Transposable elements in cancer

Kathleen H. Burns

Abstract | Transposable elements give rise to interspersed repeats, sequences that comprise most of our genomes. These mobile DNAs have been historically underappreciated — both because they have been presumed to be unimportant, and because their high copy number and variability pose unique technical challenges. Neither impediment now seems steadfast. Interest in the human mobilome has never been greater, and methods enabling its study are maturing at a fast pace. This Review describes the activity of transposable elements in human cancers, particularly long interspersed element-1 (LINE-1). LINE-1 sequences are self-propagating, protein-coding retrotransposons, and their activity results in somatically acquired insertions in cancer genomes. Altered expression of transposable elements and animation of genomic LINE-1 sequences appear to be hallmarks of cancer, and can be responsible for driving mutations in tumorigenesis.

Retrotransposition
A mechanism used by transposable elements to copy RNA intermediates to genomic DNA.

Centromeric satellites
Arrays of tandem, simple repeats at the regions of chromosomes that attach to the mitotic spindle.

Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.
kburns@jhmi.edu

doi:10.1038/nrc.2017.35
Published online 9 Jun 2017

The human genome, like that of other eukaryotes, comprises largely interspersed repeats. Essentially half our DNA is recognizable as repeated sequence derived from mobile DNA[1,2]. Relatively recent insertions result from active mobile DNAs specific to *Homo sapiens*, whereas older insertions were acquired in ancestral species and are shared between humans and other organisms[3,4]. Although they are frequently dismissed as non-functional, irrelevant 'junk' DNA, one can argue that few other processes have so profoundly affected our genetic make-up.

That there are currently active mobile DNAs in humans was fully appreciated in 1988, when researchers identified long interspersed element-1 (LINE-1; also known as L1) insertions responsible for haemophilia A[5]. Two unrelated affected individuals had independently acquired LINE-1 repeat sequences interrupting exon 14 of the factor VIII (*F8*) gene that rendered the mutated alleles incapable of coding for the coagulation factor. Today, 124 instances of genetic disease have been attributed to germline LINE-1-mediated retrotransposition[6]. Elegant experiments have demonstrated LINE-1 activity *in vitro*[7–9] and *in vivo*[10,11], and advances in genome sequencing have demonstrated that LINE-1 proteins and the sequences they mobilize are major sources of structural variation in human populations[12–19].

LINE-1 is a protein-coding transposable element that copies itself through an RNA intermediate. It is thus referred to as an autonomous (protein-coding) retrotransposon (or RNA transposon)[20]. In the germline, LINE-1 reverse transcriptase is also responsible for the retrotransposition of other RNAs. These include non-autonomous mobile elements, such as the *Alu* short interspersed element (SINE)[21] and the SINE–variable number tandem repeat (VNTR)–*Alu* (SVA) composite retroelement[22,23] (FIG. 1). LINE-1 proteins are also co-opted to deposit the occasional pseudogene[24,25] or U6 ribosomal RNA (rRNA) sequence[26–29].

Even with the recognition that LINE-1, *Alu* and SVA insertions can cause genetic disease[6], it has been possible to dismiss retrotransposable element activities as imperceptible in the life of most individuals — too infrequently active to be relevant. Cancers present an important exception. We now understand that mechanisms keeping mobile DNAs in check are sufficiently compromised in many malignancies to enable LINE-1 proteins to be expressed and mobile DNAs to become dynamic constituents of cancer genomes.

This Review will introduce mobile DNAs resident in our genomes and distinguish those types that are currently active. It will then focus on replication of the LINE-1 retrotransposon and how expression of LINE-1 and other interspersed repeats is altered in cancers. Finally, this Review discusses the dynamics of somatically acquired insertions and their roles in tumorigenesis.

