyping 26 of these elements revealed that two of these L1s are only found in Africa and that two others are absent from the H952 subset of the Human Genome Diversity Panel. These results suggest that the so-called 'hot' L1s are more abundant in the human population than previously thought, and ongoing L1 retrotransposition continues to be a major source of inter-individual genetic variation.

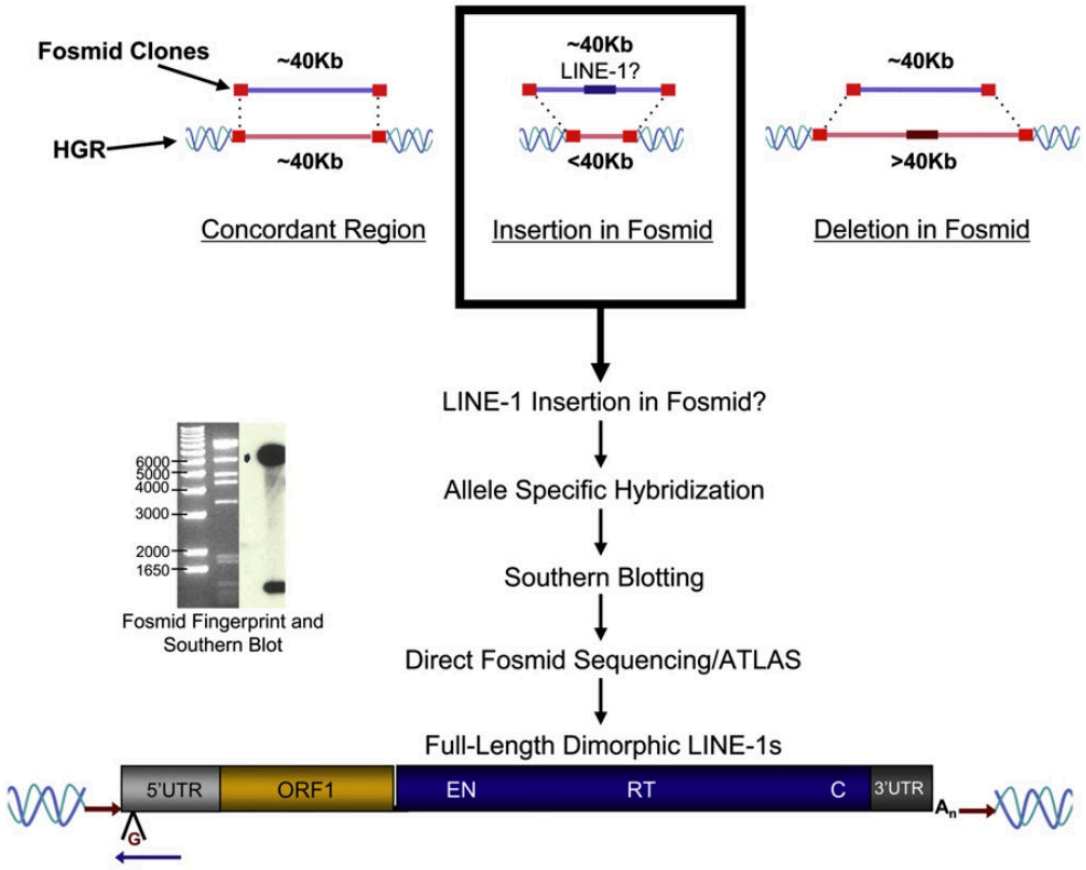


Figure 13: Fosmid sequencing protocol (312)**.** *In silico* comparison is performed for fosmid end sequences (red squares) from individual genomic libraries (blue horizontal line) and the HGR (pink horizontal line), which enables the detection of fosmids that may contain insertions or deletions with respect to the HGR. Insertion fosmids were screened by allele-specific oligonucleotide hybridization to detect Single Nucleotide Polymorphisms (SNPs) that are present in the 5' UTR of the youngest L1 elements (one discriminating character utilized, a deletion of the G residue at position 74 in recent L1s, is indicated in maroon). Putative L1HS-containing fosmids were analyzed by Southern blotting with a 5' UTR probe (blue arrow). A representative digest and Southern blot is shown. The ~6 kb band is diagnostic for the full-length L1. ATLAS and/or DNA sequencing confirmed the presence of a dimorphic, full- length L1Hs insertion.

- **Ewing PCR**

This method, developed by Adam Ewing & Haig Kazazian, aims at finding all human-specific L1 retrotransposon insertions in the genome (314). The technique is summarized in Figure 14. More generally, it allows to interrogate genomic locations of repeated sequences for which a common 3' sequence is known based on the reference genome sequence (314). By applying this method to the genome of several individuals from various regions of the world, they suggested that two individual genomes differ at an average of 285 sites with respect to L1 insertion (314).

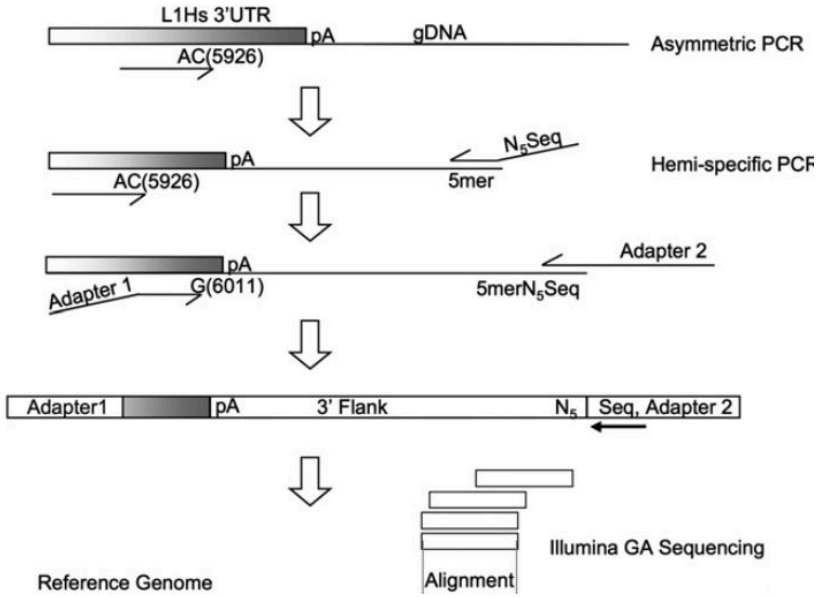


Figure 14: Ewing PCR flow chart (314)**.** This is a PCR-based enrichment method followed by single-end sequencing. Priming is achieved with a specific primer, which anneals to the 3' region of L1HS elements. Extension products are then amplified by a nested hemi-specific PCR using: 1) an L1-specific primer and a degenerate oligonucleotide with a non-matching tail; 2) a nested L1-specific primer and a tail-specific primer. Finally, deep sequencing is performed (single-end reads).

- **Ewing pipeline**

Adam Ewing has developed one of these methods, which is summarized in Figure 15. Based on his pipeline and on wet-lab validation, he discovered hundreds of L1 insertions not represented in the reference human genome assembly, many of which appear to be specific to populations or groups of populations, particularly Africans. Cross-comparison of several studies showed that on an average 27% surveyed non-reference insertions are present in only one study, indicating the low allele frequency of many retrotransoposon polymorphisms (317).

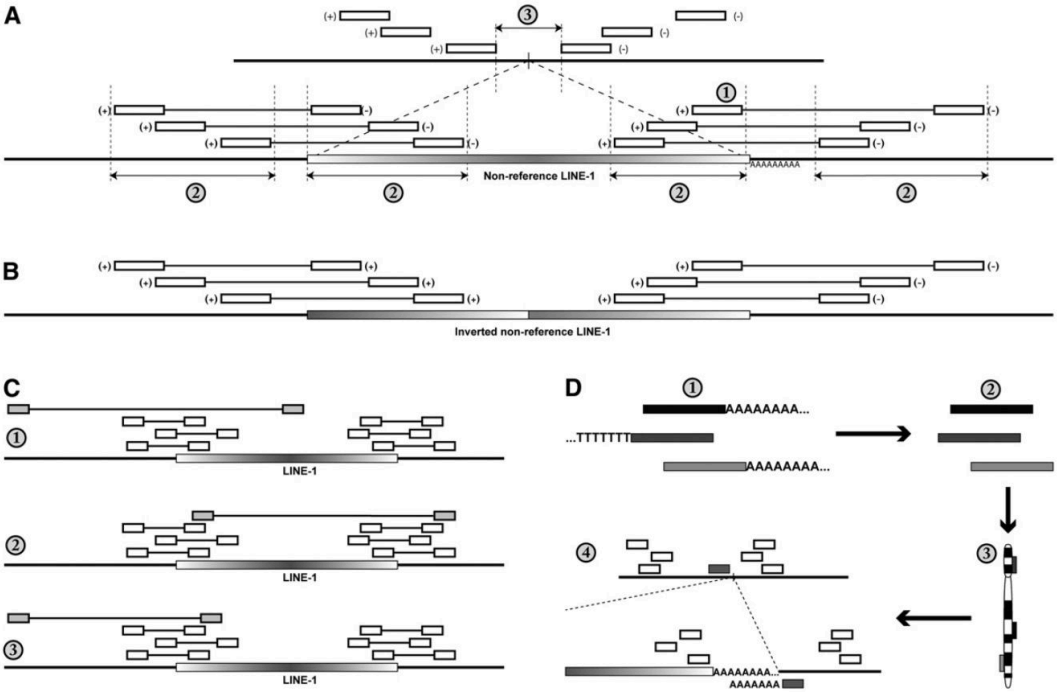


Figure 15: Ewing pipeline flow diagram (317)**.** This computational method employs the discordant read pair information. This pipeline is used to identify non-reference L1 insertions from whole genome resequencing data. First, short reads with one end in L1 and the other in the reference genome are identified and then clustered based on location on the reference genome. The 3' end must be detected for new L1 insertion. Reads are clustered within a minimum distance of <100 bp. At the 5' end, L1 insertions may be inverted which results in the reads aligning to reference L1 in the same strand at the 3' or 5' ends.

- **TranspoSeq**

TranspoSeq also employs discordant read pair information (Figure 16) and was developed as part of the Cancer Genome Atlas project (316). It was first applied in 2014 on whole genomes or exomes from 200 tumor/normal pairs across 11 tumor types. Many novel germline insertions along with 810 somatic insertions in lung squamous, head and neck, colorectal, and endometrial carcinomas were identified. They found that the overall rates of genomic rearrangement and somatic mutation are correlated with high somatic retrotransposition rates in tumors.

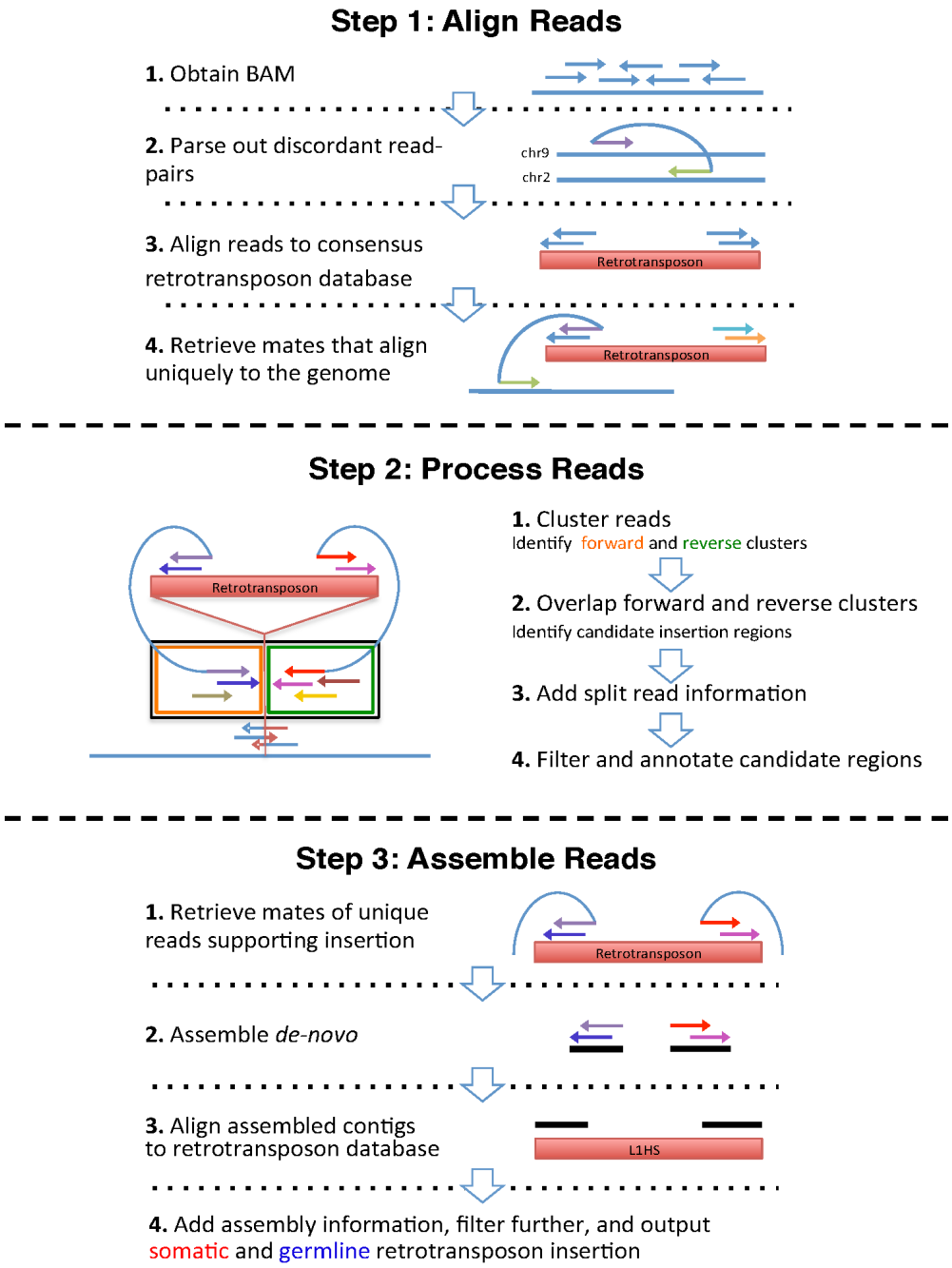


Figure 16: TranspoSeq method chart (316)**.** Discordant read pairs are clustered into reverse or forward strands. Then clusters are checked for overlaps and *de novo* assembly is performed on the loci. The resultant contigs are aligned to both genome and mobilome for annotation purpose. Finally, insertions are classified into somatic or germline based on the filtration criteria.

- **Tea (Transposable element analysis pipeline)**

Tea (transposable element analysis, Figure 17), a software developed by Eunjung Lee (26), takes advantage of discordant reads, but also of clipped reads, to precisely infer the position of TE as well as insights in the mobilization mechanism through target-site duplication or deletion annotations. Tea was applied to whole-genome sequencing data from tumor and matched normal blood samples from 43 colorectal, prostate, ovarian, multiple myeloma and glioblastoma cancer patients, revealing 194

high confidence somatic TE insertions. They found that somatic L1 insertions were enriched in genes often mutated in cancer, suggesting a functional impact on tumorigenesis. These insertions disrupted their target genes and were showing bias for DNA methylated cancer regions.

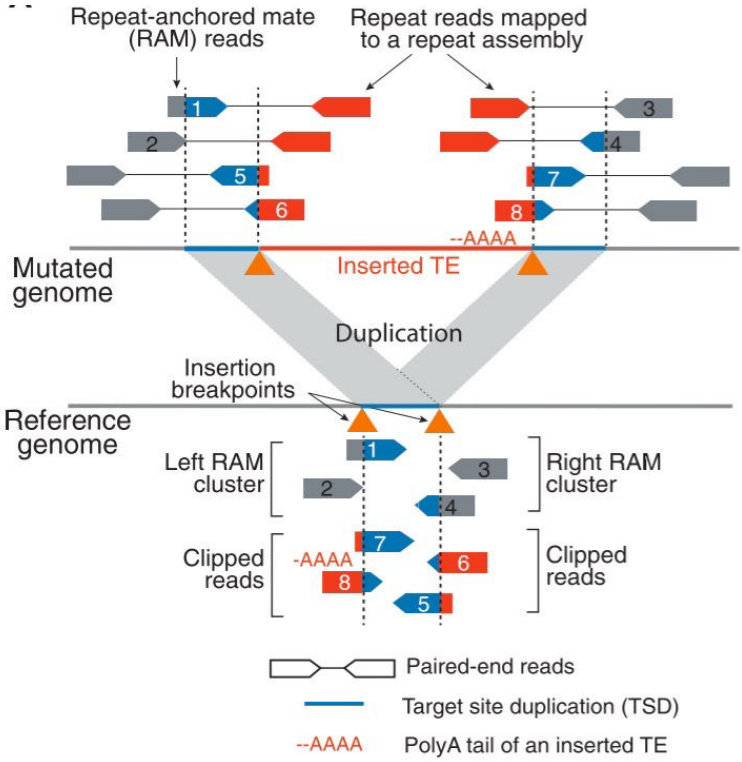


Figure 17: TEA flow chart (26)**.** Two types of supporting reads are identified in order to detect somatic insertions of TEs from paired-end read data in tumor and matched normal genomes: (i) Repeat-anchored mate (RAM) reads, in which one of read of a pair is mapped to a unique location in the genome, whereas the other is associated with a TE (reads 1 to 4), and (ii) clipped reads, which span the TE insertion breakpoints and show partial alignment to the reference or the repeat assembly (reads 5 to 8). The distances between the clipping positions and the clipped sequences are used to infer the insertion mechanism.Then, duplicated sequences at the insertion site (TSD) and the poly-A tail of the inserted TE are the other characteristics looked at.

- **TraFic (Transposome Finder in Cancer)**

TraFic (Transposome Finder in Cancer, Figure 18) is capable of finding: (i) solo L1 which are somatically retrotransposed; (ii) partnered transductions in which a unique downstream sequence has been mobilized with an L1 element (3' transductions); and (iii) orphan transductions, when only the unique sequence downstream of an active L1 is retrotransposed without cognate LINE (318). The 3' transductions (partnered or orphan) are used to identify the source element (progenitor) giving rise to somatic retrotransposition events. Hallmarks of retrotranspoition are both the integration point and the L1 source element locus. For the identification of putative solo-L1 and L1-transduction integration sites (and more generally of TE insertions), TraFic uses paired-end sequencing data. The identification of somatic TEs (solo-L1, Alu, SINE, and ERV) is performed in three steps: (i) selection of candidate reads. This module categorizes reads into 3 different types (single-end, inter-chromosome and aberrant); (ii) transposable element masking using repeatmasker database; (iii) clustering and prediction of TE integration sites; and (iv) filtering of germline events.

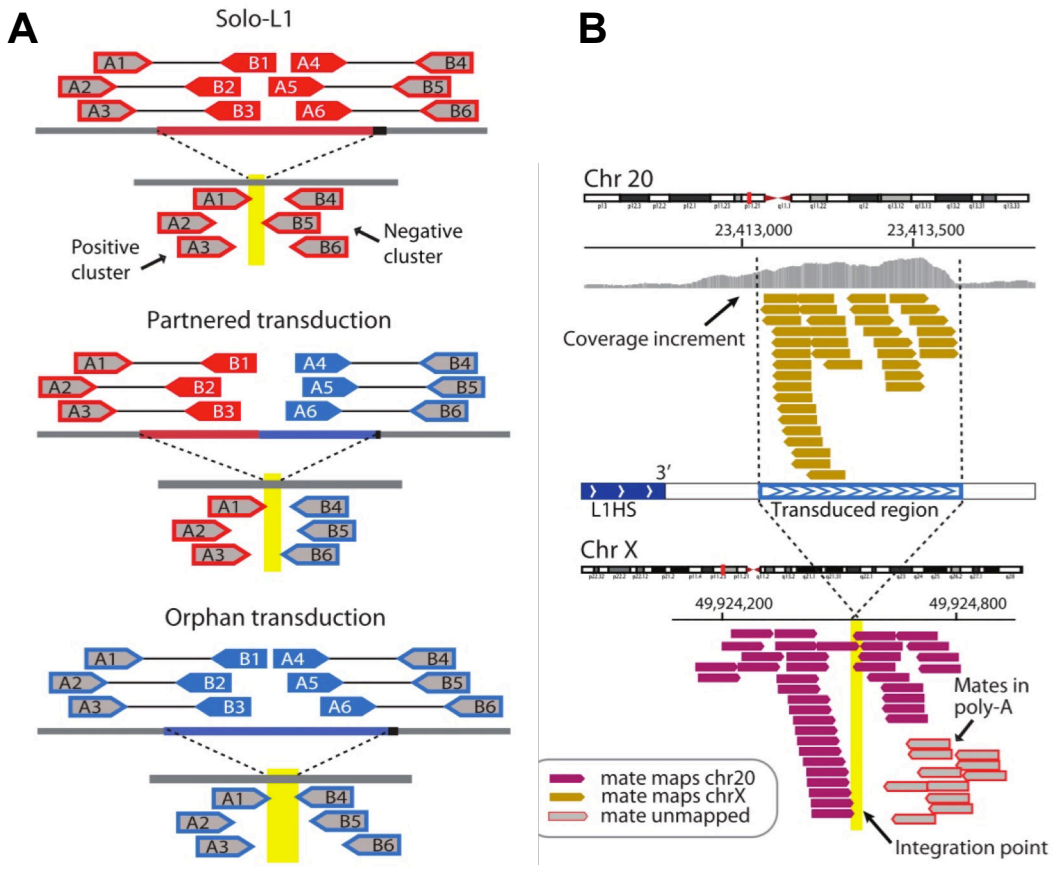


Figure 18: TraFic flow chart (318)**.** (A) Strategy followed to identify somatic solo-L1 and L1 transductions (partnered or orphan). The pipeline relies on two read clusters (positive and negative clusters) pointing to the same region of the genome where somatic element is inserted. (B) Example of a partnered 3' transduction on chromosome 20 showing coverage increment downstream of the element resulting from genome-wide amplification of the transduced material (top). Reads responsible for the coverage increment pair with different chromosomes (chromosome X, bottom). A cluster of reads around the breakpoint indicates the presence of a poly(A) tail. Other reads reveal the presence of target site duplication.

The identification of L1-mediated transductions is performed at a second stage, according to the following steps:

(i) Candidate read selection (inter-chromosome or aberrant).
(ii) Clustering and prediction of transductions if a) they share the same orientation, b) the distance relative to the nearest mapped read of the clusters is equal or less than the average read size, c) their mates are also clustered together.
(iii) Filtering of germline tranductions

Finally, TraFic estimates insertion size, reconstructs TE boundaries and then detects target site duplications.

Analysis of 290 tumors and matched normal controls across 12 cancer types by TraFic identified 2756 L1 retrotransposition events including solo L1s and 3’ transductions (318). In this study, somatic retrotransposition was detected in 53% of the patients. 24% of these events were 3′ transductions.

- **RetroSeq**

This pipeline (321) uses discordant read pairs from whole-genome paired-end sequence data to identify non-reference L1 insertions. From the input BAM format file, the software finds:

i. The discordant read pairs, which map to both reference genome and mobilome (Alu, SINE and LINE etc.). It uses either user supplied annotated TE file or aligns the reads with exonerate to the Mobilome index.
ii. Then, it clusters the discordant reads identified in the previous step at genomic locations while keeping track of the strands.
iii. Forward and reverse clusters are then merged around the potential putative break points.

This pipeline also uses the information from soft-clipped reads. Benchmarking their software using data from the 1000 Genomes Project for a CEU trio (father NA12891, mother NA12892 and the female offspring NA12878) has shown that this software was able to predict most of the Trio insertions correctly, and that its specificity and sensitivity are improved as compared to other methods such as Tangram or Tea.

- **Tangram**

Tangram (Figure 19) also uses discordant read pairs and soft-clipping information from whole-genome paired-end sequence data to identify structural variations (320). To this goal, it scans the reads against reference genome and mobilome. The read pair method collects the reads with one mate mapping on the reference genome and the other on mobilome. Then, genomic locations of these reads are clustered to locate insertion position. They use MOSAIK aligner tags to identify the type of insertion. Distance between closest mates to the real breakpoint defines the breakpoint confidence interval. In case of split read method, one of the mates is either unaligned or soft clipped with the reference genome or the mobilome. Breakpoint is determined by the alignment location of the first segment.

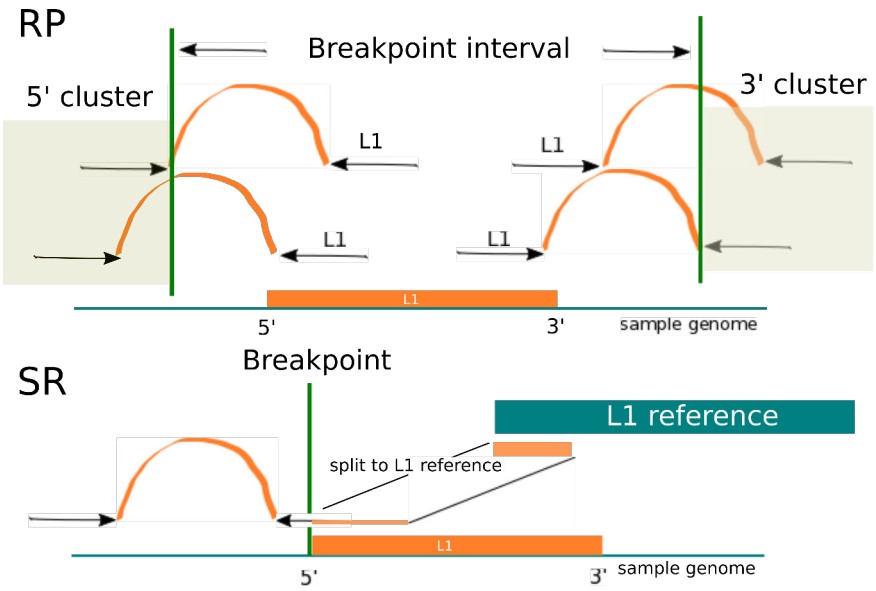


Figure 19: Tangram method (320)**.** (Top) Read pair (RP) method. Blue line with orange represents genome with mobile element Insertion. Each pair of black arrows represent a read aligned to the genome. For RP method, mates (opaque box) are collected to estimate insertion location. MOSAIK aligner provides the type of insertion (ZA Tags). The distance between two uniquely aligned mates that are closest to the real breakpoint gives the breakpoint confidence interval. (Bottom) For split read (SR) method, those read pairs are collected with one uniquely aligned to the genome and other mate is either unaligned or soft-clipped. Unaligned or soft-clipped reads are split into two segments; one of them is aligned to the normal human genome and the other to the Mobile element reference (blue). The breakpoint can be determined by the alignment location of the first segment.

- **Mobster**

Mobster (Figure 20) detects non-reference TE insertions from both whole genome and whole exome data (319) and also uses discordant read pairs and clipped reads along with mobilome data reference sequences.

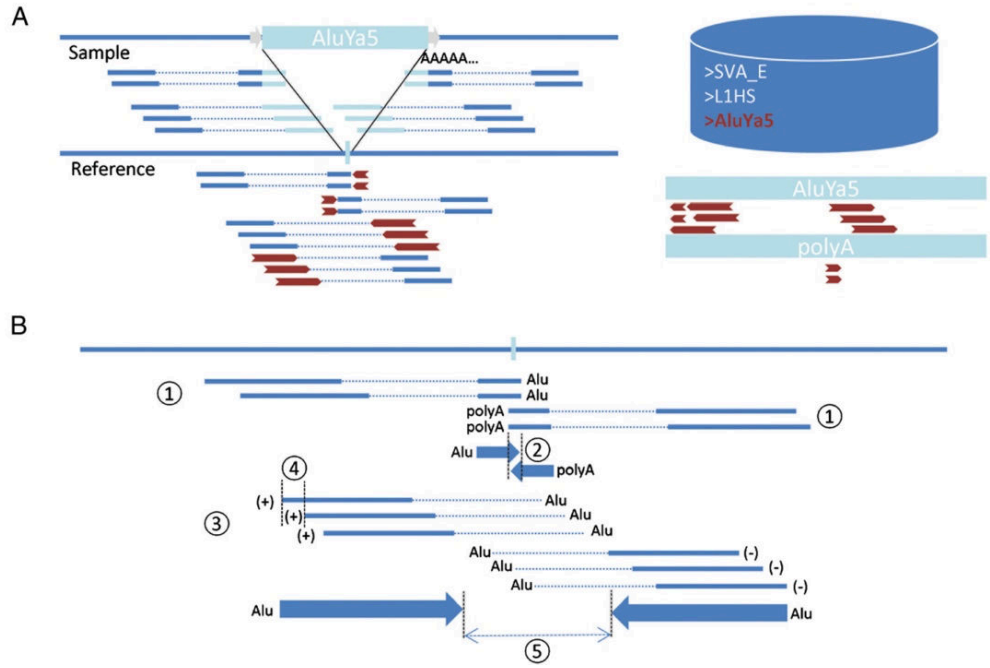


Figure 20: Mobster method (319). (A) Discordant ends and soft clipped reads are kept if one of the mates or the unclipped end is mapped uniquely to the reference genome. These reads are mapped to the mobilome and checked for having a polyA/T-tail. After mapping, reads belonging to unambiguously mapped *Alu*, L1, SVA, or HERV-K are identified. Clipped and discordant read anchors are clustered separately. (B) Criteria for selection of split anchors (1) are: (i) supported by the same TE family or same polyA/T (ii) clipped on the same side (iii) clipped within a few bp to each other. 5'- and 3'-clipped clusters (2) are indicative of the same TE insertion event if: (i) both clusters support the same TE family or one of the clusters supports the polyA/T-tail and the other cluster supports TE family (ii) overlap of max of 50 bp (for TSD) between clusters or are separated by a max of 20 bp (for deletions). Similarly, discordant pair anchors (3) are clustered if: (i) they have same strand; (ii) support the same TE family (iii) Start positions within a specified neighborhood distance (4). Discordant clusters from the forward strand 5' and reverse 3' clusters are indicative of the same TE insertion event when there is an overlap of max 50 bp or user-specified window size (5). Clipped and discordant clusters passing these criteria are merged if they overlap.

2.3.2. L1 retrotransposition occurs in germline and somatic tissues

- **L1 activity in germ cells and during embryogenesis**

Mouse models have shown that insertions could occur in the early stages of development (322–325). In 1993, Packer reported an abundant 8kb-long L1 transcript in mouse blastocysts (326). This study provided a lot of information on tissue and cells where murine L1 is expressed and potentially mobilized. L1 transcripts were detected in mouse blastocysts, indicating that L1 is expressed during early development in mice (326). L1 expression was also detected in testis

and ovary of mice, in germ cells and in some somatic cells (327, 328). In particular, full-length, sense-strand L1 RNA and L1-encoded protein were detected in the early meiotic cell types, namely leptotene and zygotene spermatocytes at postnatal day 14 of development (327).

Human L1, which is more difficult to study *in vivo*, seems to have similar patterns of activity. Several hESCs, as well as embryonic carcinoma (hEC) cell lines, accumulate L1 RNPs, the functional form of the retrotransposition machinery, and diverse L1 mRNAs, representing both young and old L1 subfamilies (287, 329, 330). L1 insertion causing chronic granulomatous disease carried by the X chromosome in a male patient demonstrated that retrotransposition could occur during maternal meiosis and confirmed mobility in germ cells (331). L1 expression has been also reported in cancer germ cells (332). Finally, in transgenic rodent models containing either human or mouse L1 elements controlled by their original promoters, L1 RNA is abundant in both germ cells (333)(324), but also in early embryos and thus nonheritable L1 retrotransposition events during embryogenesis might create genomic diversity within one individual (324). Another recent mouse model using an L1 element under the control of an inducible promoter and containing a "gene-trap" cassette, confirmed that early embryogenesis is a major window of permissiveness allowing L1 retrotransposition and results in somatic mosaicism, even when L1 expression is controlled by a heterologous promoter (334).

In conclusion, L1 retrotransposition in the germline or during early embryogenesis (before the differentiation of germ cells) acts as a source of genetic diversity within the human population. L1 insertion appears every 200 births (20, 335). Occasionally, they lead to the emergence of new genetic diseases (see § 2.3.4).

- **L1 activity in post-embryonic stages**

Morse reported the first case of L1 retrotransposition event that occurred in somatic cancer cells (336). Ever since, there is a growing body of evidence that L1 mRNA and L1 proteins can be expressed in some normal or tumor somatic cells and that L1 retrotransposition within these cells may produce somatic mosaicism. L1 mobilization in tumors will be developed later (see § 2.3.5).

Belancio *et al.* examined the presence of L1 transcripts in various human tissues (228). They were able to detect endogenous full-length L1 (FL L1) transcripts in human esophagus, prostate, stomach and placenta tissues. No full-length L1 transcripts were detected in colon, skeletal muscle, heart muscle, and brain, testis, ovaries, lung and thymus tissues. Surprisingly, transcripts corresponding to truncated L1 mRNAs were detected in all tissues examined in this study. The level of truncated transcripts, corresponding either to prematurely polyadenylated L1 mRNA or to differently spliced and polyadenylated L1 mRNAs, was especially high in testis tissue (228). Detection of high levels of L1 transcription in testis is consistent with the L1 promoter being regulated by the testis determining factor gene *SRY* (223). However, enhanced L1 RNA production, in this case, seems to entail severely restricted RNA processing. These results show that often L1 expression in somatic cells occurs in the absence of effective transposition.

In contrast to these expression studies, experiments in rats have shown that L1s may be mobilized in neural precursors. In the adult transgenic mouse brain, retrotransposition events were found in both neurogenic and non-neurogenic areas, indicating that retrotransposition may happen during both embryonic and adult neurogenesis (337). These results were later confirmed with human neural progenitor (338). The later study also observed a possible increase in L1 copy number in the genomic DNA of different area of the brain (including the hippocampus) in comparison to other somatic organs such as the heart or the liver, suggesting L1 mobility during human neurogenesis. This was confirmed a couple of years later by direct sequencing of somatic L1, Alu and SVA insertions in the brain by RC-seq (310) (see p. 43 for a description of this technique). Nevertheless, the extent of L1 mobilization is still highly debated with frequencies ranging from less than 0.6 (339) to 13.7 or even more per neuron (338, 340). Interestingly, somatic L1 retrotransposition in neural cells preferentially occurs into euchromatic regions of the genome (340). Neuronal specificity of somatic L1 retrotransposition is at least partially regulated by the *Wnt* pathway and is due to the replacement of a *Sox2/HDAC1* repressor complex by a *beta-catenin/TCF/LEF* activation complex, which leads to chromatin remodeling and transcriptional activation of L1 (230).

Overall, the activity of L1 retrotransposons during neurogenesis can create specific genetic mosaicism, which in turn may affect gene expression, neuronal function, and plasticity.

2.3.3. L1 is a source of natural variation among humans

▪ L1 as a source of insertional polymorphisms and deletions

Detailed analysis of mutational mechanisms indicates that approximately 20–30% of structural variations are caused by non-LTR-retrotransposons (20–23). *Alu*, L1, and SVA retrotransposition rates are estimated to be one in 21 births, 212 births, and 916 births, respectively. In addition, by comparing non-reference L1 elements among different individuals in 1000 Genome Project data, it was estimated that two individual genomes differ at an average of 285 sites with respect to L1 insertion (314). Each *de novo* L1 insertion represents a unique historic event. Retrotransposition is an ongoing process. Some of the polymorphic insertions are shared among different people or whole populations, whereas others might be found in only a single individual (private insertions).

Like other active retrotransposons, SVA elements, which are mobilized by L1, show inter-individual variation in humans and can be polymorphic for their absence or presence in the genome. As per the estimates 37.5% of SVA E elements and 27.6% of SVA F elements are polymorphic for their presence in the genome (341) and the average human is estimated to have 56 SVA absence/presence polymorphisms (25).
L1 can also mobilize other cellular RNAs and allows the creation of new pseudogenes (52, 53) (see § 2.2.4). Genomic deletions have also been associated with insertion events (13, 14, 262, 291, 342). Since the divergence of human and chimpanzee, more than 7000 retrotransposons have been inserted into the human

genome (343). It has been found that mobile elements are associated with approximately 0.14% of disease-causing mutations (see also §2.3.4 and §2.3.5).

- **L1 in ectopic recombination**

Many studies suggest that there is a correlation between transposable element insertions and the breakpoints of segmental duplications and SVs in the human genome (344–347). Indeed, in addition to canonical insertion events, L1 retrotransposons can also create genomic instability by several additional mechanisms, such as nonallelic homologous recombination (NAHR) (348), also called ectopic recombination. This might lead to deletion between two retrotransposons from the same family. DNA breaks in L1 sequences can also be repaired by nonhomologous end-joining (NHEJ) also leading to deletions (4, 13, 14, 262, 289, 290, 292, 349–353).

For example, 140 mobile element-mediated deletions have been identified in the human genome reference (354). Among them, 98 are Alu recombination-mediated deletions (ARMD), 9 are L1 recombination-mediated deletions (L1RMD) (354). They also identified 33 NHEJ-mediated deletions. 22 out of the 26 L1-associated NHEJ events occurred within the L1 elements. Which suggests that L1 elements could be subjected to a high frequency of DNA-double strand breaks (DSBs).

- **L1 as a source of satellites**

Non-LTR-retrotransposons can possibly give birth to the microsatellites concurrently to their integration into the genome. An analysis of microsatellites at orthologous loci in three primate genomes indicates that 26% of microsatellite births and 24% of microsatellite deaths occur within Alu and L1 sequences subsequent to retrotransposition (355).

Several studies have reported that polyA tails of retrotransposons may give rise to new microsatellites, also known as Simple Short Repeats (SSRs). Retrotransposon-derived microsatellites are created either through errors introduction during reverse transcription of the primary retrotransposon transcript or through accumulation of random mutations in the middle A-rich regions and oligo(dA)-rich tails of Alu and L1 elements after insertion (356, 357). It should be noted that two examples have been reported where the expansion of Alu-derived SSRs led to genetic diseases in humans (358, 359).

Microsatellite instability has been found to cause a variety of human diseases with over 40 neurological, neurodegenerative and neuromuscular disorders associated with trinucleotide repeat instability (360, 361). Some significant examples include Huntington's disease, Fragile X syndrome, and Friedreich's Ataxia. Due to their unique sequence composition, microsatellites can change the physical forms of DNA where they occur (362, 363), which can have implications for gene expression and genome stability. This relationship between the microsatellites and non-LTR-retrotransposons is not unidirectional. While both L1s and Alus give birth to microsatellites, especially poly(A) mononucleotide microsatellites, these

microsatellite sequences can also affect the fitness of their “parent” due to their unusually high mutation rates.

▪ L1 and DNA double-strand breaks (DSBs)

Not only L1 insertions can greatly alter the structure of our genome, but their proteins can also contribute to its dynamics. Indeed, regardless of retrotransposition, the endonuclease activity of *ORF2* might also cause DNA double strand breaks, genetic instability or chromosomal translocations (364, 365). This phenomenon does not require a retrotransposition-competent L1 element: even a defective element can express ORF2p or a fragment of this protein, which retains its endonuclease activity (366). L1-expressing cells accumulate DNA damage, which can be detected through the formation of nuclear **ϒ**-H2AX foci (365). Mutations in ORF2p endonuclease domain resulted in essentially complete loss of the **ϒ**-H2AX foci in HeLa cells. This result points out that the endonuclease of L1 is required for DNA-double strand break formation, although it may not make breaks on both strands.

The role of genetic instability in diverse phenotypes like aging, fertility, and cancer has highlighted the importance of understanding how cells can respond to endogenous sources of DNA damage. This work demonstrates that the L1 integration process produces DNA double-strand breaks.

▪ L1 shuffles our genome by transduction

L1 can co-mobilize 3’ (cis-mobilization) downstream segments near L1 insertion to new locations in tissue culture cells (44, 176). One of the main reasons for this mechanism could be the weakness of the polyadenylation signal, causing L1 transcription to use alternative polyadenylation site downstream. A hybrid transcript carrying this unique segment is reverse transcribed and re-inserted into the genome. This highlights role of L1 as a player in exon or regulatory region reshuffling (367, 368). It has been estimated that ~1% of human genome DNA has been transduced by L1 that is interestingly comparable to the exonic percentage in genome. This highlights the role of L1 in genomic plasticity by shuffling genomic DNA (27). Figure 21 shows the two possible types of transduction taking an upstream or downstream region along with them and fitting it somewhere else in the genome at the insertion point.

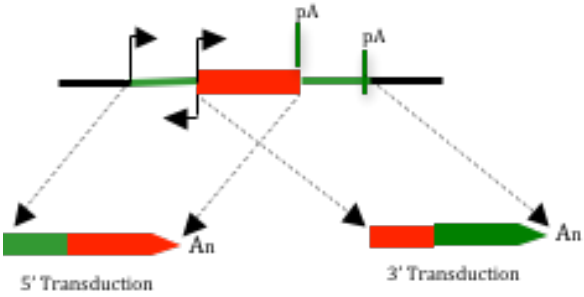


Figure 21: L1-mediated transduction. (Top) A progenitor L1 copy with an upstream promoter and a downstream polyadenylation signal. (Bottom, left) A new L1 copy with a 5' transduction. (Bottom, right) A new L1 copy with a 3' transduction. L1 sequence is shown in red color and upstream and downstream genomic sequence are indicated in green.

Three-prime transduction mediated by L1 of a novel noncoding gene into exon 67 of the dystrophin gene was observed in two studies (369, 370). L1-induced disease caused by an orphan 3' transduction was also reported recently in the dystrophin gene (371). Earlier, four cases of L1-driven insertional mutagenesis in the dystrophin gene have also been reported (33, 369, 372, 373).

Recently, Tubio using TraFiC bioinformatics pipeline (see p. 51) studied 290 cancer types and observed that L1-mediated 3' transductions occured in ~25% of analyzed cancer genomes. Orphan transductions formed almost half of all the transductions. Transduced region size was typically 1 kb, but could attain up to 12 kb (318). Only 72 germline L1 were responsible for ~95% of transductions. They identified 2 « Hot L1s » which were located at chromosome 22q12 and 6p24.1 regions. These 2 hot L1 accounted for almost a third of all somatic transductions. L1-mediated somatic transductions can shuffle coding and regulatory regions on a large scale. Around 2.3% of events of somatic transductions distributed neighboring exons or even complete genes elsewhere in the genome. For example, the whole of exon 18 of the *STK31* gene was picked up and reinserted into the *NRXN3* gene. Also, altogether, 86 somatic transductions have transported 251 transcription factor-binding sites somewhere else in the genome.

Recently, Badge modified its ATLAS method (see p. 45) to specifically use transduction specific primers. Their protocol only amplifies the loci containing transduced sequences (374). With this approach, they identified 25 L1s from three active L1 transduction lineages (L1RP, AC002980, LRE3) and showed the plasticity of the polyadenylation location within transduced family (374).

2.3.4. L1 mobilization can lead to genetic diseases

L1 continues to evolve and affect our genome by playing an instrumental role in sculpting the structure and function of our genomes. There movement can lead to sporadic cases of diseases. L1 are responsible for many insertional mutations. Nearly 65 cases of L1 induced mutations leading to genetic diseases have been described (229, 353, 375) previously. After the report of a Hemophilia-A, which was caused by a *de novo* L1 insertion (376), more than 100 cases of mobile element-

associated structural variants (MASVs) have been documented to lead to human diseases, either directly by L1 insertion, or indirectly by L1-mediated mobilization of other sequences (Alu, SVA, or pseudogenes), such as cases of Pelizaeus-Merzbacher disease, Lesch-Nyhan syndrome, Tay-Sachs disease, familial hypercholesterolemia, and Hunter syndrome (349, 377, 378). L1 insertion is probably not as random as what is generally considered. Indeed, certain genes, such as *NF1*, are hotspots of insertion of L1 because many independent insertions in this gene have been identified (379). Similarly, if two independent insertions of retrotransposons exactly the same chromosomal position have been described in the *BTK* gene, resulting in a sex-linked agammaglobulinemia (*XLA*) (380).

2.3.5. L1 activities remodel the genome of many epithelial cancers

L1 can play several roles in cancer. First, they induce germline mutations in genes favoring the appearance of tumors. They are then mobilized somatically and get involved in tumorigenesis or genetic heterogeneity of tumor cells.

L1 transcripts have been detected in different types of human cancer (e.g. testis, bladder and liver cancers) as well as in many cancer cell lines (65). Tumor-specific ORF1p protein expression was observed in many cases of breast cancer (381), or germ cell tumors (382). The tumor cells express ORF1p while the adjacent healthy tissue does not. More recently, a study showed that not only the presence of ORF1p could be a marker of tumorigenesis, but also its subcellular localization could be indicative of the prognosis of patients with breast cancer. Indeed, nuclear staining of tumor cells is correlated with decreased patient survival after the diagnosis of cancer, but also a higher probability of relapse after treatment or to a more rapid onset of metastasis (383). Finally, a study of a large number of different tumor types revealed that nearly half of all cancers specifically express ORF1p (384).

The first reported case of L1 mobilization in somatic cells was an L1 insertion into the second intron of the c-*myc* gene in breast cancer (336). It should be noted that c-*myc* is a proto-oncogene that is strongly implicated in the control of cellular proliferation, programmed cell death, and differentiation (385, 386). Miki reported the second case of cancer-related L1 insertion (387). They found an L1 element inserted into the last exon of *APC* gene. This insertion exon in the *APC* tumor suppressor gene is directly involved in tumorigenesis. More recently next-generation sequencing efforts have confirmed the presence of numerous somatic insertions in tumor cells (26, 311, 313, 318, 371). Somatic insertions present in tumors but absent from healthy tissues were also found in lung cancer (313) and colorectal cancers (371). Interestingly, some cancers appear to cause more inserts than others. For example, somatic L1 mobilization is common in colorectal cancer, but no event has been detected so far in myeloma and glioblastoma despite of several studies (26, 313). A pan-cancer study has indeed shown that L1 retrotransposition mostly occurs in cancers of epithelial origin (26). In this study, 194 somatic insertions were discovered and one-third of them are located in genes, including tumor suppressor genes. All of these data highlights some somatic mobility of retrotransposons in human tumors, but the determinant of this reactivation remains a mystery.

Alu sequence insertions, associated with cancers have also been described (388–390). Some inserts are somatic (388, 390) while others are germline events (389). Germline insertions, which are heritable, may be associated with a predisposition to cancer and cooperate with somatic insertions toward tumorigenesis. For example, RC-seq experiments performed on patients with hepatocellular carcinoma, showed that these two causes were present (311). Some patients have germline mutations in *MCC* (Mutated in colorectal cancer) while others showed tumor-specific insertions in the same gene, reinforcing the idea that L1 insertions might be driving mutations.

Alterations of DNA methylation is a common feature of tumors and comprise paradoxically two contradictory phenomenons: (i) global (genome-wide) hypomethylation, and (ii) local hypermethylation which occurs typically at CpG islands surrounding the transcriptional start regions of individual genes (391). Usually, the overall decrease in methylation found in cancer cells involves the parallel decrease in the methylation of L1 and other retrotransposons (392). Although, the molecular mechanisms underlying cancer-related loss of methylation remain largely unknown. There is a strong evidence indicating that demethylation plays an active role in cancer progression (393). Hypomethylation of L1 can be correlated with genomic instability in certain cancers, such as lung cancer (66) or to changes in the transcriptome, especially due to the expression of its bidirectional promoters (67, 68). These alterations can occur at different stages of tumorigenesis. For example, in colon and bladder cancers L1 hypomethylation appears at the early stages (394, 395) , whereas in prostate cancer only at the late stages (396). In fact, prostate cancer seems to deviate from the prevailing model of epigenetic dysregulation, in which DNA hypomethylation is involved in cancer initiation. More likely, in prostate cancer hypomethylation is involved in the formation and propagation of metastases. Noteworthy, in some of cancer cells (e.g. renal carcinomas) the decrease in L1 methylation is very slim and is unlikely to be of any significance for cancer progression (394). The cumulative effect of the expression and mobility of L1 in cancers leads us to consider using L1 as biomarkers for certain cancers (65, 383, 397).

Therefore, L1 can play several roles in cancer. First, they can induce germline mutations in genes favoring the appearance of tumors. They can also be mobilized somatically. Due to alterations of their methylation profile, they are not only involved in the genetic heterogeneity of tumor cells, but also in their epigenetic heterogeneity.

2.4. L1 insertions can reshape the human transcriptome in multiple ways

2.4.1. L1 can cause exonization or alternative splicing

L1 insertions within the host genes introduce splicing sites, which can promote exonization (creation of new exon) or alternative splicing of mRNA of particular genes. Figure 22 and Figure 23 summarize the different possible consequences of L1 insertions on gene structure, and on the formation of alternative transcripts.

Exonization is a process in which an L1 sequence inserts into an intron and part of it is retained in the mature mRNA. Cases have been found in both mouse and humans (398–400). Alu, which use the L1 machinery are also actively involved in exonization across the genome. Recently, using Individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation (iCLIP) against hnRNP C, Zarnack *et al*. documented a large number of cryptic Alu-derived exons. They found 1,318 cryptic exons that had originated from Alu in addition to 585 Ensembl annotated Alu exons. Thus, a total of 1,903 Alu exons were characterized. They concluded that hnRNP C is playing a role in protecting transcriptome from the harmful effects of aberrant Alu exonization (401).

It has been estimated that 92-94 % of human genes exhibit alternate splicing, ~86 % with a minor isoform frequency of around 15% (29). There are approximately 95% of human multi-exonic genes that are alternatively spliced (402). Studying the consensus sequence of the L1 element has revealed many acceptor or donor sites for alternative splicing. L1 can generate numerous transcripts of variable size that could possibly be due to alternative splicing of the L1 sequence, which contains cryptic acceptor and splice donor sites and some of them proven to be functional (30). Splicing can, therefore, change the L1 RNA after transcription and thus limit its impact by creating non-active RNA. On the other hand, the study of Expressed Sequence Tags (ESTs) showed that L1 splicing sites inserted into genes may be used during the maturation of gene transcripts, which is a mechanism by which L1s may contribute to the plasticity of our genome (30). A donor site appears to be mainly used in alternative splicing and is located at the position 97 in the 5' UTR. This was demonstrated by Belancio *et al.*, who screened a human database of EST for evidence of L1-mediated splicing variants. In total, they found 39 evident splicing events between an L1 SD site (at position +97 at 5' UTR) and SA sites of 21 different genes (30).

Introduction of new splicing sites by retrotransposons can result in a severe gene disruption as well as in new coding and non-coding gene creation (31–35). This is a perfect example how the same retrotransposon-induced mechanism can appear destructive or beneficial for the host.

Many diseases have been associated to this phenomenon like the case of Wilson's disease, which is caused by alternative splicing and *Alu* exonization. *ATP7B* gene gets homozygous 3039-bp deletion spanning from intron 1 to exon 2 (403). Deletion and exonization of an adjacent *AluY* in *COL4A5* gene cause Alport syndrome (404).

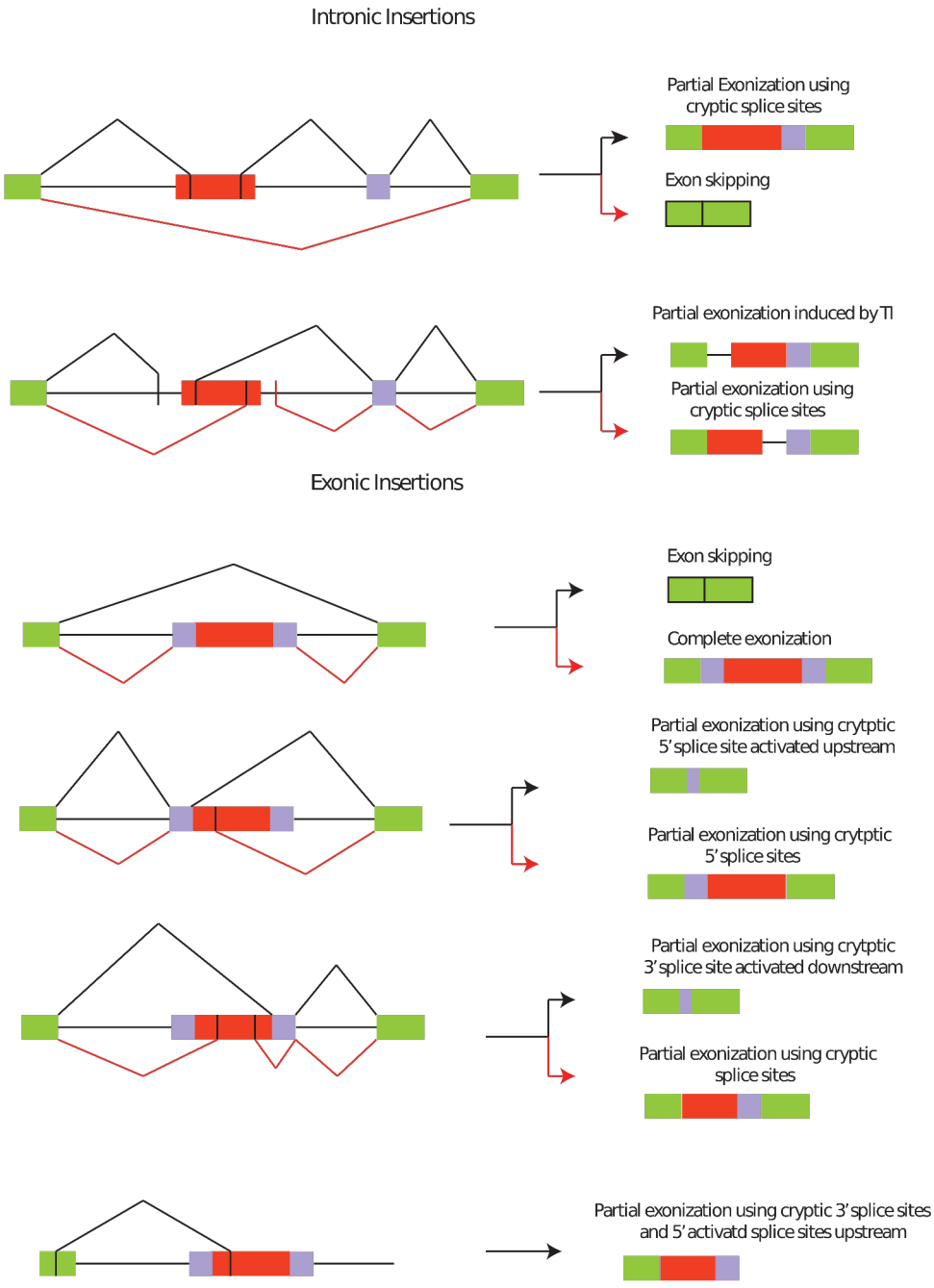


Figure 22: Splicing mechanisms due to L1 integration. Adapted from (408). Different possible outcomes of L1 insertions into genes have been depicted. (Top) Intronic L1 integrations. (Bottom) Exonic L1 integrations. L1 has been shown in red color while the diagonal lines show splicing schemes. Resulting transcript variants are shown on the right side.