The human mobilome

Much of our DNA is repetitive sequence. This includes the extensive tandem arrays of simple repeats that make up centromeric satellites and telomere ends; simple repeats that are distributed in the genome, including VNTRs; and more complex high copy number sequences. The latter repeats are dense in pericentromeric[30] and subtelomeric[31] heterochromatin but are also interspersed genome wide. They are derived from mobile DNAs or transposable elements, and are colloquially known as the mobilome.

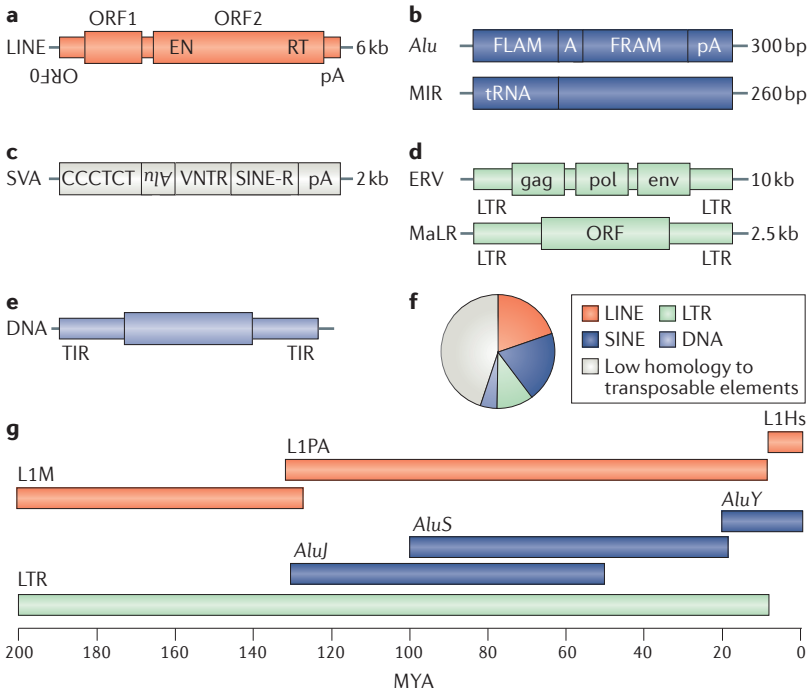


Figure 1 | **Types of mobile DNA in the human genome.** The schematic illustrates the organization of consensus sequences for the most common types of transposable element in the human genome. **a** | Long interspersed element-1 (LINE-1; also known as L1) is shown with the open reading frames (ORFs) for ORF1 protein (ORF1p) and ORF2p marked; locations of the endonuclease domain (EN) and reverse transcriptase domain (RT) of ORF2p are indicated. The element ends with a 3′ polyadenylation (poly(A)) tail (pA). **b** | LINE-1 ORF2p is responsible for mobilizing the non-autonomous *Alu* short interspersed elements (SINEs) derived from signal recognition particle 7SL RNA[169]. Mammalian-wide interspersed repeat (MIR) SINE sequences remain in high copy number in the human genome but are no longer capable of retrotransposition. **c** | Composite SINE–variable number tandem repeat (VNTR)–*Alu* (SVA) elements are also non-autonomous and mobilized by LINE-1 ORF2p. **d** | The multiple ORFs of endogenous retroviruses (ERVs; *gag*, *pol* and *env*) and the single ORF of mammalian apparent long terminal repeat (LTR) retrotransposon (MaLR) sequences are shown, although many of these insertion sites have collapsed to solo LTR sequences. **e** | DNA transposons are recognizable by terminal inverted repeats (TIRs). These sequences are no longer active and transposase ORFs are frequently not found. **f** | The pie chart shows the relative abundance of each class of mobile DNA as a proportion of total genomic DNA. Most human elements belong to class I retrotransposons (LINE, SINE and LTR) and DNA transposons are class II elements. **g** | The timeline shows approximate spans of recent transposable element activity during human evolution. In parts **a**–**d**, approximate sizes are given for an intact consensus sequence to the right of each type of mobile element. A, A-rich domain; FLAM, free left *Alu* monomer; FRAM, free right *Alu* monomer; MYA, millions of years ago; tRNA, transfer RNA.

Human reference genome assembly
A version of the human genome, for example, the human Dec. 2013 (GRCh38/hg38) assembly. Structural variants caused by mobile element insertions are not consistently incorporated.

Source element
Full-length genomic long interspersed element-1 (LINE-1) sequence capable of retrotransposition.

The mobilome comprises about half of the human reference genome assembly[2] (FIG. 1), and there is evidence that nearly two-thirds of our DNA is complex repetitive sequence that is probably the result of deterioration of recognizable mobile DNAs[32]. Transposable elements are categorized broadly as class I elements (so-called copy-and-paste retrotransposons), which use reverse transcribed RNA intermediates to produce copies of themselves, and class II elements (so-called cut-and-paste DNA transposons) that excise from a donor site to reintegrate elsewhere in the genome[33]. Most human elements are of the former class (FIG. 1f). Reverse transcription of these sequences may be interrupted such that only a fragment of the full-length sequence is incorporated, but it is not highly error prone[27]. New insertions enter the genome nearly identical in sequence to their respective source element[27]. This sequence homology is eroded over time through neutral substitutions, and antecedent elements are also bifurcated by incoming integrations[34], such that evidence of the origin of a repeat is effaced over time. Higher organisms without active mobile elements are rare[4], and our species has never known an interruption to their activity.

Autonomous mobile DNAs encode proteins needed for their movement in the genome, including those with reverse transcriptase (class I) or transposase (class II) enzymatic activity. In modern humans, one autonomous element dominates transposition, the species-specific LINE-1 subfamily, L1PA1, also known as L1Hs. LINE-1 sequences are non-long terminal repeat (non-LTR) retrotransposons. In total, LINE-1 sequences make up approximately 17% of the human genome[1]. Multiple successive primate LINE-1 subfamilies (L1PA1–17, in reverse chronological order) have been active for more than 100 million years in the evolutionary lineage that leads to humans[35,36] (FIG. 1g). That is, at multiple points in our history as a species, there has been a discrete switch from one active LINE-1 subfamily to another that diverges from the previously active element by a few nucleotides. There are hundreds of thousands of fragments of LINE-1 sequences in the genome, and most are present as 3′ fragments or have mutated open reading frames (ORFs) such that they are incapable of retrotransposition. However, approximately 100 copies of L1PA1 and L1PA2 are intact, potential source elements for new insertions. These are frequently polymorphic structural variants, with both pre-insertion (or 'empty') and insertion alleles found in human populations[14,15,37–39].

SINE sequences are non-autonomous (non-protein coding) elements that collectively make up a proportion of our total DNA comparable to that of the LINE-1 sequences[1] (FIG. 1f). These include more than one million *Alu* repeats[1]. *Alu*Y and its subfamilies, originating approximately 20 million years ago, continue to propagate in the modern human germline using the reverse transcriptase activity of LINE-1 (REFS 12,40). Like LINE-1, *Alu* elements are a significant source of structural variation in humans[12,13,16–18,41,42]. Mammalian-wide interspersed repeats (MIRs) are also very frequent SINEs in our genome, although they are no longer mobile and all copies are fixed[43,44]. In addition to *Alu*, LINE-1 protein can mobilize SVA sequences[22,23,45] — these are composite retroelements specific to humans and great apes[46]. They number approximately 2,700 copies in the human genome[1].

Besides LINE-1 sequences, the other major autonomous retroelements in humans are LTR sequences. These are structurally similar to exogenous, infective retroviruses. They include endogenous retroviruses (ERVs), mammalian LTR transposons (MLTs) and shorter mammalian apparent LTR retrotransposon (MaLR)[47] repeats. Human ERVs, specifically HERV-K, have been active most recently (FIG. 1g); although intact

elements capable of retrotransposition are not known[48], evidence of their recent activity exists in the form of small numbers of polymorphic loci in human populations. At some of these loci, three variants — 'empty' or preinsertion alleles, solo LTRs and full-length proviral sequences — segregate in extant populations[48–51].