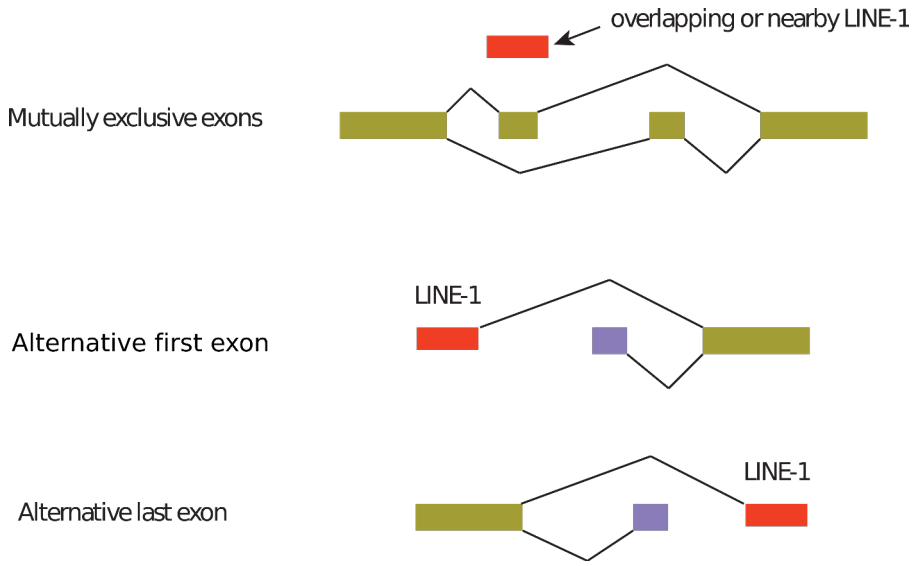


Figure 23: Additional possible consequences of L1 integration on alternative transcript formation. L1 has been shown in red color. Diagonal lines indicate the splicing scheme.

2.4.2. L1 contains cryptic polyA signals causing premature polyadenylation

RNA Pol-II transcription of L1 is negatively affected by numerous termination and polyadenylation signals present along the L1 sequence (36). Some of these sites appear to be much stronger than the relatively weak poly(A) site found at the 3' end of the L1 element (37). The L1 sequence is, therefore, a “difficult” DNA template for cellular RNA polymerase II (PolII).

Nuclear export and translation efficiency seems to be influenced by polyadenylation, which stabilizes mRNA transcripts. Human genes vastly use alternative polyadenylation sites, and transposable elements embed these signals, which suggests that TEs can influence the 3′ end processing of host gene transcripts (38).

Premature polyadenylation for a gene harboring an L1 insertion may lead to the translation of a novel isoform of the protein encoded by this gene. For example, a case of TE-induced alternative mRNA processing of the human *ATRN* gene has been described (405). *ATRN* transcripts were cleaved and polyadenylated within an L1 element that had retrotransposed into its intron. A soluble form of Attractin was encoded by the transcripts polyadenylated within the L1 element. This is a classical case of how TEs can bring about transcript diversity and directly affect cellular functions.

Experiments performed with different L1-coding plasmids confirmed that L1 sequence contains multiple cryptic polyA signals, in both sense and antisense orientations. Since different plasmid systems revealed the different strength of premature polyadenylation, this phenomenon is likely context dependent (36, 137).

2.4.3. L1 sense and antisense promoters can produce transcriptional interference and act as alternative promoters

L1 5' UTR has sense and antisense promoter activities. Consequently, its integration near or in a gene can impact the expression of this gene, for example through alternative initiation of transcription at its own promoter. The L1 promoter region contains a CpG island, which is heavily methylated in most normal tissues, which controls the transcriptional activity of this retroelement (226, 406). However, alterations of DNA methylation can lead to activate its promoter activities and their use as alternative promoters for the neighboring genes.

L1 bi-directional promoter increases the diversity of possible alternative transcripts. For example, insertion of an L1 in an intron in reverse orientation of transcription of the gene may result in gene breakage (407). In this phenomenon, the transcription of the gene leads to two transcripts: one containing exons upstream of L1 and ending at an early antisense polyadenylation site of the element and the other one starts at the second L1 antisense promoter containing exons downstream of the L1 insertion. Different transcriptional effects of a new insertion on the expression of nearby genes may occur together and lead to the synthesis of a wide variety of alternative transcripts, a phenomenon called transcriptional interference (Figure 24).

Mart Speek described for the first time the L1 antisense promoter (ASP) activity, which resides between 600-400 bases of the 5' UTR (39). He observed that cDNAs isolated from NTera2D1 cells often represent chimeric transcripts that contain 5' UTR of L1 spliced to the sequence of known genes or non-coding sequences. Up to now, the L1 ASP promoter has been shown to serve as alternative promoter for more than 40 human genes in a tissue-specific manner (39, 40, 174, 408). Therefore, L1 brings in the transcript diversity by providing alternative promoters for their host genes.

Recently, an antisense promoter was also experimentally characterized in mouse L1 retrotransposons, but located in *ORF1*, which leads as in humans to alternative transcription initiation. Indeed, ~100 novel fusion transcripts have been characterized (409).

In addition, analyses of transcription start sites (TSS) in mammals (mouse and human) by CAGE found that apart from the two known 5' UTR promoter, an additional potential alternative TSS resides in the 3'UTR of L1 (174). But the potential impact of this finding on surrounding gene expression remains unexplored.

Finally, the use of promoters embedded in TEs results in a drastic change of genic regulatory processes, particularly those mediated by epigenetic marks. The hypomethylation of some L1 sequences in tumors can lead to the reactivation of their antisense promoter contributing to disease progression. Indeed, L1 provides an alternative transcription start site for many human genes like *c-MET*, a receptor tyrosine kinase whose activation can cause tumorigenicity in a variety of tumors (39–42). Hypomethylation of an L1 copy embedded in *c-MET* intron 2 activates the

transcription of a truncated and oncogenic *c-MET* transcript in bladder cancer and has been proposed to be used as a cancer biomarker (68).

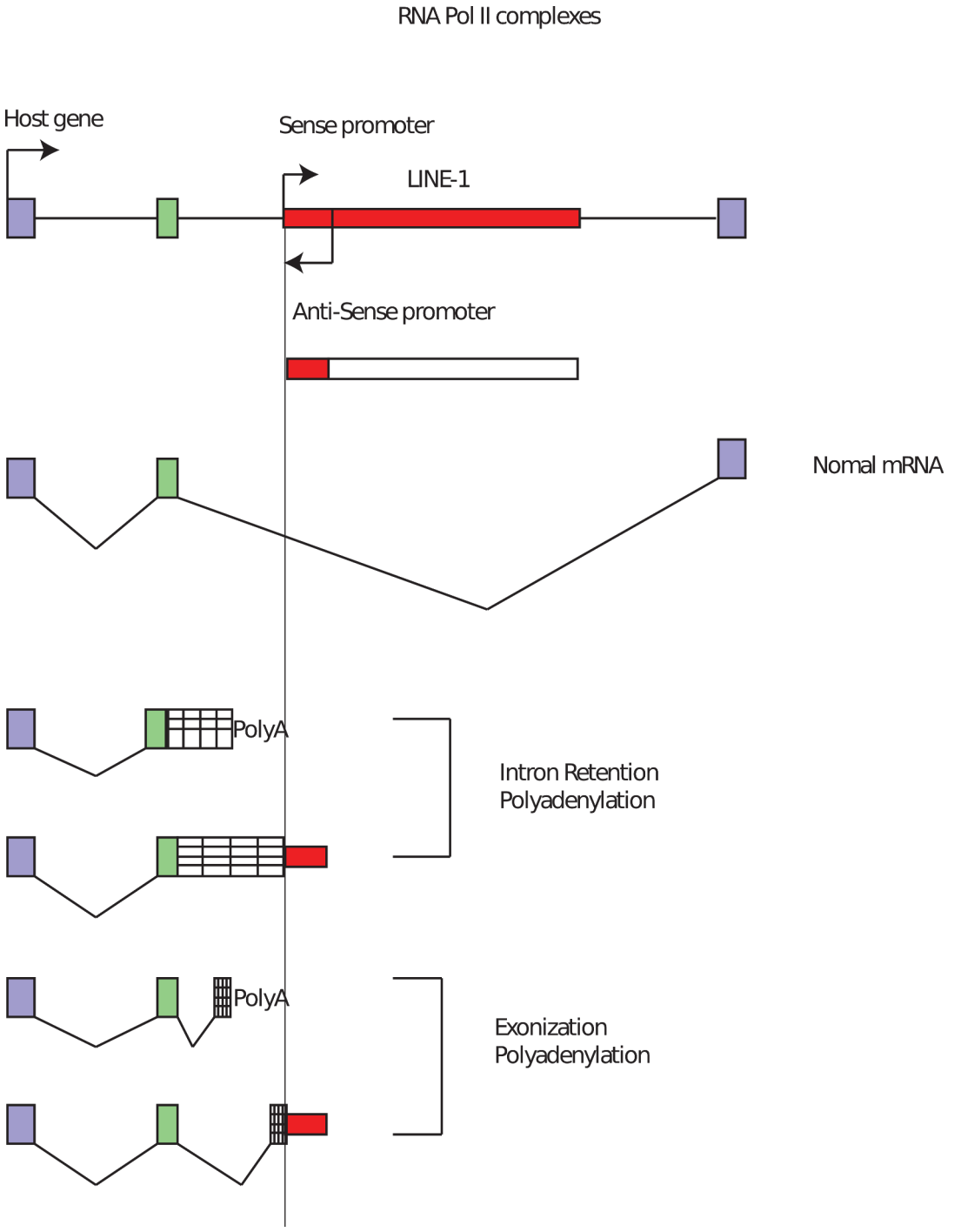


Figure 24: Transcriptional interference by L1. Adapted from (410). Sense and antisense promoter activities are major forces along with polyA and cryptic splice sites to cause intron retention. The cases shown illustrate L1 in intron causing host gene elongation by intron retention, cryptic splice sites, and polyA signal.

3. Thesis objectives

LINE-1 (L1) retrotransposons are the only autonomously active jumping genes in the human genome. They encode two proteins ORF1p and ORF2p, which associate with the L1 mRNA to form a ribonucleoprotein particle, which is considered as the core of the L1 retrotransposition machinery. L1 replicates via an RNA intermediate that is reverse transcribed into DNA at the site of insertion. A new L1 copy is produced when ORF2p nicks the genomic DNA and extends this newly formed 3′ end using the L1 mRNA as a template, a process known as target-primed reverse transcription (TPRT). The molecular determinants that influence L1 target site choice are not fully understood. This process has generated a considerable amount of structural genomic variants, which have impacted the organization of our transcriptome, through multiple mechanisms. But the comprehensive landscape of transcriptional variants due to L1 elements is currently unknown. We present here a bioinformatic work aimed at: (i) understanding the mechanisms that contribute to the distribution of new LINE-1 insertions within the genome; (ii) exploring the extent of L1-mediated genome variations; and (iii) its consequences on the diversity of human transcripts.

Towards this goal, I first tested the “snap-velcro model” *in silico*, at the genomic level. This model, based on quantitative biochemical assays, proposes that the DNA target site sequence and structure influence the reverse transcription step beyond endonuclease cleavage, and thus target site choice. I provided genomic evidence to support these in vitro findings.

Second, I developed an essential resource to explore L1-mediated genome variations by building the most comprehensive database so far of L1 insertional polymorphisms, identified in healthy or pathological human samples and published in peer-reviewed journals. This resource provides a bridge to link L1 insertional polymorphisms with phenotype or disease.

Finally, I designed and implemented a novel strategy to explore the landscape of transcript isoforms induced by L1 elements using RNA sequencing data. This work has the potential to highlight the overall impact of somatic insertions on gene expression and to help in understanding how the full set of L1 elements present in a given individual is regulated at the transcriptional level. Thus, in the longer term, this method could contribute to revealing L1-mediated mechanisms leading to transcriptome plasticity in tumor cells.

RESULTS

1. The specificity and flexibility of L1 reverse transcription priming at imperfect T-tracts

1.1. Context of the study

LINE-1 (L1) retrotransposons are the most abundant family of autonomously replicating retroelements in mammals. LINE-1 contains an internal promoter, which is located in the 5′-untranslated region and encodes two proteins, ORF1p and ORF2p, both being required for LINE-1 retrotransposition. ORF1p is an RNA-binding protein (115) and ORF2p an enzyme with endonuclease and reverse transcriptase activities (9, 10). ORF1p and ORF2p proteins associate with the L1 mRNA to form a ribonucleoprotein particle, which is considered as the core of the L1 retrotransposition machinery (11, 12). A new L1 copy is produced when ORF2p nicks the genomic DNA and extends this newly formed 3′ end using the L1 mRNA as a template, a process known as target-primed reverse transcription (TPRT) (5, 7, 10). L1 endonuclease recognizes consensus 5’ TTTT / AA 3’ sequence and induces the generation of nick. It is often preceded by long series of T, which may have the capability of hybridizing to the polyA tail of RNA L1. So we wanted to explore to which degree does the cleaved DNA need to be complementary to the poly(A) tail of the L1 RNA for efficient priming of reverse transcription. A previous technique called LEAP (L1 Element Amplification Protocol) (11) to measure the reverse transcriptase activity based on PCR amplification of reverse transcription products is more qualitative than quantitative. So new technique was developed by my fellow authors using isolated native RNPs of human cells transfected as for the LEAP technique, but to measure the efficiency of initiation of reverse transcriptase by quantifying the incorporation of labeled nucleotides during the synthesis cDNA.

We were able to show that the L1 RNP could effectively use a primer, whose terminal 4 nucleotides (Ts) anneal to polyA tail. We also observed like others did before that L1 RT could tolerate a terminal primer mismatch. The abundance of Ts in the 10 terminal nucleotides of the primer plays a role in the initiation efficiency of reverse transcription. Based on quantitative data initiation of reverse transcription obtained with more than 60 different primers, we hypothesized "snap-velcro" model to describe the specificity and flexibility of this.

My contribution in the project was to test the model by analyzing the distribution of recent L1 insertions or in vivo in the human genome. To perform this task, I developed a C++ program as described in the protocol S1. Potential (human genome) or real (recent catalogs of somatic L1 insertions in cancer genomes for which the insertion sites are annotated at nucleotide resolution) target sites with a recognizable EN target sequence were categorized based on their snap and velcro states. Followed by analysis to evaluate the respective effect of the snap and/or velcro on L1 insertion site frequencies (normalized frequency). This proved to be an extremely critical part of the project.

Finally, we showed that complementarity between DNA at the target site and L1 RNA poly(A) is important for RT priming efficiency and apart from the critical 4 terminal bases up to 10 bases influence RT priming efficiency and can compensate for the terminal mismatches.

1.2. Article-I

OPEN ACCESS Freely available online

PLOS | GENETICS

The Specificity and Flexibility of L1 Reverse Transcription Priming at Imperfect T-Tracts

Clément Monot[1,2,3☯], Monika Kuciak[1,2,3☯¤a], Sébastien Viollet[1,2,3], Ashfaq Ali Mir[1,2,3], Caroline Gabus[4¤b], Jean-Luc Darlix[4¤c], Gaël Cristofari[1,2,3]*

1 INSERM, U1081, Institute for Research on Cancer and Aging, Nice (IRCAN), Nice, France, **2** CNRS, UMR 7284, Institute for Research on Cancer and Aging, Nice (IRCAN), Nice, France, **3** University of Nice-Sophia-Antipolis, Faculty of Medicine, Nice, France, **4** Ecole Normale Supérieure de Lyon, Human Virology Department, INSERM U758, Lyon, France

Abstract

L1 retrotransposons have a prominent role in reshaping mammalian genomes. To replicate, the L1 ribonucleoprotein particle (RNP) first uses its endonuclease (EN) to nick the genomic DNA. The newly generated DNA end is subsequently used as a primer to initiate reverse transcription within the L1 RNA poly(A) tail, a process known as target-primed reverse transcription (TPRT). Prior studies demonstrated that most L1 insertions occur into sequences related to the L1 EN consensus sequence (degenerate 5′-TTTT/A-3′ sites) and frequently preceded by imperfect T-tracts. However, it is currently unclear whether—and to which degree—the liberated 3′-hydroxyl extremity on the genomic DNA needs to be accessible and complementary to the poly(A) tail of the L1 RNA for efficient priming of reverse transcription. Here, we employed a direct assay for the initiation of L1 reverse transcription to define the molecular rules that guide this process. First, efficient priming is detected with as few as 4 matching nucleotides at the primer 3′ end. Second, L1 RNP can tolerate terminal mismatches if they are compensated within the 10 last bases of the primer by an increased number of matching nucleotides. All terminal mismatches are not equally detrimental to DNA extension, a C being extended at higher levels than an A or a G. Third, efficient priming in the context of duplex DNA requires a 3′ overhang. This suggests the possible existence of additional DNA processing steps, which generate a single-stranded 3′ end to allow L1 reverse transcription. Based on these data we propose that the specificity of L1 reverse transcription initiation contributes, together with the specificity of the initial EN cleavage, to the distribution of new L1 insertions within the human genome.

Citation: Monot C, Kuciak M, Viollet S, Mir AA, Gabus C, et al. (2013) The Specificity and Flexibility of L1 Reverse Transcription Priming at Imperfect T-Tracts. PLoS Genet 9(5): e1003499. doi:10.1371/journal.pgen.1003499

Editor: Tom Eickbush, University of Rochester, United States of America

Received October 25, 2012; **Accepted** March 22, 2013; **Published** May 9, 2013

Funding: This work was financially supported by grants to GC from the Association pour la Recherche contre le Cancer (ARC, www.arc-cancer.net, Subvention Fixe #4854), the Institut National de la Santé Et de la Recherche Médicale (INSERM, www.inserm.fr), the Institut National du Cancer (INCa, www.e-cancer.fr, Avenir 2008 Grant), and the European Research Council (ERC, erc.europa.eu, #243312, Retrogenomics). CM was supported by a PhD fellowship from the French Ministry of Research and from the Association pour la Recherche contre le Cancer (ARC, www.arc-cancer.net). MK was supported by a PhD fellowship from the Ligue Nationale Contre le Cancer (www.ligue-cancer.net). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Gael.Cristofari@unice.fr

¤a Current address: University of Geneva, Faculty of Medicine, Department of Genetic Medicine and Development, Geneva, Switzerland
¤b Current address: University of Geneva, Sciences III, Department of Molecular Biology, Geneva, Switzerland
¤c Current address: Laboratoire de Biophotonique et Pharmacologie, Faculté de Pharmacie, UMR 7213 CNRS, Université de Strasbourg, Illkirch, France

☯ These authors contributed equally to this work.

Introduction

Retrotransposons are highly repetitive and dispersed sequences, accounting for almost half of our DNA [1]. These elements have the ability to proliferate in genomes through an RNA-mediated copy-and-paste mechanism, called retrotransposition. LINE-1 (L1) elements are the only autonomously active elements in humans and one of the most active elements in mice. They belong to the broad family of non-LTR retrotransposons (see [2–6] for recent reviews).

L1 retrotransposition starts with the transcription of a 6 kb L1 RNA driven by an internal Pol-II promoter [7]. After its export to the cytoplasm, the bicistronic L1 mRNA is translated into two proteins (ORF1p and ORF2p), which associate preferentially in *cis* with their encoding mRNA [8–11]. This is a critical feature of the L1 replication mechanism since it limits the association of the L1 machinery with other cellular mRNAs, including defective L1 RNA sequences, and thus increases the specificity of the reverse transcription process. The resulting complex is a stable ribonucleoprotein (RNP) thought to form the core of the retrotransposition machinery [10,12–19]. Its precise composition is currently unknown but it contains at least the L1 RNA and the ORF1p and ORF2p proteins [10,16,18,19]. The ORF1p protein is a trimeric RNA binding protein with RNA chaperone activity [20–25] and the ORF2p protein shows endonuclease (EN) and reverse transcriptase (RT) activities [26,27]. All are essential to L1 retrotransposition [16,18,28,29]. The L1 RNP is imported into the nucleus where reverse transcription and integration into the host genome take place [30].

Author Summary

Jumping genes are DNA sequences present in the genome of most living organisms. They contribute to genome dynamics and occasionally result in hereditary genetic diseases or cancer. L1 elements are the only autonomously active jumping genes in the human genome. They replicate through an RNA–mediated copy-and-paste mechanism by cleaving the host genome and then using this new DNA end as a primer to reverse transcribe its own RNA, generating a new L1 DNA copy. The molecular determinants that influence L1 target site choice are not fully understood. Here we present a quantitative assay to measure the influence of DNA target site sequence and structure on the reverse transcription step. By testing more than 65 potential DNA primers, we observe that not all sites are equally extended by the L1 machinery, and we define the rules guiding this process. In particular, we highlight the importance of partial sequence complementarity between the target site and the L1 RNA extremity, but also the high level of flexibility of this process, since detrimental terminal mismatches can be compensated by an increasing number of interacting nucleotides. We propose that this mechanism contributes to the distribution of new L1 insertions within the human genome.

The current model for non-LTR retrotransposon integration, named target-primed reverse transcription (TPRT), was originally deduced from biochemical studies on the insect R2Bm element [31]. This retrotransposon encodes a single protein with EN and RT activities and integration of new copies occurs at a specific and defined position in the rDNA [31,32]. The TPRT process is initiated by the formation of a nick in the genomic double-stranded DNA target. Then the R2 RT extends the newly formed 3′OH using the R2 RNA as a template [27,31,33–35]. Priming of reverse transcription occurs without any complementarity between the R2 RNA template and the DNA target site [36,37]. Non-LTR retrotransposons can be divided into several clades, which differ considerably in the machinery that they encode (single or multiple ORFs, restriction-like or APE-endonuclease, RNaseH or not, etc…) [38]. Despite these differences, cell culture-based retrotransposition assays and analyses of novel or recent integration sites have revealed the same overall requirement for EN and RT activities, supporting the TPRT model [28,39–43]. Intriguingly, non-LTR retrotransposon 3′ ends and preintegration sites often exhibit partial sequence identity, suggesting that annealing of the target site DNA to the RNA template might be a necessary step to prime reverse transcription, in contrast to R2 [40–43]. This step could significantly influence the genomic distribution of these elements, by imposing additional constraints after the initial endonuclease cleavage.

As regards L1, conclusive evidence on whether primer-template complementarities are required for efficient reverse transcription initiation is lacking. Most L1 pre-integration sites contain an EN recognition sequence (5′-TTTT/A-3′) and are often preceded by T-tracts of variable length [1,27,44–50]. Thus, in theory, the region covering the EN consensus and its upstream sequence has the ability to base-pair with the L1 poly(A) tail and to promote reverse transcription initiation. Nevertheless, target sites frequently contain nucleotides other than Ts, sometimes at the 3′ terminal end of the nicked DNA, which could severely impair interaction with the L1 RNA and extension by L1 RT. On the other hand, isolated recombinant L1 ORF2p produced in insect cells was found to equally extend any linear DNA substrate *in vitro*, without apparent sequence or structure requirement, or any need for primer-template complementarity [33]. Likewise, native L1 RNPs enriched from cells are able to extend oligonucleotides ending with terminal mismatches [10,51], indicating that complementarity base-pairing between the 3′ end of the target DNA and the L1 RNA template is not an absolute requirement. But Kulpa and Moran also observed that primer sequence could influence RT initiation [10]. A common limitation of these previous studies was the use of PCR-based assays, which precluded a quantitative comparison of priming efficiencies and might lead to the detection of marginal products.

Here, we addressed the question whether - and to which degree - the liberated 3′-hydroxyl extremity on the genomic DNA needs to be accessible and complementary to the poly(A) tail of the L1 RNA for efficient priming of reverse transcription. To achieve this goal, we validated a direct L1 extension assay (DLEA) to quantitatively measure the ability of native L1 RNPs to initiate reverse transcription. Then we systematically assayed more than 65 DNA substrates varying in sequence and structure, allowing us to define the preferential rules of L1 reverse transcription priming. Our results clarify the importance of base-pairing between the L1 RNA template and the target site DNA for this process and demonstrate its exceptional flexibility.

Results

A direct L1 extension assay (DLEA) to study the initiation of reverse transcription by native L1 RNPs

To test the DNA primer requirements for initiating L1 reverse transcription, we set up a direct L1 extension assay (DLEA), which would avoid PCR and therefore would allow us to quantitate L1 priming efficiencies. The L1 retrotransposition machinery is notoriously difficult to express and to detect in most experimental systems. To obtain sufficient amounts of L1 RNPs for direct detection, we modified the protocol developed by Kulpa and Moran [10] by transiently overexpressing the canonical human L1.3 element [28] (referred thereafter as hL1) or a codon-optimized murine L1spa element (Orfeus [52], referred thereafter as mL1) in HEK293T cells, followed by a 3-day selection of transfected cells. HEK293T cells are transfected with much higher efficiency and express higher levels of transgenes than the HeLa cells, which were used in the original protocol. Then we prepared native L1 RNPs from cell extracts by sucrose cushion ultracentrifugation as previously reported (Figure 1A) [10]. In parallel, we prepared RNPs from empty vector-transfected cells or with a point mutation in the RT active site (D702A for hL1 and D709A for mL1, referred thereafter as RT* L1) as negative controls. We detected the mORF1p protein in RNP preparation from mL1-transfected cells but not from hL1 or empty vector-transfected cells by immunoblotting (Figure 1B, compare lanes 1–3 with 4–5). Similarly hORF1p levels were much higher in hL1-transfected cells than in vector control cells (Figure 1B, lanes 2–3). However long exposure revealed low levels of endogenous hORF1p in all RNP preparations (Figure 1B, lanes 1 and 4–5). To evaluate the presence of L1 RT activity and L1 RNA associated with ORF1p in the RNP preparations, we used the L1 element amplification protocol (LEAP) in which the L1 RT first extends a primer and the resulting cDNA is subsequently amplified by PCR [10]. The PCR primers are anchored in the tail of the RT primer and in the Neomycin-resistance genetic marker inserted in the transfected L1 3′ UTR. Therefore only products produced from the transfected L1 element can be amplified. Since hL1 and mL1 share the same genetic marker, the same primers can be used for both elements. As expected from previous work [10,18], we detected L1 RT

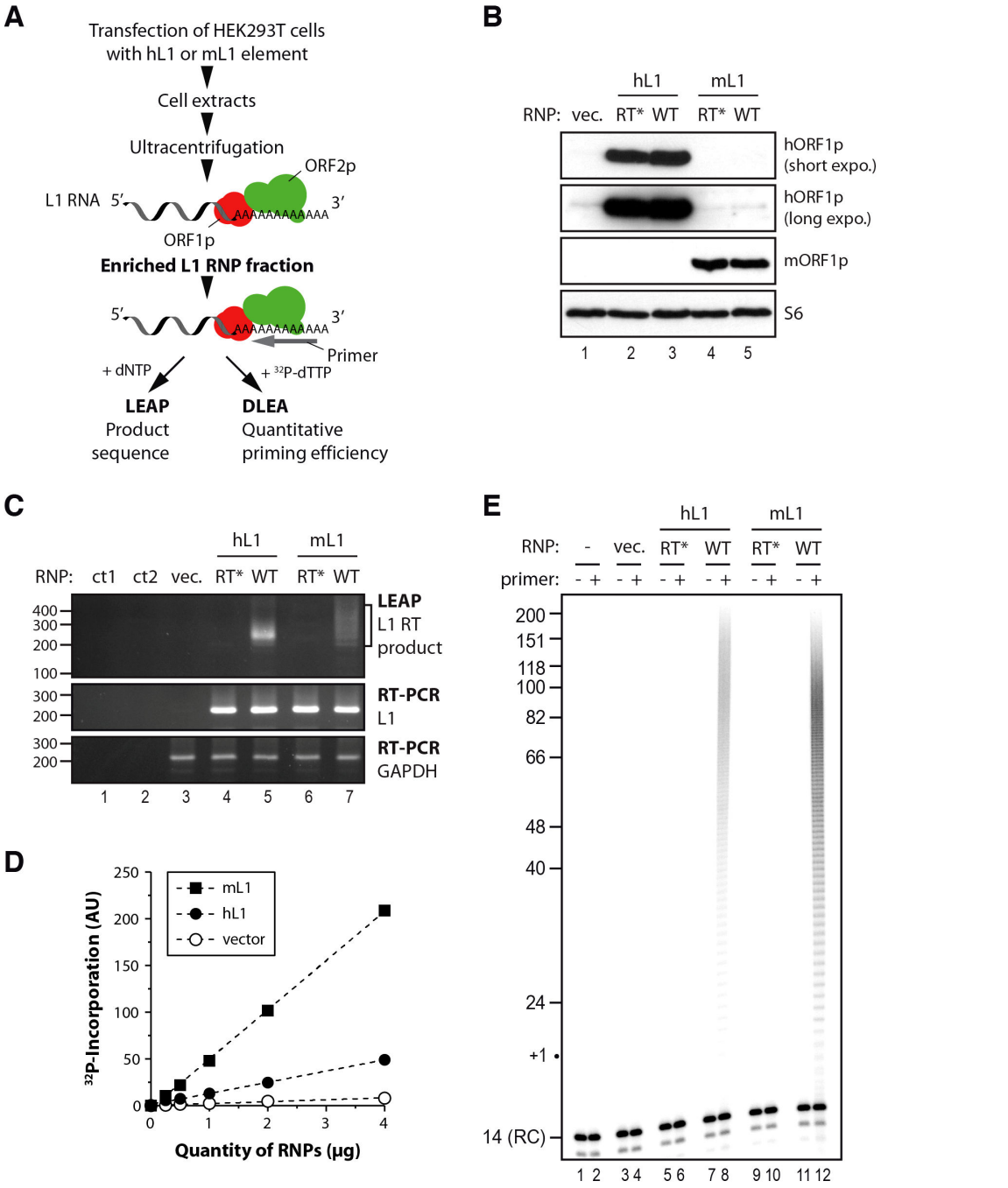


Figure 1. Initiation of L1 reverse transcription by native L1 RNPs. (A) Outline of the experimental procedure. LEAP, L1 element amplification protocol; DLEA, Direct L1 Extension Assay (B) Immunoblotting of human ORF1p (top 2 panels) or murine ORF1p (panel 3 from the top) in RNPs (16 µg) prepared from cells transfected with empty vector (lane 1), RT* hL1 (lane 2), wild-type hL1 (lane 3), RT* mL1 (lane 4), wild-type mL1 (lane 5). Ribosomal S6 protein was detected using an anti-S6 antibody and was used as an RNP loading control (bottom panel). (C) Detection of L1 RT activity by LEAP (top panel) and of L1 RNA by conventional RT-PCR (middle panel) in RNP preparations. GAPDH RNA is a cellular RNA used as a loading control for all RNPs (bottom panel). Annotations are the same as in (B). ct1, a control for the PCR step without cDNA; ct2, a control for the RT step without RNP or RNP-extracted RNA. The LEAP product is a diffuse smear starting from 207 bp (bracket). (D) Standard curve of murine (black square) or human (black circles) L1 RNP DNA polymerase activity, showing linear conditions, compared to vector control RNP (empty circles). Note that the

intrinsic activities of mL1 and hL1 RNPs cannot be directly compared due to potential differences in their levels of expression. (E) Direct L1 extension assay (DLEA) with or without a $(dT)_{18}$ primer in the presence of α-^{32}P-dTTP (even and odd lanes, respectively). Sucrose cushion fractions prepared from human (lanes 5–8) or murine (lanes 9–12) L1-transfected cells or vector-transfected cells prepared in parallel (lanes 3–4) were used as a source of RNPs. Trace amounts of a 14-nt 5′ end-labeled oligonucleotide was added *after* the reaction as a recovery control (denoted RC). RT*, RT-defective L1 RNP; WT, wild-type L1 RNP.
doi:10.1371/journal.pgen.1003499.g001

activity only in the RNP prepared from wild-type hL1 or mL1, but not in the vector or RT-defective L1 transfected cells (Figure 1C, top panel, compare lanes 5 and 7 with 3–4 and 6), even if the L1 RNA is present (Figure 1C, middle panel). Sequencing of the LEAP products confirmed that hL1 or mL1 RNA was reverse transcribed. This indicated that RNPs produced in our experimental conditions contain the core of the L1 machinery and used L1 RNA as a template. Previous studies have shown that L1 RNPs enriched on sucrose cushion as prepared here co-fractionate with many other cellular RNPs, including ribosomes [10,16]. However, the L1 RNA is reverse transcribed at least 100 times more efficiently than other co-fractionating abundant cellular RNAs [10], a property known as L1 cis-preference [8,9].

We reasoned that if L1 RNPs were active enough we should detect the extension of an $oligo(dT)_{18}$ primer in the presence of radiolabelled ^{32}P-dTTP. This reaction would mimic the initiation step of L1 reverse transcription, which starts at the poly(A) tail of the L1 RNA. After a 4 min incubation at 37°C, we purified the reaction products and resolved them on sequencing gels. A short end-labeled oligonucleotide was added *after* the reaction as a recovery control (RC). No or minimal extension was detected in vector or RT-defective controls consistent with the presence of only minimal amounts of endogenous hL1 activity in RNP preparations (Figure 1E, lanes 3–6 and 9–10, and Figure 1D). In contrast when wild-type hL1 or mL1 element was transfected we could easily detect the incorporation of radiolabelled dTMPs (Figure 1D and Figure 1E, lanes 8 and 12). Importantly, the amount of product formed was linearly dependent on the amount of L1 RNPs (Figure 1D), showing that the levels of primer extension could be quantitatively measured under the reaction conditions employed (linear phase, also known as initial velocity phase). We focused our work on reverse transcription initiation by using short extension times (4 min) and by adding only ^{32}P-dTTP to the reaction and no other dNTP. In these experimental conditions, the products were short enough to be resolved on sequencing gels and we could follow the extension at the nucleotide resolution. The linear phase ranged from 0.2–0.25 µg up to 4 µg of RNPs, which indicates a dynamic range between 10- and 20-fold (data not shown). We chose to use 2 µg of RNPs, at the upper end of the linear range, for all following experiments and to set to 100% the level of extension obtained with an $oligo(dT)_{18}$ primer under these conditions. Based on the dynamic range of the initial RNP titration, primer extension efficiencies as low as 5% should therefore be reliably quantified. The products are heterogeneous in length, consistent with the expected products of poly(A) reverse transcription and range from 19 nucleotides (nt) to approximately 150 nt (Figure 1E, lanes 8 and 12).

To further confirm that the ladder observed results directly from the reverse transcriptase activity of the transfected L1 element, we performed additional controls. RNase treatment reduced primer extension to undetectable levels (Figure S1A, compare lanes 2 and 3), showing that the detected DNA polymerase activity is RNA-dependent. If the reaction is conducted in the presence of RT inhibitors known to inhibit L1 retrotransposition and recombinant L1 RT activity [53–55] such as AZT or d4T, DNA polymerization is abolished (Figure S1B, compare lanes 2 and 3–4). No extension was detected in these experimental conditions with radiolabelled dATP, dGTP or dCTP in agreement with the reverse transcription of the poly(A) sequence (data not shown). When extension time was prolonged to 1 h (Figure S1C), the reaction was not in its linear phase anymore (and the assay was no longer quantitative). Products were longer than the maximum poly(A) length in mammals (~250 nt), which is likely to result from L1 RT slippage in the poly(A) track as recently reported *in vivo* [56]. If all four dNTPs were present in the reaction, high molecular weight products appeared, consistent with reverse transcription ongoing beyond the L1-poly(A) boundary (Figure S1D) and in agreement with the LEAP results (Figure 1C).

Altogether these results show that DLEA detects *bona fide* initiation of reverse transcription by native mammalian L1 RNPs through the direct incorporation of radiolabeled dTMP in a primer extension reaction. Importantly, DLEA is quantitative since it demonstrates a linear relationship between the signal and RNP quantities under the reaction conditions employed.

Efficient extension of single-stranded DNA by the L1 RNP requires at least 4 terminal matching bases

In contrast to most DNA polymerases, it was previously demonstrated that the hL1 RNP is able to extend a terminal mismatched base pair using a PCR-based assay followed by sequencing of the products [10]. To determine more quantitatively the efficiency of extension of such mismatched primers, we changed the last nucleotides of the $oligo(dT)_{18}$ primer to a non-T nucleotide in order to prevent base-pairing of the primer 3′ end to the L1 poly(A) tail (Figure 2A). Although decreased as compared to the $oligo(dT)_{18}$ primer, the hL1 RNP can extend a primer with a single or double terminal mismatch (V_1 and V_2, Figure 2B, lanes 3–4; V = not T) or with a mismatch at the penultimate position (VN, 15% of the $oligo(dT)_{18}$ extension, not shown), in agreement with previous reports [10,51]. In contrast, if the primer ends with more than two mismatched nucleotides (V_3 to V_6), DNA polymerization becomes undetectable under the employed reaction conditions (Figure 2B, lanes 5–7). Similarly, the hL1 RNP is not able to efficiently use an unrelated oligonucleotide ending with three Gs (the T7 promoter primer, noted R, Figure 2A) as a primer for its reverse transcription (Figure 2B, lane 8).

Next, we measured the influence of each individual terminal base on primer extension. Although all terminal mismatches reduced the efficiency of reverse transcription initiation to some extent, a terminal G was the most detrimental, whereas a C or an A was better tolerated (Figure 3). Thus the levels of extension of a T-tract is dependent on the nature of its 3′ terminal base with the following preference: T>C>A>G.

To further characterize the need for terminal matching nucleotides in the priming of hL1 reverse transcription, we added an increasing number of Ts to the R primer (T_1 to T_6). Initiation of reverse transcription is robustly detected only when the single-stranded primer ends with at least 4 Ts and trace activity can already be detected with 3 terminal Ts (Figure 2B, lanes 11–13). We obtained similar results with mL1 RNPs (Figure 2C, lanes 1–7 and Figure S2).

In order to compare the properties of the native L1 RNPs with a retroviral RT, we tested the ability of recombinant Avian Myeloblastosis Virus (AMV) RT to prime reverse transcription

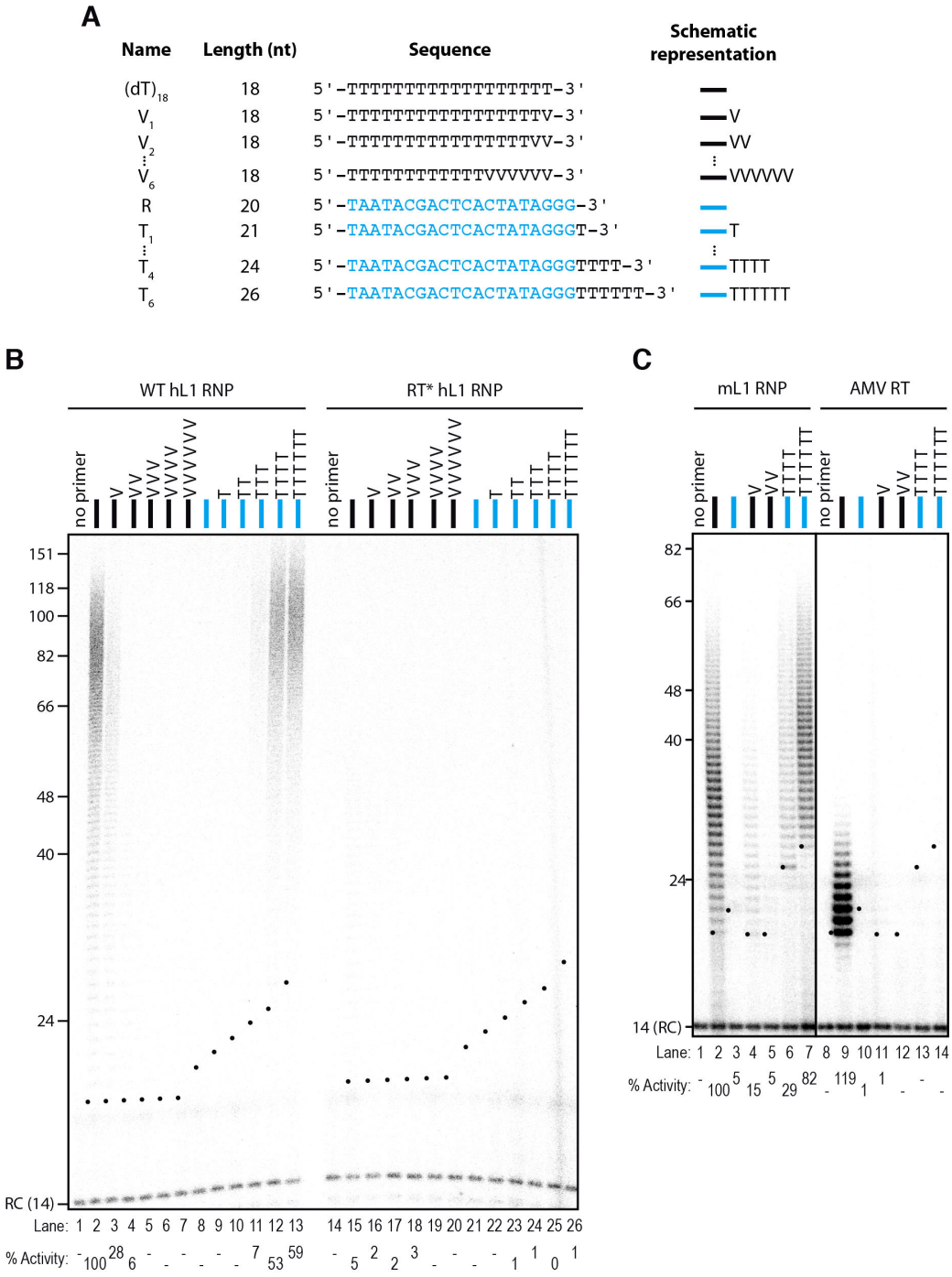


Figure 2. The L1 RNP preferentially extends primers ending with at least 4 Ts. (A) Scheme of the primers used. The oligonucleotide shown in blue and named R corresponds to the T7 promoter primer chosen as an unrelated sequence. V is the IUPAC nucleotide symbol for A, G or C but not T. (B) DLEA showing the extension of single-stranded primers by hL1 RNPs in the presence of α-^{32}P-dTTP. (C) Comparison of the mouse L1 RNP and AMV RT for their ability to extend single-stranded primers in the presence of α-^{32}P-dTTP. Experimental conditions were as in Figure 1. As a template, poly(rA) was added to the reaction performed with the AMV RT. Lanes 1–7 and 8–14 are from the same gel. RC denotes a 14 nt recovery control added after the reaction but before DNA purification. The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies. Quantification of primer extension (% Activity) was relative to levels of extension obtained with oligo(dT)$_{18}$.
doi:10.1371/journal.pgen.1003499.g002

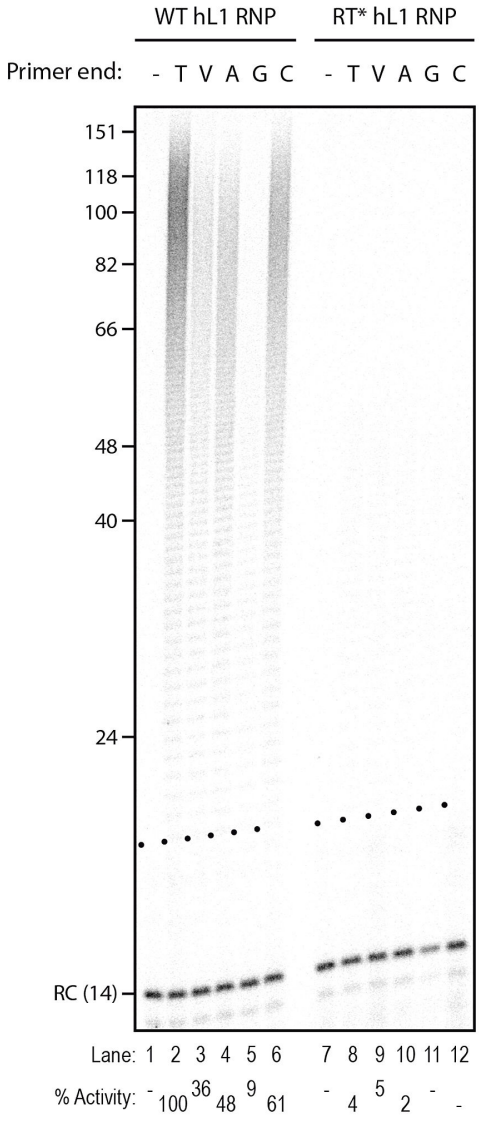


Figure 3. Influence of the terminal nucleotide on primer extension by L1 RNP. DLEA showing the extension of single-stranded primers by hL1 RNPs in the presence of α-^{32}P-dTTP. All primers are oligo(dT)$_{17}$-X oligonucleotides, where X corresponds to the nucleotide indicated above the lanes. V is the IUPAC nucleotide symbol for A, G or C but not T. (−) is a control without primer. Experimental conditions were as in Figure 1. RC denotes a 14 nt recovery control added after the reaction but before DNA purification. The black dots on the left side of each lane indicate the expected start of reverse transcription. Quantification of primer extension (% Activity) was relative to levels of extension obtained with oligo(dT)$_{18}$ (lane 2).
doi:10.1371/journal.pgen.1003499.g003

under identical experimental conditions. In these experiments, exogenous poly(rA) was added as a template together with quantities of the AMV RT that lead to similar levels of extension as the L1 RNP using the (dT)$_{18}$ primer (Figure 2C, compare lanes 2 and 9). Under these experimental conditions, reverse transcription by AMV RT was not primed by oligonucleotides ending with terminal mismatches (Figure 2C, compare lanes 4–5 to 11–12) or by oligonucleotides ending with 4 or 6 Ts (Figure 2C, compare lanes 6–7 to 13–14). These observations suggest that limited base-pairing interactions between the primer and the template might be stabilized by the L1 RNP, through direct binding of ORF1p or ORF2p to the single-stranded DNA. In addition, the extension products of the (dT)$_{18}$ oligonucleotide obtained with the AMV RT are much shorter than those obtained with the L1 RNP. This might suggest that the L1 RNP is more processive than the AMV RT and/or that the L1 RNP has a higher affinity for dTTP than AMV RT as shown for the R2 element [57,58]. However, since the templates used are not strictly similar, it is difficult to draw definitive conclusions on this aspect.

It was previously reported that a nuclease activity in the RNP preparations could process primers before their extension [51]. Thus, in principle, it is possible that primers ending with terminal mismatches are first processed to eliminate the mismatch(es) and then extended. Against this possibility, the majority of the products observed in sequencing gels start at the expected +1 position or above (Figure 2 and Figure S2). As an additional control, we performed LEAP reactions using primers ending with the same sequence as depicted in Figure 2A. We could amplify, clone and sequence products with up to 3 terminal mismatches (Figure S3A). Although a small percentage of processed primers were found (7 out of 160 sequences in total), the majority of the mismatches were directly extended (Figure S3C). Thus differences of extension are not due to differential processing of the primers. We note that the levels of the nuclease activity responsible for primer processing, which co-fractionates with L1 RNPs in sucrose gradients, might dependent on the cell type used to prepare RNPs. Using the same RACE primer ending with VN, Kulpa *et al.* observed processing in 33/81 (39%) of the analyzed clones obtained with HeLa cells, while Kopera *et al.* found 5/45 (11%) of processed primers in CHO-derived cell lines. In comparison, we obtained 2/70 (3%) clones showing a processed primer with RNPs prepared from HEK293T cells.

Altogether these observations show that native L1 RNPs efficiently prime reverse transcription at DNA ending with 4–6 terminal matching nucleotides, although it can accommodate terminal mismatches with lower priming efficiencies.

The L1 RNP extends primers mimicking *bona fide* insertion sites with variable efficiencies

L1 EN-mediated nicking at a consensus target site produces a 3′-OH DNA ending with four Ts [27,44]. This is consistent with our observation that the L1 RT can extend primers ending with as little as four Ts. However, L1 integration sites often contain degenerate L1 EN recognition sites that differ from the consensus recognition sequence [1,46,47]. This prompted us to analyze the ability of native hL1 RNPs to extend primers which mimic *bona fide* insertion sites. We designed 35 primers corresponding to previously published insertion sites recovered from new hL1 retrotransposition events obtained in cultured cells [46]. The sequence and the original name of each recovered clone is indicated in Figure 4A. Levels of extension were normalized to those obtained with the primer LOU541 (clone 10BglIIL1.3), which corresponds to a (dT)$_{20}$ oligonucleotide.

We observed that all sites are not equally extended (see Figure 4A). The levels of extension range between 7% (LOU535) and 120% (LOU552). The best primer is 17-fold more extended than the least-efficient primer. Even if we know that these target sites were used *in vivo* without processing [46], we choose six of them differing from each other by the position or the nature of the mismatched nucleotides to perform LEAP (Figure S3B) and we

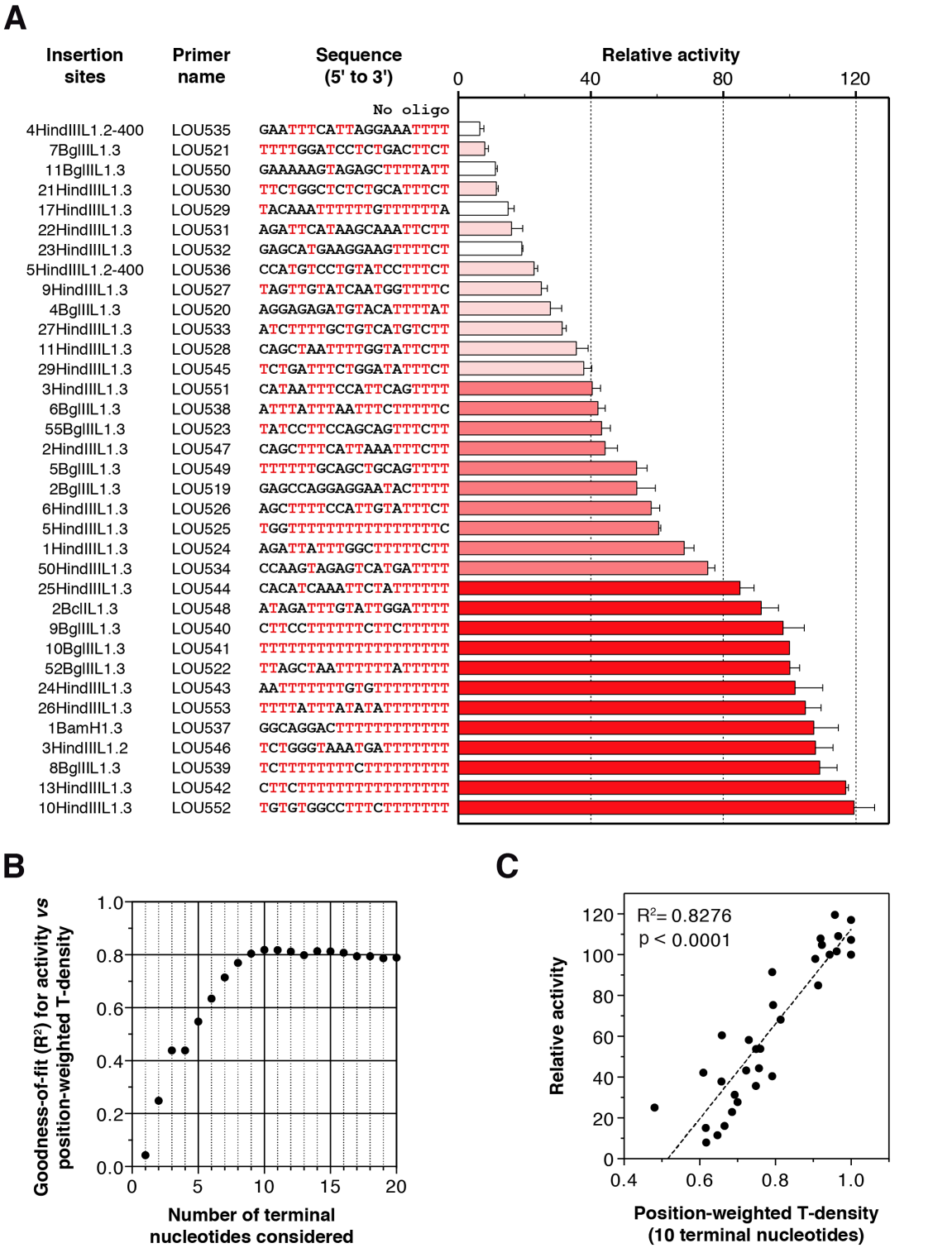

A
Insertion sites
Primer name
Sequence (5' to 3')
Relative activity
0
40
80
120
No oligo
4HindIIIL1.2-400 LOU535 GAATTTCATTAGGAAATTTT
7BglIIL1.3 LOU521 TTTTGGATCCTCTGACTTCT
11BglIIL1.3 LOU550 GAAAAAGTAGAGCTTTTATT
21HindIIIL1.3 LOU530 TTCTGGCTCTCTGCATTTCT
17HindIIIL1.3 LOU529 TACAAATTTTTTGTTTTTTA
22HindIIIL1.3 LOU531 AGATTCATAAGCAAATTCTT
23HindIIIL1.3 LOU532 GAGCATGAAGGAAGTTTTCT
5HindIIIL1.2-400 LOU536 CCATGTCCTGTATCCTTTCT
9HindIIIL1.3 LOU527 TAGTTGTATCAATGGTTTTC
4BglIIL1.3 LOU520 AGGAGAGATGTACATTTTAT
27HindIIIL1.3 LOU533 ATCTTTTGCTGTCATGTCTT
11HindIIIL1.3 LOU528 CAGCTAATTTTGGTATTCTT
29HindIIIL1.3 LOU545 TCTGATTTCTGGATATTTCT
3HindIIIL1.3 LOU551 CATAATTTCCATTCAGTTTT
6BglIIL1.3 LOU538 ATTTATTTAATTTCTTTTTC
55BglIIL1.3 LOU523 TATCCTTCCAGCAGTTTCTT
2HindIIIL1.3 LOU547 CAGCTTTCATTAAATTTCTT
5BglIIL1.3 LOU549 TTTTTTGCAGCTGCAGTTTT
2BglIIL1.3 LOU519 GAGCCAGGAGGAATACTTTT
6HindIIIL1.3 LOU526 AGCTTTTCCATTGTATTTCT
5HindIIIL1.3 LOU525 TGGTTTTTTTTTTTTTTTTC
1HindIIIL1.3 LOU524 AGATTATTTGGCTTTTTCTT
50HindIIIL1.3 LOU534 CCAAGTAGAGTCATGATTTT
25HindIIIL1.3 LOU544 CACATCAAATTCTATTTTTT
2BclIL1.3 LOU548 ATAGATTTGTATTGGATTTT
9BglIIL1.3 LOU540 CTTCCTTTTTTCTTCTTTTT
10BglIIL1.3 LOU541 TTTTTTTTTTTTTTTTTTTT
52BglIIL1.3 LOU522 TTAGCTAATTTTTTATTTTT
24HindIIIL1.3 LOU543 AATTTTTTTGTGTTTTTTTT
26HindIIIL1.3 LOU553 TTTTATTTATATATTTTTTT
1BamH1.3 LOU537 GGCAGGACTTTTTTTTTTTT
3HindIIIL1.2 LOU546 TCTGGGTAAATGATTTTTTT
8BglIIL1.3 LOU539 TCTTTTTTTTCTTTTTTTTT
13HindIIIL1.3 LOU542 CTTCTTTTTTTTTTTTTTTT
10HindIIIL1.3 LOU552 TGTGTGGCCTTTCTTTTTTT
B
Goodness-of-fit (R²) for activity vs position-weighted T-density
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Number of terminal nucleotides considered
C
R²= 0.8276
p < 0.0001
Relative activity
120
100
80
60
40
20
0
0.4
0.6
0.8
1.0
Position-weighted T-density (10 terminal nucleotides)

Figure 4. Extension of primers mimicking *bona fide* human L1 insertion sites by the human L1 RNP. (A) Relative extension of primers as measured by DLEA. Extension of each primer was normalized to the extension levels obtained with the $(dT)_{20}$ primer (LOU541 corresponding to the 10BglIIIL1.3 insertion site). This ratio, expressed as a percentage, was designated as 'Relative activity'. Bars were color-coded and sorted according to the efficiency of priming (red, activity ≥80%; medium red, 40%≤Activity<80%; light red, activity <40%; white, primers excluded from the correlation analyses due to hairpin formation). Bars indicate the mean and error bars the S.E.M. (n = 3). The name of the insertion sites correspond to the recovered clones from cultured cells published in [46]. (B) A role for the primer terminal nucleotides in hL1 RNP reverse transcription priming. For each *n* between 1 and 20, the correlation between activity and position-weighted T-density of the terminal *n* nucleotides was calculated. The goodness-of-fit (R^2) only marginally changes when $n>10$, indicating that the terminal 10 nucleotides are the most relevant determinants for priming efficiency. Note that the 4th bases at the 3′ terminus in all the primers of this set are coincidentally identical (T). For this reason, R^2 is identical for $n = 3$ and $n = 4$. See the 'Results' and 'Material and Methods' sections for a detailed definition of the position-weighted T-density. (C) An example of correlation between the density of Ts close to the 3′ end of the primer (position-weighted T-density) and the efficiency of reverse transcription priming (for $n = 10$). For the graph shown in (B) and (C), primers which could fold into a structured hairpin (white bars in A) were excluded from the analysis (see Figure 6, Figure 7, Figure 8 for a detailed analysis of primer structure on reverse transcription efficiency).
doi:10.1371/journal.pgen.1003499.g004

sequenced the products. Again we found a small number of processed primers (~5%), but the majority of products result from the direct extension of mismatched primers (Figure S3).

We categorized primers based on their potential of extension (Figure 4A; 0–40%, light red; 40–80%, medium red; 80–120%, dark red). Four primers have the ability to form stable hairpins (Figure 4A, white bars), and were excluded from further analyses since hairpin formation is dependent on primer length, which was arbitrarily chosen (the specific impact of primer structure on L1 RT initiation is presented at the end of the 'Results' section). Top ranking primers (dark reds) all end with at least 4 Ts, often more, and are extremely rich in Ts, in agreement with the results presented in Figure 2. Interestingly, primers with a mismatch in the last critical four nucleotides are more efficiently extended if they are preceded by a T-rich upstream sequence. For example, primers LOU525, LOU527 and LOU538 all end with 5′-TTTC-3′ and their respective levels of extension are LOU527<‴ LOU538<LOU525, which roughly follows the number of Ts close to the 3′ end. This suggests a compensation mechanism allowing the extension of primers ending with suboptimal sequences.

To address the significance of this phenomenon more quantitatively, we calculated for each oligonucleotide two parameters: (i) the density of Ts (number of Ts/length of the oligonucleotide), which simply reflects the abundance of Ts in the primer, and (ii) the position-weighted T-density, which is similar but the weight of each T is inversely proportional to the distance from the 3′ end (see Material and Methods section for more details). Using linear regression, we found that the activity correlates significantly with both parameters ($p = 0.0002$ and $p<0.0001$, respectively) but the goodness-of-fit is much better with the position-weighted T-density than with the T-density ($R^2 = 0.7895$ *vs* 0.3950, not shown). To evaluate the number of terminal nucleotides that contribute to priming efficiency, we further correlated the priming efficiency with position-weighted T-density, taking into account a variable number of terminal nucleotides. The goodness-of-fit (R^2) increases steadily up to 10 considered nucleotides and then reaches a plateau (Figure 4B). Considering nucleotides beyond position 10 (from the 3′ primer end) does not improve the correlation. The correlation between priming efficiency and the position-weighted T-density when only the last 10 nucleotides are considered is plotted in Figure 4C ($R^2 = 0.8276$).

In conclusion, we have demonstrated biochemically that complementarity between the L1 poly(A) tail and the last 10 nucleotides of the target DNA plays a role in extension at the target site, the last 4 nucleotides being the most critical. Suboptimal primers with a mismatch in their last 4 nucleotides are extended with a lower efficiency, which can be partially compensated by increasing the number of Ts in the upstream sequence.

The "snap-velcro" model and supportive evidence

To illustrate these findings, we propose that the four terminal bases of the primer, which overlap with the EN nuclease recognition sequence, act as a specific snap and the upstream six bases act as a weaker velcro strap (Figure 5A). When the snap is closed (perfect terminal matches, EN consensus sequence), initiation is efficient, but is enhanced if the velcro strap (upstream bases) is also tightly fastened. Inversely, if the snap is open (terminal mismatches), extension occurs preferentially if this is compensated by a tightly fastened velcro strap. The rational to distinguish snap and velcro regions is to highlight the preponderant role of the terminal nucleotides, which is also reflected in the position-weighted T-density mode of calculation.

To test this model, we determined for each primer whether the snap is open or closed and whether the velcro strap is loosely or tightly fastened. A snap was considered closed only if the 3′ end of the primer was $(T)_4$. The velcro strap was considered as tightly fastened if the position-weighted T-density score of this region was at least half of its maximum value (see Materials and Methods section for the precise definition of these states). Then for each group we calculated the mean efficiency of extension by the hL1 RNP (Figure 5B, data from Figure 4A). In agreement with the model, tightly fastened velcro improves the extension of target sites with a snap closed and partially rescue those with a snap open. Both snap and velcro contribute extremely significantly to the differences of extension between primers ($p<0.0001$, two-way ANOVA).

A testable prediction of this model is that, *in vivo*, at the genomic level, L1 elements would more frequently insert at putative EN recognition sites with a closed snap and a tightly fastened velcro strap; and that a tightly fastened velcro would favor insertions as compared to similar sites with an open velcro. To test this model, we searched in the human reference genome (hg19) for the position of all potential EN targets: R/TTTT, which corresponds to a closed snap; or R/VTTT, R/TVTT, R/TTVT and R/TTTV, which correspond to open snaps (R = purine, V = not T). For each of them, we extracted the 10 nucleotides upstream of the nick position and categorized each on the basis of its snap/velcro status to obtain the exact frequency of each category in hg19. Then we extracted the exact insertion sites for all the L1HS polymorphic insertions present in dbRIP [59] or in recent catalogs of somatic L1 insertions in cancer genomes [60,61] for which the insertion sites are annotated at nucleotide resolution. Since some insertions occurred through an EN-independent mechanism, we only kept sites with a recognizable EN target (R/TTTT, R/VTTT, R/TVTT, R/TTVT, R/TTTV, as above). We categorized these sites based on their snap/velcro status. First, we determined the distribution of these categories in the human reference genome (hg19, Figure 5C) or its repeat-masked counterpart (hg19 RM, Figure 5C) and we compared it to that

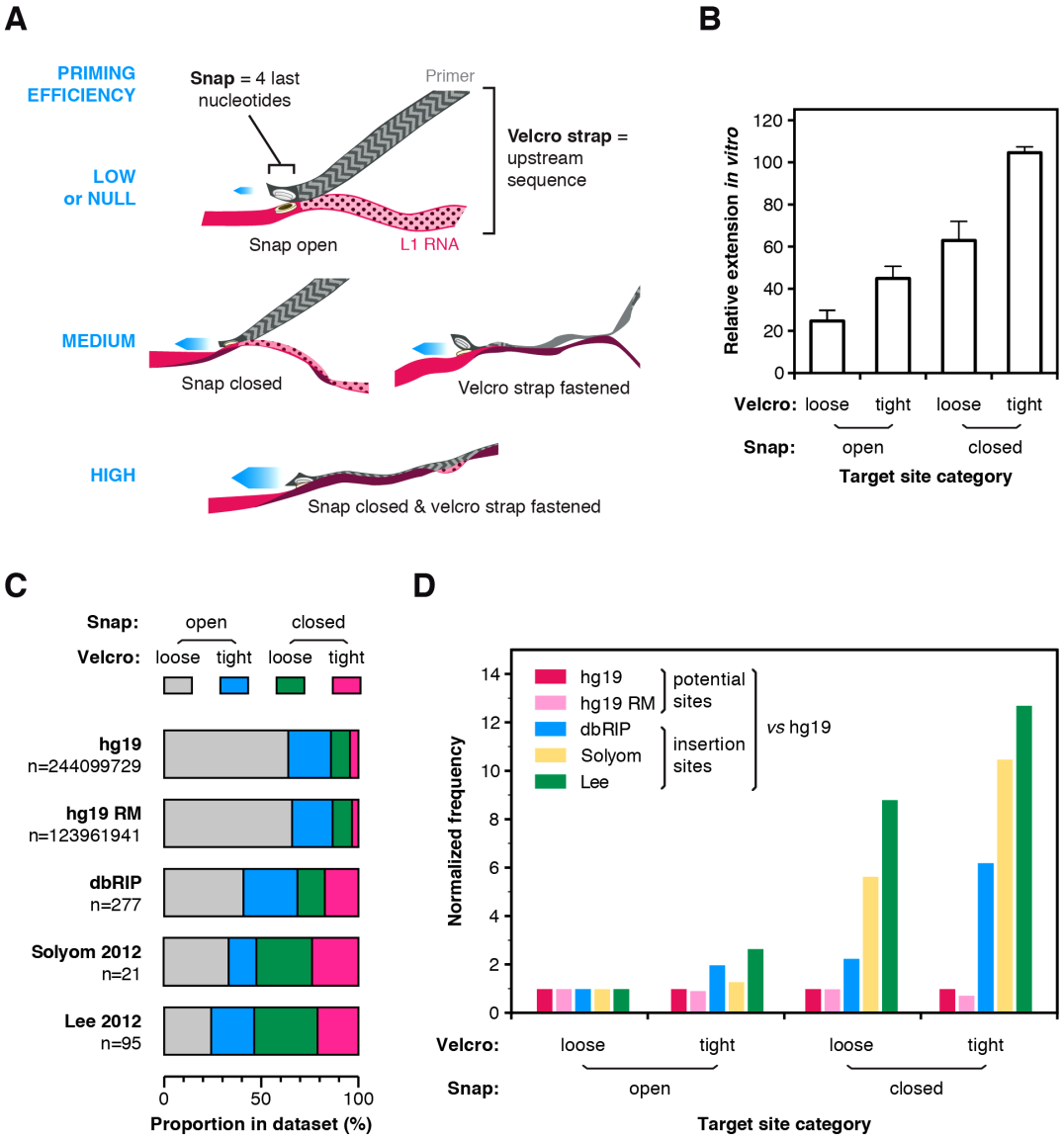


Figure 5. The snap-velcro model and supporting biochemical and genomic evidence. (A) A snap-velcro model for priming of L1 reverse transcription. The snap represents the 4 last nucleotides of the primer. It is considered as closed if it ends with 4 Ts (perfect terminal match) and as open if it contains a mismatch in the last 4 Ts. The velcro represents the 6 upstream bases. It is considered as tightly fastened only if the position-weighted T-score of this region is at least 50% of the maximum score. Otherwise, it is considered as loosely or not fastened. When the snap is closed and the velcro is tightly fastened, reverse transcription is high (bottom). If the snap is open *or* if the velcro is loosely fastened, reverse transcription priming is reduced (middle). Finally, if the snap is open *and* the velcro loosely fastened, reverse transcription priming is low or null (top). (B) *In vitro* efficiency of reverse transcription priming by the hL1 RNP depending of the snap and velcro status. Bars indicate the mean and error bars the S.E.M. Data are from Figure 4A, white bars excluded (see legend Figure 4). Both snap and velcro contribute extremely significantly to the differences of extension between primers ($p<0.0001$, two-way ANOVA). (C) Proportion of sites in the snap and velcro categories for the human genome (hg19), the repeat-masked human genome (hg19 RM) and in polymorphic L1 insertion datasets (dbRIP, Solyom 2012 and Lee 2012). Note that the proportion of sites falling in each of the snap-velcro category is significantly different in the L1 insertion datasets (dbRIP, Solyom 2012 and Lee 2012) as compared to the proportions found in hg19 or repeatmasked hg19 (Chi-square test, two-tailed $P<0.0001$). (D) Human L1s preferentially insert into target sites with snap closed and velcro fastened. Potential (hg19 or hg19 RM) or real (dbRIP or Lee 2012) target sites with a recognizable EN target sequence were categorized based on their snap and velcro states. The frequency of each category for each dataset was calculated and divided by the frequency of the corresponding category in the reference genome hg19 (enrichment). For each dataset, enrichment was further normalized to the enrichment of the "open snap/loose velcro" category to evaluate the respective effect of the snap and/or velcro on L1 insertion site frequencies (normalized frequency). The raw data for panels C and D are compiled in Table S2.
doi:10.1371/journal.pgen.1003499.g005

of L1 insertions in each dataset (dbRIP, Solyom and Lee, Figure 5C). Strikingly, the proportion of L1 insertions in sites with closed snap and/or tightly fastened velcro was significantly increased as compared to their proportion in the human genome (Chi-square test, $p<0.0001$ for all insertion datasets). As an additional analysis, we calculated the frequency of each category in a given L1 insertion datasets as compared to their frequency in the human genome. We normalized this enrichment relative to the insertion sites with an open snap and a loosely fastened velcro strap. As shown in Figure 5D, L1 insertions are more frequent at sites with a closed snap or a tightly fastened velcro, and even more frequent at sites having both. Consistent with the *in vitro* data, given a snap status, insertions are more frequent at sites with a tightly fastened velcro than with a loosely fastened velcro. Other studies have previously reported that T-richness extends beyond four nucleotides upstream of the cleavage site [48,50]. Our analysis differs from these previous observations in that each position is not considered independently from the others. Altogether the distribution of polymorphic L1 insertions *in vivo* is consistent with the snap-velcro model at the genomic level, but it should also be stressed that, *in vivo*, other determinants are likely to influence L1 insertion profiles.

Extension of dsDNA by the L1 RNP

An alternative pathway of L1 integration uses preformed double-stranded DNA lesions instead of EN-mediated cleavage. To determine whether the L1 RNP is able to directly initiate reverse transcription at blunt DNA ends, we designed model hairpins ending with four or six Ts at their 3′ terminus (Figure 6A, primers H and H-ext). Notably, we used hairpins instead of two separate DNA strands to exclude the possibility that remaining free single-stranded primers could be extended (Figure 6A).

The expected start position of each extension product (+1), which depends on primer length (see Figure 6A), is indicated by a black dot on the left side of each lane. Although we can readily detect elongation of the single-stranded ext-$(dT)_{18}$ primer (Figure 6B, lane 2), no mL1-specific extension was observed with these blunt substrates (Figure 6B, compare lane 2 to 3–4). The radiolabeled molecules detected below the +1 of the reverse transcription (Figure 6B, between 40 and 56 nt and Figure 7B, below 40 nt) result from contaminating activities, which co-fractionate with the mL1 RNP in the sucrose cushion (see below for a detailed characterization). In addition, we asked whether the mL1 RNP could access and extend a stretch of 4 Ts embedded in a duplex DNA. No extension was observed when we used various hairpins with 3′ recessed ends ending with 4 Ts (Figure 6A, 5′TT-H, 5′GC-H, 5′CTGC-H and Figure 6B, compare lanes 5–7 to 12–14). Identical results were obtained with hL1 RNPs (Figure S4A).

Since L1 elements are believed to integrate into double-stranded genomic DNA and L1 RNPs can efficiently extend single-stranded oligonucleotides (see above), we reasoned that L1 RNPs might be able to prime DNA synthesis on double-stranded primers ending with a 3′ overhang. To test this hypothesis we designed model hairpins extended by a 3′ overhang of increasing size (Figure 7A, primers H_0 to H_6). In contrast to reactions performed with blunt or 3′-recessed hairpin substrates, initiation of mL1 reverse transcription is easily detected as soon as the 3′ overhang reaches a length of 6 nt, as shown by the mL1-specific ladder which appears above 50 bp (Figure 7B, compare lane 8 to 3–7 and 19). Increasing the length of the overhang to 8 nt slightly increases the levels of reverse transcription, which indicates that a 6 nt 3′ overhang is necessary and sufficient for efficient extension by the mL1 RNP. In the experiments using single-stranded substrates, we demonstrated that 4 matching bases at the 3′ end of the substrate are sufficient to prime reverse transcription at detectable levels. This is also true for 3′ overhang hairpins, since a hairpin with a 6- or 8-nucleotide 3′ overhang but ending with only 4 Ts is extended, although to lower levels than a similar single-stranded primer ending with 4Ts (Figure 7B, lanes 9–10 and Figure S2, lane 12). Identical results were obtained with hL1 RNPs (Figure S4B).

As mentioned above, incubation of L1 RNP fractions with hairpin primers and ^{32}P-dTTP results in labeled products, which are shorter than the expected +1 of the reverse transcription reaction (Figure 6B and Figure S4A, between 40 and 56 nt and Figure 7B and Figure S4B, below 40 nt). These products are also detected at similar levels with RT-defective L1 RNP preparations (Figure 6B, lanes 9–14 and Figure 7B, lanes 14–22) and with RNPs prepared from vector-transfected cells (data not shown), suggesting that they result from contaminating cellular activities, which co-fractionate with the L1 RNP in the sucrose cushion. To verify this hypothesis, we further purified the mL1 RNPs by immunoprecipitation using an antibody raised against the mORF1p protein (Figure 8A and 8B), and then we performed reverse transcription reactions on the beads. As a negative control, we performed the immunoprecipitation with the preimmune serum. First, we could directly detect the mL1 RT activity in the immunoprecipitated complex (Figure 8C, compare lanes 8 and 14), reinforcing the notion that the L1 RNA, ORF1p and ORF2p form a stable complex [18]. Second, the immunopurified mL1 RNP extends the H_6 hairpin primer with a 3′ overhang but not the blunt or 3′-recessed primers (Figure 8C, compare lanes 9–12 and 15–18). Third, the short products formed upon incubation with the sucrose cushion mL1 RNP preparation disappear if the mL1 RNP is further purified by immunoprecipitation (Figure 8C, compare lanes 3–6, dashed boxes, and 15–18). Altogether these observations confirm that the bands below the +1 are indeed nonspecific products resulting from cellular contaminating activities and that the ladder-like products above ~50 nt are *bona fide* L1 RNP reverse transcription products.

Based on these data we conclude that native L1 RNPs preferentially extend DNA substrates ending with at least 4 Ts and a 6-nt single-stranded 3′ overhang, but does not efficiently extend blunt or 3′-recessed double-stranded DNA substrates.

Discussion

Although L1 elements are responsible for a very large part of mammalian genomes and are an important source of genetic diversity and diseases [60,62–66], detailed molecular mechanisms of their replication remain poorly studied at the biochemical level. We have developed here a direct L1 extension assay (DLEA) to explore the impact of primer sequence and structure on reverse transcription initiation by native L1 RNPs (Figure 1 and Figure S1). The DLEA protocol differs from previous approaches [10,33,51,55,67] because it combines native L1 RNP purification from cell extracts, by sucrose cushion ultracentrifugation or immunopurification (Figure 8), with the direct detection of extension products. Since it does not require a PCR amplification step, the DLEA allows quantitative comparisons of priming efficiencies for a large variety of substrates with different sequences and structures. A limitation of this assay is the absence of sequence information on the product. Therefore we complemented DLEA data with LEAP amplification and sequencing.

By testing more than 65 different primers, including many that mimic *bona fide* L1 insertion sites recovered from cultured cells, we could define the rules of L1 reverse transcription initiation with an unprecedented resolution: (i) partial sequence complementarity

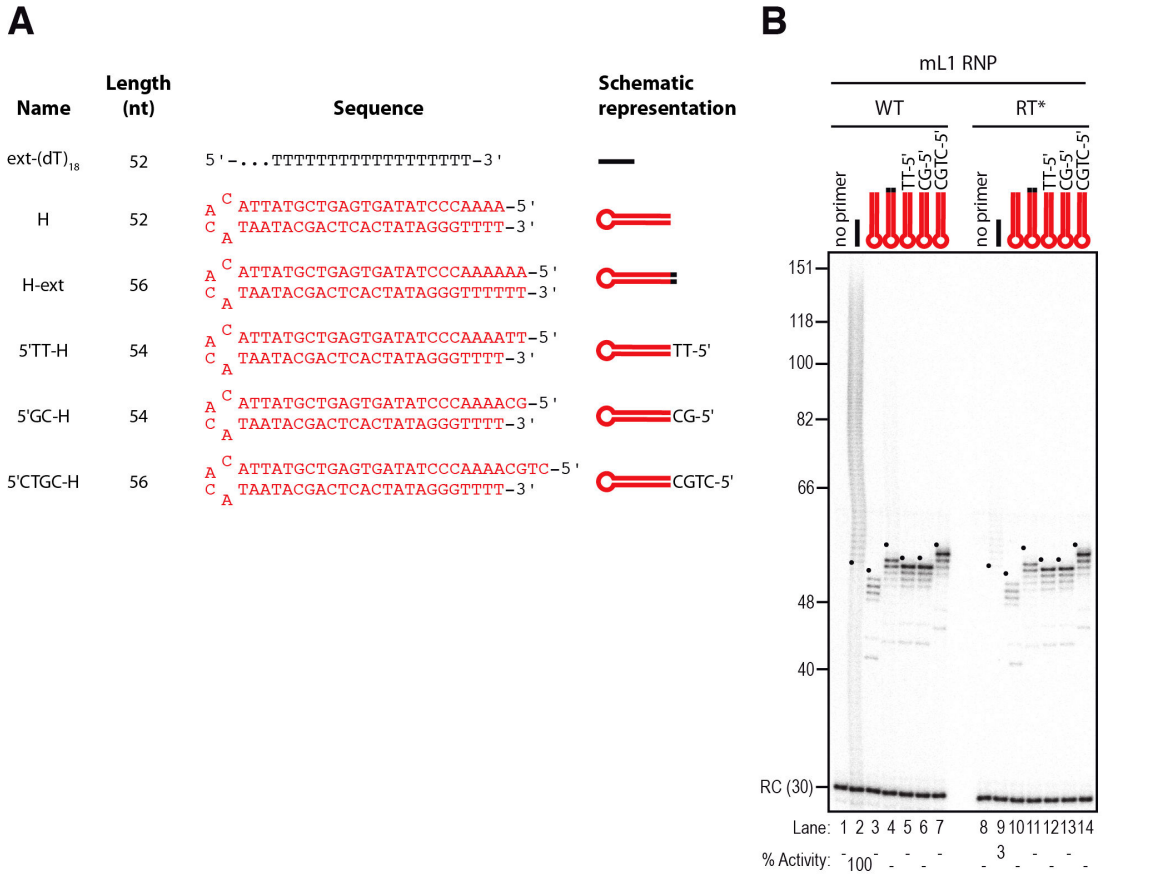


Figure 6. Double-stranded primers with blunt or 3′-recessed are not efficiently extended by mL1 RNPs. (A) Scheme of the primers used. (B) DLEA showing the absence of extension of double-stranded primers with blunt or 3′ recessed ends in the presence of α-^{32}P-dTTP. Note that the only products observed with hairpin primers (lanes 3–7) result from contaminating cellular activities (see text and Figure 8 for further characterization). RC denotes a 30 nt recovery control added after the reaction but before DNA purification. Quantification of primer extension (% Activity) was relative to levels of extension obtained with ext-$(dT)_{18}$ (lane 2). The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies. Results obtained with hL1 RNPs were identical and are shown in Figure S4.
doi:10.1371/journal.pgen.1003499.g006

between the 10 terminal nucleotides of the target site and the L1 RNA poly(A) tail impact reverse transcription initiation (Figure 2 and Figure S2, and Figure 4); (ii) four terminal Ts are sufficient to promote efficient extension of the target DNA (Figure 2 and Figure S2); (iii) the L1 RNP can tolerate a mismatch in the crucial last 4 nucleotides if it is compensated by an increased number of matching nucleotides upstream of these bases (Figure 2, Figure S2 and Figure 4); (iv) the preferred terminal base is T>C>A>G (Figure 3). Based on these quantitative data, we propose a 'snap-velcro' model to illustrate the high level of flexibility of the L1 RNP toward primer use (Figure 5A). This model identifies two distinct regions in the cleaved target DNA: (i) the terminal 3′ four nucleotides (snap), which correspond to the EN recognition site, and are also essential to reverse transcription initiation; and (ii) the upstream six nucleotides (velcro), which enhance reverse transcription efficiency and compensate potential mismatches in the snap region, when rich in Ts.

Studying the properties of L1 RNPs *in vitro* provides detailed molecular insights into specific steps of the retrotransposition process. This is a useful complement to retrotransposition cellular assays, which offer a more global view of this mechanism. Nevertheless, a number of differences between the *in vitro* and *in vivo* situations, and between endogenously and ectopically expressed L1, should be emphasized. First, reverse transcription initiation is uncoupled from the cleavage of the target DNA, in primer extension assays such as LEAP or DLEA. Thus, we cannot completely exclude that L1 RNPs would utilize a different priming mechanism in the context of a L1 TPRT reaction. Likewise, it is possible that the detected activity results from a minor fraction of the RNPs, which can only extend exogenous primers. This situation is reminiscent of L1 reverse transcription initiation at existing DNA lesions as hypothesized for EN-independent integration events [51,68–70]. Second, due to read-through transcription, L1 RNAs expressed from endogenous loci sometimes contain a first poly(rA) sequence, which is transcribed by RNA-Polymerase II from the L1 poly(dA) tail and can occasionally be imperfect, followed by a downstream genomic sequence, and ending with a perfect poly(rA) tail generated by Poly(A)-Polymerase [71,72]. Theoretically, alternative nucleotides present in such internal and imperfect poly(A) sequences could match

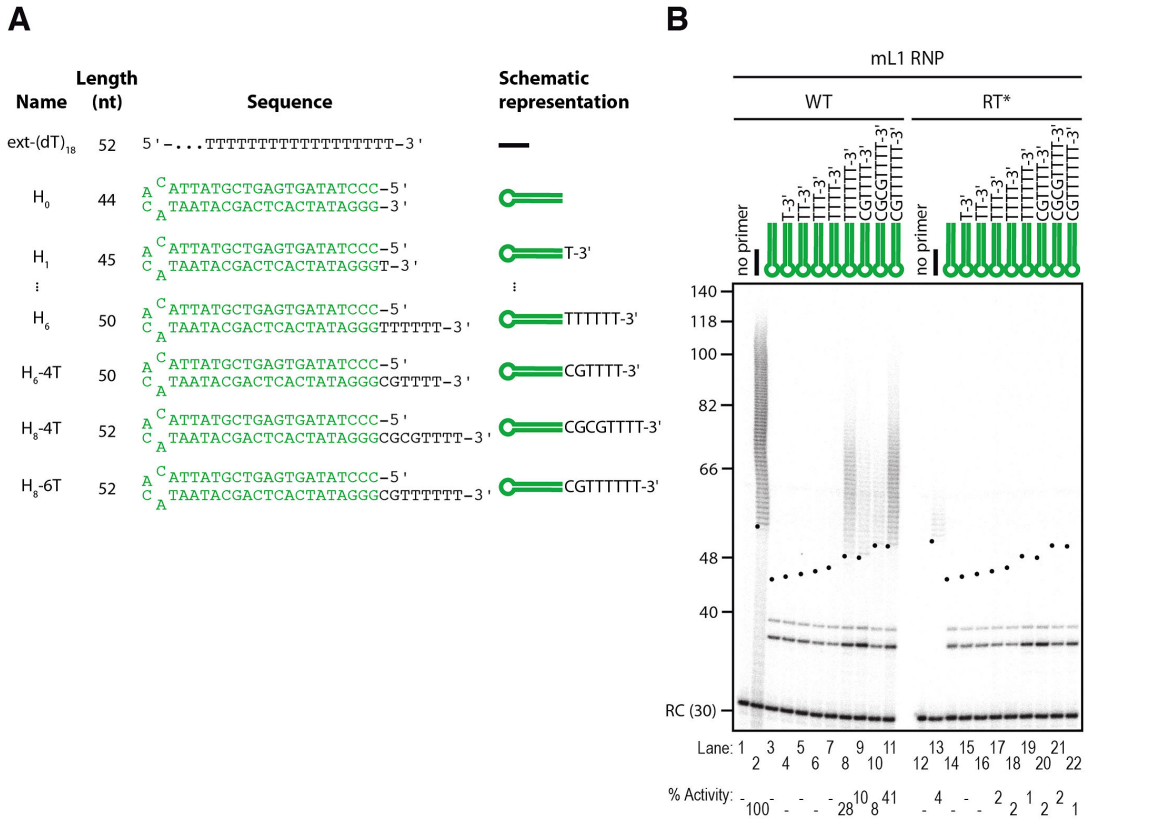


Figure 7. The L1 RNP preferentially extends double-stranded DNA with a 3′ overhang. (A) Scheme of the primers used. (B) Extension by mL1 RNPs of double-stranded primers ending with a 3′ overhang in the presence of α-^{32}P-dTTP. Note that the doublet below 40 nt observed in lanes 3–11 and 14–22 results from contaminating cellular activities (see text and Figure 8 for further characterization). RC denotes a 30 nt recovery control added after the reaction but before DNA purification. Quantification of primer extension (% Activity) was relative to levels of extension obtained with ext-(dT)$_{18}$ (lane 2). The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies. Results obtained with hL1 RNPs were identical and are shown in Figure S4.
doi:10.1371/journal.pgen.1003499.g007

perfectly to degenerate endonuclease sites, such that mismatches between primer and template would be less frequent. In contrast, L1 RNA polyadenylation in ectopically expressed constructs is generally driven by the strong SV40 polyadenylation sequence and by Poly(A)-Polymerase leading to perfect poly(rA) tails. Finally, our data suggest that target site choice is dictated not only by the specificity of the first EN cleavage, but also by the efficiency of RT priming after nicking. Interestingly, an engineered L1 endonuclease with relaxed sequence specificity *in vitro* has been described [73]. *In vivo*, L1 elements carrying this endonuclease variant still integrate in extended T-rich sequences, which shows that additional factors other than the EN specificity contribute to L1 insertion profile *in vivo*. Our data suggest that primer-template complementarity might be one of these factors, by promoting the initiation of reverse transcription, but it is also very likely that additional partners or inhibitors influence L1 targeting *in vivo*, modulating or relaxing EN or RT specificity. Indeed, L1 insertions occasionally take place at sites that do not strictly follow the rules described here (Figure 5C, and [46,47,49,51,69]), suggesting that primers for which we cannot detect extension by DLEA might actually be L1 substrates. From our data we can only conclude that they are extended *in vitro* at least 10–20 fold less efficiently than the best target sites that were used as references in our assays.

In contrast to the L1 RNP, R2 reverse transcriptase does not require sequence matching to prime DNA synthesis and does not require a 3′ overhang [74]. This might be related to the fact that specific structures in the R2 RNA allow the R2 RT to position and guide the exact start of reverse transcription at the cleavage site [36]. In this configuration, primer-template annealing is no longer a requirement to position the primer at the end of the template. Biochemical studies with non-LTR retrotransposon RT from other clades will be necessary to determine, which of these two situations is the rule and the exception.

The current model of L1 retrotransposition, which has been largely inspired by studies on the R2 element, starts with a nick in the target DNA followed by the extension of this nick. Our data indicate that extension by the L1 RNP is efficient on single-stranded DNA substrates, but inefficient when the 3′ OH is embedded in duplex DNA, either at a blunt end or at a 3′ recessed end (Figure 6B and Figure S4A). In contrast, it efficiently initiates reverse transcription on double-stranded DNA molecules ending with a 3′ single-stranded overhang (Figure 7B and Figure S4B). Thus, our results suggest an additional step in the retrotranspo-

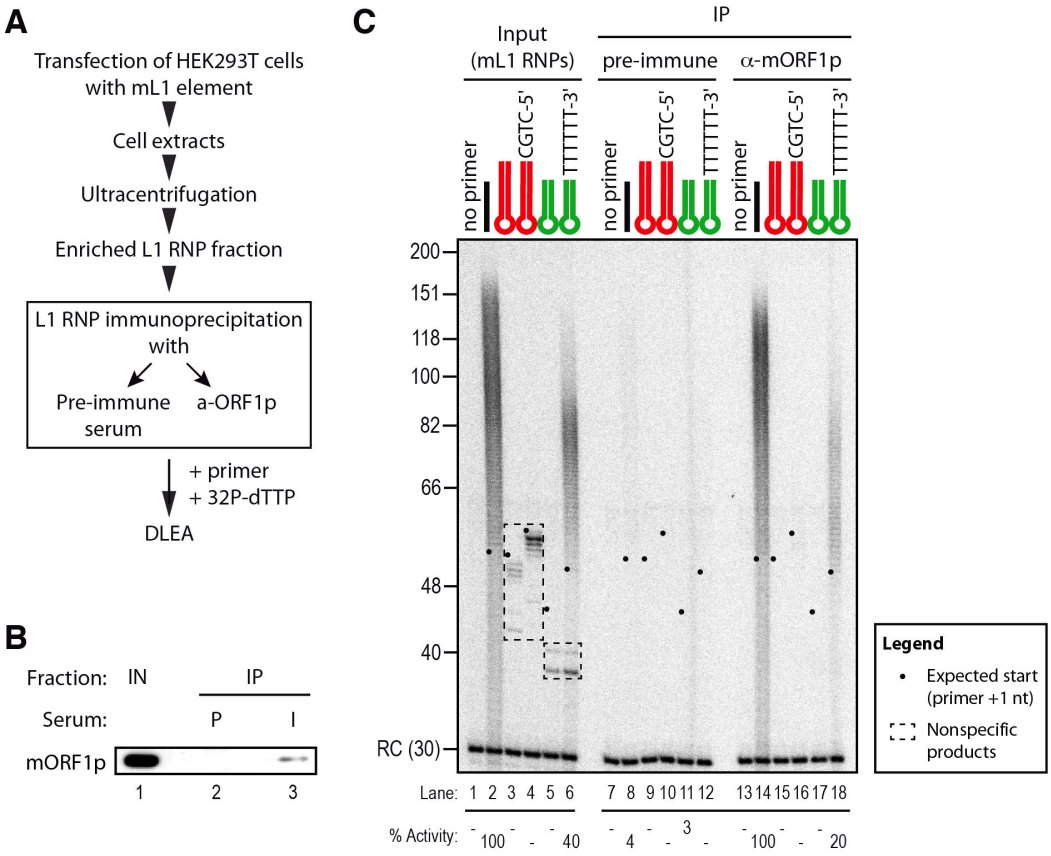


Figure 8. Priming of reverse transcription by immunopurified mL1 RNP. (A) Outline of the experimental procedure. (B) Immunoblot of the mL1 RNP immunoprecipitation (IP). IPs were performed on mL1 RNP preparations (Input, IN, lane 1) using preimmune (P, lane 2) or mORF1p-immune (I, lane 3) sera. Blot was probed with the anti-mORF1p serum. (C) Primer extension assay performed with mL1 RNPs (lanes 1–6), beads of the preimmune serum IP (lanes 7–12) or beads from the anti-mORF1p serum IP (lane 13–18). Note that the products suspected to be nonspecific (dashed boxes, lanes 3–6) indeed result from contaminating cellular activities and disappear upon immunoprecipitation, while the specific reverse transcription products are still observed (lanes 14 and 18). RC denotes a 30 nt recovery control added after the reaction but before DNA purification. Quantification of primer extension (% Activity) was relative to levels of extension obtained with ext-$(dT)_{18}$ (lane 2 for Input and lane 14 for IP). The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies.
doi:10.1371/journal.pgen.1003499.g008

sition process, which generates a single-stranded 3′ end from a blunt end or from a nick to allow L1 reverse transcription. We envisage two ways in which this 3′ overhang could be established. In the first model, the L1 endonuclease directly generates a double-strand break with staggered cuts instead of acting sequentially on one strand and then on the other strand only after minus strand cDNA synthesis. Consistently, recombinant L1 endonuclease can linearize plasmid DNA *in vitro* [27] and ectopic L1 expression results in the activation of a DNA damage response in cultured cells [75,76]. In the second model, an unidentified machinery could promote unwinding of the nicked DNA or permit strand-exchange between the duplex DNA and the RNA moiety of the L1 RNP. The ORF1p protein has been proposed to play such a role through its nucleic acid chaperone activity [20,24]. Indeed, nucleic acid chaperone activities promote reverse transcription in retroviruses and LTR-retrotransposons through several mechanisms, including primer annealing to the template RNA [77–80]. All the experiments described here use native L1 RNP preparations, which contain ORF1p (Figure 1 and Figure 8). However, in our experimental conditions, we were unable to detect extension of blunt or 3′ recessed double-stranded substrates. Thus, if such a DNA remodeling machinery is involved, it has to be of cellular origin. Nevertheless, it should be noted that, in primer extension assays, as performed in LEAP or DLEA experiments, the initiation of reverse transcription is uncoupled from the cleavage of the target DNA, in contrast to the TPRT process. Thus, we cannot completely exclude that the L1 RNP would utilize a different priming mechanism in the context of a L1 TPRT reaction.

The requirement of a 3′ overhang could also be relevant to alternative L1 integration pathways. Indeed, L1s can initiate reverse transcription at preformed DNA lesions or at telomeric ends and thus insert into the genome independently of their EN activity [51,68–70]. EN-independent retrotransposition was only observed in cell lines deficient in the nonhomologous end-joining (NHEJ) pathway [68]. Interestingly, binding of NHEJ components to DNA ends interferes with end resection [81]. As a result of this competition, end resection (the first step of homologous recombination) is increased in NHEJ-deficient cell lines. Thus, we

speculate that EN-independent retrotransposition might require the 5′ to 3′ end resection step, which initiates HR, to generate a 3′ overhang suitable for L1 reverse transcription initiation. The link between end resection factors (such as the MRN complex, CtIP, Exo1, BLM, Dna2, etc.) and the ability of L1 to engage in EN-independent insertions will be an important direction for future studies. Similarly, the L1 RNP is also able to prime cDNA synthesis at dysfunctional telomeres in NHEJ-deficient hamster cells [51,69]. Telomeres end with a 3′ overhang [82,83], the formation of which is highly regulated and involves a specialized set of factors [84]. Telomeres can also be extended by a specialized cellular RNP with reverse transcriptase activity, called telomerase [85,86]. Like L1, telomerase requires a 3′ single-stranded overhang to extend double-stranded DNA [87]. Thus our observations reinforce the notion that these two endogenous reverse transcriptases, which are evolutionary related [88–90], share common mechanistic properties [51].

In conclusion, our data demonstrate that partial sequence complementarity between the target site and the L1 RNA facilitates L1 reverse transcription priming and highlight the flexibility of the L1 RT. Interestingly, EN cleavage and RT priming appear to target the same TTTT sequence, suggesting that these two L1 biochemical activities have co-evolved. We speculate that their exceptional flexibility has participated in the evolutionary success of the L1 family and in its wide spread distribution within mammalian genomes.

Materials and Methods

Plasmids and oligonucleotides

Plasmids JM101/L1.3 and JM105/L1.3 respectively contain WT and RT-mutated (D702A) versions of the human L1.3 element in a pCEP4 backbone (a kind gift of N. Gilbert) [9]. Plasmid pWA121 contains a codon-optimized version of the mouse L1spa element in a pCEP4-Puro backbone (a kind gift of J. D. Boeke) [91]. A fragment containing mORF2p was amplified by PCR from pWA121 using oligonucleotides LOU266 and LOU267. The purified attB PCR product was cloned into pDONR207 using BP Clonase II under the manufacturer's conditions (Gateway system, Life Technologies) to obtain plasmid pVan239. A point mutation in the RT domain (D709A) was introduced in this construct using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies) and the DNA primer pair LOU419-LOU420 to generate pVan330 (mORFeus RT*). The RT* mutation introduces a new SacII restriction site in ORF2, allowing quick screening of the mutation. The latter was confirmed by sequencing. A SdaI-NruI DNA fragment containing part of ORF2p from this entry clone was inserted back into the original pWA121 plasmid digested by the same enzymes. A full list of the oligonucleotides used in this study is provided as Table S1.

Antibodies

Peptides corresponding to the C-termini of mouse (N-CNQYKNGNNALEKTRR-C) or human (N-CERNNRYQPL-QNHAKM-C) ORF1p were synthesized and coupled to the KLH protein as a carrier. The first cysteine (underlined) is not present in the ORF1p sequence but was added for the coupling reaction with the carrier protein. KLH-coupled peptides were used to immunize rabbits (Eurogentec). For immunoblotting the mORF1p antiserum (SE-0560), the hORF1p antiserum (SE-6798), and the S6 protein antibody (Cell signaling, #2217) were used at a dilution of 1:2000.

Oligonucleotide purification

One hundred micrograms of each lyophilized oligonucleotide was dissolved in 10 µl of 98% deionized formamide, 1 mM EDTA, 0.01% (w/v) xylene cyanol and 0.01% (w/v) bromophenol blue and resolved in 10% polyacrylamide-urea denaturing gels. Full length oligonucleotides were visualized by UV shadowing, excised from the gel and eluted overnight at 37°C in 0.3 M sodium acetate, 0.1% SDS and 10 mM $MgCl_2$. Eluted oligonucleotides were precipitated with ice-cold ethanol (3v). After centrifugation for 30 min at 4°C at 16'000 g, the pellets were washed with 70% ethanol, air-dried and dissolved in 10 mM Tris-HCl pH 8.0, 1 mM EDTA.

Production of L1 RNPs in human cells

L1 RNPs were produced in HEK293T cells grown in Dulbecco's Modified Eagle Medium (DMEM, Life Technologies) containing 2 mM L-Glutamine, 4500 mg/L D-Glucose, 1 mM Sodium Pyruvate, 10% (v/v) fetal bovine serum (Life Technologies) and 100 units/mL penicillin/streptomycin (Life Technologies). Cells were plated at 3×10^6 cells per 10 cm Petri dish. Twenty-four hours after plating, the cells were transfected with 24 µg of plasmid DNA (see plasmids above) per dish using the calcium phosphate method. Growth medium was changed 5 hours later. One day post-transfection, cells were split into two plates in growth medium supplemented with 1.5 µg/mL puromycin (mORFeus, Life Technologies) or 100 µg/mL hygromycin (L1.3, Life Technologies). Cells were collected 4 days post-transfection by trypsinization, pooled and washed in PBS. Cell pellets were lysed in 500 µL of CHAPS lysis buffer (10 mM Tris-HCl [pH 7.5], 1 mM $MgCl_2$, 1 mM EGTA, 0.5% (w/v) CHAPS, 10% (v/v) Glycerol, supplemented before use with Complete EDTA-free protease inhibitors cocktail (Roche) and 1 mM DTT). After incubation at 4°C for 15 min, cell debris was removed by spinning down extracts at 4°C for 10 min at 16'000 g. Supernatants were transferred to clean tubes and 500 µL of lysis buffer were added to each of them.

Partial purification of L1 RNP by sucrose cushion and ultracentrifugation

L1 RNPs were prepared as previously described [10]. In brief, a sucrose cushion was prepared with 8.5% and 17% (w/v) sucrose in 20 mM Tris-HCl [pH 7.5], 80 mM NaCl, 5 mM $MgCl_2$, 1 mM DTT and Complete EDTA-free protease inhibitors cocktail (Roche). For each sucrose cushion, 1 mL of cell lysates, prepared as described above, was used. Samples were centrifuged for 2 h at 178'000 g at 4°C and the pelleted material was resuspended in 100 µL H_2O. Total protein concentration was determined by Bradford assay (Biorad). The samples were diluted in 50% (v/v) glycerol, quick frozen in liquid nitrogen and stored at −80°C until use.

Immunoprecipitation of L1 RNP

Protein A-Sepharose beads (Sigma) were blocked overnight at 4°C in PBS containing 0.5 mg/mL of bovine serum albumin (BSA) and washed twice in 1 mL of IP buffer (10 mM Tris-HCl [pH 7.5], 150 mM NaCl). Eight microliters of preimmune or anti-mORF1p serum were bound to 70 µl of blocked beads for 3 h at 4°C. For each immunoprecipitation, 200 µL of L1 RNPs (2 µg/µL) were diluted 1:1 (v/v) in IP buffer. The RNPs were precleared with blocked beads for 1 h at 4°C and incubated for 3 h at 4°C with antibody-bound beads on a rotating wheel. After 4 washes in IP buffer, the bead slurry was split equally into 7 tubes (6 for RT reactions and 1 for immunoblotting). Beads were pelleted for

5 min at 4°C at 750 g, supernatants were removed and the RT reaction mixture was directly added to the beads (see below).

Direct L1 extension assay (DLEA)

Reverse transcriptase assays were carried out for 4 min at 37°C in 25 µL reactions containing 2 µg of RNPs, 400 nM of primer, 50 mM Tris-HCl [pH 7.5], 50 mM KCl, 5 mM $MgCl_2$, 10 mM DTT, 0.05% (v/v) Tween-20 and 10 µCi of α-^{32}P-dTTP (3000 Ci/mmol, PerkinElmer). In reactions using the Avian Myeloblastosis Virus RT (AMV RT, Promega), the RNPs were replaced by 0.04 U of AMV RT and 250 ng of poly(rA) template (Roche). Reactions were stopped by the addition of 8.3 mM EDTA and 0.83% SDS final. Trace amounts of a ^{32}P-labelled 14- or 30-mer DNA oligonucleotide were added as recovery control (noted RC (14) or RC (30) in the figures). Products were purified by phenol-chloroform extraction and ethanol precipitation with 10 µg of glycogen as a carrier and 0.1 mM sodium acetate [pH 5.2]. DNA pellets were resuspended in 98% deionized formamide containing 10 mM EDTA, 0.02% (w/v) xylene cyanol and 0.02% (w/v) bromophenol blue, heated to 95°C for 5 min, and analyzed on 13% polyacrylamide-urea sequencing gels. After drying, gels were exposed to a PhosphorImager screen.

For primers used in Figure 4, we first resolved the products on sequencing gels to verify that the profiles of the products were similar to those obtained with other linear oligonucleotides and that nonspecific products were not generated. In a second time, to facilitate quantification of a large number of reactions performed in parallel, we spotted 5 µL of each reaction onto DE-81 paper immediately after the 4 min incubation, in triplicate. DE-81 paper is an ion exchange paper, which retains the incorporated nucleotides, but not the free dNTPs. Papers were next washed 5 times with 200 mL of 2x saline-sodium citrate (SSC) solution and exposed to a PhosphorImager screen. We tested the complete set of primers three times.

For gel or spot quantification, the reaction without primer obtained with a given RNP preparation was used as background and was subtracted from the reaction with primers. Only the signal above the primer size was quantified for the hairpin oligonucleotides.

RNase treatment and reverse transcriptase inhibitors

To determine whether ^{32}P incorporation was RNase sensitive (Figure S1A), we incubated reaction mixes in the presence of 30 µg of RNase A and 150 U of RNase I (New England BioLabs), or of 40 U of RNasin (Promega) as a negative control, for 1 h at 37°C before adding ^{32}P-dTTP and primer. RT inhibitors (AZT and d4T, also known as Stavudin) as triphosphate derivatives were obtained from Biocentric. They were added to reactions at a final concentration of 10 µM (Figure S1B).

L1 element amplification protocol (LEAP)

LEAP was performed as previously described [10] with only minor modifications. Briefly, L1 reverse transcription was carried out for 1 h at 37°C in 50 µL reactions containing 0.75 µg L1 RNP (50% (v/v) glycerol), 50 mM Tris-HCl [pH 7.5], 50 mM KCl, 10 mM DTT, 5 mM $MgCl_2$, 0.05% (v/v) Tween-20, 20 U RNasin (Promega), 200 µM dNTP, and 0.4 µM LEAP primer. Eventually, unextended primers were eliminated through an S-400HR size-exclusion spin column (GE Healthcare). Reverse transcription products (1 µL of the LEAP reaction) were PCR-amplified in 50 µL reactions containing 1 U of Platinum Taq DNA Polymerase (Life technologies), 0.2 µM of primers LOU851 and LOU312, 200 µM dNTP, 3 mM $MgCl_2$ in the Platinum Taq buffer. A first step at 94°C for 2 min was followed by 35 cycles of [30 s at 94°C, 30 s at 60°C and 30 s at 72°C]. The final extension was at 72°C for 5 min. PCR products were analyzed by 2% agarose gel electrophoresis in 1x TBE. Gels were stained by SYBR Safe (Life technologies) or ethidium bromide. LEAP products were gel-purified with a gel extraction kit (Macherey Nagel) and cloned into the pGEM-T-easy vector (Promega), according to manufacturer's protocol. Clones from isolated colonies were sequenced by GATC. Regions with low quality (Phred<Q20) were trimmed or filtered out using Geneious 5.

RNA isolation and conventional RT–PCR

Total RNA was extracted from 30 µg of L1 RNP using TRIzol extraction (Molecular Research Center Inc) following the manufacturer's instruction. RNA was resuspended in 20 µL of milliQ water and quantified by Nanodrop. One microgram of RNA was digested by 1 U of RNase-free RQ1 DNase (Promega) in 10 µL reaction in the manufacturer's buffer at 37°C for 30 min. DNase was heat-inactivated for 10 min at 65°C. Then, cDNA synthesis was performed at 50°C for 1 h in 20 µL reactions containing 6 µL of the DNase reaction, 200 U of SuperScript III Reverse Transcriptase (Life technologies), 500 µM dNTP, 50 pmol of RACE primer, 40 U RNAseOUT (Life technologies), 50 mM Tris-HCl [pH 8.0], 75 mM KCl, 3 mM $MgCl_2$ and 5 mM DTT. Primer pairs used for PCR were LOU851/LOU312 (mOrfeus or L1.3) or LOU852/LOU312 (GAPDH). PCR products were resolved by 2% agarose gel electrophoresis in 1x TBE.

T-density and position-weighted T-density

The *T-density* is calculated by dividing the number of Ts in the oligonucleotide by the length of the oligonucleotide. The *position-weighted T-density* gives more weight to Ts which are close the 3′ extremity of the primer. The weight is inversely proportional to the distance from the 3′ end.

For example:

Primer LOU519 has a *position-weighted T-count* equal to:

$$1+(1/2)+(1/3)+(1/4)+(1/7)=2.23$$

Primer LOU541 has a *position-weighted T-count* equal to:

$$1+(1/2)+(1/3)+\ldots+(1/18)+(1/19)+(1/20)=3.60$$

The *position-weighted T-density* of a given primer is calculated by dividing the *position-weighted T-count* of this primer to the maximum *position-weighted T-count*. Thus the *position-weighted T-density* of LOU519 is equal to 2.23/3.60 = 0.62 and the position-weighted T-density of LOU541 is equal to 3.60/3.60 = 1

Snap and velcro definitions

The snap is considered open if the 4 terminal nucleotides contain a non-T nucleotides and closed if the last four nucleotides are 4 Ts. We calculated a *position-weighted T-count* for the upstream 6 nucleotides (velcro region) and we divided it by the maximum value (1/5)+(1/6)+…+(1/10) = 0.84563492 to obtain the velcro *position-weighted T-density*. We consider a velcro as fastened if its *position-weighted T-density* is ≥0.5 (half of the maximum) and opened otherwise.

Analysis of snap/velcro category enrichment in genomic datasets

All putative integration sites with a perfect or degenerate EN recognition sequence (from 3′ to 5′, R/TTTT, R/VTTT, R/

TVTT, R/TTVT, R/TTTV) were recovered from both strands of the reference human genome (hg19) or from its repeatmasked version (hg19 RM). For each putative EN site, snap and velcro status were defined as described above. The C++ program used to achieve this task is available in Protocol S1. Polymorphic L1 insertions were extracted from dbRIP [59] or from cancer genome whole-genome sequences [60,61]. Only insertion sites with an identifiable EN recognition site as defined above were kept for the analysis. This filtering step was necessary to eliminate internal initiation events most likely related to EN-independent insertions or other forms of structural variation and insertion sites which position was not precise at nucleotide resolution. Raw data are provided in Table S2. For each dataset, we calculated the frequency of each category and we normalized first to hg19 count and second to the "open snap/tightly fastened velcro" category to evaluate the effect of a closed snap and/or velcro. We compared observed (polymorphic L1 insertions) and expected (hg19) frequencies by Chi-squared test. We used the Graphpad Prism 6.00 software for Mac for all statistical analyses.