Finally, DNA transposons, which move by excision and reintegration in the genome, the so-called cut-and-paste elements, now constitute less than 5% of human DNA[1] (FIG. 1f). Recognizable by short (14–25 bp) terminal inverted repeats (TIRs), these include Tigger and Charlie DNA sequences and some other medium reiterated sequences (MERs)[52]. These are no longer active as such in humans, although there are examples of human genes that are co-opted DNA transposons. Recombinase activating gene (RAG) proteins that mediate V(D)J recombination for antigen detection by lymphocytes have been borrowed from DNA transposons as a means of generating antibody diversity[53–58]. The genes encoding histone–lysine *N*-methyltransferase SET domain and mariner transposase fusion protein (*SETMAR*), *P*-element-derived thanatos-associated protein 9 (*THAP9*) and piggyBac transposable element-derived protein 5 (*PGBD5*) have also been distantly incorporated during human evolution and retain enzymatic functions of the parent transposable element[59–61].

Mobile element expression in cancer

Promoter activity. The first kilobase of a full-length LINE-1 sequence encodes a bidirectional, CpG-rich internal RNA polymerase II (RNApolII) promoter. DNA methylation is considered an important mechanism for the silencing of genomic transposable elements, and the LINE-1 promoter sequence is typically methylated[62]. Germline-expressed PIWI-interacting RNAs (piRNAs) are crucial in establishing this methylation[63], and a wide array of Kruppel-associated box (KRAB) domain-containing zinc-finger proteins (KZFPs) bind to transposable elements and promote DNA methylation and histone 3 lysine 9 trimethylation (H3K9me3)-dependent heterochromatin formation[64–68]. LINE-1 methylation is compromised in the genome-wide hypomethylation that characterizes many forms of human malignancy[69], and genome-wide LINE-1 promoter hypomethylation correlates with retrotransposition[70]. Few studies have focused on individual, active LINE-1 loci in human tumours, but these reinforce a model whereby LINE-1 promoter hypomethylation enables expression and retrotransposition[71,72]. Other studies have related collective or presumably representative LINE-1 promoter hypomethylation in tumours to clinical measures. For example, LINE-1 promoter hypomethylation has been associated with genomic instability[73] and poor prognosis[74] in non-small-cell lung cancer (NSCLC), poor outcomes in colon cancer[75] and oesophageal squamous cell carcinoma[76], decreased overall survival and drug resistance in young patients with breast cancer[77], recurrence of hepatocellular carcinoma after resection[78], and aggressive histology, poorer progression-free intervals and poorer survival in ovarian cancer[79].

Retrotransposon RNA. LINE-1 RNA is notoriously difficult to study. Because LINE-1 sequences are pervasive in the genome and are common components of introns, they are transcribed as 'read-through' coupled to the expression of many genes. There is also evidence that LINE-1 RNA is a long-lived component of structural heterogeneous nuclear RNAs (hnRNAs)[80]. These relatively abundant species of LINE-1 RNA can obscure the presence of the so-called unit transcript (that is, the 6-kb, full-length retrotransposon intermediate)[81–84]. Only very recently have rigorous RNA sequencing (RNA-seq)-based studies been used to identify LINE-1 unit transcripts; these require a tailored analysis, integrating these reads with other data. For a specific sample to be characterized, genomic sequences of full-length, potentially transcribed LINE-1 loci are needed, as internal sequence variants or unique sequences flanking insertion sites are used in these RNA-seq-based approaches. Chromatin immunoprecipitation (ChIP) reads indicating RNApolII occupancy and histone modifications (for example, H3K4 trimethylation or H3K27 acetylation) immediately 5′ of a LINE-1 (REF. 85), unambiguous mapping of RNA reads from 5′ rapid amplification of cDNA ends (5′ RACE) sequences[86] or read-through transcripts extending beyond the 3′ polyadenylation (poly(A)) tails of a LINE-1 insertion can be used to infer transcription of specific, full-length LINE-1 loci[87,88]. These advances provide truly new capabilities to characterize LINE-1 RNA expression. Their wider application is likely to produce catalogues of somatically expressed LINE-1 in various human cancers.