Supporting Information

Figure S1 Additional characterization of the L1 RNP RT activity by DLEA. (A) RNA-dependent DNA polymerase activity of L1 RNPs. Murine L1 RNPs were incubated for 1 h at 37°C in the presence (lane 3) or in the absence (lane 4) of RNases before the start of the reaction. (B) RT inhibitors prevent primer extension by L1 RNPs. Reactions were performed with mL1 RNPs in the presence of thymidine analogs (10 μM of azidothymidine triphosphate AZTTP, denoted by A, lane 3; 10 μM of 2,3-didehydro-3-deoxythymidine triphosphate d4TTP, denoted by D, lane 4), or in the presence of water as a negative control (lane 2). (C) Time-course of $(dT)_{18}$ primer extension by hL1 RNP. (D) Formation of long cDNA species upon addition of all four dNTPs. Reactions were performed with hL1 RNPs in presence of α-^{32}P-dTTP and a $(dT)_{18}$ primer, with (lanes 3 & 6) or without (lanes 1–2 & 4–5) cold dATP, dCTP and dGTP (dVTP, IUPAC nomenclature).
(TIF)

Figure S2 The murine L1 RNP preferentially extends primers ending with at least 4 Ts. DLEA showing the extension of single-stranded primers by mL1 RNPs in the presence of α-^{32}P-dTTP. RC denotes a 14 nt recovery control added after the reaction but before DNA purification. The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies. Quantification of primer extension (% Activity) was relative to levels of extension obtained with oligo$(dT)_{18}$. Primers are identical to Figure 2.
(TIF)

Figure S3 LEAP with hL1 RNPs and mismatched primers. (A) Primers with terminal mismatches. LEAP was performed with RNPs prepared from hL1-transfected cells (top panel), from vector-transfected cells (middle panel), or without RNPs (bottom panel). Primers are identical to those used in Figure 2, except that they have a 5′ extension to anchor the PCR (see Table S1 for sequence). (B) Primers mimicking L1 integration sites. LEAP was performed with RNPs prepared from hL1-transfected cells (top panel), from vector-transfected cells (middle panel), or without RNPs (bottom panel). Primers are identical to those used in Figure 4, except that they have a 5′ extension to anchor the PCR (see Table S1 for sequence). (C) LEAP products from (A) and (B) were gel purified, cloned and sequenced. For each oligonucleotide, the top sequence and number of clones correspond to the extension of unprocessed primer, whereas other sequences correspond to the extension of processed primers.
(TIF)

Figure S4 Human L1 RNPs preferentially extends double-stranded DNA with a 3′ overhang. (A) Absence of extension by hL1 RNPs of double-stranded primers with blunt or 3′-recessed end in the presence of α-^{32}P-dTTP. Note that the products observed with hairpin primers (lanes 3–7) result from contaminating cellular activities (see main text and Figure 8). (B) Extension by hL1 RNPs of double-stranded primers ending with a 3′ overhang in the presence of α-^{32}P-dTTP. Note that the doublet below 40 nt observed in lanes 3–11 and 14–22 results from contaminating cellular activities (see text and Figure 8 for further characterization). RC denotes a 30 nt recovery control added after the reaction but before DNA purification. The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies. Results obtained with mL1 RNPs were identical and are shown in Figure 6, Figure 7, Figure 8.
(TIF)

Protocol S1 Source code of the software used to find putative endonuclease sites in the human genome and to calculate their associated snap/velcro scores.
(GZ)

Table S1 List of oligonucleotides used in this study.
(XLSX)

Table S2 Data used to calculate genomic enrichment of L1 insertions depending on the snap-velcro status of the target. The table sheets are the following: (hg19) For each potential L1 EN target site present in hg19, the snap status was defined and the position-weighted A density was calculated. Sites with position-weighted A density equal to or above 0.5 were considered as having a closed velcro strap. (hg19 RM) Same as above but with a repeatmasked hg19 reference genome. (dbRIP sequences) L1HS dbRIP entries used in Figure 5C and 5C and their snap/velcro status. (dbRIP counts) Number of dbRIP entries in each category. (dbRIP weblogo) Weblogo of the junction sequence (−2/+10) for dbRIP entries. (Lee2012 sequences) L1HS somatic insertions in cancer used in Figure 5C and 5C and their snap/velcro status. (Lee2012 counts) Number of L1HS somatic insertions in each category. (Lee2012 weblogo) Weblogo of the junction sequence (−2/+10) for Lee2012 entries. (Solyom2012 sequences) L1HS somatic insertions in colon cancer used in Figure 5C and 5C and their snap/velcro status. (Solyom2012 counts) Number of L1HS somatic insertions in each category. (Solyom2012 weblogo) Weblogo of the junction sequence (−2/+10) for Solyom2012 entries.
(XLSX)

Acknowledgments

We thank T. Ohlmann for hosting us before our move to University of Nice-Sophia Antipolis, E. van Obberghen and N. Gautier for giving us access to a radioelement manipulation room during the transition, M.-J. Giraud-Panis for managing the radioelement manipulation room, and IRCAN Genomics Core Facility for providing access to PhosphorImager and Nanodrop. We are grateful to Michael Chang and Chrysa Latrick for critical reading of the manuscript and to anonymous reviewers for helpful suggestions. We thank Nicolas Gilbert, John V. Moran, and Jef D. Boeke for plasmids and/or protocols, and Aurore-Cécile Valfort for generating the mouse Orfeus RT point mutation. We are grateful to Minzhi Xu-Olivetti for artwork.

Author Contributions

Conceived and designed the experiments: CM MK SV GC. Performed the experiments: CM MK SV. Analyzed the data: CM MK SV AAM GC. Contributed with preliminary experiments: CG J-LD. Contributed reagents/materials/analysis tools: AAM. Wrote the paper: GC.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
2. Goodier JL, Kazazian HH (2008) Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell 135: 23–35.
3. Belancio VP, Hedges DJ, Deininger P (2008) Mammalian non-LTR retrotransposons: for better or worse, in sickness and in health. Genome Res 18: 343–358.
4. Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10: 691–703.
5. O'Donnell KA, Burns KH (2010) Mobilizing diversity: transposable element insertions in genetic variation and disease. Mob DNA 1: 21.
6. Beck CR, Garcia-Perez JL, Badge RM, Moran JV (2011) LINE-1 Elements in Structural Variation and Disease. Annu Rev Genomics Hum Genet 12: 187–215.
7. Swergold GD (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. Mol Cell Biol 10: 6718–6729.
8. Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24: 363–367.
9. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, et al (2001) Human L1 retrotransposition: cis preference versus trans complementation. Mol Cell Biol 21: 1429–1439.
10. Kulpa DA, Moran JV (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol 13: 655–660.
11. Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran JV (2006) Unconventional translation of mammalian LINE-1 retrotransposons. Genes Dev 20: 210–224.
12. Martin SL (1991) Ribonucleoprotein particles with LINE-1 RNA in mouse embryonal carcinoma cells. Mol Cell Biol 11: 4804–4807.
13. Hohjoh H, Singer MF (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. EMBO J 15: 630–639.
14. Kolosha VO, Martin SL (1997) In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. Proc Natl Acad Sci U S A 94: 10155–10160.
15. 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.
16. Kulpa DA, Moran JV (2005) Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. Hum Mol Genet 14: 3237–3248.
17. Goodier JL, Zhang L, 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.
18. Doucet AJ, Hulme AE, Sahinovic E, Kulpa DA, Moldovan JB, et al (2010) Characterization of LINE-1 ribonucleoprotein particles. PLoS Genet 6: e1001150. doi:10.1371/journal.pgen.1001150
19. Goodier JL, Mandal PK, Zhang L, Kazazian HH (2010) Discrete subcellular partitioning of human retrotransposon RNAs despite a common mechanism of genome insertion. Hum Mol Genet 19: 1712–1725.
20. 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.
21. 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.
22. Martin SL, Branciforte D, Keller D, Bain DL (2003) Trimeric structure for an essential protein in L1 retrotransposition. Proc Natl Acad Sci U S A 100: 13815–13820.
23. 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.
24. Martin SL (2010) Nucleic acid chaperone properties of ORF1p from the non-LTR retrotransposon, LINE-1. RNA Biol 7: 67–72.
25. 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.
26. Mathias SL, Scott AF, Kazazian HH, Boeke JD, Gabriel A (1991) Reverse transcriptase encoded by a human transposable element. Science 254: 1808–1810.
27. Feng Q, Moran JV, Kazazian HH, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87: 905–916.
28. Moran JV, Holmes SE, Naas TP, DeBerardinis RJ, Boeke JD, Kazazian HH (1996) High frequency retrotransposition in cultured mammalian cells. Cell 87: 917–927.
29. Martin SL, Cruceanu M, Branciforte D, Wai-Lun Li P, Kwok SC, et al (2005) LINE-1 retrotransposition requires the nucleic acid chaperone activity of the ORF1 protein. J Mol Biol 348: 549–561.
30. Kubo S, Seleme MC, Soifer HS, Perez JL, Moran JV, et al (2006) L1 retrotransposition in nondividing and primary human somatic cells. Proc Natl Acad Sci U S A 103: 8036–8041.
31. 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.
32. Xiong YE, Eickbush TH (1988) Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. Cell 55: 235–246.
33. Cost GJ, Feng Q, Jacquier A, Boeke JD (2002) Human L1 element target-primed reverse transcription in vitro. EMBO J 21: 5899–5910.
34. Christensen SM, Ye J, Eickbush TH (2006) RNA from the 5′ end of the R2 retrotransposon controls R2 protein binding to and cleavage of its DNA target site. Proc Natl Acad Sci U S A 103: 17602–17607.
35. Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134: 221–234.
36. 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.
37. 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.
38. Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements. Mol Biol Evol 16: 793–805.
39. Kajikawa M, Okada N (2002) LINEs mobilize SINEs in the eel through a shared 3′ sequence. Cell 111: 433–444.
40. Osanai M, Takahashi H, Kojima KK, Hamada M, Fujiwara H (2004) Essential motifs in the 3′ untranslated region required for retrotransposition and the precise start of reverse transcription in non-long-terminal-repeat retrotransposon SART1. Mol Cell Biol 24: 7902–7913.
41. Anzai T, Osanai M, Hamada M, Fujiwara H (2005) Functional roles of 3′-terminal structures of template RNA during in vivo retrotransposition of non-LTR retrotransposon, R1Bm. Nucleic Acids Res 33: 1993–2002.
42. Ichiyanagi K, Nakajima R, Kajikawa M, Okada N (2007) Novel retrotransposon analysis reveals multiple mobility pathways dictated by hosts. Genome Res 17: 33–41.
43. Dong C, Poulter RT, Han JS (2009) LINE-like retrotransposition in Saccharomyces cerevisiae. Genetics 181: 301–311.
44. 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.
45. Ostertag EM, Kazazian HH (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res 11: 2059–2065.
46. Gilbert N, Lutz-Prigge S, Moran JV (2002) Genomic deletions created upon LINE-1 retrotransposition. Cell 110: 315–325.
47. Symer DE, Connelly C, Szak ST, Caputo EM, Cost GJ, et al (2002) Human l1 retrotransposition is associated with genetic instability in vivo. Cell 110: 327–338.
48. Szak ST, Pickeral OK, Makalowski W, Boguski MS, Landsman D, Boeke JD (2002) Molecular archeology of L1 insertions in the human genome. Genome Biol 3: research0052.
49. Gilbert N, Lutz S, Morrish TA, Moran JV (2005) Multiple fates of L1 retrotransposition intermediates in cultured human cells. Mol Cell Biol 25: 7780–7795.
50. 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.
51. Kopera HC, Moldovan JB, Morrish TA, 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.
52. Han JS, Boeke JD (2004) A highly active synthetic mammalian retrotransposon. Nature 429: 314–318.
53. Jones RB, Garrison KE, Wong JC, Duan EH, Nixon DF, Ostrowski MA (2008) Nucleoside analogue reverse transcriptase inhibitors differentially inhibit human LINE-1 retrotransposition. PLoS ONE 3: e1547. doi:10.1371/journal.-pone.0001547
54. Kroutter EN, Belancio VP, Wagstaff BJ, Roy-Engel AM (2009) The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition. PLoS Genet 5: e1000458. doi: 10.1371/journal.p-gen.1000458
55. Dai L, Huang Q, Boeke JD (2011) Effect of reverse transcriptase inhibitors on LINE-1 and Ty1 reverse transcriptase activities and on LINE-1 retrotransposition. BMC Biochem 12: 18.

56. Wagstaff BJ, Hedges DJ, Derbes RS, Campos Sanchez R, Chiaromonte F, et al (2012) Rescuing Alu: Recovery of New Inserts Shows LINE-1 Preserves Alu Activity through A-Tail Expansion. PLoS Genet 8: e1002842. doi:10.1371/journal.pgen.1002842
57. Bibillo A, Eickbush TH (2002) High processivity of the reverse transcriptase from a non-long terminal repeat retrotransposon. J Biol Chem 277: 34836–34845.
58. Jamburuthugoda VK, Eickbush TH (2011) The reverse transcriptase encoded by the non-LTR retrotransposon R2 is as error-prone as that encoded by HIV-1. J Mol Biol 407: 661–672.
59. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat 27: 323–329.
60. Lee E, Iskow R, Yang L, Gokcumen O, Haseley P, et al (2012) Landscape of somatic retrotransposition in human cancers. Science 337: 967–971.
61. Solyom S, Ewing AD, Rahrmann EP, Doucet T, Nelson HH, et al (2012) Extensive somatic L1 retrotransposition in colorectal tumors. Genome Res 22: 2328–2338.
62. Akagi K, Li J, Stephens RM, Volfovsky N, Symer DE (2008) Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition. Genome Res 18: 869–880.
63. Ewing AD, Kazazian HH (2010) High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res 20: 1262–1270.
64. Beck CR, Collier P, Macfarlane C, Malig M, Kidd JM, et al (2010) LINE-1 Retrotransposition Activity in Human Genomes. Cell 141: 1159–1170.
65. Huang CR, Schneider AM, Lu Y, Niranjan T, Shen P, et al (2010) Mobile interspersed repeats are major structural variants in the human genome. Cell 141: 1171–1182.
66. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, et al (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. Cell 141: 1253–1261.
67. 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.
68. Morrish TA, Gilbert N, Myers JS, Vincent BJ, Stamato TD, et al (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. Nat Genet 31: 159–165.
69. Morrish TA, Garcia-Perez JL, Stamato TD, Taccioli GE, Sekiguchi J, Moran JV (2007) Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. Nature 446: 208–212.
70. 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.
71. Pickeral OK, Makałowski W, Boguski MS, Boeke JD (2000) Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res 10: 411–415.
72. Goodier JL, Ostertag EM, Kazazian HH (2000) Transduction of 3′-flanking sequences is common in L1 retrotransposition. Hum Mol Genet 9: 653–657.
73. 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.
74. Bibillo A, Eickbush TH (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. J Biol Chem 279: 14945–14953.
75. 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.
76. 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.
77. Cristofari G, Gabus C, Ficheux D, Bona M, Le Grice SF, Darlix JL (1999) Characterization of active reverse transcriptase and nucleoprotein complexes of the yeast retrotransposon Ty3 in vitro. J Biol Chem 274: 36643–36648.
78. Cristofari G, Ficheux D, Darlix JL (2000) The GAG-like protein of the yeast Ty1 retrotransposon contains a nucleic acid chaperone domain analogous to retroviral nucleocapsid proteins. J Biol Chem 275: 19210–19217.
79. Cristofari G, Bampi C, Wilhelm M, Wilhelm FX, Darlix JL (2002) A 5′-3′ long-range interaction in Ty1 RNA controls its reverse transcription and retrotransposition. EMBO J 21: 4368–4379.
80. Cristofari G, Darlix JL (2002) The ubiquitous nature of RNA chaperone proteins. Prog Nucleic Acid Res Mol Biol 72: 223–268.
81. Kass EM, Jasin M (2010) Collaboration and competition between DNA double-strand break repair pathways. FEBS Lett 584: 3703–3708.
82. Makarov VL, Hirose Y, Langmore JP (1997) Long G tails at both ends of human chromosomes suggest a C strand degradation mechanism for telomere shortening. Cell 88: 657–666.
83. McElligott R, Wellinger RJ (1997) The terminal DNA structure of mammalian chromosomes. EMBO J 16: 3705–3714.
84. Wu P, Takai H, de Lange T (2012) Telomeric 3′ Overhangs Derive from Resection by Exo1 and Apollo and Fill-In by POT1b-Associated CST. Cell 150: 39–52.
85. Greider CW, Blackburn EH (1987) The telomere terminal transferase of Tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. Cell 51: 887–898.
86. Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. Science 276: 561–567.
87. Lingner J, Cech TR (1996) Purification of telomerase from Euplotes aediculatus: requirement of a primer 3′ overhang. Proc Natl Acad Sci U S A 93: 10712–10717.
88. Eickbush TH (1997) Telomerase and retrotransposons: which came first? Science 277: 911–912.
89. Nakamura TM, Cech TR (1998) Reversing time: origin of telomerase. Cell 92: 587–590.
90. Gladyshev EA, Arkhipova IR (2007) Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. Proc Natl Acad Sci U S A 104: 9352–9357.
91. An W, Davis ES, Thompson TL, O'Donnell KA, Lee CY, Boeke JD (2009) Plug and play modular strategies for synthetic retrotransposons. Methods 49: 227–235.

SUPPLEMENTARY FIGURES

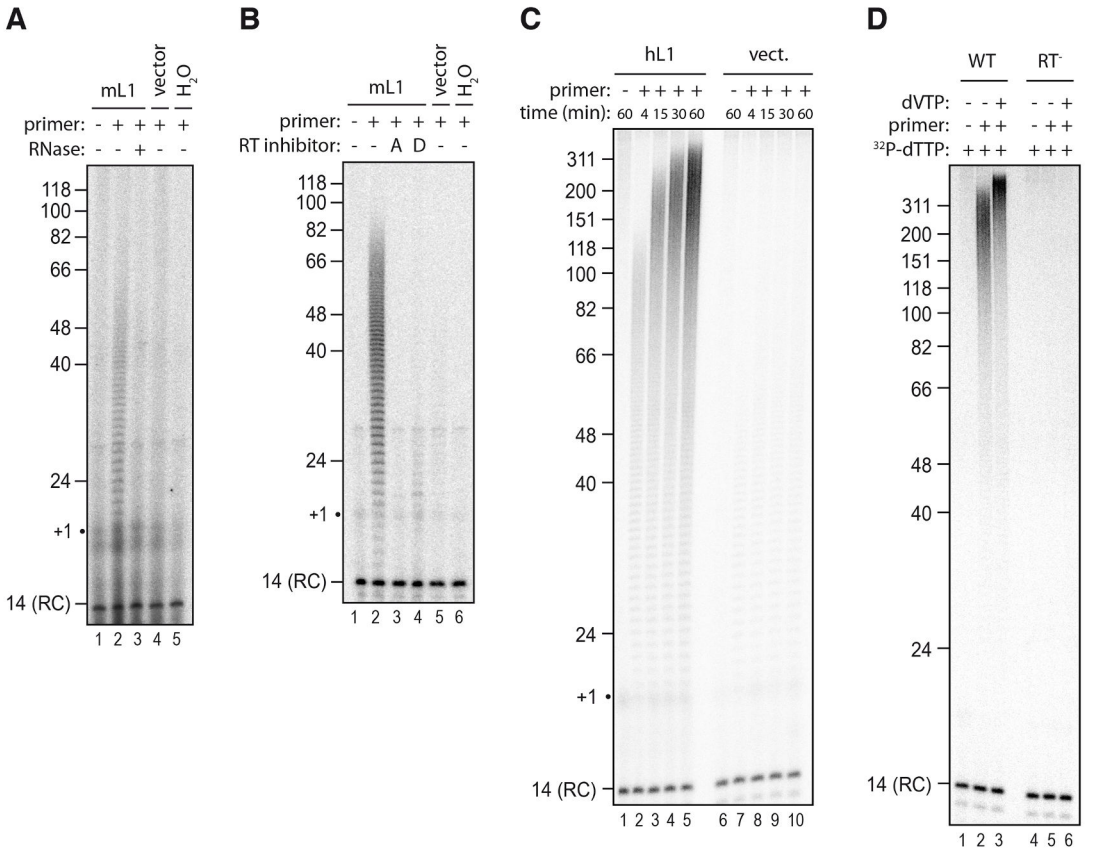


Figure S1: Additional characterization of the L1 RNP RT activity by DLEA.
(A) RNA-dependent DNA polymerase activity of L1 RNPs. Murine L1 RNPs were incubated for 1 h at 37°C in the presence (lane 3) or in the absence (lane 4) of RNases before the start of the reaction. (B) RT inhibitors prevent primer extension by L1 RNPs. Reactions were performed with mL1 RNPs in the presence of thymidine analogs (10 µM of azidothymidine triphosphate AZTTP, denoted by A, lane 3; 10 µM of 2,3-didehydro-3-deoxythymidine triphosphate d4TTP, denoted by D, lane 4), or in the presence of water as a negative control (lane 2). (C) Time-course of (dT)18 primer extension by hL1 RNP. (D) Formation of long cDNA species upon addition of all four dNTPs. Reactions were performed with hL1 RNPs in presence of α-32P-dTTP and a (dT)18 primer, with (lanes 3 & 6) or without (lanes 1–2 & 4–5) cold dATP, dCTP and dGTP (dVTP, IUPAC nomenclature).

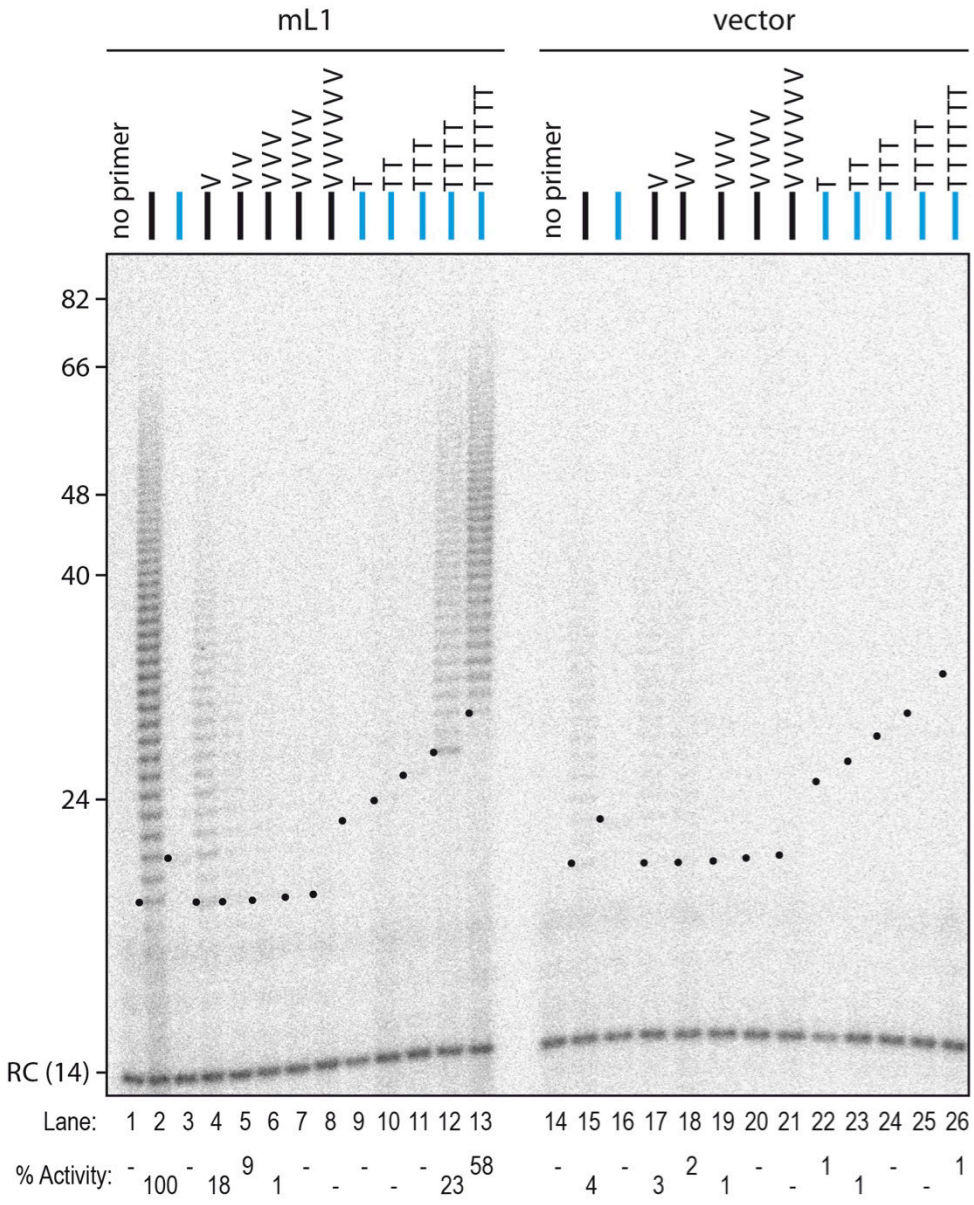
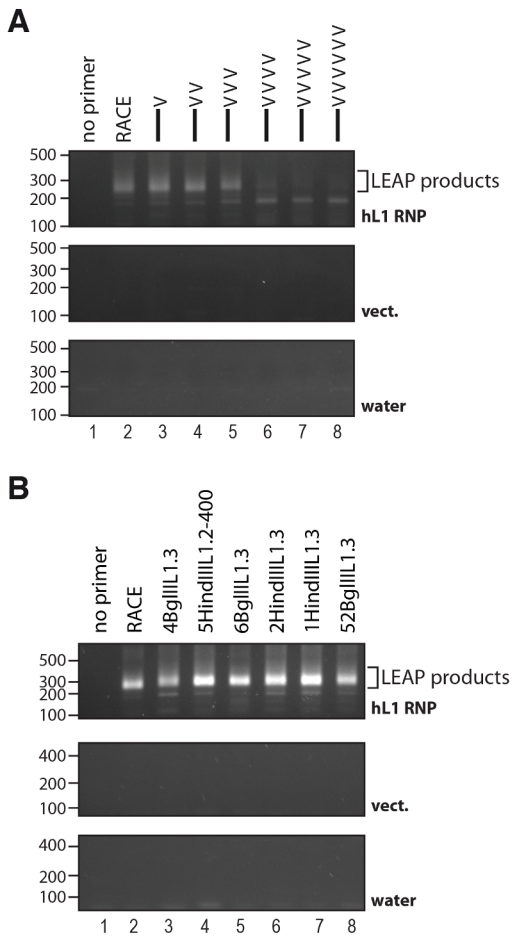


Figure S2. The murine L1 RNP preferentially extends primers ending with at least 4 Ts. *DLEA showing the extension of single-stranded primers by mL1 RNPs in the presence of α-32P-dTTP. RC denotes a 14 nt recovery control added after the reaction but before DNA purification. The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies. Quantification of primer extension (% Activity) was relative to levels of extension obtained with oligo(dT)18. Primers are identical to Figure 2.*



C

Primer	Sequence	Number of clones
RACE	...AGGTTTTTTTTTTTTVN	68/70
	...AGGTTTTTTTTTTTT	2/70
V1	...AGGTTTTTTTTTTTV	37/38
	...AGGTTTTTTTTTTT	1/38
V_2	...AGGTTTTTTTTTTVV	23/24
	...AGGTTTTTTTTTTV	1/24
	...AGGTTTTTTTTTT	0/24
V_3	...AGGTTTTTTTTTVVV	25/28
	...AGGTTTTTTTTTVV	1/28
	...AGGTTTTTTTTTV	1/28
	...AGGTTTTTTTTT	1/28
4BglIIL1.3	...AGGAGGAGAGATGTACATTTTAT	23/26
	...AGG	3/26
5HindIIIL1.2-400	...AGGCCATGTCCTGTATCCTTTCT	23/24
	...AGG	1/24
6BglIIL1.3	...AGGATTTATTTAATTTCTTTTTC	26/29
	...AGGATTTATTTAATTTCTTTTT	3/29
2HindIIIL1.3	...AGGCAGCTTTCATTAAATTTCTT	23/23
1HindIIIL1.3	...AGGAGATTATTTGGCTTTTTCTT	25/26
	...AGGG	1/26
52BglIIL1.3	...AGGTTAGCTAATTTTTTATTTTT	28/28

Figure S3.LEAP with hL1 RNPs and mismatched primers.

(A) Primers with terminal mismatches. LEAP was performed with RNPs prepared from hL1-transfected cells (top panel), from vector-transfected cells (middle panel), or without RNPs (bottom panel). Primers are identical to those used in Figure 2, except that they have a 5' extension to anchor the PCR (see Table S1 for sequence). (B) Primers mimicking L1 integration sites. LEAP was performed with RNPs prepared from hL1-transfected cells (top panel), from vector-transfected cells (middle panel), or without RNPs (bottom panel). Primers are identical to those used in Figure 4, except that they have a 5' extension to anchor the PCR (see Table S1 for sequence). (C) LEAP products from (A) and (B) were gel purified, cloned and sequenced. For each oligonucleotide, the top sequence and number of clones correspond to the extension of unprocessed primer, whereas other sequences correspond to the extension of processed primers.

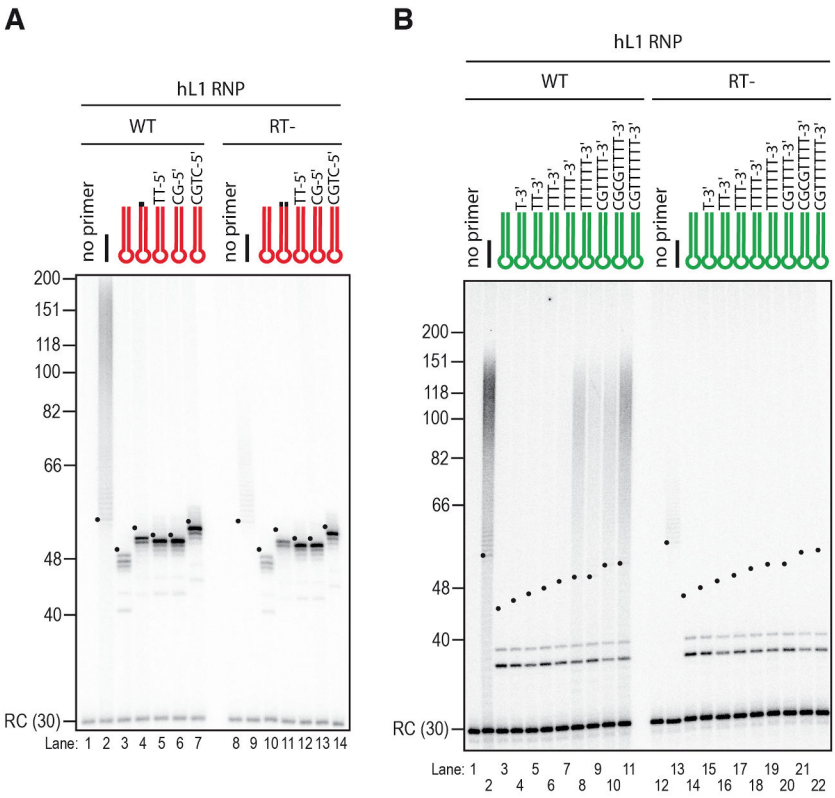


Figure S4.Human L1 RNPs preferentially extends double-stranded DNA with a 3′ overhang.

(A) Absence of extension by hL1 RNPs of double-stranded primers with blunt or 3'-recessed end in the presence of α-32P-dTTP. Note that the products observed with hairpin primers (lanes 3–7) result from contaminating cellular activities (see main text and Figure 8). (B) Extension by hL1 RNPs of double-stranded primers ending with a 3' overhang in the presence of α-32P-dTTP. Note that the doublet below 40 nt observed in lanes 3–11 and 14–22 results from contaminating cellular activities (see text and Figure 8 for further characterization). RC denotes a 30 nt recovery control added after the reaction but before DNA purification. The black dots on the left side of each lane indicate the expected start of reverse transcription. Their position varies since primer length varies. Results obtained with mL1 RNPs were identical and are shown in Figure 6, Figure 7, Figure 8.

Table S1. List of oligonucleotides used in this study

Primer	Name in the figures	Description	Sequence (from 5' to 3')
LOU266		Gateway ORF2p entry clone, forward primer	GGGGACAACTTTGTACAAAAAAGTTGGCATGCCCCCCCC TGACCACCAAGA
LOU267		Gateway ORF2p entry clone, reverse primer (with stop codon)	GGGGACAACTTTGTACAAGAAAGTTGGTTAGTAGCCGC TGATCAGGCTGT
LOU419		Directed mutagenesis mL1 (D709A), forward primer	GATCAGCCTGTTCGCCGCGGACATGATCGTGTACAT
LOU420		Directed mutagenesis mL1 (D709A), reverse primer	ATGTACACGATCATGTCCGCGGCGAACAGGCTGATC
LOU398	$(dT)_{18}$	Linear RT primer	TTTTTTTTTTTTTTTTTT
T7 prom	R	Linear RT primer	TAATACGACTCACTATAGGG
LOU425	T_1	Linear RT primer	TAATACGACTCAACTATAGGGGT
LOU426	T_2	Linear RT primer	TAATACGACTCAACTATAGGGGTT
LOU427	T_3	Linear RT primer	TAATACGACTCAACTATAGGGGTTT
LOU428	T_4	Linear RT primer	TAATACGACTCAACTATAGGGGTTTT
LOU429	T_6	Linear RT primer	TAATACGACTCAACTATAGGGGTTTTTT
LOU440	V_6	Linear RT primer	TTTTTTTTTTTTVVVVVV
LOU441	V_4	Linear RT primer	TTTTTTTTTTTTTTVVVV
LOU442	V_3	Linear RT primer	TTTTTTTTTTTTTTTVVV
LOU443	V_2	Linear RT primer	TTTTTTTTTTTTTTTTVV
LOU444	V_1	Linear RT primer	TTTTTTTTTTTTTTTTTV
LOU1249		Linear RT primer	TTTTTTTTTTTTTTTTTA
LOU1250		Linear RT primer	TTTTTTTTTTTTTTTTTG
LOU1251		Linear RT primer	TTTTTTTTTTTTTTTTTC
LOU430	H_0	Hairpin RT primer (blunt)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGG
LOU431	H_1	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGT
LOU432	H_2	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGTT
LOU433	H_3	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGTTT
LOU434	H_4	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGTTTT
LOU435	H_6	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGTTTTTT
LOU467	H_6-4T	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGCGTTTT
LOU468	H_8-6T	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGCGTTTTTT
LOU469	H_8-4T	Hairpin RT primer (3' overhang)	CCCTATAGTGAGTCGTATTACACATAATACGACTCACTA TAGGGCGCGTTTT
LOU436	H	Hairpin RT primer (blunt)	AAAACCCTATAGTGAGTCGTATTACACATAATACGACTC ACTATAGGGTTTT
LOU470	H-ext	Hairpin RT primer (blunt)	AAAAAACCCTATAGTGAGTCGTATTACACATAATACGAC TCACTATAGGGTTTTTT

LOU437	5'TT-H	Hairpin RT primer (3' recessed)	TTAAAACCCTATAGTGAGTCGTATTACACATAATACGAC TCACTATAGGGTTTT
LOU438	5'GC-H	Hairpin RT primer (3' recessed)	GCAAAACCCTATAGTGAGTCGTATTACACATAATACGAC TCACTATAGGGTTTT
LOU439	5'CTGC-H	Hairpin RT primer (3' recessed)	CTGCAAAACCCTATAGTGAGTCGTATTACACATAATACG ACTCACTATAGGGTTTT
RBD3	RC (14)	Recovery control	TACGTTCTATGCTA
LOU491	RC (30)	Recovery control	GCCGCCGGGATCACTCTCGGCATGGACGAG
LOU519	2BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	GAGCCAGGAGGAATACTTTT
LOU520	4BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	AGGAGAGATGTACATTTTAT
LOU521	7BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TTTTGGATCCTCTGACTTCT
LOU522	52BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TTAGCTAATTTTTTATTTTT
LOU523	55BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TATCCTTCCAGCAGTTTCTT
LOU524	1HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	AGATTATTTGGCTTTTTCTT
LOU525	5HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TGGTTTTTTTTTTTTTTTC
LOU526	6HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	AGCTTTTCCATTGTATTTCT
LOU527	9HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TAGTTGTATCAATGGTTTTC
LOU528	11HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CAGCTAATTTTGGTATTCTT
LOU529	17HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TACAAATTTTTTGTTTTTTA
LOU530	21HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TTCTGGCTCTCTGCATTTCT
LOU531	22HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	AGATTCATAAGCAAATTCTT
LOU532	23HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	GAGCATGAAGGAAGTTTTCT
LOU533	27HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	ATCTTTTGCTGTCATGTCTT
LOU534	50HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CCAAGTAGAGTCATGATTTT
LOU535	4HindIIIL1.	Linear RT primer derived	GAATTTCATTAGGAAATTTT

	2-400	from Gilbert *et al.* Cell, 2002 (L1 insertion site)	
LOU536	5HindIIIL1.2-400	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CCATGTCCTGTATCCTTTCT
LOU537	1BamH1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	GGCAGGACTTTTTTTTTTTT
LOU538	6BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	ATTTATTTAATTTCTTTTTC
LOU539	8BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TCTTTTTTTTCTTTTTTTTT
LOU540	9BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CTTCCTTTTTTCTTCTTTTT
LOU541	10BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TTTTTTTTTTTTTTTTTTTT
LOU542	13HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CTTCTTTTTTTTTTTTTTT
LOU543	24HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	AATTTTTTTGTGTTTTTTTT
LOU544	25HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CACATCAAATTCTATTTTTT
LOU545	29HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TCTGATTTCTGGATATTTCT
LOU546	3HindIIIL1.2	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TCTGGGTAAATGATTTTTTT
LOU547	2HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CAGCTTTCATTAAATTTCTT
LOU548	2BclIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	ATAGATTTGTATTGGATTTT
LOU549	5BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TTTTTTGCAGCTGCAGTTTT
LOU550	11BglIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	GAAAAAGTAGAGCTTTTATT
LOU551	3HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	CATAATTTCCATTCAGTTTT
LOU552	10HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TGTGTGGCCTTTCTTTTTTT
LOU553	26HindIIIL1.3	Linear RT primer derived from Gilbert *et al.* Cell, 2002 (L1 insertion site)	TTTTATTTATATATTTTTTT
RACE	RACE	Primer for LEAP reaction	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTT

		from Kulpa *et al.* Nat Struct Mol Biol, 2006	TTTTTVN
LOU863	4BglIIL1.3	Primer for LEAP reaction derived from 4BglIIL1.3 and LOU312	GCGAGCACAGAATTAATACGACTCACTATAGGAGGAGA GATGTACATTTTAT
LOU864	5HindIIIL1.2-400	Primer for LEAP reaction derived from 5HindIIIL1.2-400 and LOU312	GCGAGCACAGAATTAATACGACTCACTATAGGCCATGT CCTGTATCCTTTCT
LOU865	6BglIIL1.3	Primer for LEAP reaction derived from 6BglIIL1.3 and LOU312	GCGAGCACAGAATTAATACGACTCACTATAGGATTTATT TAATTTCTTTTTC
LOU866	2HindIIIL1.3	Primer for LEAP reaction derived from 2HindIIIL1.3 and LOU312	GCGAGCACAGAATTAATACGACTCACTATAGGCAGCTTT CATTAAATTTCTT
LOU867	1HindIIIL1.3	Primer for LEAP reaction derived from 1HindIIIL1.3 and LOU312	GCGAGCACAGAATTAATACGACTCACTATAGGAGATTAT TTGGCTTTTTCTT
LOU868	52BglIIL1.3	Primer for LEAP reaction derived from 52BglIIL1.3 and LOU312	GCGAGCACAGAATTAATACGACTCACTATAGGAGATTAT TTGGCTTTTTCTT
LOU962	V_1	Primer for LEAP reaction derived from RACE	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTT TTTTV
LOU963	V_2	Primer for LEAP reaction derived from RACE	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTT TTTVV
LOU964	V_3	Primer for LEAP reaction derived from RACE	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTT TTVVV
LOU965	V_4	Primer for LEAP reaction derived from RACE	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTT TVVVV
LOU966	V_5	Primer for LEAP reaction derived from RACE	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTT VVVVV
LOU967	V_6	Primer for LEAP reaction derived from RACE	GCGAGCACAGAATTAATACGACTCACTATAGGTTTTTTV VVVVV
LOU312		Linker RACE primer from Kulpa *et al.* Nat Struct Mol Biol, 2006	GCGAGCACAGAATTAATACGACT
LOU851		Sense L1 3'end primer for LEAP from Kulpa *et al.* Nat Struct Mol Biol, 2006	GGGTTCGAAATCGATAAGCTTGGATCCAGAC
LOU852		Sense GAPDH 3' end primer for RT-PCR from Kulpa *et al.* Nat Struct Mol Biol, 2006	GACCCTCACTGCTGGGGAGTCC
SP6	SP6 primer	Universal primer for sequencing	ATTTAGGTGACACTATAG

Table S2. Data used to calculate genomic enrichment of L1 insertions depending on the snap-velcro status of the target.

The table sheets are the following: (hg19) For each potential L1 EN target site present in hg19, the snap status was defined and the position-weighted A density was calculated. Sites with position-weighted A density equal to or above 0.5 were considered as having a closed velcro strap. (hg19 RM) Same as above but with a repeatmasked hg19 reference genome. (dbRIP sequences) L1HS dbRIP entries used in Figure 5C and 5C and their snap/velcro status. (dbRIP counts) Number of dbRIP entries in each category. (dbRIP weblogo) Weblogo of the junction sequence (−2/+10) for dbRIP entries. (Lee2012 sequences) L1HS somatic insertions in cancer used in Figure 5C and 5C and their snap/velcro status. (Lee2012 counts)

Number of L1HS somatic insertions in each category. (Lee2012 weblogo) Weblogo of the junction sequence (−2/+10) for Lee2012 entries. (Solyom2012 sequences) L1HS somatic insertions in colon cancer used in Figure 5C and 5C and their snap/velcro status. (Solyom2012 counts) Number of L1HS somatic insertions in each category. (Solyom2012 weblogo) Weblogo of the junction sequence (−2/+10) for Solyom2012 entries.
Protocol S1.

Source code of the software used to find putative endonuclease sites in the human genome and to calculate their associated snap/velcro scores.

```
/*****************************************************************************************************************
 * Copyright (c) 2013, Institute for Research on Cancer and Aging, Nice (IRCAN), INSERM U1081 - CNRS UMR 7284, University of Nice - Sophia-Antipolis, Faculty of Medicine, Nice, FRANCE
 * All rights reserved.
 * Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:
 * - Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
 * - Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
 *- Neither the name of the Institute for Research on Cancer and Aging, Nice (IRCAN), INSERM U1081 - CNRS UMR 7284, University of Nice - Sophia-Antipolis nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.
 * THIS SOFTWARE IS PROVIDED BY THE IRCAN AND CONTRIBUTORS ``AS IS'' AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED
 * WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE IRCAN AND CONTRIBUTORS BE LIABLE FOR ANY
 * DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES;
 * LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT
 * (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.
 * Contributors: Ashfaq Ali Mir
 * Affiliation:  Laboratory of Gael Cristofari
 * Contact:    Gael.Cristofari@unice.fr
 * Date:       03 April 2013
 * Version:   Stable 1.0
 * Language:   C++
 * Name:   GenomeAnalyzerSnapVelcroV1.0.cpp
 * Description:   This software calculates weighted scores for Snap Velcro classified strings across genome
 * Input Format: Text file with single without spaces or newline (containing nucleic acid characters)
 * Usage:   ./GenomeAnalyzerSnapVelcroV1.0 InputFile.txt
 * Manual:    associated Readme.txt
 *****************************************************************************************************************/
#include<iostream>
#include<fstream>
#include<string>
#include<set>
#include<map>
using namespace std;


int GetIntVal(string strConvert) {
        int intReturn;
        intReturn = atoi(strConvert.c_str());
        return(intReturn);
}

void chg2String(char *ln){
string str;
str=ln;
cout<<"line[]'s length is "<<str.length()<<endl;
cout<<str<<endl;
}
multimap<int, int> mm1;
   multimap<int, int>::iterator it1;


      int flanking = 0;
      int back = 0;
      int front = 0;
```

```
        int location = 0;

int main(int argc ,char* argv[]){

    ifstream myFile(argv[1]);
    string line;
    string forward;
    if(! myFile){
        cout << "Error opening the File" << endl;
        exit(1);
    }

    while(! myFile.eof()){
      getline(myFile, line);
        for(int i=0; i < line.size(); i++){

        int count = 0;
        int A_count_H = 0;
        int A_count_8 = 0;
        int A_count_4 = 0;
        float A_w1 = 0.000f;
        float A_w2 = 0.000f;
        float A_w3 = 0.000f;
        float A_w4 = 0.000f;
        float A_w5 = 0.000f;
        float A_w6 = 0.000f;
        float A_w7 = 0.000f;
        float A_w8 = 0.000f;
        float A_w9 = 0.000f;
        float A_w10 = 0.000f;
        string A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12;
        float TWeight_8 = 0.000f;
        float TWeight_4 = 0.000f;


      forward = line.substr(i,12);

                    for (int j=0; j < forward.size(); j++)
                    {

                      if(j == 0){

                        A1 = forward[j];

                      }
                      if(j == 1){

                        A2 = forward[j];

                      }
                      if(j == 2){
                        A3 = forward[j];

                      if(A3.compare("A")==0){
                            A_count_8++;
                            A_count_H++;
                            A_w1 = 1.000f;

                            }
                            }
                       if(j == 3){
```

```
                A4 = forward[j];

                if(A4.compare("A")==0){
                A_count_8++;
                A_count_H++;
                A_w2 = 0.500f;

                }
                }
if(j == 4){

                A5 = forward[j];

                if(A5.compare("A")==0){
                    A_count_8++;
                    A_count_H++;
                    A_w3 = 0.333f;

                }
                }
                if(j == 5){

                A6 = forward[j];

                if(A6.compare("A")==0){
                     A_count_8++;
                     A_count_H++;
                     A_w4 = 0.250f;

                            }
                            }
                            if(j == 6){


                            A7 = forward[j];

                            if(A7.compare("A")==0){
                            A_count_4++;
                            A_count_8++;
                            A_w5 = 0.200f;

                            }

                            }
                            if(j == 7){

                            A8 = forward[j];

                            if(A8.compare("A")==0){
                                        A_count_4++;
                                        A_count_8++;
                                        A_w6 = 0.167f;
                            }

                            }
                            if(j == 8){

                                        A9 = forward[j];

                            if(A9.compare("A")==0){
                                        A_count_4++;
                                        A_count_8++;
```

```
                                        A_w7 = 0.143f;
                                    }

                                    }
                                    if(j == 9){

                                        A10 = forward[j];

                                    if(A10.compare("A")==0){
                                        A_count_4++;
                                        A_count_8++;
                                        A_w8 = 0.125f;

                                    }
                                     }
                                    if(j == 10){

                                        A11 = forward[j];

                                    if(A11.compare("A")==0){
                                        A_count_4++;
                                        A_count_8++;
                                        A_w9 = 0.111f;

                                    }

                                     }
                                    if(j == 11){

                                        A12 = forward[j];

                                    if(A12.compare("A")==0){
                                                A_count_4++;
                                                A_count_8++;
                                                A_w10 = 0.100f;

                                        }
                                         }

                                     }
                        TWeight_8 = A_w1 + A_w2 + A_w3 + A_w4 + A_w5 + A_w6 + A_w7+ A_w8 + A_w9 + A_w10;
                        TWeight_4 = A_w5 + A_w6 + A_w7+ A_w8 + A_w9 + A_w10;
                        string Motif = A1 + A2 + A3 + A4 + A5 + A6 + A7 + A8 + A9 + A10 + A11 + A12;

            if((A2.compare("T")==0 || A2.compare("C")==0)){

                        if(A_count_H == 3){

cout << TWeight_8 << "\t" << TWeight_4 << "\t" << A_count_8 << "\t" << A_count_4 << "\t" << "1" << endl;     //   open   =   1

                                    }
                        else if(A_count_H == 4) {

cout << TWeight_8 << "\t" << TWeight_4 << "\t" << A_count_8 << "\t" << A_count_4 << "\t" << "0" << endl;     //   closed   =   0;


                                    }
                                }
                                 }
    }
```

```
return 0;
}

################ CODE ENDS
```
****

README: Usage specification

Input File Format Specifications:

- This C++ program takes an input text file containing a single string (without space or newline)
- The file should contain nucleic acid characters like A / C / G / T and can tolerate masked characters (e.g., X / N), which are skipped.

Output Format:

A_weighted(+1_+10)<tab>A_weighted(+5_+10)<tab>A_Count(+1_+10)<tab>A_Count(+5_+10)<tab>Snap(Snap open = 1 / Snap close = 0)

where:

- A_weighted(+1_+10) refers to the position-weighted A count for nucleotides +1 to +10 after endonuclease site
- A_weighted(+5_+10) refers to the position-weighted A count for nucleotides +5 to +10 after endonuclease site (velcro score)
- A_Count(+1_+10) refers to the number of A for nucleotides +1 to +10 after endonuclease site
- A_Count(+5_+10) refers to the number of A for nucleotides +5 to +10 after endonuclease site
- Snap refers to the status of the snap region

Example of output:

2.45	0.367	6	2	0
1.283	0.2	4	1	1

Steps to run the program in a Unix shell :

- Compile program command : g++ GenomeAnalyzerSnapVelcroV1.0.cpp -o GenomeAnalyzerSnapVelcroV1.0
- Program running command : ./GenomeAnalyzerSnapVelcroV1.0 InputFile.txt > OutputFile.txt
- Extract the count for each A_weighted(+5_+10) in OutputFile.txt : awk '{print $2}' OutputFile.txt | sort -k 2n | uniq -c > OutputFile2.txt

Specifications :

- This program has been written, tested and compiled on 64 bit machine (Mac Os 10.6.4)
- gcc version 4.2.1 (Apple Inc. build 5666)

Additional information :

- each chromosome and each strand have to be treated one by one

2. euL1db: the European database of L1HS retrotransposon insertions in humans

2.1. Context of the study

Retrotransposons constitute almost half of our genome. They are mobile genetics elements—also known as jumping genes—but only the L1HS subfamily of Long Interspersed Nuclear Elements (LINEs) has retained the ability to jump autonomously in modern humans. The role of retrotransposition as a source of genetic diversity and diseases in humans has been shown by many studies. Advances in deep-sequencing technologies have shed a new light on the extent of L1-mediated genome variations. They have also lead to the discovery that L1-HS is not only able to mobilize in the germline - resulting in inheritable genetic variations - but can also jump in somatic tissues, such as embryonic stem cells, neuronal progenitor cells, or in many cancers.

Most retrotransposition events is the consequence of highly active, or 'hot', L1-HS loci that constitute a small minority of total active L1-HS elements, with many of these being population-specific elements or unique to a particular individual, also known as private copies. Therefore understanding the link between L1-HS insertion polymorphisms and phenotype or disease requires a comprehensive view of the different L1HS copies present in given individuals.

There were few resources before euL1db like dbRIP and dbVar/DGVa (411), which contain a minute set of L1 data in a non-specific way and lacked recent L1 insertions including the one from 1000 genomes project. Therefore, there was a need for a comprehensive resource with exhaustive and most suitable data structure for human specific L1 insertion data.

euL1db provides a curated and comprehensive summary of L1 retrotransposon insertion polymorphisms (RIPs) identified in healthy or pathological human samples and published in peer-reviewed journals. An important feature of euL1db is that insertions can be retrieved at a sample-by-sample level to facilitate correlations between the presence/absence of an L1 insertion with a specific phenotype or disease.

euL1db allows the user to search, browse, compare, submit, download and visualize the L1 insertion data. The user can also perform batch querying and look for overlapping insertions within the query gene / insertion lists vs euL1db. Insertion data can be retrieved at the study, insertion, sample, individual and family levels.

As a lead author, I did most of the work, which included design, and implementation of the data structure for efficient data processing. Other tasks included setting up the web server, relational database system and Java programming of different modules using JSP, JSTL, JDBC, JSF, Servlets, Beans, Ajax and other related web technologies. Therefore, making this database an efficient client-server technology

application was quite challenging. Apart from that I was also involved in data curation process.

Therefore, the purpose of euL1db is to provide centralized and user-friendly access to known germline and somatic L1HS insertions, which will be critical to elucidate the physiological or pathological impact of novel L1HS insertions. This resource will be useful in a large variety of fields such as human genetics, neurosciences or cancer genomics.

2.2. Article-II

Nucleic Acids Research Advance Access published October 28, 2014

doi: 10.1093/nar/gku1043

euL1db: the European database of L1HS retrotransposon insertions in humans

Ashfaq A. Mir[1,2,3], Claude Philippe[1,2,3] and Gaël Cristofari[1,2,3,*]

[1]INSERM, U1081, Institute for Research on Cancer and Aging of Nice (IRCAN), F-06100 Nice, France, [2]CNRS, UMR 7284, Institute for Research on Cancer and Aging of Nice (IRCAN), F-06100 Nice, France and [3]Faculty of Medicine, Institute for Research on Cancer and Aging of Nice (IRCAN), University of Nice-Sophia-Antipolis, F-06100 Nice, France

Received August 25, 2014; Revised October 5, 2014; Accepted October 10, 2014

ABSTRACT

Retrotransposons account for almost half of our genome. They are mobile genetics elements—also known as jumping genes—but only the L1HS subfamily of Long Interspersed Nuclear Elements (LINEs) has retained the ability to jump autonomously in modern humans. Their mobilization in germline—but also some somatic tissues—contributes to human genetic diversity and to diseases, such as cancer. Here, we present euL1db, the European database of L1HS retrotransposon insertions in humans (available at http://euL1db.unice.fr). euL1db provides a curated and comprehensive summary of L1HS insertion polymorphisms identified in healthy or pathological human samples and published in peer-reviewed journals. A key feature of euL1db is its sample-wise organization. Hence L1HS insertion polymorphisms are connected to samples, individuals, families and clinical conditions. The current version of euL1db centralizes results obtained in 32 studies. It contains >900 samples, >140 000 sample-wise insertions and almost 9000 distinct merged insertions. euL1db will help understanding the link between L1 retrotransposon insertion polymorphisms and phenotype or disease.

INTRODUCTION

Repetitive DNA accounts for half of our genome. Most of these repeats are retrotransposons, i.e. mobile genetic elements, which proliferate through an RNA-mediated copy-and-paste mechanism, called retrotransposition. A tiny fraction of human retrotransposons is still able to autonomously generate new copies in modern humans (1). These active elements all belong to the L1HS subfamily (HS stands for human-specific), a subgroup of the L1 (Long Interspersed Nuclear Element-1 or LINE-1) clade of non-Long Terminal Repeat (LTR) retrotransposons found in vertebrates, plants and fungi. The L1 retrotransposon machinery is also able to mobilize *in trans* non-autonomous retrotransposons belonging to the Short Interspersed Nuclear Element (SINE) class (*Alu*, SVA); or cellular RNAs (U6, mRNA), which results in processed pseudogene formation (see (2–4) for recent reviews). Other transposable elements are molecular fossils and do not mobilize in modern humans.