The LINE-1 promoter is bidirectional: promoter activity in the sense direction transcribes the two ORFs of LINE-1 and produces the RNA intermediate for retrotransposition. The 5′ end of LINE-1 also has antisense promoter (ASP) activity capable of initiating transcripts that oppose the LINE-1 orientation. Use of splice donors in the antisense LINE-1 sequence together with downstream splice acceptor sites from gene exons can lead to chimeric transcripts[89–91]. At the *MET* locus, which encodes the hepatocyte growth factor receptor tyrosine kinase, such transcripts originate in an intronic L1PA2 element and splice to exons coding the C-terminal end of MET. This truncated RNA isoform has been described in several types of cancer, including in the blast phase of chronic myeloid leukaemia (CML) and in colon, bladder and oesophageal cancers[92–97]. Retrotransposon sequence incorporated into L1-ASP transcripts can also provide the start of an ORF (ORF0) such that a subset of these chimeric RNAs have the potential to code for fusion proteins[91,98]. Thus, genomic LINE-1 sequences have the potential to alter the regulation and coding potential of cellular RNAs.

Although LTR retrotransposons have been rendered immobile in humans, their sequences have persistent, and sometimes co-opted, regulatory functions[99]. Compared with LINE-1 sequences, these are more straightforward to study as there are fewer polymorphic variants excluded from the human reference genome assembly and the sequences are sufficiently divergent to allow for unambiguous short read alignments.

PIWI-interacting RNAs
(piRNAs). Small RNAs (26–31 nucleotides) that bind to PIWI family proteins and have key roles in silencing retroelements.

Short read alignments
Alignments of short reads from massively parallel sequencing (MPS) studies to a reference sequence; for example, to the human reference genome assembly.