A full-length human L1 is ~6.0 kb in length, contains an internal promoter located in the 5′-untranslated region and encodes two proteins, ORF1p and ORF2p, both being required for L1 retrotransposition. ORF1p is an RNA-binding protein (5) and ORF2p an enzyme with endonuclease and reverse transcriptase activities (6,7). These proteins associate with the L1 mRNA to form a ribonucleoprotein particle, which is considered as the core of the L1 retrotransposition machinery (8,9). A new L1 copy is produced when ORF2p nicks the genomic DNA and extends this newly formed 3′ end using the L1 mRNA as a template, a process known as target-primed reverse transcription (TPRT) (7,10,11). This process results in a short duplication of the target site (TSD, target-site duplication). Abortive retrotransposition often leads to 5′ truncated L1 copies (12,13). Some L1 insertions exhibit both a 5′ truncation and an inversion, due to twin priming (14). Finally, L1 insertions can also contain 5′- or 3′-transductions corresponding to genomic sequences immediately upstream or downstream their progenitor copies. Such events originate from the retrotransposition of L1 transcripts generated from upstream promoters or ending downstream of the L1 sequence due to the weakness of the natural L1 polyadenylation signal (13,15,16). L1 target site preference is currently not fully defined, but both the endonuclease consensus sequence and the ability of the target site to partially anneal to the L1 mRNA poly(A) tail contribute to this process (7,17–19).

*To whom correspondence should be addressed. Tel: +33 4 93 37 70 87; Fax: +33 4 93 37 70 92; Email: Gael.Cristofari@unice.fr

In the past 5 years, advances in deep-sequencing technologies have shed a new light on the extent of L1-mediated genome variation (20,21). L1HS represents ~3.3 Mb of the reference human genome (~0.1%). These L1 copies are often referred to as 'reference L1HS elements'. However, each individual has additional L1HS copies not present in the reference genome, referred to as 'non-reference L1HS elements', which contribute to our genetic diversity (22–27). On the average, two human individual genomes differ at 285 sites with respect to L1 insertion presence or absence (27). These recent studies have also led to the discovery that L1HS is not only able to mobilize in the germline—resulting in inheritable genetic variations (3,28,29)—but can also jump in some somatic tissues, such as brain (30–32) or in many cancers (26,33–39). Most retrotransposition events are the consequence of highly active, or 'hot', L1HS loci that constitute a small minority of full-length L1HS elements, with many of these being population-specific or even unique to a particular individual (private copies) (1,24). Therefore, understanding the link between L1HS insertion polymorphisms and phenotype or disease requires a comprehensive view of the different L1HS copies present in given individuals.

euL1db provides a curated and comprehensive summary of L1 retrotransposon insertion polymorphisms (RIPs) identified in healthy or pathological human samples and published in peer-reviewed journals. A sample is defined here as the primary biological material (e.g. tissue biopsy, blood, cell or cell line) from which a genomic DNA preparation was obtained and a sequencing library prepared. An important feature of euL1db is that insertions can be retrieved at a sample-by-sample level to facilitate correlations between the presence/absence of an L1 insertion with a specific phenotype or disease.

DATABASE STRUCTURE AND CONTENT

The euL1db database is organized in several tables, which are interconnected in a dynamic way, through the MySQL relational database management system. A simplified view of the object relationships is depicted in Figure 1 and a more detailed view of the underlying database structure is shown in Supplementary Figure S1.

The 'Study' table contains information about the study in which L1HS insertions were cataloged and mapped. Typically, a study will correspond to a single publication. Each study uses a coherent set of methods and analyses. Because these parameters determine to a large extent the variability that exists between data sets, all data in euL1db are organized by study. The 'Individual' table relates to the source individuals from whom the samples were originally taken from. The same individual might have been subjected to multiple analyses, possibly in different studies (e.g. Figure 1, individual 1, present in study 1 and 2). When available, euL1db stores the gender, the geographical origin, potential familial links with other individuals in euL1db and health information. All individuals from the 1000 Genomes Project have been incorporated in euL1db, even though only a small portion has been analyzed for L1HS content. This was necessary to maintain the family architecture and to facilitate future updates. The 'Sample' table describes the primary biological sample taken from a given individual and from which L1HS insertions were cataloged and mapped. When available, euL1db stores the anatomical and potential pathological data, and whether it was prepared from a single-cell or from multiple cells. Potential relationships between samples are also recorded (e.g. normal-tumor pairs). Importantly, a given sample can only be linked to a single study, and is given a unique ID. Unanalyzed individu-

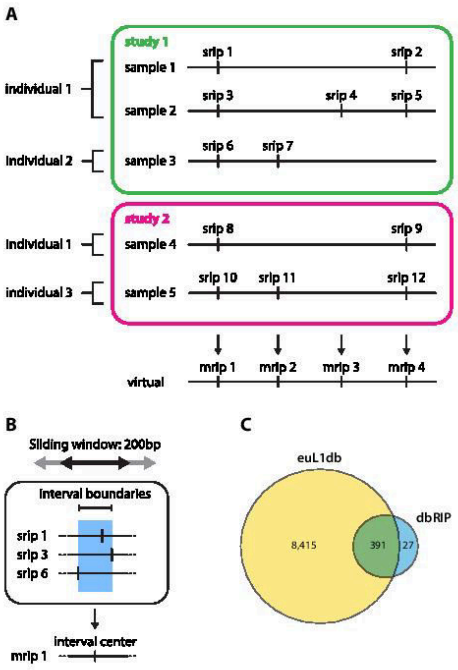


Figure 1. Database organization, data model and content. (**A**) Relationship between euL1db objects. euL1db is organized by study. Each study contains one or more samples. A sample originates from a single individual. Individuals can be analyzed in multiple studies. An SRIP (sample retrotransposon insertion polymorphism) is a real insertion detected in a given sample and has a unique ID prefixed by srip. Several samples from different individuals might possess an SRIP at the same genomic location. A private L1HS insertion will correspond to an SRIP only found in samples of the same individual. Inversely, an L1HS insertion which is fixed in the human population will appear as an SRIP at the same location in all the genome-wide samples of euL1db. Thus SRIP are highly redundant. In contrast, MRIP (meta-retrotransposon insertion polymorphisms) are virtual insertions obtained by merging overlapping or close SRIP, which are likely to correspond to the same retrotransposition event. Thus MRIP are non-redundant. (**B**) Approach used in euL1db to define unique L1HS insertion events. Nearby SRIPs are merged into a single MRIP if they satisfy all the following requirements: (i) they are located within 200 bp of each other, (ii) they share the same strand orientation, and (iii) they are all germline. Somatic retrotransposition events are unique by nature, and are not merged with germline events, nor merged together. Therefore, somatic SRIPs give rise to MRIPs containing only a single SRIP. (**C**) Overlap between euL1db and dbRIP. Numbers correspond to MRIP records in euL1db and to L1HS records in dbRIP (transposable elements not belonging to the L1HS subfamily were not taken into account to draw this Venn diagram).

als from the 1000 Genomes Project are not linked to any sample.

L1HS insertions found in a given sample are cataloged as 'sample retrotransposon insertion polymorphism' or SRIP. An SRIP is defined minimally by its genomic coordinates and is linked to a unique sample (Figure 1A). Additional optional information might include its genomic strand, its internal sequence, the length and sequence of its TSD or deletion, the presence of a 5′- or 3′-transduction, the presence of a 5′ inversion, the size of its downstream poly(A) sequence, its coordinates relative to the Repbase L1HS consensus sequence (and the positions of an inversion, if present), its allele frequency, if it is a somatic or a germline insertion, and its integrity (i.e. full length, 5′-truncated, 3′-truncated or internal fragment). Each SRIP is given a unique ID in euL1db, which is prefixed by srip (e.g. srip34564). Because several SRIP might actually correspond to the same original insertion event, some have identical or close genomic coordinates (e.g. srip 1, 3, 6, 8 and 10 in Figure 1A). To reduce this redundancy and to facilitate comparisons within and across studies, a set of virtual insertions named 'meta-retrotransposon insertion polymorphism' or MRIP has been computationally generated (Figure 1A and B). An MRIP refers to a unique genomic interval, which contains overlapping or close SRIP, likely corresponding to the same original insertion event. In practice, nearby SRIPs are merged into a single MRIP if they satisfy all the following requirements: (i) they are located within 200 bp of each other, (ii) they share the same strand orientation, and (iii) they are all germline insertions. Somatic retrotransposition events are unique by nature, and are not merged with germline events, nor merged together. Therefore, somatic SRIPs give rise to MRIPs containing only a single SRIP. Using a 200-bp window around SRIP rather than precise coordinates was necessary since different methods and studies have variable accuracy in defining the precise location of L1HS insertions. The rational for choosing the size of this window is detailed in the Supplementary Methods. Each MRIP is given a unique ID in euL1db, which is prefixed by mrip (e.g. mrip1234). Although the probability of finding two independent germline insertion events in the same 200-bp window is extremely low, it is not null. The euL1db framework allows users to compare annotations provided for each SRIP within a given MRIP. Depending on the study, SRIP annotations may include the length and/or sequence of the TSD, the reverse-transcribed L1 sequence or other additional potential rearrangements (inversion, transduction). In a situation where distinct insertion events were wrongly combined in a single MRIP, discrepancies in the SRIPs annotations could alert the user that caution should be taken. This also applies for the most extreme case, i.e. two independent insertion events occurring at the same exact nucleotide. Since reference L1HS insertions are virtual insertions derived from a consensus reference sequence and not from a biological sample, we have chosen to include them in a distinct table, entitled the 'Reference' table, and to assign them an ID prefixed by ref (e.g. ref123). This table is used internally to determine whether a given SRIP or MRIP actually corresponds to a reference L1HS insertion, and to annotate each record. The total number of SRIPs and MRIPs included for each study is graphed in Supplementary Figure S3.

Table 1. euL1db content statistics

Record type	Number of records
Studies	32[a]
Samples	943
Individuals	741
Families	50[b]
SRIP	142,495
MRIP	8991
Reference L1HS	1545

[a]Out of 32 studies, 10 used high-, 1 medium- and 21 low-throughput approaches.
[b]With at least two individuals analyzed.

In addition to these main tables, euL1db uses a 'Method' table, which contains the methods used to call SRIPs in the different studies, and a 'Family' table, which classifies the familial relationships between euL1db individuals (mostly from the 1000 Genomes Project). An individual without known relative in euL1db is not linked to any family.

The data contained in euL1db originate from peer-reviewed publications and have been manually curated and entered. The source of data and the curation process are detailed in Supplementary Methods and in Supplementary Table S1. The reference L1HS insertions have been processed from the UCSC RepeatMasker track table. The summary statistics at the time of writing are displayed in Table 1.

DATA ACCESS

euL1db can be interrogated through a user-friendly Web Server (http://euL1db.unice.fr). A set of detailed tutorials and examples of use are accessible from the 'Help' tab. The detailed description of the Web Server architecture is described in Supplementary Figure S1.

There are several ways to query euL1db: (i) by searching SRIP or MRIP located in a single locus (genomic region, gene) or in a single individual ('Search' tab); (ii) by browsing the different tables and using filters to select a specific subset of data across and within studies, families, individuals, samples, insertions ('Browse' tab); (iii) by batch query using a list of multiple loci (genomic coordinates) or genes (gene names) ('Utilities' tab).

Users can choose to display L1 insertions as SRIP or MRIP in (i) graphical- (UCSC genome browser, dbVar genome browser) (40,41); (ii) tabular- (sortable html tables); or (iii) text-formats (including in standard BED format for subsequent analyses with other tools). Tables can be customized to display the information of interest for the user.

RELATIONSHIP AND DIFFERENCES WITH OTHER DATABASES

Several resources are related to—but distinct from—euL1db. Repbase is a database of consensus repetitive DNA sequence and as such does not contain any localization information (42). One of its entries is the L1HS consensus sequence and has been subsequently used to annotate the human reference genome and to identify the genomic loci corresponding to L1HS elements

(Smit, A.F.A., Hubley, R. & Green, P. *RepeatMasker Open-3.0*. 1996–2010 <http://www.repeatmasker.org>). This information is available through the RepeatMasker table of the UCSC Genome Browser (40). The reference L1HS elements included in the 'Reference' table of euL1db have been processed and annotated using the latter. dbRIP was an early effort to catalog and annotate polymorphic retrotransposon insertions in humans (43). In contrast to dbRIP, euL1db stores data in a sample-wise manner and contains the most recent data sets obtained by high-throughput sequencing, including those from the 1000 Genomes Project. Although dbRIP could not be directly included in euL1db since samples are not documented in dbRIP, 94% of dbRIP L1HS records have an MRIP equivalent in euL1db (Figure 1C). dbRIP has also unique features since it contains non-autonomous human retrotransposons such as Alu or SVA sequences and not only L1HS insertions. As a particular case of structural variation, L1 retrotransposon insertions are also documented in dbVar/DGVa or DGV as mobile element insertions (MEI) (41). The data structure logics in dbVar/DGVa and euL1db are comparable (sample-wise variants and merged variants). euL1db is specialized for L1HS insertions, while dbVar/DGVa can include any type of structural variants, including MEI. However, the set of information for L1HS insertions provided by euL1db is much more exhaustive, and only a single study (out of 32 at the time of writing) stored in euL1db was also deposited in dbVar/DGVa.

CONCLUDING REMARKS

High-throughput sequencing technologies have considerably fostered the study of L1-mediated genomic variation and its impact on human health. We anticipate that this trend will continue in the next years, particularly with the availability of long-reads sequencing approaches, which might greatly facilitate the detection of L1HS insertions and their accurate positioning on the genome by generating reads that span the entire element and both flanking regions. In this respect, euL1db database and server have been tailored to support a considerable increase of SRIP, while keeping a fast-response time. To summarize, euL1db provides a centralized and user-friendly access to known germline and somatic L1HS insertions, which will be critical to elucidate the physiological or pathological impact of novel L1HS insertions. This resource will be useful in a large variety of fields such as human genetics, neurosciences or cancer genomics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We are grateful to the L1 community and to the 1000 Genomes Consortium for data sharing. The authors also thank R. Manas and M. Ducellier (University of Nice-Sophia Antipolis) for assistance with the IT infrastructure. We thank Aurelien Doucet for critical reading.

FUNDING

Institut National de la Santé Et de la Recherche Medicale and the Institut National du Cancer [2009–340 to G.C.]; the European Research Council [243312 to G.C.]; Agence Nationale pour la Recherche [ANR-11-LABX-0028–01 to G.C.]. Funding for open access charge: European Research Council [243312 to G.C.].
Conflict of interest statement. None declared.

REFERENCES

1. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V. and Kazazian,H.H.J. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl. Acad. Sci. U.S.A.*, **100**, 5280–5285.
2. Burns,K.H. and Boeke,J.D. (2012) Human transposon tectonics. *Cell*, **149**, 740–752.
3. Hancks,D.C. and Kazazian,H.H. (2012) Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.*, **22**, 191–203.
4. Beck,C.R., Garcia-Perez,J.L., Badge,R.M. and Moran,J.V. (2011) LINE-1 elements in structural variation and disease. *Annu. Rev. Genomics Hum. Genet.*, **12**, 187–215.
5. Hohjoh,H. and Singer,M.F. (1997) Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J.*, **16**, 6034–6043.
6. Mathias,S.L., Scott,A.F., Kazazian,H.H., Boeke,J.D. and Gabriel,A. (1991) Reverse transcriptase encoded by a human transposable element. *Science*, **254**, 1808–1810.
7. Feng,Q., Moran,J.V., Kazazian,H.H. and Boeke,J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905–916.
8. Kulpa,D.A. and Moran,J.V. (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat. Struct. Mol. Biol.*, **13**, 655–660.
9. Doucet,A.J., Hulme,A.E., Sahinovic,E., Kulpa,D.A., Moldovan,J.B., Kopera,H.C., Athanikar,J.N., Hasnaoui,M., Bucheton,A. *et al.* (2010) Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet.*, **6**, e1001150.
10. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
11. Cost,G.J., Feng,Q., Jacquier,A. and Boeke,J.D. (2002) Human L1 element target-primed reverse transcription in vitro. *EMBO J.*, **21**, 5899–5910.
12. Gilbert,N., Lutz-Prigge,S. and Moran,J.V. (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell*, **110**, 315–325.
13. Symer,D.E., Connelly,C., Szak,S.T., Caputo,E.M., Cost,G.J., Parmigiani,G. and Boeke,J.D. (2002) Human l1 retrotransposition is associated with genetic instability in vivo. *Cell*, **110**, 327–338.
14. Ostertag,E.M. and Kazazian,H.H. (2001) Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.*, **11**, 2059–2065.
15. Moran,J.V., DeBerardinis,R.J. and Kazazian,H.H.J. (1999) Exon shuffling by L1 retrotransposition. *Science*, **283**, 1530–1534.
16. Goodier,J.L., Ostertag,E.M. and Kazazian,H.H. (2000) Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.*, **9**, 653–657.
17. Repanas,K., Zingler,N., Layer,L.E., Schumann,G.G., Perrakis,A. and Weichenrieder,O. (2007) Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res.*, **35**, 4914–4926.
18. Monot,C., Kuciak,M., Viollet,S., Mir,A.A., Gabus,C., Darlix,J.L. and Cristofari,G. (2013) The specificity and flexibility of L1 reverse transcription priming at imperfect T-tracts. *PLoS Genet.*, **9**, e1003499.
19. Viollet,S., Monot,C. and Cristofari,G. (2014) L1 retrotransposition: the snap-velcro model and its consequences. *Mob. Genet. Elem.*, **4**, e28907.

20. O'Donnell,K.A. and Burns,K.H. (2010) Mobilizing diversity: transposable element insertions in genetic variation and disease. *Mob. DNA*, **1**, 21.
21. Ray,D.A. and Batzer,M.A. (2011) Reading TE leaves: new approaches to the identification of transposable element insertions. *Genome Res.*, **21**, 813–820.
22. Ewing,A.D. and Kazazian,H.H. (2011) Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.*, **21**, 985–990.
23. Stewart,C., Kural,D., Strömberg,M.P., Walker,J.A., Konkel,M.K., Stütz,A.M., Urban,A.E., Grubert,F., Lam,H.Y., Lee,W.P. *et al.* (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.*, **7**, e1002236.
24. Beck,C.R., Collier,P., Macfarlane,C., Malig,M., Kidd,J.M., Eichler,E.E., Badge,R.M. and Moran,J.V. (2010) LINE-1 retrotransposition activity in human genomes. *Cell*, **141**, 1159–1170.
25. Huang,C.R., Schneider,A.M., Lu,Y., Niranjan,T., Shen,P., Robinson,M.A., Steranka,J.P., Valle,D., Civin,C.I., Wang,T. *et al.* (2010) Mobile interspersed repeats are major structural variants in the human genome. *Cell*, **141**, 1171–1182.
26. Iskow,R.C., McCabe,M.T., Mills,R.E., Torene,S., Pittard,W.S., Neuwald,A.F., Van Meir,E.G., Vertino,P.M. and Devine,S.E. (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, **141**, 1253–1261.
27. Ewing,A.D. and Kazazian,H.H. (2010) High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.*, **20**, 1262–1270.
28. Kazazian,H.H., Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G. and Antonarakis,S.E. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, **332**, 164–166.
29. Kaer,K. and Speek,M. (2013) Retroelements in human disease. *Gene*, **518**, 231–241.
30. Coufal,N.G., Garcia-Perez,J.L., Peng,G.E., Yeo,G.W., Mu,Y., Lovci,M.T., Morell,M., O'Shea,K.S., Moran,J.V. and Gage,F.H. (2009) L1 retrotransposition in human neural progenitor cells. *Nature*, **460**, 1127–1131.
31. Baillie,J.K., Barnett,M.W., Upton,K.R., Gerhardt,D.J., Richmond,T.A., De Sapio,F., Brennan,P.M., Rizzu,P., Smith,S., Fell,M. *et al.* (2011) Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, **479**, 534–537.
32. Evrony,G.D., Cai,X., Lee,E., Hills,L.B., Elhosary,P.C., Lehmann,H.S., Parker,J.J., Atabay,K.D., Gilmore,E.C., Poduri,A. *et al.* (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, **151**, 483–496.
33. Shukla,R., Upton,K.R., Muñoz-Lopez,M., Gerhardt,D.J., Fisher,M.E., Nguyen,T., Brennan,P.M., Baillie,J.K., Collino,A., Ghisletti,S. *et al.* (2013) Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell*, **153**, 101–111.
34. Solyom,S., Ewing,A.D., Rahrmann,E.P., Doucet,T., Nelson,H.H., Burns,M.B., Harris,R.S., Sigmon,D.F., Casella,A., Erlanger,B. *et al.* (2012) Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.*, **22**, 2328–2338.
35. Lee,E., Iskow,R., Yang,L., Gokcumen,O., Haseley,P., Luquette,L.J., Lohr,J.G., Harris,C.C., Ding,L., Wilson,R.K. *et al.* (2012) Landscape of somatic retrotransposition in human cancers. *Science*, **337**, 967–971.
36. Miki,Y., Nishisho,I., Horii,A., Miyoshi,Y., Utsunomiya,J., Kinzler,K.W., Vogelstein,B. and Nakamura,Y. (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.*, **52**, 643–645.
37. Rodić,N. and Burns,K.H. (2013) Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? *PLoS Genet.*, **9**, e1003402.
38. Goodier,J.L. (2014) Retrotransposition in tumors and brains. *Mob. DNA*, **5**, 11.
39. Helman,E., Lawrence,M.S., Stewart,C., Sougnez,C., Getz,G. and Meyerson,M. (2014) Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.*, **24**, 1053–1063.
40. Karolchik,D., Barber,G.P., Casper,J., Clawson,H., Cline,M.S., Diekhans,M., Dreszer,T.R., Fujita,P.A., Guruvadoo,L., Haeussler,M., *et al.* (2014) The UCSC genome browser database: 2014 update. *Nucleic Acids Res.*, **42**, D764–D770.
41. Lappalainen,I., Lopez,J., Skipper,L., Hefferon,T., Spalding,J.D., Garner,J., Chen,C., Maguire,M., Corbett,M., Zhou,G. *et al.* (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res.*, **41**, D936–D941.
42. Jurka,J., Kapitonov,V.V., Pavlicek,A., Klonowski,P., Kohany,O. and Walichiewicz,J. (2005) Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.*, **110**, 462–467.
43. Wang,J., Song,L., Grover,D., Azrak,S., Batzer,M.A. and Liang,P. (2006) dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.*, **27**, 323–329.

euL1db: the European database of L1HS retrotransposon insertions in humans

Ashfaq A. Mir, Claude Philippe and Gaël Cristofari

SUPPLEMENTARY DATA

Supplementary Data online contains a "Supplementary Methods" section, 2 Supplementary Figures and 1 Supplementary Table.

Supplementary Methods

Window size selection to build MRIP records. Since each study use a distinct set of methods to detect L1 insertions and specific algorithms to define the junctions between an L1 copy and its flanking sequence, identical L1 insertions might be reported with slightly different coordinates in different datasets. For this reason, each MRIP was built by merging SRIPs that fall into a 200-bp sliding window. To choose the size of this window, we evaluated the distribution of the SRIP-to-SRIP distances in euL1db. We first extracted non-redundant SRIPs from each Study. Practically, if a Study includes multiple SRIPs with the same exact genomic coordinates, we kept only one of them. This step prevents Studies with a large number of samples to be overrepresented in the distribution. Then, we used Bedtools (1) to form clusters of SRIPs falling in a 1 kb-window (bedtools cluster with option –d 1000). Clusters with a single SRIP were excluded. Finally, we calculated the distance between each pair of non-redundant SRIP pair falling in a given cluster (using bedtools closest with option -d). Strand orientation was not taken into account since strand information is missing for some SRIPs. The distribution of the SRIP-to-SRIP distances is graphed in Supplementary Figure 2 and indicates that 95% of the SRIP are less than 200-bp distant (median: 7bp; 75% quartile: 27 bp; 95% percentile: 214 bp). We choose the 200 bp value to keep acceptable insertion site accuracy and to avoid splitting a unique biological event into artificially distinct records.

Data curation and inclusion. Data included in euL1db were often directly extracted from the main or supplementary data of the original publications. When positions were reported in hg18 reference genome coordinates, we translated them into GRCh37/hg19 coordinates using the liftOver tool of the UCSC Genome Browser website (available at: https://genome.ucsc.edu/cgi-bin/hgLiftOver) (2). Only L1HS insertions were kept. Putative insertions that were tested by PCR and/or Sanger sequencing but failed these additional validations were excluded. Some studies required additional processing, which are detailed below. The sources of high-throughput data included in euL1db at the time of writing are summarized in Supplementary Table 1.

The original Beck *et al.* 2010 publication (3) only reports a short DNA sequence corresponding to the preintegration site, and its chromosome number and cytogenetic band. We remapped the preintegration site sequence to reference genome hg19 with BWA (4) to obtain the precise genomic

coordinates. For sequences with multiple possible positions (MAPQ=0), we selected the position consistent with the reported cytogenetic band.

The original Iskow *et al.* 2010 publication (5) provides the DNA sequence downstream of each putative L1 insertions obtained by 454 sequencing (in opposite orientation relative to L1) and the genomic coordinates of the BLAT best hit after mapping them on the hg18 reference genome. To obtain the strand information of these putative insertions, we remapped the published DNA sequences (1389 in total) to hg19 using BWA. In a first step, we used bwa mem (bwa mem with options -t4 -M). We recovered 980 uniquely mapped positions consistent with the original BLAT analysis; 285 unmapped sequences; and 124 sequences mapping to multiple positions or corresponding to chimeric sequence. The latter were discarded and not included in euL1db. In a second step, unmapped sequences from the first step were mapped again using an algorithm with increased sensitivity for short sequences (bwa aln with options -l12 -o2 / bwa samse with –n 10 option), allowing us to recover an additional set of 116 uniquely mapping sequences consistent with the original BLAT coordinates. In total, 1095 high-confidence insertions out of 1389 sequences from the 454 Iskow experiments have been included in euL1db. The published table reporting the results of ABI Sanger sequencing experiments includes the coordinates of the flanking region sequenced but not the DNA sequence itself, nor its orientation. We used the middle of this segment as a coordinate for the insertion point and we could not deduce strand information.

The Baillie *et al.* publication (6) contains two distinct sets of data obtained by RC-seq: (i) germline polymorphic retrotransposon insertions discovered in a pool of genomic DNA isolated from the blood of several individuals; and (ii) putative somatic and germline retrotransposon insertions discovered in genomic DNA isolated from different brain regions of three individuals. Because many of the putative somatic insertions are only supported by a single sequencing read, we only kept in euL1db somatic insertions that have been validated by PCR and/or Sanger sequencing. In contrast, germline insertions were more robustly identified (several sequencing reads in multiple independent libraries) and were kept, whether further tested by PCR or not. Only L1HS elements were included (L1-Ta and L1-Pre-Ta).

Similarly the Solyom *et al.* article (7) reports insertions found in multiple colon cancer samples or their matched normal tissues. This was achieved by a combination of L1-seq and RC-seq approaches. For the L1-seq approach (their Sup. Tab. 1A and 1B), we kept as germline insertions, those found in both normal and tumor tissues and, as somatic insertions, those found only in one tissue and validated by PCR and/or Sanger sequencing (their Sup. Tab. 3). For the RC-seq data (their Sup. Table 2), we only kept those tagged as "high confidence" and we further added/removed insertions validated/invalidated by PCR and/or Sanger sequencing data (their Sup. Tab. 3). Here as well, only L1HS elements were included.

Some studies have searched L1 insertions in pooled samples (obtained from different individuals, cell lines, etc…) as a facilitated mean to find common inherited L1 polymorphisms. In these cases,

insertions cannot be linked to samples/individuals. Such samples were recorded in euL1db but tagged as “Mix” in their sample relationship field.

Supplementary References

1. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

2. Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res*, **42**, D764-D770.

3. Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M. and Moran, J.V. (2010) LINE-1 Retrotransposition Activity in Human Genomes. *Cell*, **141**, 1159-1170.

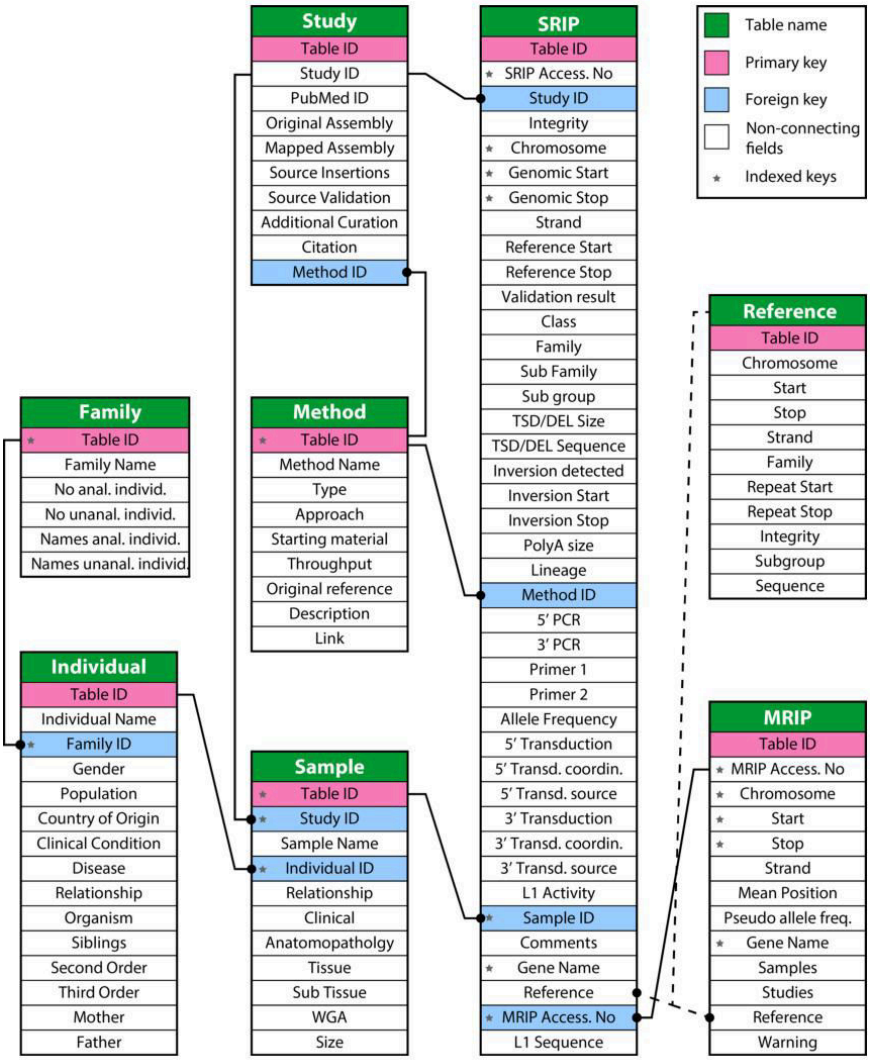
4. Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**, 589-595.

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

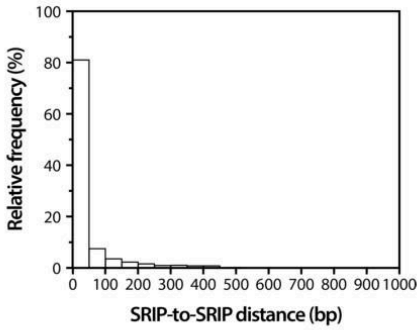
6. Baillie, J.K., Barnett, M.W., Upton, K.R., Gerhardt, D.J., Richmond, T.A., De Sapio, F., Brennan, P.M., Rizzu, P., Smith, S., *et al.* (2011) Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, **479**, 534-537.

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

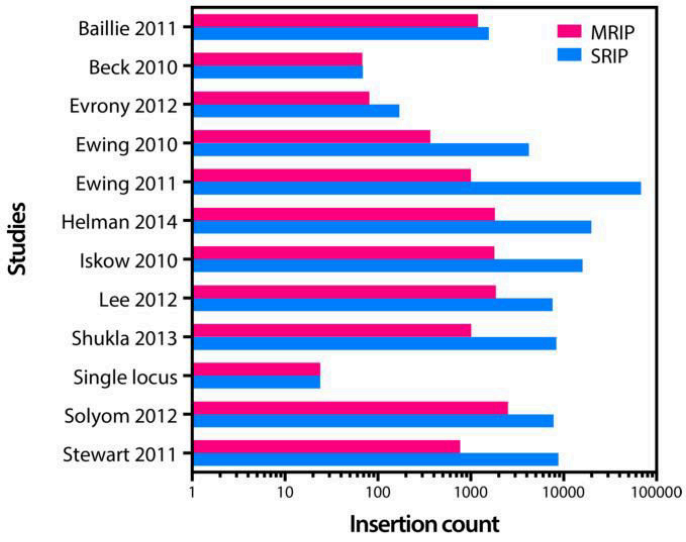
Supplementary Figures



Supplementary Figure 1. euL1db table structure. Data have been organized in 8 Meta tables (names in green) tailored to allow efficient data mining and reduce processing time by built-in functions. Meta data tables distribute the information into chunks of systematic and interconnected data records. Primary keys are shown in pink, foreign keys are shown in blue and are linked by plain lines to their original table. Indexed keys are marked with a star. Fields related to a single table are presented with a white background and carry information specific to each stored data type. Meta tables form a light layer of information network for zooming down to a particular detail without processing all of the data. The "reference" fields of the SRIP and MRIP tables are filled non-dynamically (dashed line) with the "Reference" table, which contains the reference L1HS insertions. This data structure allows maximum amount of information to be stored without compromising processing time and memory utilization by interrogating external programs.



Supplementary Figure 2. Distribution of SRIP-to-SRIP distances in euL1db. Since 95% of SRIP are less than 200 bp distant from each others, we choose this window size to merge SRIP into MRIP (see Fig. 1B). The detailed procedure is explained in Supplementary Methods.



Supplementary Figure 3. Count and origin of L1HS insertions included in euL1db.

Supplementary Table

Study.ID	PMID	Source of insertions data	Source of validation data	Coordinates liftover	Additional processing
Baillie2011	22037309	Tables S4, S5	Tables S6, S7	-	See text
Beck2010	20602998	Table S2	Table S2	hg18 to hg19	See text
Evrony2012	23101622	Table S3	Table S3	-	-
Ewing2010	20488934	Table S1	Table S1	hg18 to hg19	-
Ewing2011	20980553	Table S4	-	hg18 to hg19	-
Helman2014	24823667	Tables 2, 3, 5*	Table 1	hg18 to hg19 **	
Iskow2010	20603005	Tables S1, S2	Tables S1, S2	hg18 to hg19	See text
Lee2012	22745252	Tables S2, S6, S8	Tables S3, S4, S7	hg18 to hg19	-
Shukla2013	23540693	Table S3	Tables S5, S6	-	-
Solyom2012	22968929	Tables 1, S1A, S1B, S2	Table S3	-	See text
Stewart2011	21876680	Table S1	Table S1	hg18 to hg19	DEL excluded[§]

Supplementary Table 1. Source data for high- and medium-troughput L1 mapping studies.
*, Table numbers are conflicting in the Supplementary Data of this article. **, Only for 3 tissues (COAD, READ, LAML). Other samples were already mapped to the hg19 assembly. §, In this study, mobile element insertion (MEI) annotated as DEL are insertions present in the reference genome, but absent from a particular sample.

3. A computational approach to reveal the landscape of transcriptional isoforms induced by L1 elements in human cells

3.1. Context of the study

LINE-1 (L1) retrotransposons constitute almost 17% of our genome and constitute the most abundant family of autonomously replicating retroelements in mammals. They had a significant impact in the organization and functioning of the mammalian genomes by continuously amplifying over the last ~170 million years (2–4). L1 element replicates via an RNA intermediate that is copied into genomic DNA at the site of insertion. Detailed analysis of mutational mechanisms indicates that approximately 20–30% of structural variations are caused by non-LTR retrotransposons (20–23). L1 elements can affect our genome in many ways. Firstly, if an insertion takes place within an exon it can modify the coding sequence. L1 3' transduced sequence can fit into new sites, which can either be used as an exon, or it can modulate the gene expression by providing regulatory sequences. It has been estimated that ~1% of human genome DNA has been transduced by L1 that is interestingly comparable to the exonic percentage in the genome.

Studying the consensus sequence of the L1 element has revealed many acceptor or donor sites for alternative splicing. L1 can generate numerous transcripts of variable size that could possibly be due to alternative splicing of the L1 sequence, which contains cryptic acceptor site and splice donor and some of them proven to be functional (30). Splicing can, therefore, change the L1 RNA after transcription and thus limit its impact by creating non-active RNA.

A premature polyadenylation of the transcript of the gene harboring L1 insertion may lead to translation of new isoform of the protein encoded by this gene. L1s contains antisense promoter (ASP) within their 5' UTR. This ASP provides alternative transcription start site for many human genes. L1 transcripts have been detected in different types of human cancer (e.g. testis, bladder and liver cancers) as well as in many cancer cell lines (65). The transposable elements are therefore a source of genetic variation and that these different mechanisms have helped to change the regulation of genome transcriptomics (70).

Therefore, there is a need for such a method which can help us to get a set of L1 chimeric transcripts within a given sample and also tell us which type of alternate splicing events could possibly be there due to newly integrated LINE-1 at a particular locus in the genome.

One of the ways to identify hallmarks of actively jumping LINE-1 insertions is by using split and discordant read pairs, which contain a piece of L1 in the RNA-seq

data. There are many published approaches, which use this information to detect novel non-reference L1 Insertions from RNA-seq data like Tea, RetroSeq, and Mobster etc. However, there is no approach till date, which can pinpoint transcriptional isoforms due to newly, integrated LINE-1 elements.

Therefore, we have developed a novel computational approach, which uses discordant and split read pair information from RNA-seq data and couples it with two-tier ultra sensitive transcriptome assembly both with and without chimeric reads to identify the L1 chimeric transcripts. We also locate the antisense transcripts and then annotate the assembled chimeric transcripts for different alternate splicing events, which might be due to LINE-1.

As a lead author for this unpublished work, I did most of the work, which included theoretical aspect, development of the algorithm, setting up of the software application and writing the analysis scripts for different modules.

The extent of L1 chimeric transcript formation and the landscape of the affected genes remain unexplored. Our work will shed light on the following questions in the long run: 1- what proportion of L1 copies lead to tumor-specific L1 chimeric transcripts? 2- what are the dominant forms of transcript alternations resulting from L1 element in cancer transcriptomes? 3- Do L1 chimeric transcripts give rise to novel isoforms of cancer-related genes?

Finally, It is currently unknown how the full set of L1 elements present in a given individual is regulated at the transcriptional level, and which cellular or genomic environment is permissive for their expression. In addition, although many somatic insertions have been described in several tumor types, their overall impact on gene expression has been only poorly explored. Our software will shed light on these two aspects.

3.2. Article-III

A computational approach to reveal the landscape of transcriptional isoforms induced by LINE-1 elements in human cells

Ashfaq A. Mir[1,2,3] and Gaël Cristofari[1,2,3,*]

[1]INSERM, U1081, Institute for Research on Cancer and Aging of Nice (IRCAN), F-06100 Nice, France,

[2]CNRS, UMR 7284, Institute for Research on Cancer and Aging of Nice (IRCAN), F-06100 Nice, France and

[3]Faculty of Medicine, Institute for Research on Cancer and Aging of Nice (IRCAN), University of Nice-Sophia-Antipolis, F-06100 Nice, France

* To whom correspondence should be addressed.
Tel: +33 4 93 37 70 87; Fax: +33 4 93 37 70 92
Email: Gael.Cristofari@unice.fr

Keywords: Retrotransposons; Antisense transcription; Chimeric transcripts; RNA-seq; Epigenetics; Noncoding RNA; Alternative splicing; Somatic mutations; Bioinformatics

ABSTRACT

LINE-1 (L1) retrotransposons, which compose almost 17% of our genome, are the only active and autonomous family of mobile genetic elements in humans. They are mobilized in germ cells - but also in some somatic tissues. They contribute to human genetic diversity and can occasionally lead to disease, such as cancer. L1 reactivation can drive genomic instability through novel somatic insertions, which can be directly mutagenic by disrupting genes or regulatory sequences. In addition, L1 sequences contain many regulatory cis-elements (sense and antisense promoters, polyadenylation signals, cryptic splicing sites). Therefore, L1 insertions near a gene or within intronic sequences can also produce more subtle genic alterations. This phenomenon is not limited to tumor-specific L1 insertions: even the de-repression or activation of existing and inherited L1 copies in tumors can contribute to cancer progression by altering the expression of their neighboring genes, notably by generating L1 chimeric transcripts. Here, we present a new RNA-seq analysis pipeline that can: (i) identify L1 chimeric transcripts; (ii) annotate *de novo* assembled chimeric transcripts for different alternative splicing events; and (iii) locate anti-sense transcripts. This method could find 3189 chimeric transcripts in Breast cancer cell line (MCF7) and 2957 chimeric transcripts in human embryonic cancer cell line 2102Ep (min transcript length 500 bp). This work will help in understanding the mechanisms leading to transcriptome plasticity in tumor cells and will provide a rational basis for the use of retrotransposon chimeric transcripts as biomarkers.

INTRODUCTION

LINE-1 (L1) retrotransposons constitute the most abundant family of autonomously replicating retroelements in mammals. Their continuous amplification over the last ~170 million years (Myr) has had a significant impact on the organization and function of mammalian genomes (2–4). L1 element replicates via an RNA

intermediate which is copied into genomic DNA at the site of insertion (5–7). L1 also mobilizes in trans short-interspersed elements (SINEs), such as *Alu* or SVA sequences. Mutational mechanisms indicate that approximately 20–30% of structural variations are caused by non-LTR retrotransposons (20–23). *Alu*, L1, and SVA retrotransposition rates are estimated to be one in 21 births, 212 births, and 916 births, respectively. L1 element may be a source of variability for the genome through various mechanisms. First, an L1 element or a SINE can insert within an exon and modify the coding sequence of a gene (376). Second, L1 has the ability to generate 3' transductions and can, therefore, fit into new sites and copy sequences from their original locus. The transduced sequence may have several effects: it can be used either as an exon, or it can modulate gene expression by providing regulatory sequences at the site of new insertion (16, 17). It has been estimated that ~1% of human genome DNA has been transduced by L1, a proportion comparable to the percentage of exons in our genome. This highlights the role of L1 in the genomic plasticity by shuffling genomic DNA (27). L1-mediated 3' transductions which are in the downstream sequence comprise ~25% of tumors in cancer genomes as per the analysis by Tubio (318).

In addition, L1 intronic insertions can significantly alter transcript splicing through (i) intron retention, a process by which, an entire intron sequence is maintained in the mature transcript; or (ii) exonization of an intronic region, or (iii) by exon skipping (410). It has been estimated that 92-94 % of human genes exhibit alternative splicing, ~86 % with a minor isoform frequency of around 15% (29). L1 can generate numerous transcripts of variable size that could possibly be due to alternative splicing of the L1 sequence, which contains cryptic acceptor site and splice donor and some of them have been proven to be functional (30). There are approximately 95% of human multi-exonic genes that are alternatively spliced (402). Introduction of new splicing sites by retrotransposons can result in a severe gene disruption as well as in new coding and non-coding gene creation (31, 33–35, 229). Splicing can, therefore, change the L1 RNA after transcription and thus limit its impact by creating non-active RNA. On the other hand, the study of ESTs (Expressed-sequenced tags) showed that the L1 splicing sites inserted into genes may be used during the maturation of gene transcripts, which is a mechanism by which L1s may contribute to the plasticity of our transcriptome (30).

RNA Pol-II transcription of LINE-1 is negatively affected by numerous termination and polyadenylation signals present along the L1 sequence (36). Some of these sites appear to be much stronger than the relatively weak polyA site found at the 3' end of the LINE-1 element (37). The L1 sequence is, therefore, a "difficult" DNA template for cellular RNA polymerase II (Pol II). Nuclear export and translation efficiencies are influenced by polyadenylation, which stabilizes mRNA transcripts. Human genes vastly use alternative polyadenylation sites, and transposable elements embed these signals, which suggests that TEs can influence the 3′ end processing of host gene transcripts (38).

It has ben suggested that up to 18% of human genes have alternative promoters (170). L1 bi-directional promoter greatly contributes to this diversity of alternative transcript initiation sites. For example, insertion of an L1 in an intron in the reverse

orientation of transcription of the gene may result in a gene breakage phenomenon (407). L1s contains antisense promoter (ASP) within their 5' UTR. This ASP provides alternative transcription start site for several human genes like *c-MET*, a receptor tyrosine kinase whose activation can cause tumorigenicity in a variety of tumors (39–42). In a similar way, L1 ASP has been shown to serve as an alternative promoter for more than 40 human genes in a tissue-specific manner (39, 40, 174, 408).

Transposable elements play a critical role in engineering transcriptional networks, permitting coordinated gene expression, and facilitating the evolution of novel physiological processes. TE-derived exons are tissue specific and L1 expression is not uniform throughout the body adding another layer of complexity to our transcriptome (28). TE-derived retrogenes act as an evolutionary toolbox to promote transcript diversity (55–58). Retrogenes embedded within host gene introns can influence transcription and cause premature upstream transcript polyadenylation. This mechanism can indirectly influence mRNA processing, and the landscape of alternative transcripts. Small RNAs also play crucial role in the complex regulatory network of gene expression in all organisms (412).

Transposable elements can provide ready to use transcription factor binding sites, which expend the physiological or pathological conditions in which a target gene can be regulated (43–46). It has been already shown that binding sites for 5 transcription factors (ESR1, TP53, POU5F1, SOX2, and CTCF) are embedded within many families of transposable elements (175).

Altogether, transposable elements drive the evolution of our transcriptome, and can even create new genic networks by bringing regulatory elements that can eventually respond to similar biological signals.

To obtain a comprehensive view of TE impact on the landscape of human alternative transcripts, we have developed a method for precisely identifying chimeric transcript isoforms resulting from TE structural variations in RNA-Seq data. We applied this approach to L1 elements and to two different cell lines naturally expressing high levels of L1. We identified 3189 chimeric transcripts in MCF7 cells, isolated from breast cancer, and 2957 chimeric transcripts in 2102Ep embryonal carcinoma cells (with a minimal transcript length of 500 bp).

MATERIALS AND METHODS

Algorithm

Briefly, our approach uses discordant and split read pair information to identify L1 chimeric transcripts. The originality of our method is that we combine this information with two successive *de novo* transcriptome assemblies. The first one is built with all RNA-seq reads, while the second one is constructed without the chimeric reads identified in the first step. Finally, we compare the two assemblies to identify isoforms directly created by the presence of L1 insertions.

Step-1: Super read creation (Optional):

This is an optional step in the pipeline. However, super read creation, which is used in many *de novo* genome assembly algorithmic techniques, improves the quality of transcriptome assembly especially in case of poorly sized fragment libraries and also of short reads with different lengths (413), even when genome-guided approach is used in StringTie, our best choice of transcriptome assembler (414). Super read creation has been shown to allow error free assembly of longer scaffolds. To test the ability of super reads to improve chimeric transcript detection, we implemented algorithmic techniques from *de novo* genome assembly to transcriptome assembly, using the super-read module of MaSuRCA genome assembler (413), which extends every read in both directions as long as this extension is unique. The "*superreads.pl*" script identifies pairs of reads that belong to the same super-read, and then extracts the sequence containing the pair plus the sequence between them; i.e., the entire sequence of the original DNA fragment (Figure 1).

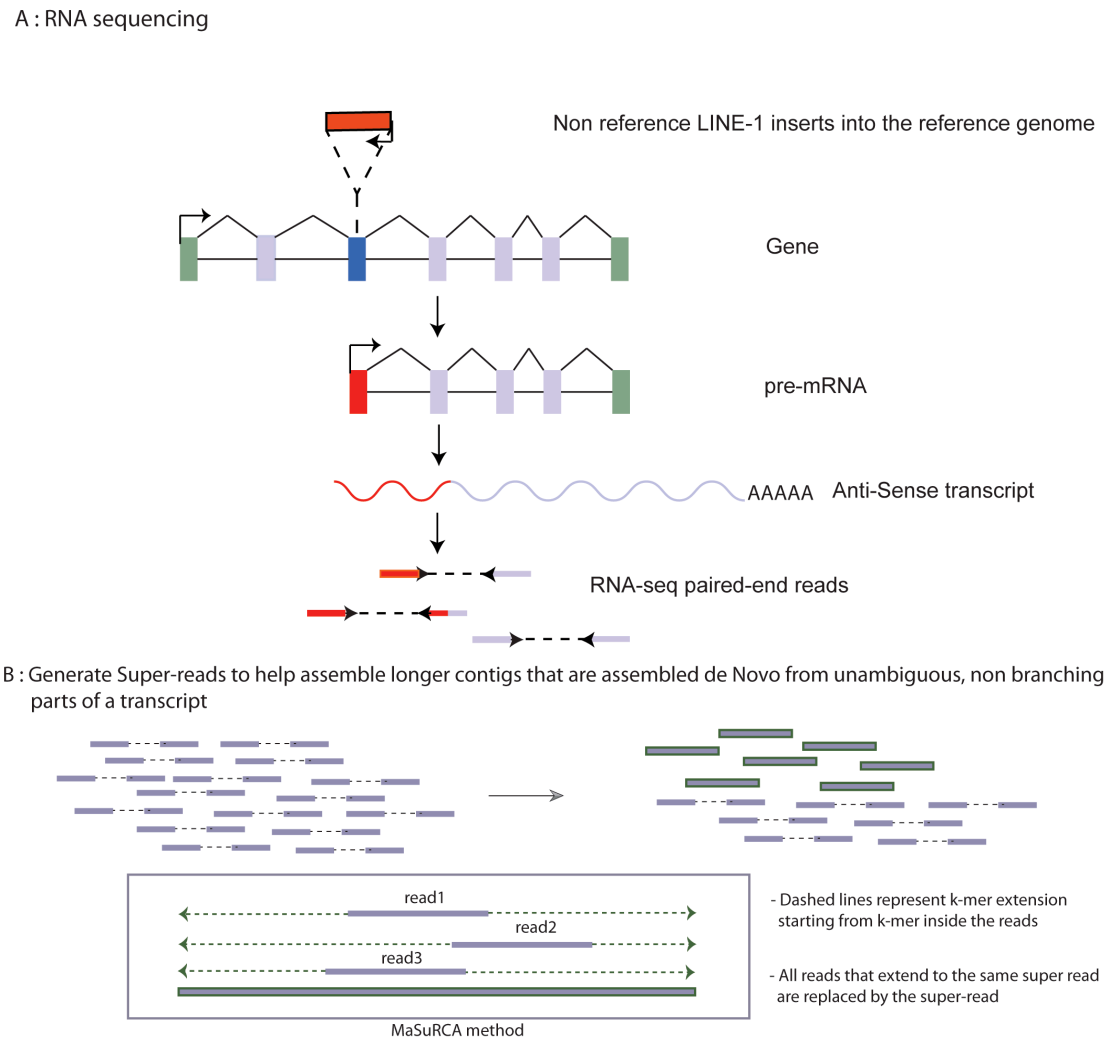


Figure 1: RNA-seq and super read creation.
(A) Scheme showing the insertion of a novel L1 into a gene followed by RNA-sequencing. Thus mate pairs capture fragment of this novel insertion either as discordant or split reads.

(B) MaSuRCA script is used to generate super reads. The box below depicts how unique ends are extended to form super reads.

Step-2: Identify chimeric reads in RNA-seq reads:

Read pairs are first mapped against a set of LINE-1 consensus sequence and then against human genome reference by HISAT (415). Discordant and split read pairs are identified (Figure 2). Exon aware spliced alignment is performed in the case of genome mapping. This step is performed to identify chimeric reads.

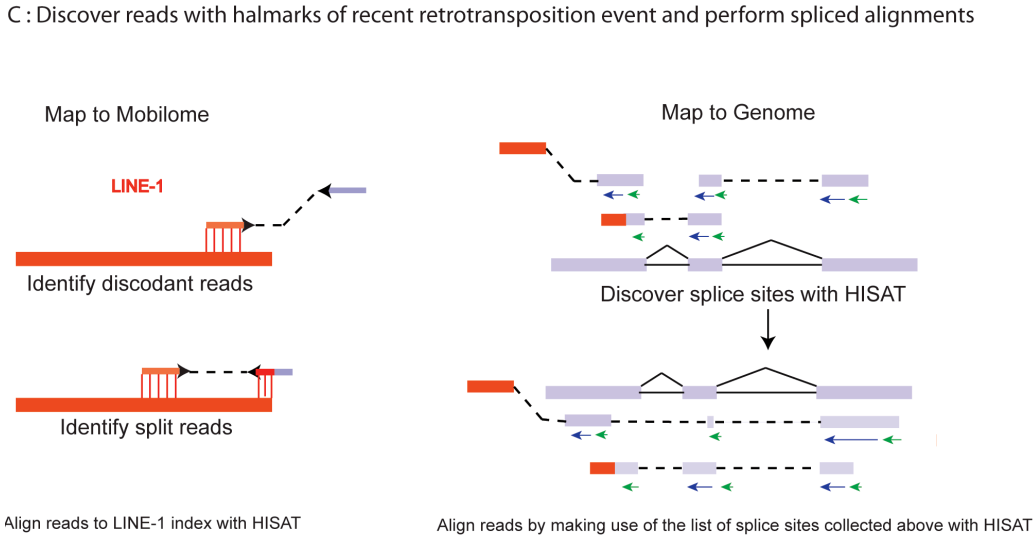


Figure 2: Read mapping against mobilome and genome.
The left side of the figure shows mapping of reads on LINE-1 index and identifying discordant and split reads and red color shows the piece of L1. The right side of the figure shows read alignment using splice junction information and red color shows the piece of L1. Red color depicts the unmapped piece of L1.

The choice of read mapper is critical for transcriptomic studies. Tophat2 is widely used since it allows spliced alignments (416). However, it is limited by its inability to perform read soft-clipping. Recently, Tophat2 creators have released a new read mapper software, called HISAT (415), which combines these two features. Moreover, HISAT is >50 times faster than Tophat2 and is comparable with other methods like GSNAP, STAR, MapSplice, and SMALT etc., but requires much less memory. Therefore, we chose to use HISAT.

Step-3: Identify the location of Chimeric reads in the reference genome (Figure 3).

The chimeric reads discovered in the previous step by mapping on a mobilome subset are used to identify the location of their remaining part in the reference genome.

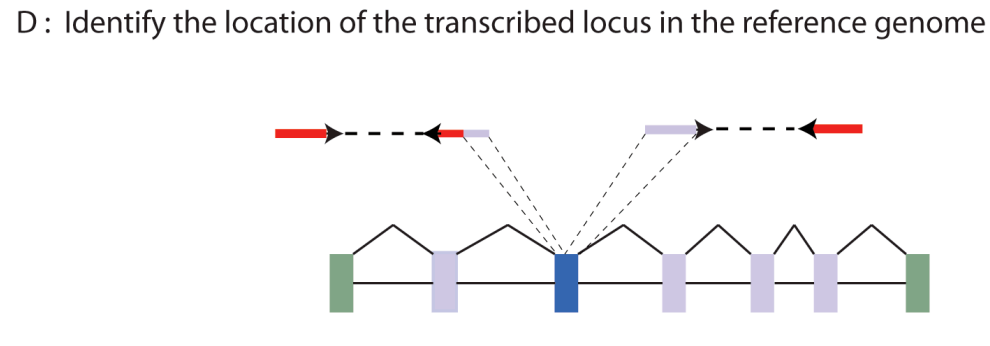


Figure 3: Identify the genomic part of chimeric reads.
Chimeric reads containing L1 regions (shown in red) are mapped against the reference genome.