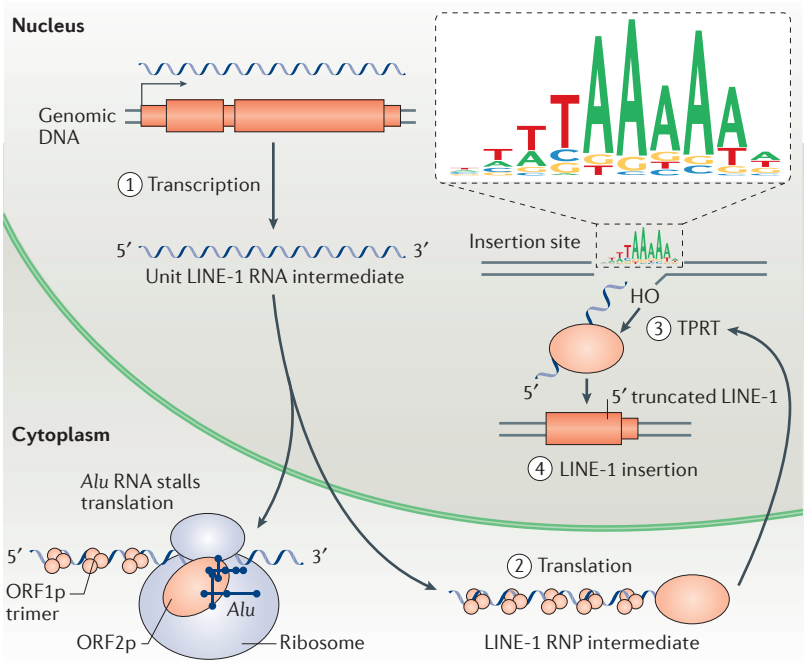


Figure 2 | **Mobilization of the LINE1 retrotransposon.** Long interspersed element-1 (LINE-1) is a retrotransposon, which means that its replication cycle involves an RNA intermediate. LINE-1 is also an autonomous element, meaning that it encodes its own protein machinery for reverse transcription and integration of its sequence into the genome. Most genomic copies of LINE-1 do not have intact copies of the open reading frames (ORFs) needed for retrotransposition; however, around 100 copies do have ORF1 protein (ORF1p)- and ORF2p-coding sequences and are retrotransposition competent. The order of events during the mobilization of a LINE-1 element, which results in a 5′ truncated insertion is: step **1**, transcription in the nucleus and export of the RNA; step **2**, translation and ribonucleoprotein (RNP) assembly in the cytoplasm; step **3**, reverse transcription of the RNA by ORF2p-mediated target primed reverse transcription (TPRT); and step **4**, resolution of the double-stranded insert. LINE-1 transcription is controlled by RNA polymerase II (RNApolII) and regulated by the methylation status of the CpG-rich internal promoter. ORF1p is a major component of the LINE-1 RNP complexes that form. It is thought that nuclear import is required for ORF2p to access the genome. ORF2p has endonuclease activity that nicks the DNA at an insertion site (the consensus sequence of insertion sites identified from the The Cancer Genome Atlas (TCGA) survey study is shown) and leaves a 3′OH group. This in turn primes the reverse transcription reaction mediated by a distinct domain of ORF2p. Resolution of the intermediate likely involves host DNA repair proteins. In addition, the highly structured *Alu* RNA can compete with LINE-1 RNA for newly translated ORF2p by ribosome stalling.

Target primed reverse transcription
(TPRT). The molecular process executed by long interspersed element-1 (LINE-1) open reading frame 2p (ORF2p), using a strand of target site DNA to prime reverse transcription of an RNA.

Cis-preference
The tendency of long interspersed element-1 (LINE-1)-encoded open reading frame 2p (ORF2p) to associate with and reverse transcribe the RNA strand that encoded the protein. This opposes other RNA species that co-opt ORF2p by association in *trans*.

ERV1, ERVL and MaLR class remnants have been shown to provide promoter sequences for long non-coding RNAs (lncRNAs) and chimeric gene transcripts expressed in cancers. Examples of the latter include macrophage colony stimulating factor 1 receptor (*CSF1R*)[100] and interferon regulatory factor 5 (*IRF5*) RNA isoforms[101] in Hodgkin lymphoma driven by ERV promoter sequences, Erb-b2 receptor tyrosine kinase 4 (*ERBB4*; also known as *HER4*) transcripts initiated from intronic LTRs and coding for N-terminally truncated protein products in anaplastic large cell lymphoma[102] and a LTR2–fatty acid binding protein 7 (FABP7) fusion protein in diffuse large B cell lymphoma[103,104].

LINE-1 and LTR transcription also has the potential to result in double-stranded RNA (dsRNA). Origins may include lncRNA originating from a transposable element promoter that overlaps and opposes the direction of a cellular mRNA[105], individual transposon loci with bidirectional transcription[89,106] and, potentially, hybridizations of highly similar, complementary RNA strands encoded from different loci. These dsRNA species can instigate gene silencing, for example, at the metastasis suppressor tissue factor pathway inhibitor 2 (*TFPI2*) locus in breast and colon cancers[105]. They can also activate cellular antiviral response pathways in ovarian cancer and melanoma treated with DNA methyltransferase inhibitors[107].