Additionally, to be more stringent and select the chimeric reads only from the significant clusters, the criteria below can be applied on each read cluster (optional):
1 - All the reads should have the same strand within the cluster.
2 - All the reads within the cluster should support the same mobile element of origin.
3 - All the reads within the cluster should originate from the same side of the mobile element of origin.

Step-4: *De Novo* transcriptome assembly (Figure 4).

First, an alignment file (BAM) containing all the reads mapped on the human genome are assembled *de novo* using the StringTie assembler (414). Then a second *de novo* assembly is performed after removing the chimeric reads from the BAM file. This step helps to identify the transcripts formed by chimeric reads at a particular locus within the reference genome.

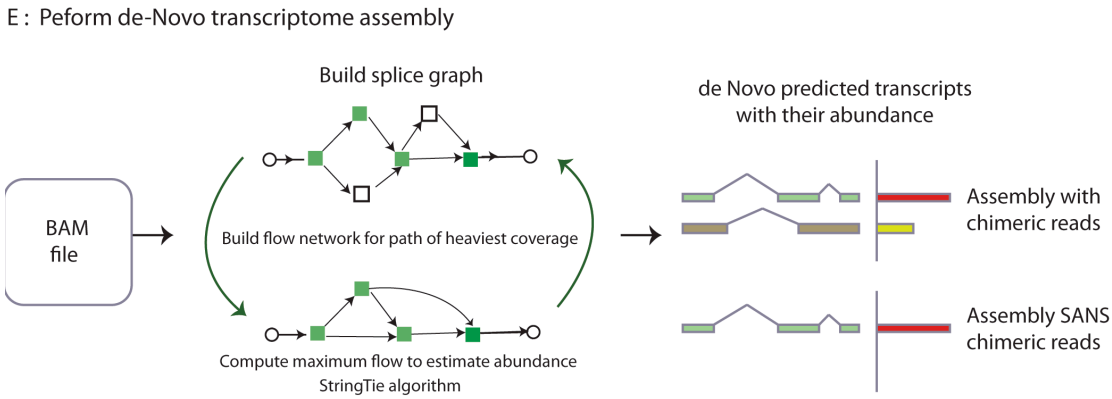


Figure 4: Two-tier transcriptome assembly.
Using BAM file as an input transcriptome assemblies are performed with and without chimeric reads.

The quality of assembly can adversely affect the final results. The recently published transcriptome assembler StringTie (414) seems to outperform Cufflinks assembler (417). StringTie has also been shown to perform much better than other assemblers

such as Trinity, IsoLasso, Traph and Scripture (414). Highly covered regions and the regions where introns have been retained have posed a tough challenge for the transcriptome assemblers so far and StringTie has been shown to assemble these regions convincingly (414).

Step-5: Comparing the set of transcripts obtained by *de novo* assembly with the known reference transcript annotation datasets (Figure 5).

This step uses Cuffcompare (417) to compare and tag known and unknown transcripts using reference transcript annotations.

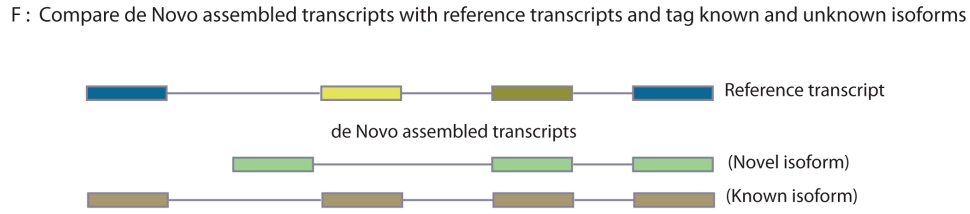


Figure 5: Transcript tagging.
Reference transcript is shown at the top with de-novo assembled transcripts below.

Step-6: Identify the chimeric transcripts (Figure 6).

In this step, L1-created transcripts isoforms are identified by comparing transcripts assembled with and without chimeric reads (in GTF format).

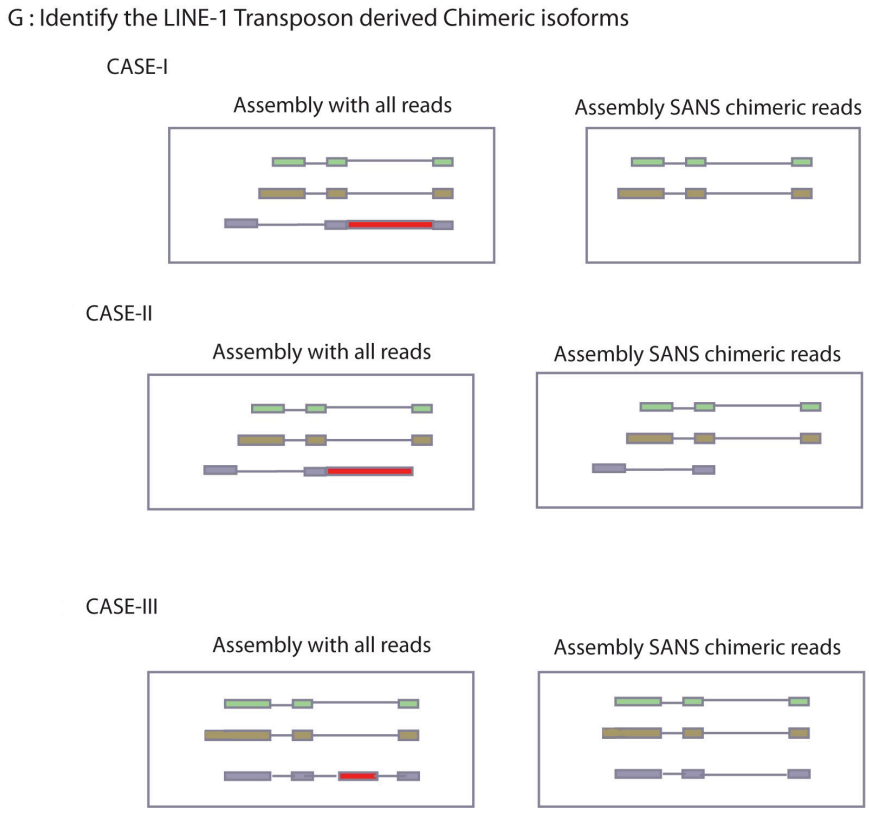


Figure 6: Cases for identifying L1 chimeric isoforms.
This is a direct method to detect chimeric transcripts by comparing the two-tier assemblies. Case-I: Isoform with a (red) piece of L1 which appeared before (left side box) disappears after removing chimeric reads in the second level assembly (right side box). Case-II: Isoform with a (red) piece of L1 is significantly reduced in length after second level assembly. Case-III: Alternative splicing event disappears in an isoform with a (red) piece of L1 after the second level of assembly (shown on the right side box).

Step-7: Annotate the assembled transcripts for alternative splicing events (Figure 7).

Once L1 chimeric isoforms have been identified, we annotate the nature of the alternative transcript using the SUPPA software (418). The types of events included are:
1- Alternative 5' splice sites
2- Alternative 3' splice sites
3- Intron retention
4- Exon skipping
5- Mutually exclusive exons

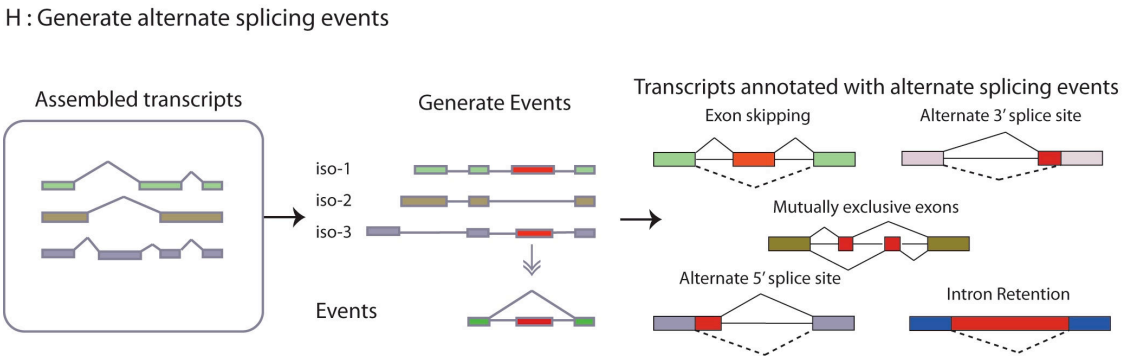


Figure 7: identify alternative splicing events.
This step uses a GTF file as input, generates the different possible events to explain transcript variations, and provides a functional annotation on the type of events, which are shown on the extreme right side. L1 fragments are shown in red.

The recently published algorithm SUPPA seems to be more efficient for alternative splice isoform detection when compared to other choices such as rMATS, Splicing compass, PASA and SplicingTypesAnno (418). Processing speed of SUPPA is relatively fast. Other alternative splice event detection algorithms based on artificial intelligence models like SpliceGrapher (419) are limited in the case of novel splice sites detection especially when it comes to unassembled or novel genomes.

Step-8: Identify antisense transcripts.

In this step, L1 chimeric transcripts found in the previous steps are checked for nearest full-length LINE-1 at exact ends but in the opposite orientation using the BEDTools suite (420).

Sequence and annotation databases

1- Mobilome sequences and annotation tracks: Mobilome sequence database index was created from the REPBASE database version "*RepBase20.07".* Taking all the 82 L1 subfamily consensus sequences (421). L1 annotation file was made by merging the known L1 annotation track from RepeatMasker with euL1db (422) and novel insertions from our lab (ATLAS-Seq method - unpublished).

2- Reference genome: Reference genome sequence database index was created from the UCSC database version hg19 for *Homo sapiens.* Transcript level annotation file was taken from GENCODE version 19 (423) in GTF format.

3- RNA-seq data: Illumina strand-specific, 2x150 bp paired-end RNA sequencing was performed by Beckman Genomics using whole cell poly(A)+ RNA isolated from human embryonal carcinoma cells (2102Ep) and human breast adenocarcinoma cells (MCF7). In the data pre-processing step, Trimmomatic (424) was used for adapter trimming and to remove low quality bases.

RESULTS

Influence of super-read assembly and soft-clipping on chimeric transcript discovery

Our pipeline identifies L1 chimeric *transcripts* present in a given sample. This is achieved by performing two parallel *de novo* transcriptome assemblies using the same RNA-seq data; the first with all reads, and the second excluding L1 chimeric *reads*. Transcript isoforms, which disappear, have significantly reduced length, or exhibit altered splicing in the second assembly as compared to the first assembly, are putative L1-related chimeric transcripts.

Given the abundance of LINE-1 element in the human genome, this method has the advantage to provide a higher level of evidence than just correlating the presence or absence of an LINE-1 element with the detection of a specific alternative transcript. However, an intrinsic limitation of this approach is that it can only detect transcripts with a detectable LINE-1 fragments in the mature transcript. Thus some L1-mediated alterations of RNA transcripts, such as exon skipping events, cannot be detected in principle.

As mentioned in the 'Method' section, super-read creation and soft-clipping are expected to strongly influence the detection of L1 chimeric transcripts. Therefore, we compared the number of putative chimeric transcript detected using diverse combinations of these options. For all, minimum transcript size was set to 500 bp. As shown in Table 1, in all settings, we detected hundreds of potential isoforms in the two cancer cell line samples analyzed.

Super-reads step	Soft-clipping	MCF7	2102Ep
yes	yes	2860	2009
no	yes	3189	2957
no	no	2480	2197

Table 1: **Total counts of L1 chimeric transcripts found in two different cancer cell lines.**

The setting resulting in the highest number of putative chimeric transcripts was using soft-clipping but not super-read creation. This was surprising, but our RNA-seq data were already obtained with relatively long reads (2x150 bp). Thus, super-read creation might not provide an advantage on this type of data. It might still be helpful to assemble shorter reads, which are of varying lengths. Indeed, the developers of this technique suggested that super-read creation can be seen as a "data debugging" technique (413). We kept it as an optional step, which can be applied depending on data read length and quality, but all analyses presented below have been performed without this option (but with soft-clipping), unless otherwise stated.

Benchmarking with known cell-type specific chimeric transcript datasets

There is currently no gold-standard dataset for L1-mediated alternative transcripts that has been established. Therefore, to evaluate the performance of our pipeline, we compared its output with known chimeric transcripts published for human breast adenocarcinoma cells (MCF7) published in 2009 by Cruickshanks *et al.* (67) and for human embryonal carcinoma cells (2102Ep) in 2011 by Macia *et al.* (69). These two publications used low-throughput modified RACE protocols to identify transcripts generated from L1 antisense promoter (ASP). Of note, they only consider a subset of potential variants (those generated from ASP), and they are far from being exhaustive. Thus, it is hard to evaluate the accuracy of our method without experimental validation or a gold-standard dataset. To calculate its specificity and sensitivity, we need to first determine the rates of true positives, true negatives, false positives and false negatives. However, Cruickshanks and Macia datasets are not exhaustive enough so we can only calculate the true positives and false negatives using these datasets.

Sensitivity (or true positive rate, TPR) can be calculated with the following formula:

$$TPR = True\ Positives / (True\ Positives + False\ Negatives).$$

False negative rate (FNR) can be calculated with the following formula:

$$FNR = False\ Negatives / (True\ Positives + False\ Negatives).$$

Once applied to the two datasets shown in Figures 8 and 9, we obtain:

Human breast adenocarcinoma cells (MCF7)

TPR = TP / (TP + FN) = 8 / (8 + 3) = 0.727
FNR = FN / (TP + FN) = 3 / (8 + 3) = 0.273

Human embryonal carcinoma cells (2102Ep)

TPR = TP / (TP + FN) = 45 / (45 + 16) = 0.738
FNR = FN / (TP + FN) = 16 / (45 + 16) = 0.262

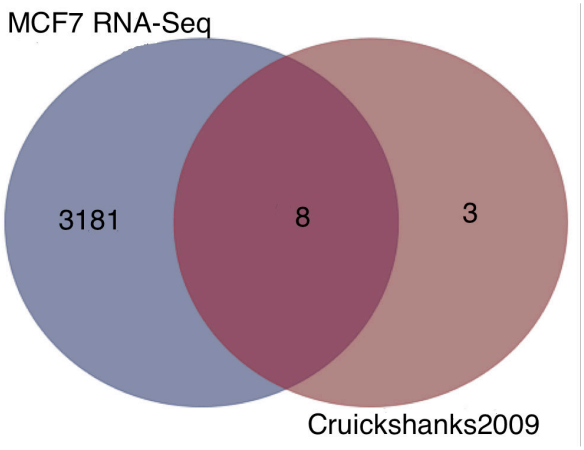


Figure 8: Comparison with MCF7 data from Cruickshanks *et al.* (67)
The Venn diagram shows the extent of overlap between chimeric transcripts detected by our pipeline on MCF7 RNA-seq data and the chimeric transcripts previously published in the same cell line by Cruickshanks et al. (67)

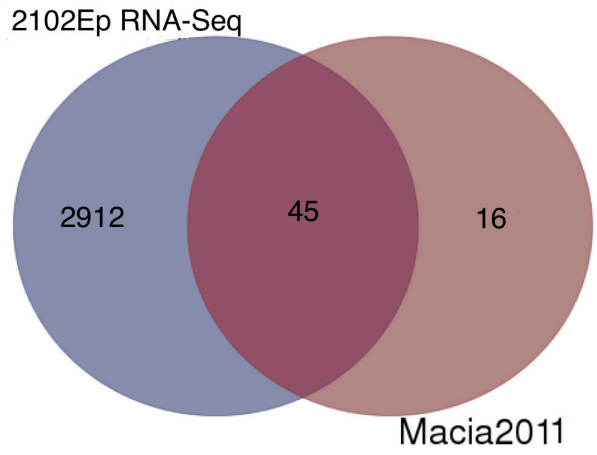


Figure 9: Comparison with 2102Ep data from Macia *et al.* (69)
The Venn diagram shows the extent of overlap between chimeric transcripts detected by our pipeline on 2102Ep RNA-seq data and the chimeric transcripts previously published in the same cell line by Macia et al. (69)

Our pipeline is able to find the majority (~72%) of already discovered chimeric transcripts for these two cell types. When we checked for the missed cases we found that the chimeric reads were present in 90% of them, but their abundance was too low to impact transcriptome assembly at the isoform level in the following step of the pipeline.

Additional validations

To further validate our results we checked for overlaps with another published dataset, obtained from *in silico* screening of expressed-sequence tags (ESTs), and thus originating from a broad range of cell types, unrelated to MCF7 or 2102Ep cells (408). In this study, the authors characterized chimeric mRNAs corresponding to sense or antisense strands of human genes and showed that the L1 ASP is capable of functioning as an alternative promoter. Examples of such chimeric transcripts include genes *KIAA1797*, *CLCN5*, or *SLCO1A2*.

First, we compared our datasets with their cases of L1 ASP-driven transcription (their Table 2). We were able to identify 88% of their cases in one or both of our datasets. The remaining could be false negative or transcripts actually not expressed in the two considered cell lines.

Second, we compared our datasets with the "*Catalogue of genes affected by transposable elements*" (pC-GATE, https://sites.google.com/site/tecatalog/), a database created and maintained by Dixie Mager lab (425). It enlists all known genes, which expression is potentially affected by transposable elements in a broad range of organisms. We filtered their data to only keep human genes cases influenced by an L1 copy and compared with our datasets (Figure 11). Our method was able to detect only 9% of pC-GATE records. However, it is important to stress that pC-GATE entries were found computationally by EST screening and not experimentally confirmed. As for Matlik datasets, the cellular origin of transcripts is very broad and not necessarily overlapping with our cell types.

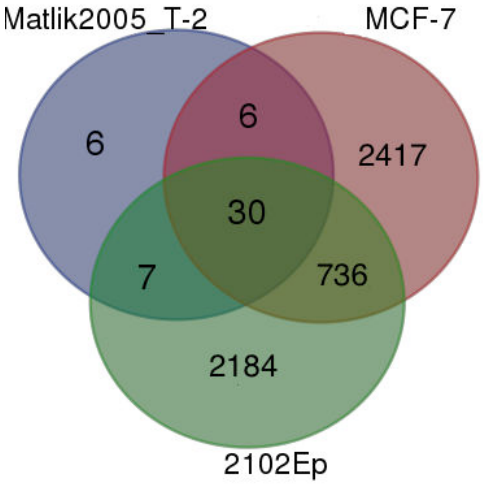


Figure 10: Comparison with L1 ASP-driven transcription data from Matlik *et al.* (408)
The Venn diagram above shows an overlap between chimeric transcripts detected by our pipeline "2102Ep" and "MCF7" against the chimeric transcripts found by Matlik2005 as per their data in table-2 of the publication.

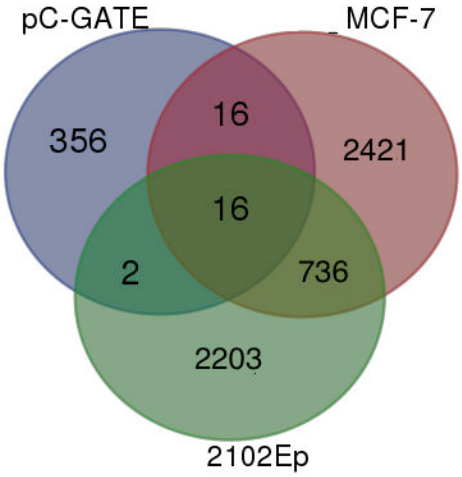


Figure 11: Comparison with pC-GATE data from Rebollo *et al.* (425)**.**
We could find some overlapping cases with our data from the putative exapted TEs very often retrieved from genome-wide analysis and stored in the pC-GATE database. The comparison only includes LINE-1-related records of this database.

L1-mediated alternative splicing

To annotate the assembled chimeric transcripts for different alternative splicing events, we used SUPPA (418), which can detect alternative 5' splice sites, alternative 3' splice sites, intron retention, exon skipping and mutually exclusive exons.
To determine the overall possible impact of L1-driven alternative splicing events we first generated all the possible events in the whole transcriptome (including, but not restricted to, L1 chimeric transcripts). Next we annotated the events found only within the putative L1 chimeric transcripts (Table 2).

Event type	MCF7			2102Ep		
	Total events	Chimeric events	% L1 chimeric events	Total events	Chimeric events	% L1 chimeric events
Alternative 5' splice site	1384	20	1,45	970	8	0,82
Alternative 3' splice site	1372	27	1,97	1099	5	0,45
Alternative 1st exon	1324	24	1,81	1147	9	0,78
Alternative last exon	326	5	1,53	289	3	1,04
Intron retention	2430	52	2,14	1477	23	1,56
Exonization or exon skipping	2974	52	1,75	2391	7	0,29
Mutually exclusive exons	115	2	1,74	83	0	0
TOTAL	9925	182	1,83	7456	55	0,74

Table 2: **Percentage of alternative splicing events due to L1 *vs* total alternative splicing events within sample.**
In this table, the number of alternative splicing events has been counted for two different cell lines (MCF7 and 2102Ep). "Total events" represent the number of all type of events found in the transcriptome (including - but not restricted to - L1-mediated events). "Whereas "Events Chimeric" represents alternative splicing events only due to L1 within chimeric transcripts. Percentage refers to the contribution of events within L1 chimeric transcripts compared to the total events found.

SUPPA was able to assign a clear alternative splicing event to only a small fraction of the chimeric transcripts (182 out of 3189 for MCF7, and 55 out of 2957 for 2102Ep cells). Whether other events are too complex to be annotated by SUPPA, or whether they correspond to completely new transcripts, remains to explore. From the SUPPA-annotated events, intron retention forms the majority of L1-driven alternative splicing events and, mutually exclusive exons events are the less abundant ones. However, these numbers might be underestimates due to the intrinsic limitation of our approach to require an L1 fragment to be included in the mature transcript for its detection as an alternative transcript.

Discovery of transcripts expressed from L1 antisense promoter (ASP)

Transcripts whose expression is driven by L1 ASP are in the opposite orientation as compared to LINE-1 and should contain a small (antisense) portion of the 5' UTR, which contains the L1 promoter region. To be more strict, we took the entire set of chimeric transcript isoforms discovered by our pipeline (listed in Table-1) (minimum transcript length >= 500) and checked them against a custom database containing all the full length LINE-1 element of UCSC repeatmasker track (426). In total, this represents a collection of 7360 full-length LINE-1 elements. When we scanned all the chimeric transcripts against this database, we found a number of anti-sense transcripts (Table 3).

Super-Reads step	Soft-clipping	MCF7	2102Ep
yes	yes	46	42
no	yes	21	29
no	no	71	211

Table 3: **Counts of all antisense transcripts found in cell line with different mapping strategies.**
In this table number of antisense transcripts found using strict criteria have been listed. Whether super-read creation or soft-clipping steps were performed or not is also indicated.

Older L1 subfamilies also contribute to chimeric transcripts

We evaluated the contribution of the different L1 subfamilies to L1 chimeric transcripts based on the source of reads in the mobilome-mapping step. The human-specific L1HS subfamily is a major source driving the L1 chimeric transcripts, but older primate-specific subfamilies also participate to a small portion of the chimeric transcripts (Table 4).

Sub family	#Transcripts in MCF7	% in MCF7
L1HS	2757	79.6 %
L1PA2	97	2.8 %
L1PA3	85	2.5 %
L1PA4	80	2.3 %
L1PA7	64	1.8 %
L1PA5	58	1.7 %
L1PA6	44	1.3 %
L1PA10	42	1.2 %
L1PB1	35	1.0 %

Table 4: **L1 subfamily contribution to the chimeric transcripts in MCF7 cells.**
Only the top L1 subfamilies, contributing to more than 1% of L1 chimeric transcripts, are displayed.

This suggests that apart from the actively jumping youngest L1HS subfamily, older subfamilies can also alter their genic environment.

L1 chimeric transcripts were found mostly within protein coding genes

While checking for the gene types where the chimeric transcripts were located, we found that they were mostly protein-coding genes (Figure 13). Examples of striking cases leading to alternative splicing or antisense transcripts are shown in Supplementary Figures S1 to S7. All these cases are very convincingly pinpointing to chimeric transcripts due to LINE-1 structural variations using their hallmarks in chimeric discordant and split reads pairs.

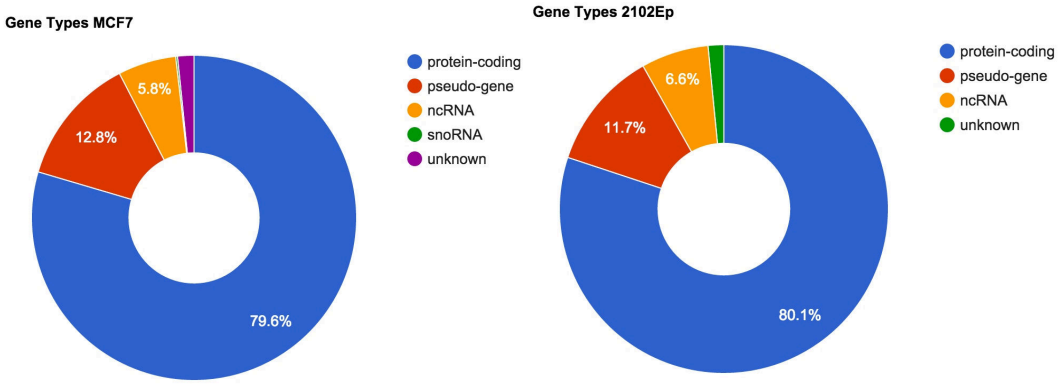


Figure 13: Gene types annotation for L1 chimeric transcripts.
Gene types where chimeric transcripts were located for MCF7 (left) and 2102Ep (right) cells.

This shows that L1 chimeric transcripts are in an environment where they can drive changes in gene expression by affecting the normal host gene transcription.
When we checked for the overlapping gene between the two cell lines, we could observe an overlap of 199 genes. These genes were mostly protein coding genes with a high number of chimeric transcripts.

Figure 14: Common genes between MCF7 and 2102Ep cell lines.
Overlapping counts of genes containing chimeric transcripts between MCF7 and 2102Ep cell (with super-reads step).

Contribution to novel exons in the genome

When we compared the assembled GTF files with Cuffcompare (417), taking both with and without chimeric read assemblies into consideration, we observed that 1.0% novel exons in MCF7 and 0.8% in 2102Ep were contributed by the chimeric reads. Also, 1.3% of all novel transcripts in the genome appeared to be linked to the contribution of chimeric reads. It also came out that 24 multi-exon transcripts were contributed by chimeric reads. L1 chimeric reads contributed 5793 and 5410 novel splice sites in MCF7 and 2102Ep cell lines respectively. This confirms that L1 can directly contribute to transcript sequence.

Chimeric transcripts in known genes with diseases caused by retroelements

For curiosity we checked our chimeric gene lists against the list of known genes with human disease-causing insertions (375) and we found hits for both the cell lines. We could find chimeric transcripts within 3 genes in 2102Ep (*CLCN5*; *DMD*; *F8*). In MCF7, we could identify 2 genes (*DMD* and *F8*).

Cases of Duchenne muscular dystrophy caused by L1 insertions in the Dystrophin gene (*DMD*) leading to exon skipping have been reported (33, 369, 371, 427). Similarly, Dent's disease cases are due to an Alu insertion causing exon skipping (428, 429). Finally, a Hemophilia A case was the first example of L1-mediated human genetic disease and was the result of an exonic insertion in the *F8* gene coding for the coagulation factor VIII (376).

To look for a general landscape of cancer-causing genes, we screened our gene lists against the Candidate Cancer Gene database (http://ccgd-starrlab.oit.umn.edu/about.php) (430). We could identify 181 and 137 candidate cancer genes among 663 and 528 genes overlapping with L1 chimeric transcripts, in MCF7 and 2102Ep cells, respectively. The cancer types potentially impacted are of broad origin, including blood cancer, colorectal cancer, liver cancer, lung cancer, nervous system cancer, pancreatic cancer, sarcoma and skin cancer.

Therefore, our method can help in finding the L1 chimeric transcripts, which might possibly give rise to novel isoforms of cancer-related genes.

DISCUSSION AND CONCLUSION

There are many published approaches so far, which use discordant and split read pairs containing LINE-1 sequence information to detect novel non-reference L1 insertions in whole-genome sequencing or whole-exome sequencing data, such as TranspoSeq (316), Tea (26), TraFic (318), RetroSeq (321), Tangram (320) or Mobster (319) . However, there is no approach till date, which can computationally pinpoint transcriptional isoforms due to LINE-1 elements. So, here, we present a novel method to precisely identify structural variation in human transcriptomes resulting from LINE-1 insertions, notably in cancer. Both old elements present in the reference genome and highly polymorphic non-reference young elements are captured in discordant and split read pairs. Then this information of chimeric reads is combined with the power of two-tier ultra sensitive *de novo* transcriptome assembly to detect the chimeric transcripts. We also included recently developed assembly techniques, called super-read creation, which generates longer contigs from unambiguous, non-branching parts of a transcript. The advantage of this optional step remains to be demonstrated.

Preliminary *in silico* benchmarking indicates that our method is able to detect a majority of already known chimeric transcripts found to be expressed in MCF7 and 2102Ep cell lines. However in the absence of gold-standard datasets, extensive wet-lab experiments will be required to define the false discovery rate of this approach and to fine-tune the different steps of the algorithm. Further analyses revealed that the pipeline described here is able to detect a broad range of events previously reported in the literature, particularly L1 antisense promoter functioning as an alternative promoter and alternative splicing, intron retention being the most

prominent. Apart from the youngest L1HS subfamily, the contribution of older subfamilies to create L1 chimeric transcripts could also be detected. We also found that the majority of the genes overlapping with L1 chimeric transcripts are protein coding, with a minority of snRNA, snoRNA and pseudo-genes. We could also detect novel coding sequences and splice sites associated with the presence of L1 chimeric reads in transcriptome assembly.

A major limitation of our approach is that it can only detect isoforms, which incorporate a detectable LINE-1 fragments in the mature transcript. Thus, particular splicing events, such as exon skipping without the simultaneous inclusion of an L1 fragment in the mature transcript, cannot be detected in principle. As underlined before, the false discovery rate of the approach could not be defined without additional experimental validation. Bona fide L1 transcripts ending in the downstream flanking genomic sequence, due to its weak polyadenylation signal is expected to be an abundant source of L1 chimeric transcripts. Therefore, it would be of interest to include a transcript annotation method able to identify these particular types of events among the L1 chimeric transcripts.

Apart from computational challenges there might also be some limitations due to RNA sequencing technologies. RNA-seq relies on cDNA synthesis and on multiple ligation steps for library preparation, which can be a source of experimental artifacts. For example, the generation of spurious second-strand cDNAs can create problems for strand-specific RNA-seq. Template switching during cDNA synthesis or fragment-fragment ligation can cause problems in exon-exon boundary and true chimeric transcript identification. It might become feasible in the future to overcome some of the above-stated limitations using direct sequencing of RNAs (DRS) (431).

L1 retrotransposon expression has been proposed both as a potential biomarker of cancer prognosis and as the starting point of L1-mediated genome instability in tumors. The expression of L1 elements might drive - or contribute to - cancer genome instability through new somatic insertions in a subset of permissive tumor types, but also through the expression of chimeric transcripts. However, it is currently unknown whether, in these permissive tumors, all L1 copies or only a small number of copies, located in a favorable genomic environment, are reactivated. Furthermore, the extent of L1-chimeric transcript formation and the landscape of the affected genes remain unexplored. Our work will shed light on the following questions: 1- what proportion of L1 copies lead to tumor-specific L1 chimeric transcripts? 2- what are the dominant forms of transcript alternations resulting from L1 element in cancer transcriptomes? 3- do L1 chimeric transcripts give rise to novel isoforms of cancer-related genes? To answer these questions, a useful additional module could allow sample-to-sample comparison of transcript variants to identify those being tumor-specific.

On the longer term, this approach will provide a conceptual and computational framework, which could be applied to larger datasets, such as those provided by the International Cancer Genome Consortium, to help in understanding the mechanisms leading to transcriptome plasticity in tumor cells and to provide a rational basis for the use of retrotransposon chimeric transcripts as biomarkers.

A computational approach to reveal the landscape of transcriptional isoforms induced by LINE-1 elements in human cells

Ashfaq A. Mir and Gaël Cristofari

SUPPLEMENTARY METHODS

Dependencies to install and run the pipeline:

A- DATA

1- A LINE-1 Index should be generated using either data from REPBASE (http://www.girinst.org/repbase/) or RepeatMasker (https://genome.ucsc.edu/cgi-bin/hgTables) files in FASTA format.
2- Transcript annotation files should be downloaded from GENCODE (http://www.gencodegenes.org/) in GTF format.
3- Known full-length LINE-1 annotation files can be downloaded from UCSC table browser (https://genome.ucsc.edu/cgi-bin/hgTables) and also from euL1db download page (http://eul1db.unice.fr/db/Data.jsp) in BED format.
4- RNA-seq data can be either downloaded from CGHUB (https://cghub.ucsc.edu/) or your own in-house data in FASTQ format.
5- Human genome index can be built from FASAT format files downloaded from either NCBI / UCSC or you can even use pre-build index from HISAT website (https://ccb.jhu.edu/software/hisat/index.shtml), which is in HISAT index format.

B-SOFTWARES

1- HISAT (https://ccb.jhu.edu/software/hisat/index.shtml)
2- StringTie (http://ccb.jhu.edu/software/stringtie/)
3- SUPPA (https://bitbucket.org/regulatorygenomicsupf/suppa)
4- MaSuRCA (http://www.genome.umd.edu/masurca.html) &
(http://ccb.jhu.edu/software/stringtie/dl/superreads.pl)
5- CuffCompare (http://cole-trapnell-lab.github.io/cufflinks/)
6- BEDtools (https://github.com/arq5x/bedtools2)
7- BAMtools (https://github.com/pezmaster31/bamtools)
8- SAMtools (https://github.com/samtools/samtools)
9- Scripts for data processing (scripts provided with the pipeline)

C-PROCEDURE

Before starting the pipeline procedure, data should be cleaned from adaptor sequences, ribosomal RNA and other possible contamination.

Step-1: Create super reads (Optional)

The usage of the superreads.pl script is documented below.
Usage: superreads.pl <pair_read1_fastq> <pair_read2_fastq> <masurca_directory> [options]*

Arguments:

The first two arguments of the superreads.pl script is files in the fastq format containing the sequences of the first and second read in each fragment, respectively. They can either plain text fastq files or compressed (with gzip or bzip2) files. The third argument represents the directory where the MaSuRCA package was installed on your system.

Options:

-t <num_threads>	Sets the number of threads to use. Default: 10.
-j <jf_size>	MaSuRCA requires the Jellyfish program to run, and this parameter sets the Jellyfish hash size. Please see the MaSuRCA documentation for more information about how to choose this parameter. Default: 2500000000.
-s <step>	As it progresses, the superreads.pl script prints the steps it successfully completed. If, for any reason, the assembly process is stopped, you don't need to redo all the successfully completed steps, and you can restart the script at the first step it didn't complete. Default: 1.
-r <paired_read_prefix>	Sets the prefix for the paired reads as required by MaSuRCA. Default:pe.
-f <fragment_size>	Specifies the mean library insert length. Default: 300.
-d <standard_deviation>	Specifies the standard deviation of the library insert length. If the standard deviation is not known, set it to approximately 15% of the mean. Default:
-l <super_reads_file_name>	Specifies the name for the assembled

	super-reads file. Default:LongReads.fq.
-u <not_assembled_reads_prefix>	Specifies the prefix for the unassembled reads file names. By default, it appends ".notAssembled.fq.gz" to the initially paired files.

Source documentation for superread.pl script (http://ccb.jhu.edu/software/stringtie/)
The output files of the superreads.pl script (the assembled super-reads file, and the two files containing the unassembled paired reads) can be then aligned to a reference genome with your read mapper of preference. For instance, you can align them with HISAT like this:

Usage: hisat [options]* PE_reads_1.notAssembled.fq.gz, LongReads.fq PE_reads_2.notAssembled.fq.gz

Step-2: Generate known Splice sites

In this step you need to download transcript annotation files from GENCODE (http://www.gencodegenes.org/) and then create the known splice sites file using the command below:

Usage: python extract_splice_sites.py genes.gtf > splicesites.txt

This utility is provided with the HISAT software. Remember that this file needs to be used only in genome mapping with HISAT.

Step-3: Align RNA-Seq data against first against Mobilome and then against Human genome:

hisat –x Reference_Index –phred33 –fr –very-sensitive-local –known-splicesite- infile genocode_ss.txt -1 file1.fastq -2 file2.fastq –S output.sam

It is to be noted that known splice site file should be given only in case of genome mapping.

More details can be found here: (https://ccb.jhu.edu/software/hisat/manual.shtml).

Step-4: Get Chimeric read locations in the genome

1- Sort mapped BAM files:
Usage: *samtools view -bS file.sam | samtools sort - file_sorted*

2- Index sorted BAM files:
Usage: *samtools index test_sorted.bam*

3- Extract chimeric reads from the mobilome BAM file using bamtools using a JSON

script:

Usage: bamtools filter -in Mobilome_sorted.bam –out Mobilome_filtered.bam -script criteria.json

The JSON script format can be like:

```
{
  "filters" :
  [
{
    "id" : "splitread",
    "cigar" : "*S*"
},
{
    "id" : "discordant",
    "isMapped" : "true",
    "isMateMapped" : "false"
}
  ]
 }
```

4- Generate BED files from mapped BAM files for reads mapped on the genome and on the mobilome:

Usage: bedtools bamtobed -i Mobilome_filtered.bam > Mobilome_filtered.bed

Usage: bedtools bamtobed -i Genome_Mapped.bam > Genome_Mapped.bed

5- Intersect the overlapping read names (split and discordant reads) we got from mobilome mapping with genome mapped reads to get their locations in the reference genome.

BAM files are indexed by chromosomal positions. Therefore, extracting read names from a few GB file can be extremely time-consuming. Thus, an easier and faster way could be to extract reads based on read names, could be to use a shell command like this:

Usage: awk 'FNR==NR{a[$4]++;next}a[$4]' Mobilome_filtered.bed Genome_Mapped.bed > Chimeric_read_pairs.bed

The command above creates an array of read names (4th column in BED files) for both the files and checks for string matches.

Step-5: Generate BAM file excluding the chimeric reads

Once we have got the Chimeric read file then we need to create another BAM file filtering out the chimeric reads from it. This can be done in the similar way as explained in step-5. Again BAMTOOLS are extremely inefficient for extracting by read name list.

Usage: awk 'FNR==NR{a[$4]++;next}!a[$1]' Chimeric_read_pairs.bed Original_Genome_mapped.sam > Sans_Chimeric_Reads_Genomic.sam

The command above creates an array of read names (4th column in BED files) for chimeric reads and 1st column in SAM format file and checks for string matches.

Then, we need to again convert the SAM file into BAM and then sort and index it like in steps-1 and 2.

Step-6: Perform two tier transcriptome assemblies

Once both sorted and indexed BAM files with and without chimeric reads were obtained, we perform 2 de-novo transcriptome assemblies, using StringTie.

Usage: stringtie Mapped_genome_with_Chimeric_Reads.bam –out with_chimeric.gtf –x chrM
stringtie Mapped_genome_without_Chimeric_Reads.bam –out without_chimeric.gtf –x chrM

More details about StringTie assembler usage can be found here: http://ccb.jhu.edu/software/stringtie/

Step-7: Compare assembled GTF files with reference annotation

Assembled transcript files were compard to the reference transcript annotation databases such as GENCODE.gtf

This can be achieved by using the CuffCompare utility from the Cufflinks assembler.

Usage: cuffcompare –r Reference_GENCODE_transcripts.gtf assembled_Transcripts.gtf –o outputprefix

Further details can be found here: http://cole-trapnell-lab.github.io/cufflinks/cuffcompare/index.html

Step-8: Generate alternative splicing events

Then, we can use SUPPA to generate alternative splicing events from assembled transcript files

Usage: python suppa.py generateEvents –i assembled_transcripts.gtf –o output_file.gtf –e SE MX RI SS FL

Further details can be found here: https://bitbucket.org/regulatorygenomicsupf/suppa

Step-9: Generate full set of chimeric transcripts for a sample

To generate all the chimeric transcripts present in our data, we need to run the below command:

The script for this is provided with our pipeline.

Usage: ./getEvents with_chimeric_Reads.gtf without_chimeric_Reads.gtf File_Prefix

Step-10: Generate alternative splicing events due to LINE-1

To generate all the alternative splicing events, we need to run the below command:

The script for this is also provided with the pipeline.

Usage: ./getEvents with_chimeric_Reads.gtf without_chimeric_Reads.gtf file_Prefix Total_Event_in_genome_file.gtf event_Prefix

Step-11: Get all anti-sense transcripts in the sample

To generate the antisense transcript file, we need to run the below command:
The script for this is also provided with the pipeline.

Usage: ./getPolyA-AntiSense.sh Chimeric_Transcripts_File Full_Length_L1_File file_Prefix

SUPPLEMENTARY CASES

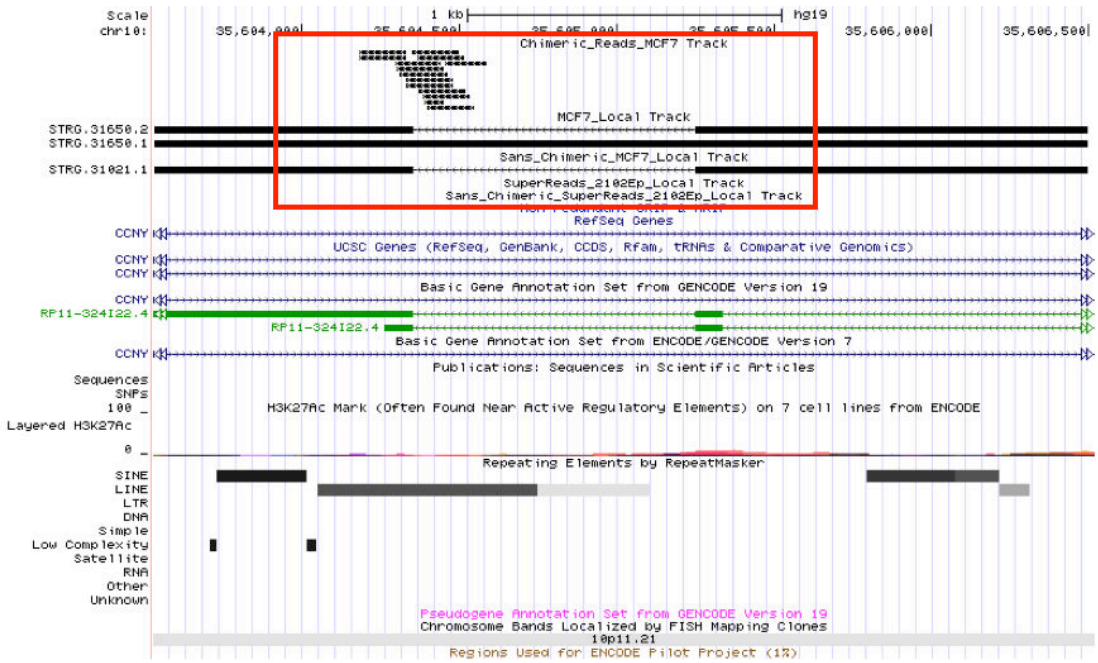


Figure S1: Intron retention case for MCF7 chimeric transcripts
UCSC genome browser screenshot from pipeline results on MCF7 cell line with and without chimeric reads for CCNY gene. It can be clearly observed that Intron retention isoform (STRG.31650.1) in the 2nd track disappears after excluding chimeric reads in the 2nd level assembly track shown 3rd track below it. Chimeric reads can be observed above in the 1st track. The LINE-1 element can be seen in the last track below.

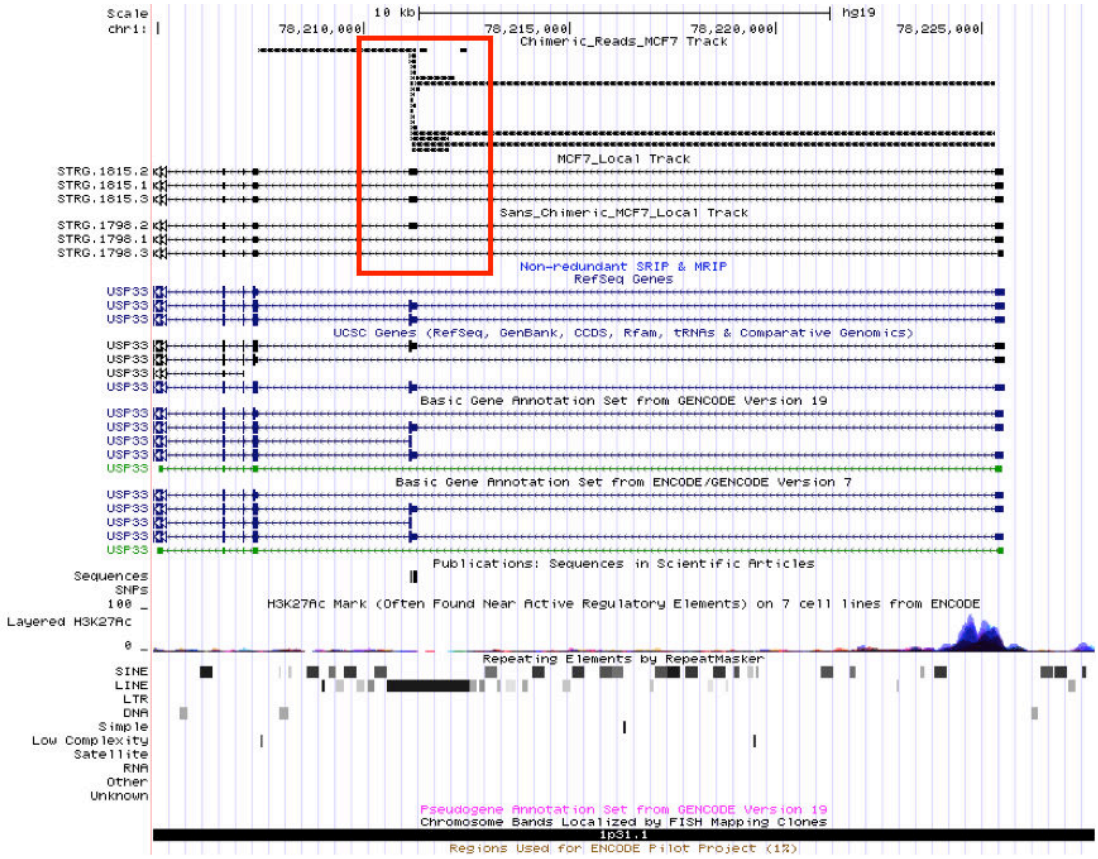


Figure S2: Exonization / exon skipping case for MCF7 chimeric transcripts
UCSC genome browser screenshot from pipeline results on MCF7 cell line with and without chimeric reads for USP33 gene. Exonization / exon skipping can be clearly observed in the isoform (STRG.1815.3) in 2nd, the exon In the 3rd track disappears after excluding chimeric reads in the 2nd level assembly. Chimeric reads can be seen in the 1st track and LINE-1 can be seen in the last track below.

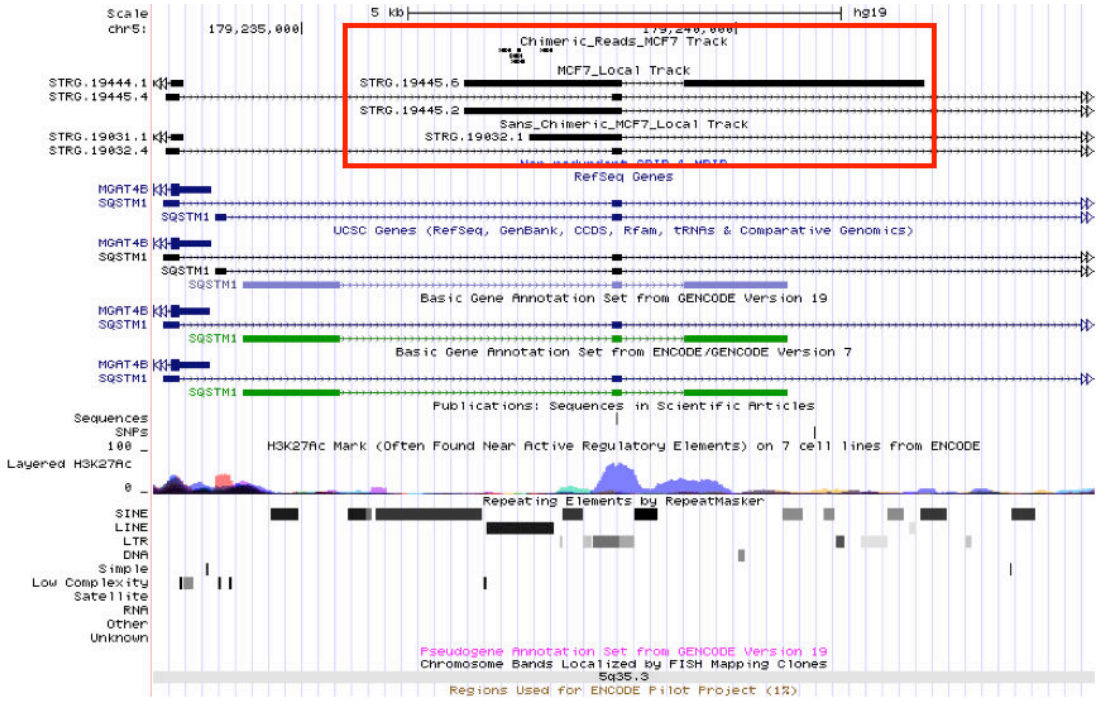


Figure S3: Alternative 1st exon case for MCF7 chimeric transcripts

UCSC genome browser screenshot from pipeline results on MCF7 cell line with and without chimeric reads for SQSTM1 gene. It can be clearly observed in the isoforms (STRG.19445.2 and STRG.19445.6), which show alternative 1st exons, disappear after excluding chimeric reads in the 2nd level assembly track shown below (3rd track). Chimeric reads can be seen above alternative 1st exon in the first track and LINE-1 can be seen in the last track below.

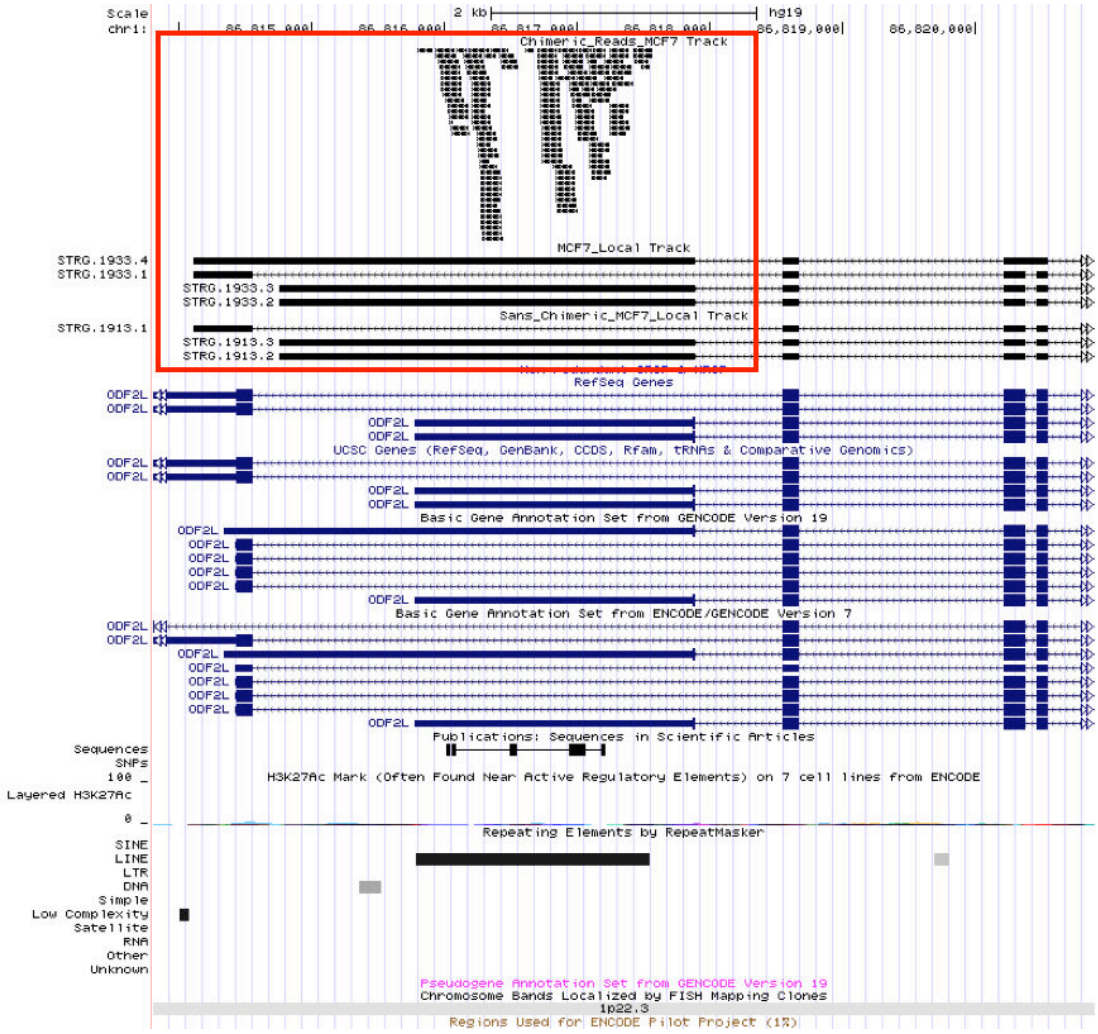


Figure S4: Alternative last exon case for MCF7 chimeric transcripts

UCSC genome browser screenshot from pipeline results on MCF7 cell line with and without chimeric reads for ODF2L gene. It can be clearly observed in the isoform (STRG.1933.4), which shows an alternative last exon in the 2nd track. This isoform disappears after excluding chimeric reads in the 2nd level assembly track shown below (3rd track). Chimeric reads can be seen above alternative last exon in the first track and LINE-1 can be seen in the last track.

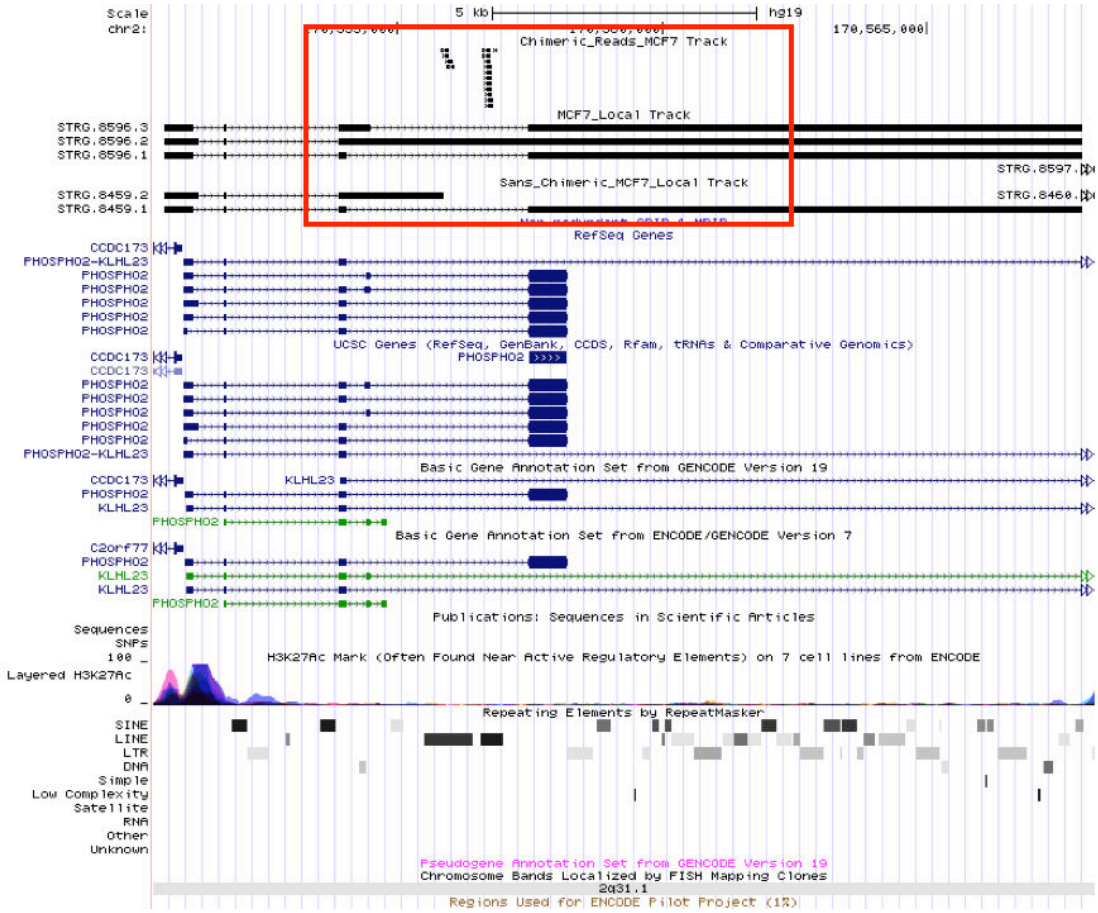


Figure S5: Alternative 5' splice site case for MCF7 chimeric transcripts
UCSC genome browser screenshot from pipeline results on MCF7 cell line with and without chimeric reads for PHOSPHO2 gene. It can be clearly observed that the isoform (STRG.8596.3), which was showing alternative 5' splice site (track 2) disappears after excluding chimeric reads in the 2nd level assembly track shown below (track 3). Chimeric reads can be seen in the first track and LINE-1 can be seen in the last track below.

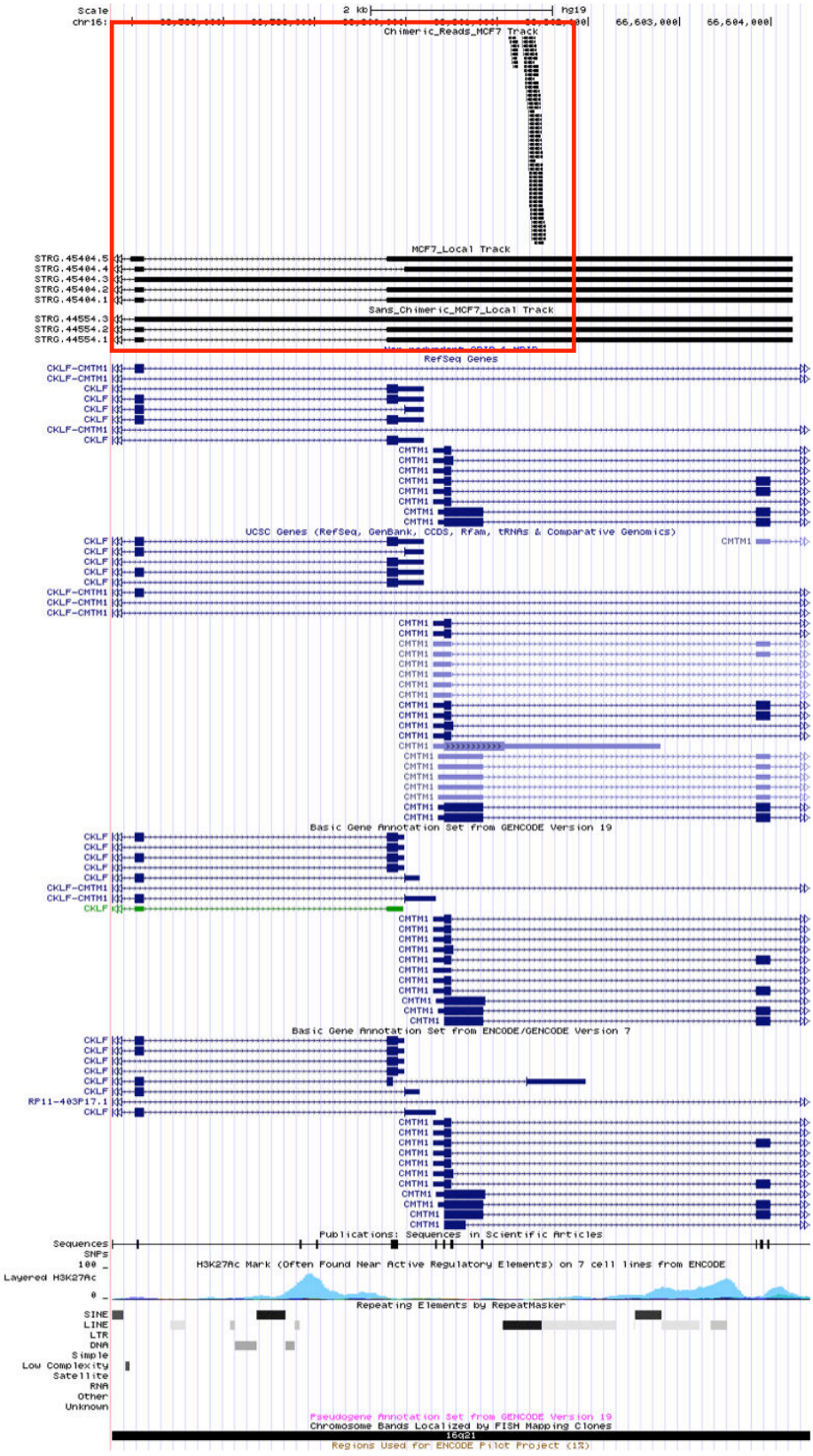


Figure S6: Alternative 3' splice site case for MCF7 chimeric transcripts
UCSC genome browser screenshot from pipeline results on MCF7 cell line with and without chimeric reads for CKLF-CMTM gene. It can be clearly observed that the isoforms (STRG.45404.5 and STRG.45404.4), which were showing alternative 5' splice sites (track 2) disappears after excluding chimeric reads in the 2nd level assembly track shown below (track 3). Chimeric reads can be seen in the first track and LINE-1 can be seen in the last track.

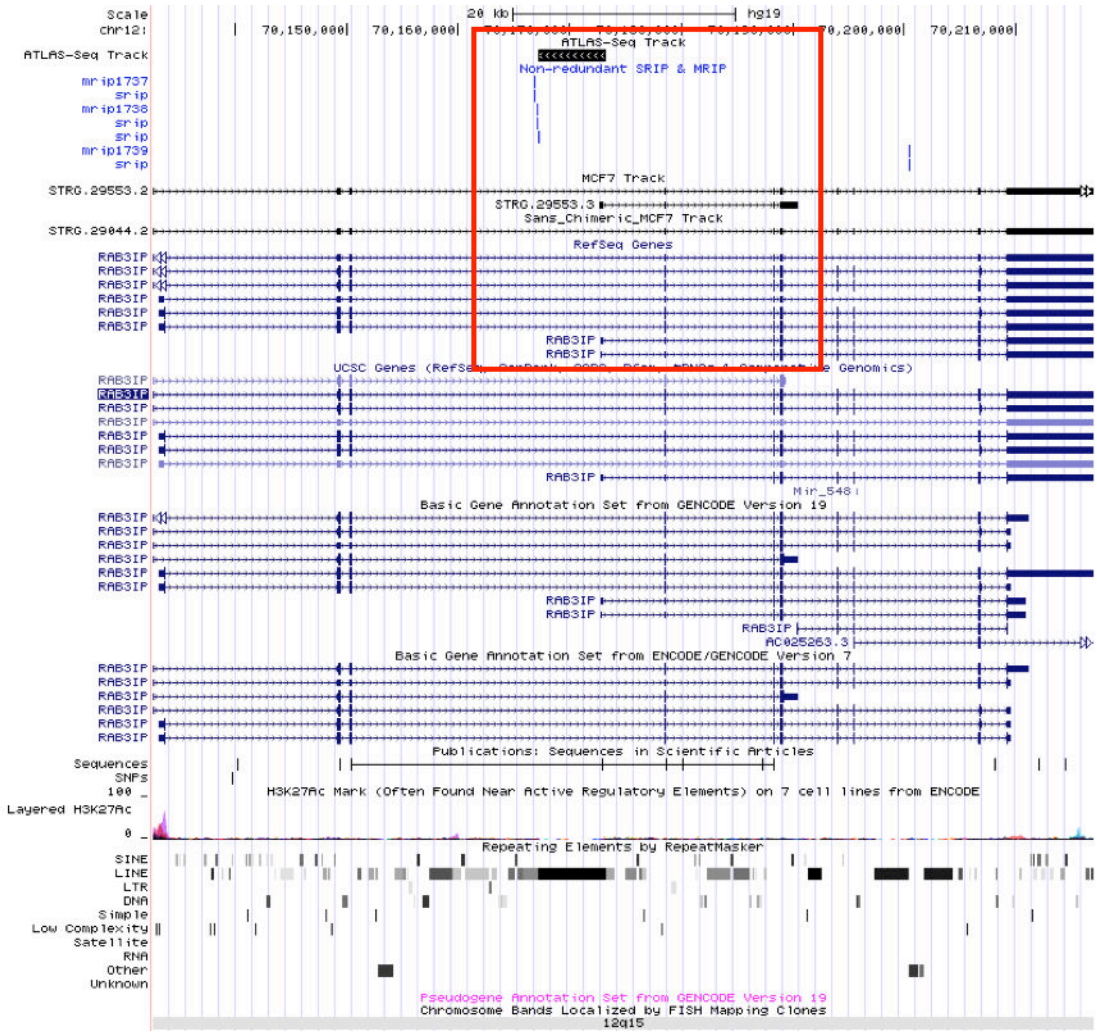


Figure S7: Antisense transcript case for MCF7 chimeric transcripts
UCSC genome browser screenshot from pipeline results on MCF7 cell line with and without chimeric reads for RAB3IP gene. It can be clearly observed the isoform (STRG.295553.3) shown in the 2nd track, which was present before disappears after excluding chimeric reads in the 2nd level assembly shown below in 3rd track. It is to be noted that it starts at the end of a full-length L1 from in-house ATLAS-Seq method detected new insertion and also the reference L1HS insertion (shown in the 1st track).

DISCUSSION

The expression of L1 elements might drive or contribute to the instability of cancer genomes through new somatic insertions in a subset of permissive tumor types and through the expression of chimeric transcripts (67, 68, 432–434). However, the extent of L1 chimeric transcript formation and the landscape of affected genes remain unexplored. More specifically, we wanted to address the following questions: (i) what proportion of L1 copies and which copies lead to tumor-specific L1 chimeric transcripts? (ii) Do L1 chimeric transcripts give rise to novel isoforms of cancer-related genes?

Since many L1 copies, especially from the youngest L1HS subfamily, are polymorphic insertions absent from the human reference genome, it is essential to have a genome-wide view of their position within the human genome, as a first step in understanding their impact on the transcriptome. Therefore, we started this research program by building the euL1db database (422), which provides a curated and comprehensive summary of L1HS insertion polymorphisms identified in healthy or pathological human samples and published in peer-reviewed journals. Next, we developed a novel computational method to detect L1 chimeric transcripts using RNA sequencing data. euL1db, by providing markers of recent polymorphic events, can help in identifying the overall transcriptional consequences of young and recently jumping active retrotransposons insertions. This turned out to be particularly important for the detection of antisense transcripts. Overall, we developed a computational framework dedicated to investigate at the genome-wide level L1-mediated structural variations of the human genome and transcriptome.

1. euL1db provides a comprehensive resource for curated human-specific L1 allowing sample level retrieval of insertion data.

1.1. Rational behind euL1db characteristics

euL1db has been developed to provide the most comprehensive and curated data on human-specific retrotransposon insertion polymorphisms (RIPs), identified in healthy or pathological human samples. Samples could be a tissue, cell or cell line or blood. Among the most important feature of euL1db is that insertions can be retrieved at the sample level. This can greatly help in correlations between presence or absence of an insertion with a specific disease or phenotype. This is also particularly useful when additional genomic data, such as RNA-seq, are available from the same samples, as for insertions discovered in the frame of the 1000 genome project (1000 GP) or of The Cancer Genome Atlas (TCGA), since it should allow studies aimed at correlating the presence or absence of a specific insertion with a specific genomic feature.

euL1db stores pathological and anatomical data. Also, if it was prepared from multiple of single cells. The relationship between samples is also recorded as (e.g., normal/tumor pairs). Every sample is associated with a unique study and a unique ID.

Many L1HS insertions are unique to an individual or population or might be shared among relatives. Therefore, it is important to organize the data in a manner that shows the relationship of an insertion with a sample, individual, family or population, as achieved in euL1db. Data have been organized into many tables, which are interconnected in a dynamic way based on the primary data keys (see Supplementary Figure S1 from Article-II for further details). This can be very useful, for example, to perform analyses on familial trios (father, mother, and child), which are available from the web interface through the family browser. Then information at the family, individual, sample or insertion level can be easily retrieved.

euL1db is a curated repository to ensure data quality. The curation method is described and available for each study in the “curation” tab. We also provide additional quality information, such as cases of conflicting annotations between distinct studies, which have been tagged with a “caution” flag.

euL1db provides access to different levels of information by providing different browsers to the end user like study browser, sample browser, insertion browser, family browser, individual browser and genome browser which has been dynamically connected to UCSC genome browser to profit from their rich datasets.

euL1db also has created two layers for insertion data. Firstly, sample-level retrotransposon insertion polymorphisms, named SRIP, which are real insertions detected in a given sample with a unique ID and meta-retrotransposon insertion polymorphisms, named MRIP, which are virtual group of SRIP likely representing a unique retrotransposition event. While building virtual insertions, germline insertions are grouped into unique non-overlapping ranges within the genome, whereas, overlapping somatic insertions are not merged into MRIP because they represent unique and new events by themselves.

1.2. euL1db limitations and future technical developments.

Large data transfer and processing through Hyper Text Transfer Protocol (HTTP), as currently implemented in euL1db, can lead to excessive server loads and inability to process web interface-driven protocols, limiting possibilities of dynamic integration of the data. Therefore, to further build a more global application, a valuable update would be to add a representational state transfer (REST) API to euL1db. This new protocol does not always communicate via Hyper Text Transfer Protocol (HTTP), which is slow for data transfer. This could enhance the performance and scalability of euL1db web application and reduce its dependence on its graphical user interface. In other words, it would allow remote and programmatic access to euL1db, which could promote its use in other third-party software’s or pipelines.

Another valuable improvement would be to implement a submission module, to facilitate the upload of new data, ensuring a regular update of the database.

1.3. euL1db applications and future perspectives.

euL1db provides valuable and rich non-reference insertions data for L1 chimeric transcript detection method. This information can be useful in annotating the novel non-reference chimeric transcripts, which may be involved in different alternate splicing events like exonization, intron retention, alternate splice sites, antisense transcripts or even chimeric transcripts within introns or non-genic locations within genome. Indeed, knowing from an independent source that a polymorphic L1 overlaps with a predicted event of L1-mediated alternative transcription provides enhanced confidence in the predictive power of the transcript prediction method. Apart from this, euL1db can also help in identifying novel retrotransposon insertions by providing a pool of already existing reference and non-reference published insertion data. One such application is to prioritize putative somatic insertions in cancer or in neurological diseases, since euL1db allows the user to exclude non-reference polymorphic insertions present in the human population. euL1db provides tools for batch processing of data under the 'utility tab' to perform such tasks.

euL1db can also help population genetics studies. Information about the difference in the frequency of the same retrotransposon can be used to infer population relationship and thus, retrotransposons can act genetic markers.

After the recent developments of mega-sequencing projects like the 10k genome project, which aims to sequence genomes of 70,000 people with rare diseases and storing their familial information, we think that human-specific data is going to be produced exponentially. Given the fact that L1 retrotransposons and their implications on human health have been of intense study within the scientific community (20–27), we expect a considerable amount of human-specific L1 structural variations data to be published in a close future, reinforcing the need for a centralized catalogue of such variants, like euL1db.

2. Development of a computational method to identify transcript isoforms due to L1 elements

2.1. Accuracy of transcript variant predictions

The overall transcriptional contribution of L1 (or even transposable elements in general) in humans has been only little studied at the genome scale so far, although a few studies have highlighted their influence on the transcriptional output on the human genome (174, 227). We recently published euL1db database, which compiles more than 9000 distinct insertions described these recent years in the literature, but this resource does not report on the functional consequences of this extensive structural variation, in particular its transcriptional output (422). To tackle this biological question, we have developed a new computational method dedicated to explore the extent of transcriptional isoforms induced by LINE-1 integration in human

cells. This method takes advantage of the data generated by RNA sequencing technology to discover novel isoforms (435). Our method attempts to identify the majority of L1 chimeric transcripts within a given sample and was initially applied to datasets obtained in cell lines for which a number of L1 chimeric transcripts were previously identified by low-throughput wet-lab approaches (67, 330).

To show that L1 transcripts could be useful as markers of malignancy Cruickshancks *et al.* (67) isolated a set of L1 chimeric transcripts induced by hypomethylation of its antisense promoter (67). These chimeric transcripts are unique to breast cancer cell lines, primary tumors and colon cancer cells. Our method was able to computationally detect 73% of chimeric transcripts experimentally detected by this study using RNA-seq data obtained in the same cell line. Similarly, Macia *et al.* (69) showed the expression of transcripts driven by the L1 antisense promoter in human embryonic stem cells and embryonal carcinoma cells (2102Ep). They notices that half of the expressed copies were absent from the human reference genome and thus polymorphic in nature (330). Again, we could computationally identify 74% of the L1 chimeric transcripts detected experimentally after comparison with their data using RNA-seq data from the same cell line.

L1-chimeric transcripts found by our computational approach outnumbers by one or two orders of magnitude those found in previous studies, raising the possibility that a significant proportion could be false positives. As already underlined, the absence of established golden standard dataset prevents us to directly evaluate the accuracy of our approach and to fine-tune the parameters of each step. Therefore, experimental validation of the putative hits should be a priority in the future. Given the diversity of potential events detected, direct wet-lab experiments are preferable to *in silico* simulations. However, there are a number of other possible reasons that can explain such a considerable difference. First, most previous studies were focusing on L1 5' extremity due to technical constraints, limiting the type of analyzed events mostly to antisense promoter-driven transcription. In contrast, our computational pipeline can theoretically identify a much broader range of events, given that a fragment of L1 is incorporated in the transcript. Second, our initial analyses were applied to RNA-seq data generated from poly(A)+ RNA, which are strongly enriched for cytoplasmic mature mRNA. In contrast, both studies cited above used total RNA as starting material, with the potential to isolate non-polyadenylated, instable, and/or non-coding RNA species, which might not pass cellular quality controls. Third, we expect that a significant fraction of the L1 chimeric transcripts are actually L1 transcripts ending in their flanking sequence due to read-through transcription. These events can be useful to identify individual L1 copies, which are actively transcribed. However they are less interesting when studying the impact of L1 insertions on genic transcription. Therefore, a valuable improvement would be to implement a specific annotation scheme for this type of transcripts, which would permit to measure their proportion among L1 chimeric transcripts, and eventually to filter them out. Finally, the methods used in the previous studies were intrinsically low-throughput and they probably identify only a tiny fraction of all existing L1 chimeric transcripts. Thus, the extent of L1-mediated transcriptional variation might be much more important than previously anticipated.

Inversely, approximately one quarter of L1 chimeric transcripts found in previous studies were not computationally detected. When we checked for the missed cases, we found that the chimeric *reads* were indeed detected in 90% of the cases in the early steps of the computational process, but their number was too small to influence transcriptome assembly at the isoform level in the later steps. A possible way to circumvent this problem could be to remove some of the non-chimeric reads overlapping with the chimeric reads when performing *de novo* assembly without the chimeric reads (SANS assembly), to render their effect detectable. Another known limitation of our technique is that it can only detect isoforms, which incorporate a detectable LINE-1 fragments in the mature transcript. Although it is in principle possible to correlate the presence/absence of any type of isoforms with the presence/absence of a specific L1 copy, the use of L1 chimeric reads brings another level of evidence, beyond correlative observations, to support the direct implication of L1.

The novelty and originality of our method lie in the fact that we use two-tier *de novo* transcriptome assembly to identify L1 chimeric transcripts. Like previously described methods dedicated to the identification of DNA structural variants, we use the information contained within the discordant and split reads to identify L1 chimeric reads. However, our pipeline does not stop at this step, and uses L1 chimeric reads to fuel two *de novo* transcriptome assemblies: one without the chimeric reads and the other with all the reads including the chimeric ones. Then comparing the two assembled transcripts helps us to pinpoint the isoforms directly contributed by the L1 chimeric reads. No other published method till date is able to identify L1 chimeric transcripts or L1-mediated transcript isoforms computationally.

2.2. Technical challenges for the detection of L1-mediated transcript variants

Several technical challenges have been solved in order to correctly identifying such L1-mediated transcriptional variants.

A first bottleneck is to correctly map reads to their biological molecule of origin without loosing meaningful information. With respect to this point, mapping transcriptomics data onto a genome reference requires the ability to correctly identify exon-exon junction and to keep both ends of the junction since they will be essential for reconstructing the different transcript isoforms. Especially mapping reads that span more than 2 exons has been challenging and most of the mappers leave them unaligned or map them incorrectly. Although this type of reads were unusual in the early RNA-seq experiments due to very short read length (36 bp or 75 bp), it is much more frequent in recent datasets, such as those used in our study (2x150 bp). Because these reads are highly informative, it is essential to map them properly. We compared many mappers and finally used HISAT (415) for our pipeline because it can achieve high accuracy of mapping using hierarchical indexing for spliced alignment of transcripts and has optimized local realignment for precise exon-intron junction definition (Figure 25). In addition, HISAT is more than 50 times faster than TopHat2, a commonly used mapper, and uses much less Random Access Memory (RAM) compared to STAR, the main mapper used by the ENCODE project. Thus,

our method can run in hours and not days or weeks using minimal RAM and machine resources, which is a notable advantage.

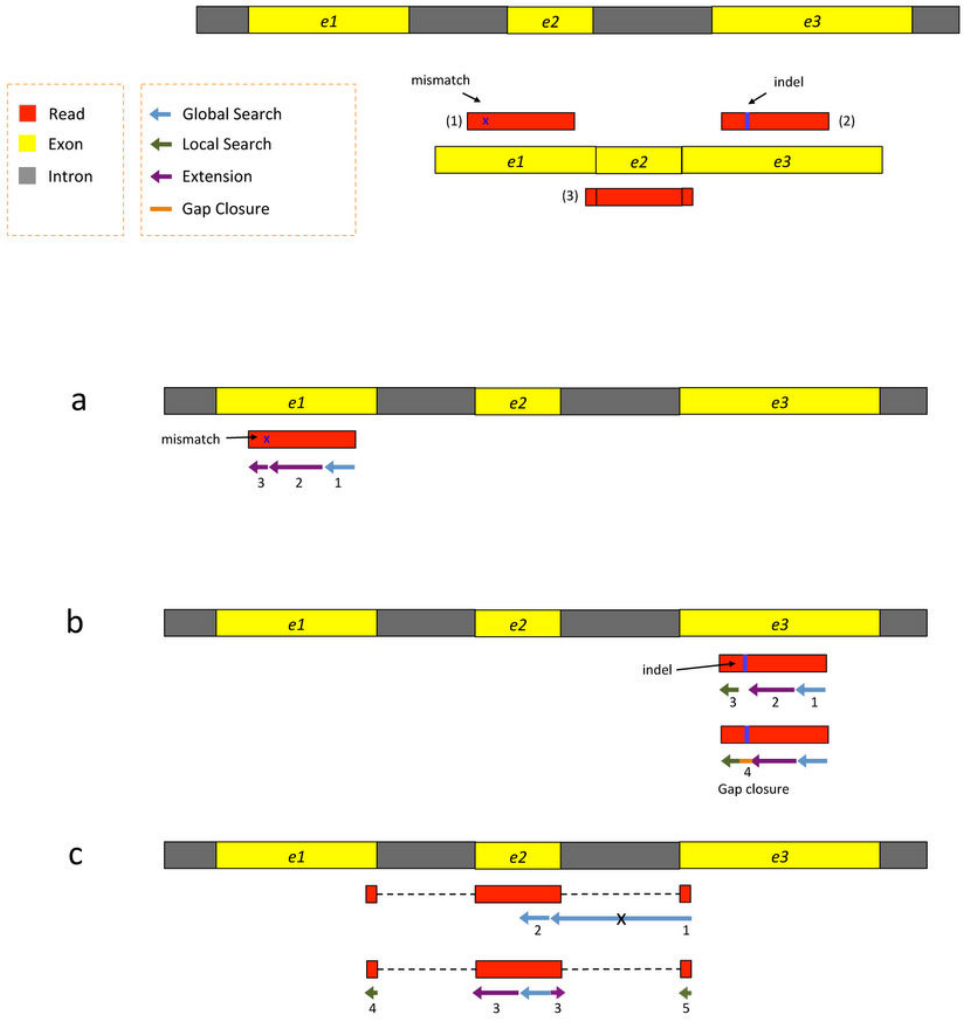


Figure 25: Handling reads spanning 3 exons by HISAT (415)**.** Reads are shown in red color. Exons in yellow and the Introns in brown. Alignment of one exonic read with one mismatch, one exonic read with an indel, and three exon-spanning reads with two small anchors on both sides. Reads are 100-bp long.

A second challenging aspect is the ability to reconstruct transcripts from aligned reads, especially to achieve assembly of complete isoforms (436). The use of StringTie assembler (414) has been critical to successfully perform transcript reconstruction. Figure 26 shows two cases of transcript reconstruction using Cufflinks (417) and StringTie (414). Cufflinks was unable to reconstruct full-length isoforms in contrast to StringTie. StringTie has also been described to perform convincing assembly of highly covered regions, which has posed a considerable challenge for transcript assembly in the past (414). One such case is that of intron retention events in nested genes. An example is shown in Figure 27.

We tried different strategies to achieve better transcriptome assembly like an optional step of super-read creation (413), a technique borrowed from *de novo* genome assembly approach but applied to enhance transcript assembly.

Theoretically, super-read creation could allow assembling error-free and longer scaffolds, specially when read length is heterogeneous. Using our relatively long read pairs, this optional step did not prove to be useful. However, additional tests would be required to conclude with RNA-seq data of different quality.

Altogether, the combination of HISAT and StringTie has been central in successfully implementing our algorithm to discover L1-mediated alternative transcripts.

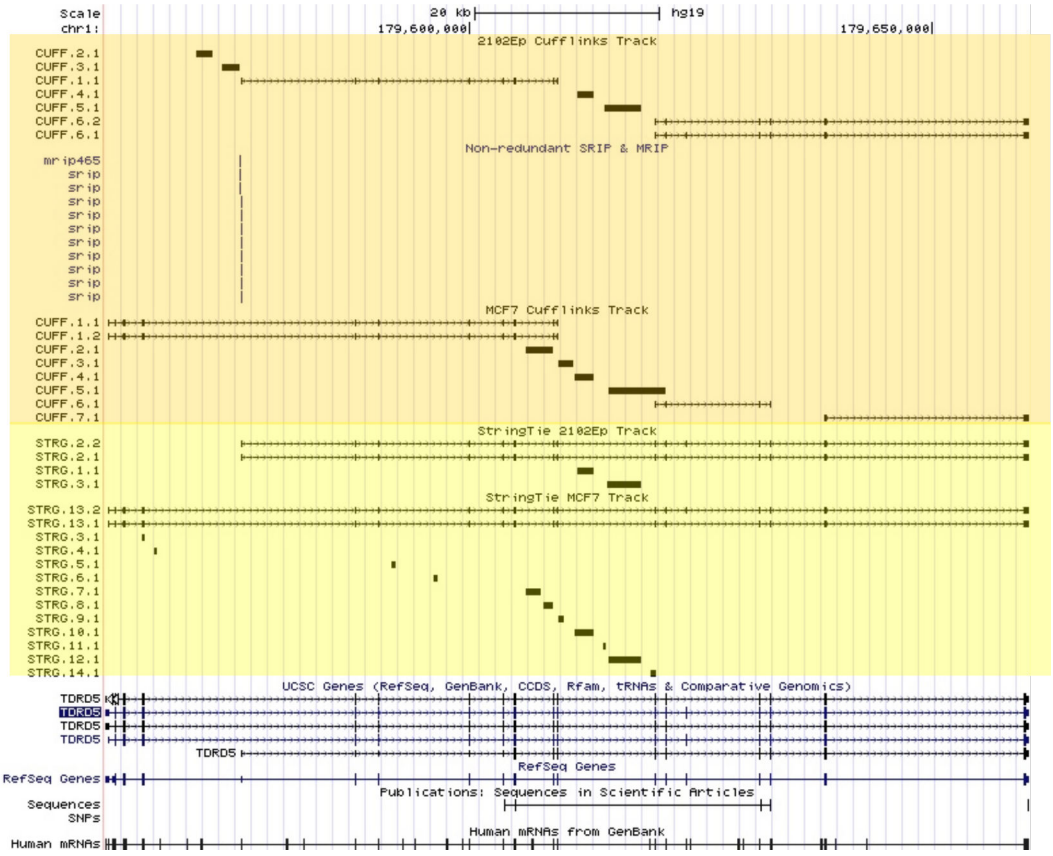


Figure 26: **Comparison of transcript reconstruction performance by StringTie and Cufflinks**. These two shaded panels show transcript reconstruction using Cufflinks (orange) and StringTie (yellow) assemblers for the *TDRD5* gene. Note that full transcripts are only assembled with StringTie, including the shorter isoforms initiated by an L1 antisense promoter in 2102Ep cells. The L1 position is shown in the 'SRIP & MRIP' lane. In contrast, Cufflinks generate several fragmented transcripts.

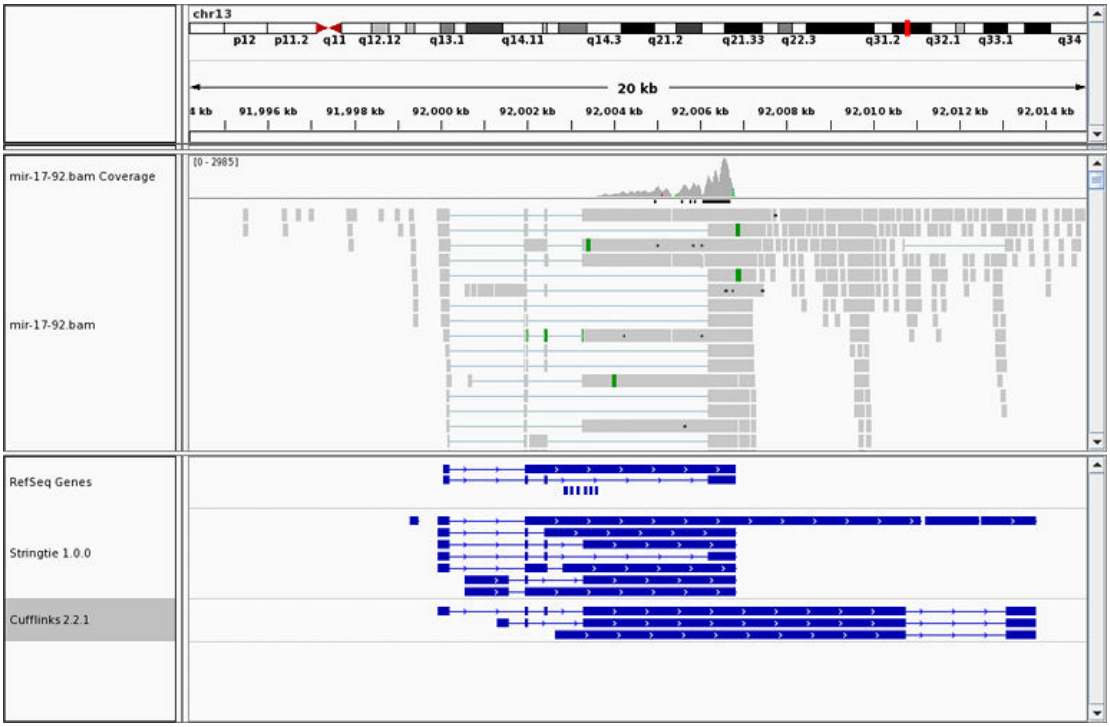


Figure 27: Intron retention detection by StringTie. Screenshot from IGV shows an example of intron retention in assembled transcripts from human kidney cell RNA-Seq data. It shows increased transcriptional activity in regions containing the miR-17-92 cluster. The 6 miRNA of the miR-17-92 cluster are encoded within the 3rd intron of the *MIR17HG* gene depicted in the RefSeq lane. Read alignments across this entire intron with nested architecture limit other assemblers from performing correct transcriptome assembly.

3. The landscape of L1-mediated structural variants of the human transcriptome.

3.1. Intron retention is the most frequently detected alternative splicing event due to L1

Among the different alternative splicing events, intron retention was found to be the most frequent. L1 intronic insertions have been shown to significantly alter transcript splicing through intron retention but also by exonization or exon skipping (375). From a biological point of view, retained introns have the potential to code for protein domains, but we found that almost 90% of retained introns occur within 3' UTRs. Intron retention has been shown to be among the most prominent alternative splicing events in general, not only due to L1 (437, 438), and this general phenomenon also preferentially impacts UTRs and more specifically 3' UTRs (438). A possible consequence of these observations, could be a modification of mRNA stability (439). Cis-acting elements within 3' UTR could affect the stability of transcript variants. This would be of particular interest to check for any relationship between this type of events and L1 orientation within the gene in future analyses.

Several other types of alternative splicing events could be detected, with the exception of mutually exclusive exons, which were rare. Another abundant type of

events detected was alternate first exons as shown in Figure S2 of Article-III for the *SQSTM1* gene and its two transcript isoforms (*STRG.19445.2 and STRG.19445.6*). In a study published by Tan *et al.*, it was shown that alternate first exons are widespread even though they were looking only for erythroid genes (440). The simplest way that alternate first exons can affect the protein isoform is by the inclusion of different start codon. This can produce a protein with different N-terminal amino acids, which can greatly alter the biological properties of this protein. Alternate first exons can also affect mRNA translation efficiency, when the 5' UTR is modified, but not the coding sequence. Alternate 5' or 3' splice sites cases have also been detected. A case of alternate 5' splice site isoforms has been shown for the *PHOSPHO2* gene (Figure S5 of Article-III). An alternate case of 3' splice site for the *CKLF-CMTM* gene is represented Figure S6 of Article-III. In general, the biological consequences of this kind of splicing could be to expend their coding capacity, which could result into alternative functional characteristics. For example, extracellular matrix proteins function in many critical processes in different tissues, and are coded by genes with multiple alternative splicing transcripts with distinct biological function (441). There have been reports of widespread exonization of transposable element sequences, including L1, which has suggested a potential for epigenetic regulation in human coding sequences (442). We also found exonization / exon skipping cases as depicted in Figure S2 (Article-III) for *USP33*. This gene has recently been found to be associated with lung cancer where it was proposed to be a candidate tumor suppressor and to be used as a prognostic marker (443).

We used SUPPA (418) utility to detect and annotate alternative splicing events. It provides us the maximum number of events associated with L1. It also allows the calculation of relative inclusion values (PSI) of alternate splicing events, which is the fraction of mRNA isoforms that includes an exon or a specific form of event (29, 444). However, the main limitation of this method is that it bases its prediction on the number of possible conformations, which might not always be the case for complex splicing events. On the contrary splicing complexity could rather be described by binary change, in such cases just one or two exon boundary changes cannot describe such a splicing situation (432). Therefore, we should enhance our alternate splice detection pipeline to find more complex events using transcript isoform changes (432).

3.2. The L1 antisense promoter provides alternative promoter activity to cellular genes

By analyzing only two cell lines, our pipeline identified many of the previously known examples of antisense transcripts generated by L1 elements. One such example is shown in Figure S7 (Article-III). This transcript (*STRG.295553.3*) in the *RAB3IP* gene is in opposite orientation as compared to L1 and starts within the 5' UTR of a full-length L1HS copy. It corresponds to a short transcript isoform of *RAB3IP*. Interestingly, this L1HS copy is in the reference genome and was independently mapped in our laboratory by ATLAS-seq in MCF7 cells and was also recorded in euL1db, highlighting the interest to combine L1 genome-wide maps or euL1db records of polymorphic insertions with transcript prediction to increase the predictive power of our approach. This transcript was further supported by a string of ESTs

containing pieces of L1 sequence (ESTs BE617461 and BE765188). *RAB3IP* is a protein-coding oncogene which has been shown to display aberrant transcription due to loss of methylation in specific intronic regions within L1 promoters, suggesting a potential role in malignancy (68, 433).

3.3. Both L1HS and older L1 subfamilies contribute to chimeric transcript formation

Our results also confirmed that apart from the youngest and actively jumping L1HS subfamily, older L1 subfamilies were forming almost 19% of source repeat elements for L1 chimeric transcripts, most prominently L1PA3, L1PA2, L1PA4, L1PA7, L1P15, L1PA6, L1PA10, and L1PB1. Because these elements are fixed in humans, it is tempting to speculate that the chimeric transcript have been positively selected for a beneficial function. It would be interesting to test if traces of such a selective process can be detected by comparing their sequences among other primates.

3.4. L1 contribute to novel exons

We could also show the direct contribution of L1 chimeric reads in the formation of novel exons (0.8% and 1.0 % for MCF7 and 2102Ep, respectively) and observed (27 and 24 in 2102Ep and MCF7, respectively) multi-exon transcripts being donated by chimeric reads (56). We also observed 1.3% and 1.6% of novel loci in MCF7 and 2102Ep cancer cell lines, respectively, being directly created by L1 chimeric reads. Apart from the exonic regions themselves, 5793 and 5410 novel splice sites in MCF7 and 2102Ep cell lines were contributed by L1 chimeric read respectively.

We were expecting to identify much more novel retrotransposon transcripts due to chimeric transcripts originating from L1 transcription bypassing L1 polyadenylation signal and ending in the flanking sequence. However, since our pipeline does not include a specific module for annotating such transcripts yet. Therefore, we cannot calculate any realistic estimates of this type of events. Almost ~1% of the human genome has been generated by transduction, a number comparable to the exonic percentage of the genome. This highlights the role of L1 in genomic plasticity by shuffling genomic DNA (27). 3' transduction has also been found to be at the origin of a significant portion of cancer genomes (318).

4. Perspectives

4.1. Identification of functionally relevant polymorphic L1 copies from RNA-seq data

Our current approach was first to identify the location of all L1 copies within a given sample, through ATLAS-seq or database comparisons (reference insertions from UCSC repeatmasker track, or polymorphic inserions from euL1db). Then we identified in RNA-seq data L1 chimeric reads and putative alternative transcripts. Finally, we could use L1 mapping information to add further support for a given transcript.

An alternative approach could be to directly use the chimeric read detection module to identify polymorphic L1HS-Ta insertions from RNA-seq data. This module can use a strategy similar to the Mobster algorithm (319), taking advantage of discordant and split read pairs to discover non-reference L1HS-Ta elements. Although, L1 mapping will not be comprehensive, a major advantage of this approach would be to only highlight L1 insertions with a potential impact on gene structure, thus functional variants. In addition, it could benefit from a vast amount of existing RNA-seq data, publicly available. This could be a general strategy for the discovery of disease-specific biomarkers for which DNA-sequencing data are not necessarily available or easily obtained.

4.2. Cancer biomarker discovery

Iskow *et al.* (313), using genomic methylation patterns, could discriminate between lung cancers with or without ongoing somatic L1 insertions, suggesting a role of hypomethylation in the activation of L1 retrotransposition in human cancers. The relationship between genomic instability, which is one of the main factors in cancer development and L1 hypomethylation has long been studied (13, 14, 228, 445). L1 hypomethylation not only leads to L1 retrotransposition, but also to the activation of L1 chimeric transcripts. The most striking example is a short isoform of the proto-oncogene *c-MET* produced upon hypomethylation and antisense promoter activation of an intronic L1 element. This isoform not only interferes with Met signaling, but can also be used to detect bladder cancer and could also be adversely induced by hypomethylating agents used as anticancer drugs (68, 446, 447).

Our pipeline has the potential to identify such biomarkers. Toward this goal, a module to compare L1-induced isoforms between two physiological situations (normal-tumor pairs for instance), would be very useful.

5. Final conclusion

In conclusion, we have developed computational tools to identify qualitative changes (alternative or novel transcript isoforms) of the human transcriptome resulting from L1 elements. These tools could be extended to other organisms and other mobile genetic elements, as far as their genomic and mobilome sequences are available. In the longer term, extending the capability of our approach to highlight quantitative changes in gene expression due to transposable elements will be the next frontier.

CONCLUSION AND PERSPECTIVES

Les rétrotransposons LINE-1 (L1) sont les seuls élément génétiques mobiles actifs et autonomes dans le génome humain. Leur réplication passe par un intermédiaire ARN et une étape de réverse transcription couplée à l'intégration dans le génome hôte. Le mécanisme qui dirige le choix du site d'intégration n'est toujours pas complètement clarifié. En se basant sur des tests quantitatifs permettant de mesurer l'efficacité de la réverse transcription de façon directe, nous avons pu évaluer l'influence de la séquence du site d'intégration et de sa structure sur l'étape de reverse transcription. En testant plus de 65 amorces différentes, nous avons observé que certains sites sont des substrats préférentiels pour l'étape de réverse transcription. Nous avons ainsi montré l'importance d'une complémentarité entre l'ADN cible et la queue poly(A) de l'ARN L1 pour un amorçage efficace de la réverse transcription. Les 4 nucléotides terminaux sont critiques, mais jusqu'à 10 nucléotides peuvent influencer ce processus, éventuellement en compensant des mésappariements terminaux. Ainsi, nous proposons que ce mécanisme puisse contribuer à la distribution des nouvelles insertions LINE-1 dans le génome humain.

Le rôle de la rétrotransposition comme source de diversité génétique, notamment de variations structurales, pouvant conduire à des maladies génétiques chez l'Homme a été montré dans plusieurs études. Les progrès des technologies de séquençage à haut-débit ont mis en lumière l'ampleur de ces variations. Ils ont également permis de découvrir que les L1s ne sont pas seulement capable de mobilisation dans la lignée germinale, aboutissant à des variations génétiques héréditaires, mais peuvent également rétrotransposer dans les tissues somatiques, comme les cellules souches embryonnaires, les cellules progénitrices neuronales ou dans plusieurs cancers. En conséquence, la compréhension du lien entre polymorphisme d'insertions et phénotype ou pathologie nécessite de disposer de répertoire précis et complet des polymorphismes d'insertion d'éléments L1 dans les génomes des individus ou des cellules concernés. Dans ce but, nous avons développé euL1db, la base de données européenne des insertions du rétrotransposon L1 humain (disponible à l'adresse http://euL1db.unice.fr), qui compile l'ensemble des insertions identifiées dans des échantillons humains sains ou pathologiques et publiées dans des journaux scientifiques, Une particularité importante d'euL1db est que les insertions peuvent être analysées au niveau de chaque échantillon pour faciliter la corrélation entre la présence/absence d'insertion L1 et un phénotype ou une maladie spécifique. euL1db fournit un accès centralisé et facilité aux insertions L1 somatiques et germinales ce qui est indispensable pour élucider l'impact physiologiques et pathologiques des nouvelles insertions. Cette ressource peut être utile dans plusieurs domaines comme la génétique humaine, les neurosciences ou la génomique du cancer.

Les insertions de L1s peuvent affecter l'expression génique de différentes manières : en changeant la séquence codante au niveau d'un exon, en s'insérant dans un intron qui sera par la suite conservé dans l'ARNm mature, par exonisation de séquences L1, par transduction de séquences codantes ou régulatrices. En effet, la rétrotransposition des L1s aboutit également à disperser un grand nombre de sites

accepteur ou donneur d'épissage présents dans la séquence des L1, dont certains sont clairement fonctionnels. L'introduction de nouveaux sites d'épissage par les rétrotransposons peut ainsi engendrer une perturbation considérable de la structure génique voire la création de nouveaux gènes codants ou non-codants. Les L1s contiennent un promoteur antisens (ASP) à leur extrémité 5' UTR qui peut conduire à des initiations alternatives de la transcription pour de nombreux gènes humains. A l'autre extrémité, les L1s peuvent également provoquer une polyadénylation précoce des transcrits dans lesquels ils sont insérés. Ces transcrits raccourcis pourront éventuellement eux-même être à l'origine d'isoformes protéiques tronqués. Ainsi, la dérépression ou l'activation de copies L1 déjà présentes et héritées dans les tumeurs peut contribuer à la progression du cancer par l'altération de l'expression des gènes situés à leur proximité, notablement en générant des transcripts L1 chimériques. Pour étudier ce processus de façon globale, nous avons développé un logiciel qui permet d'identifier les transcripts chimériques dûs aux insertions L1 à partir de données de séquençage d'ARN (RNA-seq). Ce logiciel identifie et annote les transcripts chimériques L1 en fonction du type d'épissage alternatif produit, ainsi que les transcrits antisens. Cette stratégie permet ainsi de découvrir les différents isoformes transcriptionnels induits par les éléments L1 dans les cellules humaines.

L'expression du rétrotransposon L1 a été proposée en même temps, comme un biomarqueur pronostic potentiel de nombreux types de cancer, et comme un point de départ de l'instabilité génomique dans les tumeurs. Cependant, la manière dont l'ensemble des éléments L1 présent chez un individu est régulée au niveau transcriptionnel, et le type cellulaire ou l'environnement génomique permettant son expression demeurent inconnues. De plus, plusieurs insertions somatiques ont été décrites dans plusieurs types de tumeur, mais leur impact sur l'expression des gènes n'est pas encore bien clarifié. Les outils développés lors de ce travail permettront d'éclairer ces deux aspects. Sur le long terme, cette approche apportera un cadre de travail conceptuel et technologique pour analyser des grands jeux de données, comme ceux mis à la disposition de la communauté scientifique par le consortium international de génomique du cancer (*international cancer genome consortium)*, dans le but d'améliorer notre compréhension des mécanismes menant à la plasticité du transcriptome dans les cellules cancéreuses et d'apporter une base rationnelle à l'utilisation des L1s comme biomarqueurs.

BIBLIOGRAPHY

1. Bhattacharyya MK, Smith AM, Ellis THN, Hedley C, Martin C (1990) The wrinkled-seed character of pea described by Mendel is caused by a transposon-like insertion in a gene encoding starch-branching enzyme. *Cell* 60(1):115–122.

2. Smit AF (1996) The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev* 6(6):743–748.

3. Lander ES, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.

4. Kazazian H. H. J (2004) Mobile elements: drivers of genome evolution. *Science* 303(5664):1626–1632.

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

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

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

8. Hohjoh H, Singer MF (1997) Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J* 16(19):6034–6043.

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

10. Feng Q, Moran J V, Kazazian HH, Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* 87(5):905–916.

11. Kulpa D a, Moran J V (2006) Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat Struct Mol Biol* 13(7):655–660.

12. Doucet AJ, et al. (2010) Characterization of LINE-1 ribonucleoprotein particles. *PLoS Genet* 6(10):1–19.

13. Gilbert N, Lutz-Prigge S, Moran J V (2002) Genomic deletions created upon LINE-1 retrotransposition. *Cell* 110(3):315–325.

14. Symer DE, et al. (2002) Human l1 retrotransposition is associated with genetic instability in vivo. *Cell* 110(3):327–338.

15. Ostertag EM, Kazazian H.H. J, Kazazian HH (2001) Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* 11(12):2059–2065.

16. Moran J V, DeBerardinis RJ, Kazazian H. H. J (1999) Exon shuffling by L1 retrotransposition. *Science* 283(5407):1530–1534.

17. Goodier JL, Ostertag EM, Kazazian HH (2000) Transduction of 3'-flanking sequences is common in L1 retrotransposition. *Hum Mol Genet* 9(4):653–657.

18. Repanas K, et al. (2007) Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic Acids Res* 35(14):4914–4926.

19. Viollet S, Monot C, Cristofari G (2014) L1 retrotransposition: The snap-velcro model and its consequences. *Mob Genet Elem* 4(1):e28907.

20. Xing J, et al. (2009) Mobile elements create structural variation: Analysis of a complete human genome. *Genome Res* 19(9):1516–1526.

21. Kidd JM, et al. (2010) A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell* 143(5):837–847.

22. Korbel JO, et al. (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318(5849):420–426.

23. Lam HYK, et al. (2010) Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nat Biotechnol* 28(1):47–55.

24. Stewart C, et al. (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet* 7(8):e1002236.

25. Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE (2004) Natural genetic variation caused by transposable elements in humans. *Genetics* 168(2):933–951.

26. Lee E, et al. (2012) Landscape of somatic retrotransposition in human cancers. *Science* 337(6097):967–971.

27. Pickeral OK, Makaowski W, Boguski MS, Boeke JD (2000) Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res*

10(4):411–415.

28. Djebali S, et al. (2012) Landscape of transcription in human cells. *Nature* 489(7414):101–108.

29. Wang ET, et al. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature* 456(7221):470–476.

30. Belancio VP, Hedges DJ, Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res* 34(5):1512–1521.

31. van den Hurk JAJM, et al. (2003) Novel types of mutation in the choroideremia (CHM) gene: a full-length L1 insertion and an intronic mutation activating a cryptic exon. *Hum Genet* 113(3):268–275.

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

33. Narita N, et al. (1993) Insertion of a 5’ truncated L1 element into the 3' end of exon 44 of the dystrophin gene resulted in skipping of the exon during splicing in a case of Duchenne muscular dystrophy. *J Clin Invest* 91(5):1862–1867.

34. Takahara T, et al. (1996) Dysfunction of the Orleans reeler gene arising from exon skipping due to transposition of a full-length copy of an active L1 sequence into the skipped exon. *Hum Mol Genet* 5(7):989–993.

35. Meischl C, Boer M, Ahlin A, Roos D (2000) A new exon created by intronic insertion of a rearranged LINE-1 element as the cause of chronic granulomatous disease. *Eur J Hum Genet* 8(9):697–703.

36. Perepelitsa-Belancio V, Deininger P (2003) RNA truncation by premature polyadenylation attenuates human mobile element activity. *Nat Genet* 35(4):363–366.

37. Moran J V, et al. (1996) High frequency retrotransposition in cultured mammalian cells. *Cell* 87(5):917–927.

38. Derti A, et al. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res* 22(6):1173–1183.

39. Speek M (2001) Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol* 21(6):1973–1985.

40. Nigumann P, Redik K, Mätlik K, Speek M (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*

79(5):628–634.

41. Birchmeier C, Birchmeier W, Gherardi E, Vande Woude GF (2003) Met, metastasis, motility and more. *Nat Rev Mol Cell Biol* 4(12):915–925.

42. Ma PC, Maulik G, Christensen J, Salgia R (2003) c-Met: structure, functions and potential for therapeutic inhibition. *Cancer Metastasis Rev* 22(4):309–325.

43. Thornburg BG, Gotea V, Makalowski W (2006) Transposable elements as a significant source of transcription regulating signals. *Gene* 365:104–110.

44. Polak P, Domany E (2006) Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes. *BMC Genomics* 7:133.

45. Johnson R, et al. (2006) Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res* 34(14):3862–3877.

46. Wang T, et al. (2007) Species-specific endogenous retroviruses shape the transcriptional network of the human tumor suppressor protein p53. *Proc Natl Acad Sci U S A* 104(47):18613–18618.

47. Bourque G (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Curr Opin Genet Dev* 19(6):607–612.

48. Macfarlan TS, et al. (2011) Endogenous retroviruses and neighboring genes are coordinately repressed by LSD1/KDM1A. *Genes Dev* 25(6):594–607.

49. Macfarlan TS, et al. (2012) Embryonic stem cell potency fluctuates with endogenous retrovirus activity. *Nature* 487(7405):57–63.

50. Vinckenbosch N, Dupanloup I, Kaessmann H (2006) Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A* 103(9):3220–3225.

51. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H (2005) Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* 3(11):e357.

52. Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. *Nat Genet* 24(4):363–367.

53. Wei W, et al. (2001) Human L1 retrotransposition: cis preference versus trans complementation. *Mol Cell Biol* 21(4):1429–1439.

54. Brosius J (1991) Retroposons--seeds of evolution. *Science* 251(4995):753.

55. Kaessmann H, Vinckenbosch N, Long M (2009) RNA-based gene duplication: mechanistic and evolutionary insights. *Nat Rev Genet* 10(1):19–31.

56. Brosius J (1999) Genomes were forged by massive bombardments with retroelements and retrosequences. *Genetica* 107(1-3):209–238.

57. Long M, Betran E, Thornton K, Wang W (2003) The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 4(11):865–875.

58. Emerson JJ, Kaessmann H, Betran E, Long M (2004) Extensive gene traffic on the mammalian X chromosome. *Science* 303(5657):537–540.

59. Jordan IK, Rogozin IB, Glazko G V, Koonin E V (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* 19(2):68–72.

60. Feschotte C (2008) Transposable elements and the evolution of regulatory networks. *Nat Rev Genet* 9(5):397–405.

61. Jurka J (2008) Conserved eukaryotic transposable elements and the evolution of gene regulation. *Cell Mol Life Sci* 65(2):201–204.

62. Sinzelle L, Izsvak Z, Ivics Z (2009) Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cell Mol Life Sci* 66(6):1073–1093.

63. Kapusta A, et al. (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet* 9(4):e1003470.

64. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annu Rev Biochem* 81:145–166.

65. Schulz WA (2006) L1 retrotransposons in human cancers. *J Biomed Biotechnol* 2006(1):83672.

66. Daskalos A, et al. (2009) Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *Int J Cancer* 124(1):81–87.

67. Cruickshanks H a., Tufarelli C (2009) Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. *Genomics* 94(6):397–406.

68. Wolff EM, et al. (2010) Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet*

6(4):e1000917.

69. Macia A, et al. (2011) Epigenetic control of retrotransposon expression in human embryonic stem cells. *Mol Cell Biol* 31(2):300–316.

70. Cowley M, Oakey RJ (2013) Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLoS Genet* 9(1). doi:10.1371/journal.pgen.1003234.

71. Wicker T, et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 8(12):973–982.

72. McClintock B (1950) The Origin and Behavior of Mutable Loci in Maize. *Proc Natl Acad Sci U S A* 36(6):344–355.

73. Adams JW, Kaufman RE, Kretschmer PJ, Harrison M, Nienhuis AW (1980) A family of long reiterated DNA sequences, one copy of which is next to the human beta globin gene. *Nucleic Acids Res* 8(24):6113–6128.

74. Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. *Trends Genet* 5(4):103–107.

75. Eickbush TH, Jamburuthugoda VK (2008) The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* 134(1-2):221–234.

76. Llorens C, et al. (2011) The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res* 39(Database issue):D70–4.

77. Mizuuchi M, Baker TA, Mizuuchi K (1992) Assembly of the active form of the transposase-Mu DNA complex: a critical control point in Mu transposition. *Cell* 70(2):303–311.

78. Kapitonov V V, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet* 23(10):521–529.

79. Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* 3(5):329–341.

80. Bureau TE, Ronald PC, Wessler SR (1996) A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. *Proc Natl Acad Sci U S A* 93(16):8524–8529.

81. Buisine N, Tang CM, Chalmers R (2002) Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic Neisseria sp. *FEBS Lett* 522(1-3):52–58.

82. De Gregorio E, Silvestro G, Petrillo M, Carlomagno MS, Di Nocera PP (2005)

Enterobacterial repetitive intergenic consensus sequence repeats in yersiniae: genomic organization and functional properties. *J Bacteriol* 187(23):7945–7954.

83. Ivics Z, et al. (2009) Transposon-mediated genome manipulation in vertebrates. *Nat Methods* 6(6):415–422.

84. Aronovich EL, McIvor RS, Hackett PB (2011) The Sleeping Beauty transposon system: a non-viral vector for gene therapy. *Hum Mol Genet* 20(R1):R14–20.

85. Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* 9(10):3353–3362.

86. Curcio MJ, Belfort M (1996) Retrohoming: cDNA-mediated mobility of group II introns requires a catalytic RNA. *Cell* 84(1):9–12.

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

88. Pardue M-LL, DeBaryshe PG (2003) Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. *Annu Rev Genet* 37:485–511.

89. Gladyshev EA, Arkhipova IR (2007) Telomere-associated endonuclease-deficient Penelope-like retroelements in diverse eukaryotes. *Proc Natl Acad Sci U S A* 104(22):9352–9357.

90. Malik HS, Henikoff S, Eickbush TH (2000) Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses. *Genome Res* 10(9):1307–1318.

91. Waterston RH, et al. (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

92. Volff JN, Korting C, Schartl M (2000) Multiple lineages of the non-LTR retrotransposon Rex1 with varying success in invading fish genomes. *Mol Biol Evol* 17(11):1673–1684.

93. Schnable PS, et al. (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science* 326(5956):1112–1115.

94. Huang CRL, Burns KH, Boeke JD (2012) Active transposition in genomes. *Annu Rev Genet* 46:651–675.

95. Aparicio S, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297(5585):1301–1310.

96. Lindblad-Toh K, et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* 438(7069):803–819.

97. Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem Genet* 12(4):257–269.

98. Kumar A, Bennetzen JL (1999) Plant retrotransposons. *Annu Rev Genet* 33:479–532.

99. Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. *Ann Bot* 95(1):127–132.

100. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* 41:331–368.

101. Lisch D (2013) How important are transposons for plant evolution? *Nat Rev Genet* 14(1):49–61.

102. Bessereau J-L (2006) Transposons in C. elegans. *WormBook*:1–13.

103. Kapitonov V V, Jurka J (2003) Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proc Natl Acad Sci U S A* 100(11):6569–6574.

104. Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, Gonzalez J (2011) Population genomics of transposable elements in Drosophila melanogaster. *Mol Biol Evol* 28(5):1633–1644.

105. Kidwell MG, Kidwell JF, Sved JA (1977) Hybrid Dysgenesis in DROSOPHILA MELANOGASTER: A Syndrome of Aberrant Traits Including Mutation, Sterility and Male Recombination. *Genetics* 86(4):813–833.

106. Mitra R, et al. (2013) Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. *Proc Natl Acad Sci U S A* 110(1):234–239.

107. Ray DA, et al. (2008) Multiple waves of recent DNA transposon activity in the bat, Myotis lucifugus. *Genome Res* 18(5):717–728.

108. Belshaw R, Katzourakis A, Paces J, Burt A, Tristem M (2005) High copy number in human endogenous retrovirus families is associated with copying mechanisms in addition to reinfection. *Mol Biol Evol* 22(4):814–817.

109. Coffin JM, Hughes SH, Varmus HE (1997) The Interactions of Retroviruses and their Hosts. eds Coffin JM, Hughes SH, Varmus HE (Cold Spring Harbor

(NY)).

110. Grow EJ, et al. (2015) Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* 522(7555):221–225.

111. Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* 13(1):36–46.

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

113. Kazazian H. H. J (1998) Mobile elements and disease. *Curr Opin Genet Dev* 8(3):343–350.

114. Dombroski BA, Mathias SL, Nanthakumar E, Scott AF, Kazazian HH (1991) Isolation of an active human transposable element. *Science* 254(5039):1805–1808.

115. Hohjoh H, Singer MF (1997) Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J* 16(19):6034–6043.

116. Ostertag EM, Kazazian H.H. J (2001) Twin priming: A proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res* 11(12):2059–2065.

117. Kramerov DA, Vassetzky NS (2005) Short retroposons in eukaryotic genomes. *Int Rev Cytol* 247:165–221.

118. Ohshima K, Okada N (2005) SINEs and LINEs: symbionts of eukaryotic genomes with a common tail. *Cytogenet Genome Res* 110(1-4):475–490.

119. Deragon J-M, Zhang X (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst Biol* 55(6):949–956.

120. Donehower LA, Slagle BL, Wilde M, Darlington G, Butel JS (1989) Identification of a conserved sequence in the non-coding regions of many human genes. *Nucleic Acids Res* 17(2):699–710.

121. Jurka J, Walichiewicz J, Milosavljevic A (1992) Prototypic sequences for human repetitive DNA. *J Mol Evol* 35(4):286–291.

122. Jurka J, Zietkiewicz E, Labuda D (1995) Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era.

Nucleic Acids Res 23(1):170–175.

123. Deininger PL, Jolly DJ, Rubin CM, Friedmann T, Schmid CW (1981) Base sequence studies of 300 nucleotide renatured repeated human DNA clones. *J Mol Biol* 151(1):17–33.

124. Ullu E, Tschudi C (1984) Alu sequences are processed 7SL RNA genes. *Nature* 312(5990):171–172.

125. Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* 3(5):370–379.

126. Cordaux R, Hedges DJ, Batzer MA (2004) Retrotransposition of Alu elements: how many sources? *Trends Genet* 20(10):464–467.

127. Wallace N, Wagstaff BJ, Deininger PL, Roy-Engel AM (2008) LINE-1 ORF1 protein enhances Alu SINE retrotransposition. *Gene* 419(1-2):1–6.

128. Roy-Engel AM, et al. (2002) Active Alu element “A-tails”: size does matter. *Genome Res* 12(9):1333–1344.

129. Wang W, Kirkness EF (2005) Short interspersed elements (SINEs) are a major source of canine genomic diversity. *Genome Res* 15(12):1798–1808.

130. Savage AL, Bubb VJ, Breen G, Quinn JP (2013) Characterisation of the potential function of SVA retrotransposons to modulate gene expression patterns. *BMC Evol Biol* 13:101.

131. Rais Y, et al. (2013) Deterministic direct reprogramming of somatic cells to pluripotency. *Nature* 502(7469):65–70.

132. Bantysh OB, Buzdin AA (2009) Novel family of human transposable elements formed due to fusion of the first exon of gene MAST2 with retrotransposon SVA. *Biochemistry (Mosc)* 74(12):1393–1399.

133. Hancks DC, Ewing AD, Chen JE, Tokunaga K, Kazazian HHJ (2009) Exon-trapping mediated by the human retrotransposon SVA. *Genome Res* 19(11):1983–1991.

134. Prak ET, Kazazian HHJ (2000) Mobile elements and the human genome. *Nat Rev Genet* 1(2):134–144.

135. Pace JK 2nd, Feschotte C (2007) The evolutionary history of human DNA transposons: evidence for intense activity in the primate lineage. *Genome Res* 17(4):422–432.

136. Deininger PL, Batzer MA, Hutchison CA 3rd, Edgell MH (1992) Master genes

in mammalian repetitive DNA amplification. *Trends Genet* 8(9):307–311.

137. Han JS, Szak ST, Boeke JD (2004) Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. *Nature* 429(6989):268–274.

138. Li TH, Schmid CW (2004) Alu's dimeric consensus sequence destabilizes its transcripts. *Gene* 324:191–200.

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

140. Hardies SC, Martin SL, Voliva CF, Hutchison CA 3rd, Edgell MH (1986) An analysis of replacement and synonymous changes in the rodent L1 repeat family. *Mol Biol Evol* 3(2):109–125.

141. Pascale E, Liu C, Valle E, Usdin K, Furano A V (1993) The evolution of long interspersed repeated DNA (L1, LINE 1) as revealed by the analysis of an ancient rodent L1 DNA family. *J Mol Evol* 36(1):9–20.

142. Boissinot S, Chevret P, Furano A V (2000) L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol Biol Evol* 17(6):915–928.

143. Smit AF, Tóth G, Riggs AD, Jurka J (1995) Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol* 246(3):401–417.

144. Boissinot S, Furano A V (2001) Adaptive evolution in LINE-1 retrotransposons. *Mol Biol Evol* 18(12):2186–2194.

145. Cabot EL, Angeletti B, Usdin K, Furano A V (1997) Rapid evolution of a young L1 (LINE-1) clade in recently speciated Rattus taxa. *J Mol Evol* 45(4):412–423.

146. Casavant NC, Hardies SC (1994) The dynamics of murine LINE-1 subfamily amplification. *J Mol Biol* 241(3):390–397.

147. Adey NB, et al. (1994) Rodent L1 evolution has been driven by a single dominant lineage that has repeatedly acquired new transcriptional regulatory sequences. *Mol Biol Evol* 11(5):778–789.

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

149. Scott AF, et al. (1987) Origin of the human L1 elements: Proposed progenitor

genes deduced from a consensus DNA sequence. *Genomics* 1(2):113–125.

150. Wincker P, Jubier-Maurin V, Roizes G (1987) Unrelated sequences at the 5' end of mouse LINE-1 repeated elements define two distinct subfamilies. *Nucleic Acids Res* 15(21):8593–8606.

151. Furano A V, Robb SM, Robb FT (1988) The structure of the regulatory region of the rat L1 (L1Rn, long interspersed repeated) DNA family of transposable elements. *Nucleic Acids Res* 16(19):9215–9231.

152. Padgett RW, Hutchison CA 3rd, Edgell MH (1988) The F-type 5' motif of mouse L1 elements: a major class of L1 termini similar to the A-type in organization but unrelated in sequence. *Nucleic Acids Res* 16(2):739–749.

153. Jubier-Maurin V, Cuny G, Laurent AM, Paquereau L, Roizes G (1992) A new 5' sequence associated with mouse L1 elements is representative of a major class of L1 termini. *Mol Biol Evol* 9(1):41–55.

154. Schichman SA, Adey NB, Edgell MH, Hutchison CA 3rd (1993) L1 A-monomer tandem arrays have expanded during the course of mouse L1 evolution. *Mol Biol Evol* 10(3):552–570.

155. 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(1):78–87.

156. Furano A V, Duvernell DD, Boissinot S (2004) L1 (LINE-1) retrotransposon diversity differs dramatically between mammals and fish. *Trends Genet* 20(1):9–14.

157. Pascale E, Valle E, Furano A V (1990) Amplification of an ancestral mammalian L1 family of long interspersed repeated DNA occurred just before the murine radiation. *Proc Natl Acad Sci U S A* 87(23):9481–9485.

158. Mathews LM, Chi SY, Greenberg N, Ovchinnikov I, Swergold GD (2003) Large differences between LINE-1 amplification rates in the human and chimpanzee lineages. *Am J Hum Genet* 72(3):739–748.

159. Ohshima K, et al. (2003) Whole-genome screening indicates a possible burst of formation of processed pseudogenes and Alu repeats by particular L1 subfamilies in ancestral primates. *Genome Biol* 4(11):R74.

160. Myers JS, et al. (2002) A comprehensive analysis of recently integrated human Ta L1 elements. *Am J Hum Genet* 71(2):312–326.

161. Feschotte C, Gilbert C (2012) Endogenous viruses: insights into viral evolution and impact on host biology. *Nat Rev Genet* 13(4):283–296.

162. Hua-Van A, Le Rouzic A, Boutin TS, Filee J, Capy P (2011) The struggle for life of the genome's selfish architects. *Biol Direct* 6:19.

163. Jurka J, Bao W, Kojima KK (2011) Families of transposable elements, population structure and the origin of species. *Biol Direct* 6:44.

164. Oliver KR, Greene WK (2011) Mobile DNA and the TE-Thrust hypothesis: supporting evidence from the primates. *Mob DNA* 2(1):8.

165. Oliver KR, Greene WK (2012) Transposable elements and viruses as factors in adaptation and evolution: an expansion and strengthening of the TE-Thrust hypothesis. *Ecol Evol* 2(11):2912–2933.

166. Lowe CB, Bejerano G, Haussler D (2007) Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A* 104(19):8005–8010.

167. Lowe CB, Haussler D (2012) 29 mammalian genomes reveal novel exaptations of mobile elements for likely regulatory functions in the human genome. *PLoS One* 7(8):e43128.

168. Cohen CJ, Lock WM, Mager DL (2009) Endogenous retroviral LTRs as promoters for human genes: a critical assessment. *Gene* 448(2):105–114.

169. Davuluri R V, Suzuki Y, Sugano S, Plass C, Huang TH-M (2008) The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* 24(4):167–177.

170. Landry JR, Mager DL, Wilhelm BT (2003) Complex controls: The role of alternative promoters in mammalian genomes. *Trends Genet* 19(11):640–648.

171. Larsen LK, Amri E-Z, Mandrup S, Pacot C, Kristiansen K (2002) Genomic organization of the mouse peroxisome proliferator-activated receptor beta/delta gene: alternative promoter usage and splicing yield transcripts exhibiting differential translational efficiency. *Biochem J* 366(Pt 3):767–775.

172. Schulte AM, et al. (1996) Human trophoblast and choriocarcinoma expression of the growth factor pleiotrophin attributable to germ-line insertion of an endogenous retrovirus. *Proc Natl Acad Sci U S A* 93(25):14759–14764.

173. van de Lagemaat LN, et al. (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* 19(10):530–536.

174. Faulkner GJ, et al. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* 41(5):563–571.

175. Bourque G, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res* 18(11):1752–1762.

176. Kunarso G, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet* 42(7):631–634.

177. Schmidt D, et al. (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148(1-2):335–348.

178. Davidson EH, Britten RJ (1979) Regulation of gene expression: possible role of repetitive sequences. *Science* 204(4397):1052–1059.

179. McClintock B (1984) The significance of responses of the genome to challenge. *Science* 226(4676):792–801.

180. Kidwell MG, Lisch D (1997) Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A* 94(15):7704–7711.

181. Brosius J (2003) The contribution of RNAs and retroposition to evolutionary novelties. *Genetica* 118(2-3):99–116.

182. Gentles AJ, et al. (2007) Evolutionary dynamics of transposable elements in the short-tailed opossum Monodelphis domestica. *Genome Res* 17(7):992–1004.

183. Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* 433(7025):481–487.

184. Ihmels J, et al. (2005) Rewiring of the yeast transcriptional network through the evolution of motif usage. *Science* 309(5736):938–940.

185. Rockman M V, et al. (2005) Ancient and recent positive selection transformed opioid cis-regulation in humans. *PLoS Biol* 3(12):e387.

186. Marcellini S, Simpson P (2006) Two or four bristles: functional evolution of an enhancer of scute in Drosophilidae. *PLoS Biol* 4(12):e386.

187. Tumpel S, et al. (2007) Expression of Hoxa2 in rhombomere 4 is regulated by a conserved cross-regulatory mechanism dependent upon Hoxb1. *Dev Biol* 302(2):646–660.

188. Theuns J, et al. (2000) Genetic variability in the regulatory region of presenilin 1 associated with risk for Alzheimer's disease and variable expression. *Hum*

Mol Genet 9(3):325–331.

189. Esterbauer H, et al. (2001) A common polymorphism in the promoter of UCP2 is associated with decreased risk of obesity in middle-aged humans. *Nat Genet* 28(2):178–183.

190. Bond GL, et al. (2004) A single nucleotide polymorphism in the MDM2 promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell* 119(5):591–602.

191. Polavarapu N, Marino-Ramirez L, Landsman D, McDonald JF, Jordan IK (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. *BMC Genomics* 9:226.

192. Cui F, Sirotin M V, Zhurkin VB (2011) Impact of Alu repeats on the evolution of human p53 binding sites. *Biol Direct* 6:2.

193. Marino-Ramirez L, Bodenreider O, Kantz N, Jordan IK (2006) Co-evolutionary rates of functionally related yeast genes. *Evol Bioinform Online* 2:271–276.

194. Teng L, Firpi HA, Tan K (2011) Enhancers in embryonic stem cells are enriched for transposable elements and genetic variations associated with cancers. *Nucleic Acids Res* 39(17):7371–7379.

195. Piriyapongsa J, Marino-Ramirez L, Jordan IK (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics* 176(2):1323–1337.

196. Kuang W, et al. (2009) Cyclic stretch induced miR-146a upregulation delays C2C12 myogenic differentiation through inhibition of Numb. *Biochem Biophys Res Commun* 378(2):259–263.

197. Conley AB, Miller WJ, Jordan IK (2008) Human cis natural antisense transcripts initiated by transposable elements. *Trends Genet* 24(2):53–56.

198. Jjingo D, Huda A, Gundapuneni M, Marino-Ramirez L, Jordan IK (2011) Effect of the transposable element environment of human genes on gene length and expression. *Genome Biol Evol* 3:259–271.

199. Fort A, et al. (2014) Deep transcriptome profiling of mammalian stem cells supports a regulatory role for retrotransposons in pluripotency maintenance. *Nat Genet* 46(6):558–566.

200. Rowe HM, et al. (2010) KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* 463(7278):237–240.

201. Gellersen B, Brosens IA, Brosens JJ (2007) Decidualization of the human endometrium: mechanisms, functions, and clinical perspectives. *Semin*

Reprod Med 25(6):445–453.

202. Lynch VJ, Leclerc RD, May G, Wagner GP (2011) Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. *Nat Genet* 43(11):1154–9.

203. Krull M, Petrusma M, Makalowski W, Brosius J, Schmitz J (2007) Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Res* (17):1139–45.

204. Gotea V, Makalowski W (2006) Do transposable elements really contribute to proteomes? *Trends Genet* 22(5):260–267.

205. Britten RJ (1996) Cases of ancient mobile element DNA insertions that now affect gene regulation. *Mol Phylogenet Evol* 5(1):13–17.

206. Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P (2005) Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res* 33(3):946–954.

207. Zhang J (2003) Evolution by gene duplication : an update. *Trends Ecol Evol* 18(6):292–298.

208. Betran E, Long M (2002) Expansion of genome coding regions by acquisition of new genes. *Genetica* 115(1):65–80.

209. Pan D, Zhang L (2009) Burst of young retrogenes and independent retrogene formation in mammals. *PLoS One* 4(3):e5040.

210. Badeaux MA, et al. (2013) In vivo functional studies of tumor-specific retrogene NanogP8 in transgenic animals. *Cell Cycle* 12(15):2395–2408.

211. Ohshima K, Igarashi K (2010) Inference for the initial stage of domain shuffling: tracing the evolutionary fate of the PIPSL retrogene in hominoids. *Mol Biol Evol* 27(11):2522–2533.

212. Skowronski J, Singer MF (1985) Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. *Proc Natl Acad Sci U S A* 82(18):6050–6054.

213. Skowronski J, Fanning TG, Singer MF (1988) Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol Cell Biol* 8(4):1385–1397.

214. Kurose K, Hata K, Hattori M, Sakaki Y (1995) RNA polymerase III dependence of the human L1 promoter and possible participation of the RNA polymerase II factor YY1 in the RNA polymerase III transcription system. *Nucleic Acids Res* 23(18):3704–3709.

215. Nur I, Pascale E, Furano A V (1988) The left end of rat L1 (L1Rn, long interspersed repeated) DNA which is a CpG island can function as a promoter. *Nucleic Acids Res* 16(19):9233–9251.

216. Swergold GD (1990) Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol* 10(12):6718–6729.

217. Minakami R, et al. (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(12):3139–3145.

218. Athanikar JN, Badge RM, Moran J V (2004) A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res* 32(13):3846–3855.

219. 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(11):2253–2260.

220. Yang N, Zhang L, Zhang Y, Kazazian HH (2003) An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res* 31(16):4929–4940.

221. Alexandrova EA, et al. (2012) Sense transcripts originated from an internal part of the human retrotransposon LINE-1 5' UTR. *Gene* 511(1):46–53.

222. 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(10):1697–1702.

223. Tchenio T, Casella JF, Heidmann T (2000) Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res* 28(2):411–415.

224. Woodcock DM, Lawler CB, Linsenmeyer ME, Doherty JP, Warren WD (1997) Asymmetric methylation in the hypermethylated CpG promoter region of the human L1 retrotransposon. *J Biol Chem* 272(12):7810–7816.

225. Muotri AR, et al. (2010) L1 retrotransposition in neurons is modulated by MeCP2. *Nature* 468(7322):443–446.

226. Yu F, Zingler N, Schumann G, Strätling WH (2001) Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res* 29(21):4493–4501.

227. Rangwala SH, Zhang L, Kazazian HH, Kazazian H. H. J (2009) Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol*

10(9):R100.

228. Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P (2010) Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res* 38(12):3909–3922.

229. Belancio VP, Roy-Engel AM, Deininger P (2008) The impact of multiple splice sites in human L1 elements. *Gene* 411(1-2):38–45.

230. Kuwabara T, et al. (2009) Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. *Nat Neurosci* 12(9):1097–1105.

231. Cullen BR (2000) Nuclear RNA export pathways. *Mol Cell Biol* 20(12):4181–4187.

232. Ooi SL, et al. (2001) RNA lariat debranching enzyme. *Methods Enzymol* 342:233–248.

233. Gupta K, Ott D, Hope TJ, Siliciano RF, Boeke JD (2000) A human nuclear shuttling protein that interacts with human immunodeficiency virus type 1 matrix is packaged into virions. *J Virol* 74(24):11811–11824.

234. Cullen BR (1998) Retroviruses as model systems for the study of nuclear RNA export pathways. *Virology* 249(2):203–210.

235. Whittaker GR, Helenius A (1998) Nuclear import and export of viruses and virus genomes. *Virology* 246(1):1–23.

236. Alisch RS, Garcia-Perez JL, Muotri AR, Gage FH, Moran J V (2006) Unconventional translation of mammalian LINE-1 retrotransposons. *Genes Dev* 20(2):210–224.

237. Loeb DD, et al. (1986) The sequence of a large L1Md element reveals a tandemly repeated 5’ end and several features found in retrotransposons. *Mol Cell Biol* 6(1):168–182.

238. Holmes SE, Singer MF, Swergold GD (1992) Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. *J Biol Chem* 267(28):19765–19768.

239. Leibold DM, et al. (1990) Translation of LINE-1 DNA elements in vitro and in human cells. *Proc Natl Acad Sci U S A* 87(18):6990–6994.

240. Hohjoh H, Singer MF (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J* 15(3):630–639.

241. Kolosha VO, Martin SL (1997) In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc Natl Acad Sci U S A* 94(19):10155–10160.

242. Kulpa DA, Moran J V (2005) Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum Mol Genet* 14(21):3237–3248.

243. Khazina E, Weichenrieder O (2009) Non-LTR retrotransposons encode noncanonical RRM domains in their first open reading frame. *Proc Natl Acad Sci U S A* 106(3):731–736.

244. Januszyk K, et al. (2007) Identification and solution structure of a highly conserved C-terminal domain within ORF1p required for retrotransposition of long interspersed nuclear element-1. *J Biol Chem* 282(34):24893–24904.

245. Martin SL, Branciforte D, Keller D, Bain DL (2003) Trimeric structure for an essential protein in L1 retrotransposition. *Proc Natl Acad Sci U S A* 100(24):13815–13820.

246. Martin SL, Li J, Weisz JA (2000) Deletion analysis defines distinct functional domains for protein-protein and nucleic acid interactions in the ORF1 protein of mouse LINE-1. *J Mol Biol* 304(1):11–20.

247. Khazina E, et al. (2011) Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat Struct Mol Biol* 18(9):1006–1014.

248. 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(10):8112–8117.

249. Mandal PK, Ewing AD, Hancks DC, Kazazian HH (2013) Enrichment of processed pseudogene transcripts in L1-ribonucleoprotein particles. *Hum Mol Genet* 22(18):3730–3748.

250. Cook PR, Jones CE, Furano A V. (2015) Phosphorylation of ORF1p is required for L1 retrotransposition. *Proc Natl Acad Sci* 112(14):201416869.

251. Cristofari G, Ficheux D, Darlix JL (2000) The GAG-like protein of the yeast Ty1 retrotransposon contains a nucleic acid chaperone domain analogous to retroviral nucleocapsid proteins. *J Biol Chem* 275(25):19210–19217.

252. Cristofari G, et al. (2004) The hepatitis C virus Core protein is a potent nucleic acid chaperone that directs dimerization of the viral (+) strand RNA in vitro. *Nucleic Acids Res* 32(8):2623–2631.

253. Cristofari G, Darlix J-LL (2002) The ubiquitous nature of RNA chaperone

proteins. *Prog Nucleic Acid Res Mol Biol* 72:223–268.

254. Martin SL, Bushman FD (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol* 21(2):467–475.

255. Evans JD, Peddigari S, Chaurasiya KR, Williams MC, Martin SL (2011) Paired mutations abolish and restore the balanced annealing and melting activities of ORF1p that are required for LINE-1 retrotransposition. *Nucleic Acids Res* 39(13):5611–5621.

256. Martin SL, et al. (2008) A single amino acid substitution in ORF1 dramatically decreases L1 retrotransposition and provides insight into nucleic acid chaperone activity. *Nucleic Acids Res* 36(18):5845–5854.

257. Goodier JL, Cheung LE, Kazazian HH (2013) Mapping the LINE1 ORF1 protein interactome reveals associated inhibitors of human retrotransposition. *Nucleic Acids Res* 41(15):7401–7419.

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

259. Dlakic M (2000) Functionally unrelated signalling proteins contain a fold similar to Mg2+-dependent endonucleases. *Trends Biochem Sci* 25(6):272–273.

260. Hofmann K, Tomiuk S, Wolff G, Stoffel W (2000) Cloning and characterization of the mammalian brain-specific, Mg2+-dependent neutral sphingomyelinase. *Proc Natl Acad Sci U S A* 97(11):5895–5900.

261. 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(51):18081–18093.

262. Gilbert N, Lutz S, Morrish TA, Moran J V (2005) Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol Cell Biol* 25(17):7780–7795.

263. Jurka J (1997) Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc Natl Acad Sci U S A* 94(5):1872–1877.

264. Deragon JM, Sinnett D, Labuda D (1990) Reverse transcriptase activity from human embryonal carcinoma cells NTera2D1. *EMBO J* 9(10):3363–3368.

265. Dai L, Huang Q, Boeke JD (2011) Effect of reverse transcriptase inhibitors on LINE-1 and Ty1 reverse transcriptase activities and on LINE-1

retrotransposition. *BMC Biochem* 12:18.

266. Jones RB, et al. (2008) Nucleoside analogue reverse transcriptase inhibitors differentially inhibit human LINE-1 retrotransposition. *PLoS One* 3(2):e1547.

267. Kroutter EN, Belancio VP, Wagstaff BJ, Roy-Engel AM (2009) The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition. *PLoS Genet* 5(4):e1000458.

268. 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(2):661–668.

269. Kopera HC, Moldovan JB, Morrish TA, Garcia-perez JL, Moran J V (2011) Similarities between long interspersed element-1 (LINE-1) reverse transcriptase and telomerase. *Proc Natl Acad Sci U S A* 108(51):20345–20350.

270. Pisarev A V, Shirokikh NE, Hellen CUT (2005) Translation initiation by factor-independent binding of eukaryotic ribosomes to internal ribosomal entry sites. *C R Biol* 328(7):589–605.

271. Kozak M (1989) The scanning model for translation: an update. *J Cell Biol* 108(2):229–241.

272. McMillan JP, Singer MF (1993) Translation of the human LINE-1 element, L1Hs. *Proc Natl Acad Sci U S A* 90(24):11533–11537.

273. Martin SL (2006) The ORF1 protein encoded by LINE-1: Structure and function during L1 retrotransposition. *J Biomed Biotechnol* 2006:1–6.

274. Jang SK, et al. (1988) A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J Virol* 62(8):2636–2643.

275. Komar AA, Mazumder B, Merrick WC (2012) A new framework for understanding IRES-mediated translation. *Gene* 502(2):75–86.

276. Peddigari S, Li PW-L, Rabe JL, Martin SL (2013) hnRNPL and nucleolin bind LINE-1 RNA and function as host factors to modulate retrotransposition. *Nucleic Acids Res* 41(1):575–585.

277. Dai L, Taylor MS, O'Donnell KA, Boeke JD (2012) Poly(A) binding protein C1 is essential for efficient L1 retrotransposition and affects L1 RNP formation. *Mol Cell Biol* 32(21):4323–4336.

278. Gorlich D, Kutay U (1999) Transport between the cell nucleus and the

cytoplasm. *Annu Rev Cell Dev Biol* 15:607–660.

279. Kubo S, et al. (2006) L1 retrotransposition in nondividing and primary human somatic cells. *Proc Natl Acad Sci U S A* 103(21):8036–8041.

280. Xiong YE, Eickbush TH (1988) Functional expression of a sequence-specific endonuclease encoded by the retrotransposon R2Bm. *Cell* 55(2):235–246.

281. Christensen SM, Bibillo A, Eickbush TH (2005) Role of the Bombyx mori R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res* 33(20):6461–6468.

282. Bibillo A, Eickbush TH (2004) End-to-end template jumping by the reverse transcriptase encoded by the R2 retrotransposon. *J Biol Chem* 279(15):14945–14953.

283. Bibillo A, Lener D, Klarmann GJ, Le Grice SFJ (2005) Functional roles of carboxylate residues comprising the DNA polymerase active site triad of Ty3 reverse transcriptase. *Nucleic Acids Res* 33(1):171–181.

284. Anzai T, Takahashi H, Fujiwara H (2001) Sequence-specific recognition and cleavage of telomeric repeat (TTAGG)(n) by endonuclease of non-long terminal repeat retrotransposon TRAS1. *Mol Cell Biol* 21(1):100–108.

285. Dewannieux M, Esnault C, Heidmann T (2003) LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet* 35(1):41–48.

286. Cousineau B, Lawrence S, Smith D, Belfort M (2000) Retrotransposition of a bacterial group II intron. *Nature* 404(6781):1018–1021.

287. Garcia-Perez JL, Doucet AJ, Bucheton A, Moran J V, Gilbert N (2007) Distinct mechanisms for trans-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. *Genome Res* 17(5):602–611.

288. Zingler N, et al. (2005) Analysis of 5’ junctions of human LINE-1 and Alu retrotransposons suggests an alternative model for 5'-end attachment requiring microhomology-mediated end-joining. *Genome Res* 15(6):780–789.

289. Morrish TA, et al. (2002) DNA repair mediated by endonuclease-independent LINE-1 retrotransposition. *Nat Genet* 31(2):159–165.

290. 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(11):3741–3751.

291. Callinan PA, et al. (2005) Alu retrotransposition-mediated deletion. *J Mol Biol* 348(4):791–800.

292. Srikanta D, et al. (2009) An alternative pathway for Alu retrotransposition suggests a role in DNA double-strand break repair. *Genomics* 93(3):205–212.

293. Morrish TA, et al. (2007) Endonuclease-independent LINE-1 retrotransposition at mammalian telomeres. *Nature* 446(7132):208–212.

294. Lingner J, et al. (1997) Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* 276(5312):561–567.

295. Nakamura TM, Cech TR (1998) Reversing time: origin of telomerase. *Cell* 92(5):587–590.

296. Ostertag EM, Goodier JL, Zhang Y, Kazazian HH (2003) SVA elements are nonautonomous retrotransposons that cause disease in humans. *Am J Hum Genet* 73(6):1444–1451.

297. Wagstaff BJ, et al. (2012) Rescuing Alu: Recovery of New Inserts Shows LINE-1 Preserves Alu Activity through A-Tail Expansion. *PLoS Genet* 8(8):e1002842.

298. 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(17):3386–3400.

299. Weber MJ (2006) Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet* 2(12):e205.

300. Denison RA, Van Arsdell SW, Bernstein LB, Weiner AM (1981) Abundant pseudogenes for small nuclear RNAs are dispersed in the human genome. *Proc Natl Acad Sci U S A* 78(2):810–814.

301. Van Arsdell SW, et al. (1981) Direct repeats flank three small nuclear RNA pseudogenes in the human genome. *Cell* 26(1 Pt 1):11–17.

302. Bernstein LB, Mount SM, Weiner AM (1983) Pseudogenes for human small nuclear RNA U3 appear to arise by integration of self-primed reverse transcripts of the RNA into new chromosomal sites. *Cell* 32(2):461–472.

303. Vanin EF (1985) Processed pseudogenes: characteristics and evolution. *Annu Rev Genet* 19:253–272.

304. Kabza M, Ciomborowska J, Makalowska I (2014) RetrogeneDB--a database of animal retrogenes. *Mol Biol Evol* 31(7):1646–1648.

305. Cooke SL, et al. (2014) Processed pseudogenes acquired somatically during cancer development. *Nat Commun* 5:3644.

306. Abyzov A, et al. (2013) Analysis of variable retroduplications in human populations suggests coupling of retrotransposition to cell division. *Genome Res* 23(12):2042–2052.

307. Schrider DR, et al. (2013) Gene copy-number polymorphism caused by retrotransposition in humans. *PLoS Genet* 9(1):e1003242.

308. de Boer M, et al. (2014) Primary immunodeficiency caused by an exonized retroposed gene copy inserted in the CYBB gene. *Hum Mutat* 35(4):486–496.

309. Doucet AJ, Droc G, Siol O, Audoux J, Gilbert N (2015) U6 snRNA Pseudogenes: Markers of Retrotransposition Dynamics in Mammals. *Mol Biol Evol* 32(7):1815–1832.

310. Baillie JK, et al. (2011) Somatic retrotransposition alters the genetic landscape of the human brain. *Nature* 479(7374):534–537.

311. Shukla R, et al. (2013) Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* 153(1):101–111.

312. Beck CR, et al. (2010) LINE-1 Retrotransposition Activity in Human Genomes. *Cell* 141(7):1159–1170.

313. Iskow RC, et al. (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* 141(7):1253–1261.

314. Ewing AD, Kazazian HH (2010) High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res* 20(9):1262–1270.

315. Badge RM, Alisch RS, Moran J V (2003) ATLAS: a system to selectively identify human-specific L1 insertions. *Am J Hum Genet* 72(4):823–838.

316. Helman E, et al. (2014) Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res* 24(7):1053–1063.

317. Ewing AD, Kazazian HH (2011) Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res* 21(6):985–990.

318. Tubio JMC, et al. (2014) Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* 345(6196):1251343.

319. Thung DT, et al. (2014) Mobster : accurate detection of mobile element insertions in next generation sequencing data. *Genome Biol* 15(10):1–11.

320. Lee W-P, Wu J, Marth GT (2015) Toolbox for mobile-element insertion detection on cancer genomes. *Cancer Inform* 14(Suppl 1):37–44.

321. Keane TM, Wong K, Adams DJ (2013) Genome analysis RetroSeq : transposable element discovery from next-generation sequencing data. *Bioinformatics* 1(3):389–390.

322. An W, et al. (2006) Active retrotransposition by a synthetic L1 element in mice. *Proc Natl Acad Sci U S A* 103(49):18662–18667.

323. Babushok D V, Ostertag EM, Courtney CE, Choi JM, Kazazian HH (2006) L1 integration in a transgenic mouse model. *Genome Res* 16(2):240–250.

324. Kano H, et al. (2009) L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev* 23(11):1303–1312.

325. Prak ETL, Dodson AW, Farkash EA, Kazazian HHJ (2003) Tracking an embryonic L1 retrotransposition event. *Proc Natl Acad Sci U S A* 100(4):1832–1837.

326. Packer AI, Manova K, Bachvarova RF (1993) A discrete LINE-1 transcript in mouse blastocysts. *Dev Biol* 157(1):281–283.

327. Branciforte D, Martin SL (1994) Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol Cell Biol* 14(4):2584–2592.

328. Trelogan SA, Martin SL (1995) Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proc Natl Acad Sci U S A* 92(5):1520–1524.

329. Wissing S, et al. (2012) Reprogramming somatic cells into iPS cells activates LINE-1 retroelement mobility. *Hum Mol Genet* 21(1):208–218.

330. Macia A, et al. (2011) Epigenetic Control of Retrotransposon Expression in Human Embryonic Stem Cells . *Mol Cell Biol* 31(2):300–316.

331. Brouha B, et al. (2002) Evidence consistent with human L1 retrotransposition in maternal meiosis I. *Am J Hum Genet* 71(2):327–336.

332. Bratthauer GL, Fanning TG (1992) Active LINE-1 retrotransposons in human testicular cancer. *Oncogene* 7(3):507–510.

333. Ostertag EM, et al. (2002) A mouse model of human L1 retrotransposition. *Nat Genet* 32(4):655–660.

334. O'Donnell KA, An W, Schrum CT, Wheelan SJ, Boeke JD (2013) Controlled insertional mutagenesis using a LINE-1 (ORFeus) gene-trap mouse model. *Proc Natl Acad Sci U S A* 110(29):E2706–13.

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

336. 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(6168):87–90.

337. Muotri AR, et al. (2005) Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* 435(7044):903–910.

338. Coufal NG, et al. (2009) L1 retrotransposition in human neural progenitor cells. *Nature* 460(7259):1127–1131.

339. Evrony GD, et al. (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell* 151(3):483–496.

340. Upton KRR, et al. (2015) Ubiquitous l1 mosaicism in hippocampal neurons. *Cell* 161(2):228–239.

341. Wang H, et al. (2005) SVA Elements : A Hominid-specific Retroposon Family. *J Mol Biol* 354(4):994–1007.

342. Han K, et al. (2005) Genomic rearrangements by LINE-1 insertion-mediated deletion in the human and chimpanzee lineages. *Nucleic Acids Res* 33(13):4040–4052.

343. Mills RE, et al. (2006) Recently mobilized transposons in the human and chimpanzee genomes. *Am J Hum Genet* 78(4):671–679.

344. 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(4):823–834.

345. Zhou Y, Mishra B (2005) Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A* 102(11):4051–4056.

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

347. Kim PM, et al. (2008) Analysis of copy number variants and segmental

duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history. *Genome Res* 18(12):1865–1874.

348. Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet* 18(2):74–82.

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

350. Sen SK, et al. (2006) Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* 79(1):41–53.

351. Han K, et al. (2007) Alu recombination-mediated structural deletions in the chimpanzee genome. *PLoS Genet* 3(10):1939–1949.

352. Han K, et al. (2008) L1 recombination-associated deletions generate human genomic variation. *Proc Natl Acad Sci U S A* 105(49):19366–19371.

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

354. Li J, et al. (2009) Phylogeny of the macaques (Cercopithecidae : Macaca) based on Alu elements. *Gene* 448(2):242–249.

355. Kelkar YD, Eckert KA, Chiaromonte F, Makova KD (2011) A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res* 21(12):2038–2048.

356. Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats: a source for the genesis of primate microsatellites. *Genomics* 29(1):136–144.

357. Ovchinnikov I, Troxel AB, Swergold GD (2001) Genomic characterization of recent human LINE-1 insertions: evidence supporting random insertion. *Genome Res* 11(12):2050–2058.

358. Justice CM, et al. (2001) Phylogenetic analysis of the Friedreich ataxia GAA trinucleotide repeat. *J Mol Evol* 52(3):232–238.

359. Kurosaki T, Ninokata A, Wang L, Ueda S (2006) Evolutionary scenario for acquisition of CAG repeats in human SCA1 gene. *Gene* 373:23–27.

360. Gatchel JR, Zoghbi HY (2005) Diseases of unstable repeat expansion: mechanisms and common principles. *Nat Rev Genet* 6(10):743–755.

361. Pearson CE, Nichol Edamura K, Cleary JD (2005) Repeat instability: mechanisms of dynamic mutations. *Nat Rev Genet* 6(10):729–742.

362. Eckert KA, Hile SE (2009) Every microsatellite is different: Intrinsic DNA features dictate mutagenesis of common microsatellites present in the human genome. *Mol Carcinog* 48(4):379–388.

363. Baptiste BA, et al. (2013) Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. *G3 (Bethesda)* 3(3):451–463.

364. Lin Y, Lukacsovich T, Waldman AS (1999) Multiple pathways for repair of DNA double-strand breaks in mammalian chromosomes. *Mol Cell Biol* 19(12):8353–8360.

365. Gasior SL, Wakeman TP, Xu B, Deininger PL (2006) The human LINE-1 retrotransposon creates DNA double-strand breaks. *J Mol Biol* 357(5):1383–1393.

366. Kines KJ, Sokolowski M, deHaro DL, Christian CM, Belancio VP (2014) Potential for genomic instability associated with retrotranspositionally-incompetent L1 loci. *Nucleic Acids Res* 42(16):10488–10502.

367. Boeke JD, Pickeral OK (1999) Retroshuffling the genomic deck. *Nature* 398(6723):108–109,111.

368. Eickbush T (1999) Exon shuffling in retrospect. *Science* 283(5407):1465;1467.

369. Awano H, et al. (2010) Contemporary retrotransposition of a novel non-coding gene induces exon-skipping in dystrophin mRNA. *J Hum Genet* 55(12):785–790.

370. Solyom S, et al. (2012) Pathogenic orphan transduction created by a nonreference LINE-1 retrotransposon. *Hum Mutat* 33(2):369–371.

371. Solyom S, et al. (2012) Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res* 22(12):2328–2338.

372. Yoshida K, Nakamura A, Yazaki M, Ikeda S, Takeda S (1998) Insertional mutation by transposable element, L1, in the DMD gene results in X-linked dilated cardiomyopathy. *Hum Mol Genet* 7(7):1129–1132.

373. Musova Z, et al. (2006) A novel insertion of a rearranged L1 element in exon 44 of the dystrophin gene: further evidence for possible bias in retroposon integration. *Biochem Biophys Res Commun* 347(1):145–149.

374. Macfarlane CM, et al. (2013) Transduction-specific ATLAS reveals a cohort of highly active L1 retrotransposons in human populations. *Hum Mutat* 34(7):974–985.

375. Kaer K, Speek M (2013) Retroelements in human disease. *Gene* 518(2):231–241.

376. Kazazian HH, et al. (1988) Haemophilia A resulting from de novo insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* 332(6160):164–166.

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

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

379. Wimmer K, Callens T, Wernstedt A, Messiaen L (2011) The NF1 gene contains hotspots for L1 endonuclease-dependent de novo insertion. *PLoS Genet* 7(11):e1002371.

380. Conley ME, Partain JD, Norland SM, Shurtleff SA, Kazazian HHJ (2005) Two independent retrotransposon insertions at the same site within the coding region of BTK. *Hum Mutat* 25(3):324–325.

381. Bratthauer GL, Cardiff RD, Fanning TG (1994) Expression of LINE-1 retrotransposons in human breast cancer. *Cancer* 73(9):2333–2336.

382. Su YA, Clewell DB (1993) Characterization of the left 4 kb of conjugative transposon Tn916: determinants involved in excision. *Plasmid* 30(3):234–250.

383. Harris CR, et al. (2010) Association of nuclear localization of a long interspersed nuclear element-1 protein in breast tumors with poor prognostic outcomes. *Genes Cancer* 1(2):115–124.

384. Rodić N, Burns KH (2013) Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? *PLoS Genet* 9(3):e1003402.

385. Bernard O, Cory S, Gerondakis S, Webb E, Adams JM (1983) Sequence of the murine and human cellular myc oncogenes and two modes of myc transcription resulting from chromosome translocation in B lymphoid tumours. *EMBO J* 2(12):2375–2383.

386. Hoffman B, Liebermann DA (1998) The proto-oncogene c-myc and apoptosis. *Oncogene* 17(25):3351–3357.

387. Miki Y, et al. (1992) Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res* 52(3):643–645.

388. Economou-Pachnis A, Tsichlis PN (1985) Insertion of an Alu SINE in the human homologue of the Mlvi-2 locus. *Nucleic Acids Res* 13(23):8379–8387.

389. Kloor M, et al. (2004) A large MSH2 Alu insertion mutation causes HNPCC in a German kindred. *Hum Genet* 115(5):432–438.

390. Miki Y, Katagiri T, Kasumi F, Yoshimoto T, Nakamura Y (1996) Mutation analysis in the BRCA2 gene in primary breast cancers. *Nat Genet* 13(2):245–247.

391. Ehrlich M (2002) DNA methylation in cancer: too much, but also too little. *Oncogene* 21(35):5400–5413.

392. Kaneda A, et al. (2004) Frequent hypomethylation in multiple promoter CpG islands is associated with global hypomethylation, but not with frequent promoter hypermethylation. *Cancer Sci* 95(1):58–64.

393. Wilson AS, Power BE, Molloy PL (2007) DNA hypomethylation and human diseases. *Biochim Biophys Acta* 1775(1):138–162.

394. Florl AR, Lower R, Schmitz-Drager BJ, Schulz WA (1999) DNA methylation and expression of LINE-1 and HERV-K provirus sequences in urothelial and renal cell carcinomas. *Br J Cancer* 80(9):1312–1321.

395. Suter CM, Martin DI, Ward RL (2004) Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue. *Int J Colorectal Dis* 19(2):95–101.

396. Schulz WA, et al. (2002) Genomewide DNA hypomethylation is associated with alterations on chromosome 8 in prostate carcinoma. *Genes Chromosomes Cancer* 35(1):58–65.

397. Sunami E, Vu A-T, Nguyen SL, Giuliano AE, Hoon DSB (2008) Quantification of LINE1 in circulating DNA as a molecular biomarker of breast cancer. *Ann N Y Acad Sci* 1137:171–174.

398. Nekrutenko A, Li WH (2001) Transposable elements are found in a large number of human protein-coding genes. *Trends Genet* 17(11):619–621.

399. Kim D-S, et al. (2006) LINE FUSION GENES: a database of LINE expression in human genes. *BMC Genomics* 7:139.

400. Zemojtel T, et al. (2007) Exonization of active mouse L1s: a driver of transcriptome evolution? *BMC Genomics* 8:392.

401. Zarnack K, et al. (2013) Direct competition between hnRNP C and U2AF65

protects the transcriptome from the exonization of Alu elements. *Cell* 152(3):453–466.

402. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40(12):1413–1415.

403. Mameli E, et al. (2015) Wilson's disease caused by alternative splicing and Alu exonization due to a homozygous 3039-bp deletion spanning from intron 1 to exon 2 of the ATP7B gene. *Gene* 569(2):276–279.

404. Nozu K, et al. (2014) Alport syndrome caused by a COL4A5 deletion and exonization of an adjacent AluY. *Mol Genet genomic Med* 2(5):451–453.

405. Tang W, Duke-Cohan JS (2002) Human secreted attractin disrupts neurite formation in differentiating cortical neural cells in vitro. *J Neuropathol Exp Neurol* 61(9):767–777.

406. Yoder JA, Walsh CP, Bestor TH (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet* 13(8):335–340.

407. Wheelan SJ, Aizawa Y, Han JS, Boeke JD (2005) Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res* 15(8):1073–1078.

408. Mätlik K, Redik K, Speek M (2006) L1 antisense promoter drives tissue-specific transcription of human genes. *J Biomed Biotechnol* 2006:1–16.

409. Li J, et al. (2014) An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. *Nucleic Acids Res* 42(7):4546–4562.

410. Kaer K, Branovets J, Hallikma A, Nigumann P, Speek M (2011) Intronic L1 Retrotransposons and Nested Genes Cause Transcriptional Interference by Inducing Intron Retention , Exonization and Cryptic Polyadenylation. *PLoS One* 6(10). doi:10.1371/journal.pone.0026099.

411. Lappalainen I, et al. (2013) DbVar and DGVa: public archives for genomic structural variation. *Nucleic Acids Res* 41(Database issue):D936–41.

412. Mattick JS, Makunin I V (2005) Small regulatory RNAs in mammals. *Hum Mol Genet* 14 Spec No 1:R121–32.

413. Zimin A V, et al. (2013) Genome analysis The MaSuRCA genome assembler. *Bioinformatics* 29(21):2669–2677.

414. Pertea M, et al. (2015) StringTie enables improved reconstruction of a

transcriptome from RNA-seq reads. *Nat Biotech* 33(3):290–295.

415. Kim D, Langmead B, Salzberg SL (2015) HISAT : a fast spliced aligner with low memory requirements. *Nat Methods* 12(4). doi:10.1038/nmeth.3317.

416. Kim D, et al. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* 14(4):R36.

417. Trapnell C, et al. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562–78.

418. Alamancos GP, Pagès A, Trincado JL, Eyras E (2014) SUPPA : a super-fast pipeline for alternative splicing analysis from RNA-Seq. *bioRxiv*.

419. Rogers MF, Thomas J, Reddy AS, Ben-Hur A (2012) SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol* 13(1):R4.

420. Quinlan AR, Hall IM (2010) The BEDTools manual. *Genome* 16(6):1–77.

421. Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16(9):418–420.

422. Mir a. a., Philippe C, Cristofari G (2014) euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res* 43(D1):D43–D47.

423. Harrow J, et al. (2012) GENCODE : The reference human genome annotation for The ENCODE Project. *Genome Res* 22(9):1760–1774.

424. Bolger AM, Lohse M, Usadel B (2014) Genome analysis Trimmomatic : a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120.

425. Rebollo R, Farivar S, Mager DL (2012) C-GATE - catalogue of genes affected by transposable elements. *Mob DNA* 3(1):9.

426. Karolchik D, et al. (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 42(Database issue):D764–70.

427. Holmes SE, Dombroski BA, Krebs CM, Boehm CD, Kazazian HH (1994) A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat Genet* 7(2):143–148.

428. Claverie-Martin F, Gonzalez-Acosta H, Flores C, Anton-Gamero M, Garcia-Nieto V (2003) De novo insertion of an Alu sequence in the coding region of the CLCN5 gene results in Dent's disease. *Hum Genet* 113(6):480–485.

429. Claverie-Martin F, Flores C, Anton-Gamero M, Gonzalez-Acosta H, Garcia-Nieto V (2005) The Alu insertion in the CLCN5 gene of a patient with Dent's disease leads to exon 11 skipping. *J Hum Genet* 50(7):370–374.

430. Abbott KL, et al. (2015) The Candidate Cancer Gene Database : a database of cancer driver genes from forward genetic screens in mice. *Nucleic Acids Res* 43(September 2014):844–848.

431. Ozsolak F, Milos PM (2011) Single-molecule direct RNA sequencing without cDNA synthesis. *Wiley Interdiscip Rev RNA* 2(4):565–570.

432. Sebestyen E, Zawisza MM, Eyras E, Sebesty E (2015) Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res* 43(3):1345–1356.

433. Hur K, et al. (2014) Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut* 63(4):635–646.

434. Carlini F, et al. (2010) The reverse transcription inhibitor abacavir shows anticancer activity in prostate cancer cell lines. *PLoS One* 5(12):e14221.

435. Mortazavi A, Williams BA, Mccue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5(7):1–8.

436. Steijger T, et al. (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* 10(12). doi:10.1038/nmeth.2714.

437. Kan Z, States D, Gish W (2002) Selecting for Functional Alternative Splices in ESTs Identification of Alternative Splice Patterns. *Genome Res* 12(12):1837–1845.

438. Alexandre P, Galante F, Sakabe NJO, Kirschbaum-slager N, Souza SJDE (2004) Detection and evaluation of intron retention events in the human transcriptome. *Bioinformatics* 10(5):757–765.

439. Bashirullah A, Cooperstock RL, Lipshitz HD (2001) Spatial and temporal control of RNA stability. *Proc Natl Acad Sci U S A* 98(13):7025–7028.

440. Tan JS, Mohandas N, Conboy JG (2015) High frequency of alternative first exons in erythroid genes suggests a critical role in regulating gene function. *Blood* 107(6):2557–2562.

441. Boyd CD, Pierce RA, Schwarzbauer JE, Doege K, Sandell LJ (1993) Alternate exon usage is a commonly used mechanism for increasing coding diversity within genes coding for extracellular matrix proteins. *Matrix* 13(6):457–469.

442. Huda A, Bushel PR (2013) Widespread Exonization of Transposable Elements in Human Coding Sequences is Associated with Epigenetic Regulation of Transcription. *Transcr Open Access* 1(1):1–21.

443. Wen P, et al. (2014) USP33 , a new player in lung cancer , mediates Slit-Robo signaling. *Protein Cell* 5(9):704–713.

444. Brosseau J, et al. (2010) High-throughput quantification of splicing isoforms. *RNA* 16(2):442–449.

445. Kazazian HH, Goodier JL (2002) LINE Drive : Retrotransposition and Genome Instability. *Cell* 110(3):277–280.

446. Weber B, Kimhi S, Howard G, Eden A, Lyko F (2010) Demethylation of a LINE-1 antisense promoter in the cMet locus impairs Met signalling through induction of illegitimate transcription. *Oncogene* 29(43):5775–5784.

447. Lizardi PM (2010) As we bring demethylating drugs to the clinic, we better know the DICE being cast. *Oncogene* 29(43):5772–5774.