Mobile element proteins and retrotransposition. All mobile DNA activity in modern humans can be attributed to LINE-1 proteins. LINE-1 is a 6-kb sequence with two essential protein ORFs, encoding ORF1 protein (ORF1p; also known as p40)[7,108] and ORF2p[109]. LINE sequences propagate through target primed reverse transcription (TPRT)[110,111] (FIG. 2). Briefly, LINE-1 RNA associates in ribonucleoprotein (RNP) complexes with ORF1p protein, and along with ORF2p, gains access to chromatin. TPRT hinges on two enzymatic functions of ORF2p: endonuclease and reverse transcriptase activities[112,113]. The endonuclease reaction exposes a 3′OH from the target site DNA strand for priming, and the reverse transcriptase then elongates this strand using the LINE-1 RNA as a template. Even among source elements with intact ORFs for ORF1p and ORF2p, there is some variability in promoter and protein-coding sequences, and only a small number of elements have been recognized as highly active or 'hot' by *in vitro* retrotransposition assays[37,114] (BOX 1).

ORF2p generally displays *cis*-preference for copying the LINE-1 RNA that encoded it instead of retrotransposing other LINE-1 or unrelated sequences[25]. *Alu* interspersed repeats are uniquely successful at breaking this preference, and mobilize in *trans* by co-opting LINE-1-encoded ORF2p. *Alu* are RNApolIII-transcribed sequences derived from signal recognition particle 7SL RNA[115,116], and *Alu* RNA assumes a highly structured form that stalls translation of LINE-1 at the ribosome and competes with the LINE-1 RNA to associate with ORF2p[117,118] (FIG. 2). This co-option appears commonplace in the germline, where rates of *de novo Alu* retrotransposition have surpassed those of LINE-1, although somatic *Alu* retrotransposition in cancers occurs less frequently[71,119].

ORF1p encodes an RNA binding protein[120,121] that forms trimers with avidity for the negatively charged, bendable backbone of single-stranded RNA[122,123]. ORF1p interacts with LINE-1 RNA and is required for retrotransposition of LINE-1, although its exact role is unclear. ORF1p largely localizes to the cytoplasm where it is a major constituent of LINE-1 RNPs[124–126].

LINE-1 ORF1p overexpression is a hallmark of human cancers[127–129]. ORF1p can be detected by immunohistochemistry, in more than 90% of breast and ovarian cancers, as well as nearly 90% of pancreatic cancers, 50–60% of tubular gastrointestinal tract cancers, including oesophageal cancers and colon cancers, 50% of lung cancers and 40% of prostate cancers[127]. In contrast, primary glioblastomas and low-grade B cell

lymphomas express little ORF1p[127,130]. Thus, tumour origin is an important determinant of LINE-1 expression, although the molecular mechanisms underlying this are unknown. However, tissue-specific effects are not the only determinant of ORF1p expression. Even within malignancies of one histological type, there is variation in the levels of ORF1p immunoreactivity from case to case. For example, LINE-1 ORF1p is highly expressed in approximately 20% of cases of lung adenocarcinoma, expressed at more moderate levels in approximately 30% of cases and is undetectable in the remainder[127]. Part of this may be attributed to differences in the complement of inherited, active LINE-1 inherent to each individual, although this has not been directly demonstrated. The alternative explanation is that permissiveness for LINE-1 expression reflects another genetic or epigenetic feature that is variable across these tumours. Supporting this model, p53 mutation is associated with LINE-1 ORF1p expression in lung cancers[127], and the wild-type p53 tumour suppressor represses transposable element expression in experimental systems[131]. Moreover, genomic DNA methylation profiling can be used to distinguish lung cancer cases with and without somatically acquired LINE-1 insertions[39]. These data suggest that together cellular environment and epigenetic status may supersede variations in inherited LINE-1 loci in determining ORF1p expression.

In contrast, in pancreatic ductal adenocarcinoma (PDAC), most fully developed malignancies (~90%) express LINE-1 ORF1p. This loss of retroelement control is not seen with the same uniformity early in the disease; only 11% of the histological precursor lesions, pancreatic intraepithelial neoplasias (PanINs), express LINE-1 ORF1p[127]. Although ORF1p expression levels vary between individuals, there is little heterogeneity within a single tumour or between different sites of disease in the same patient[132].

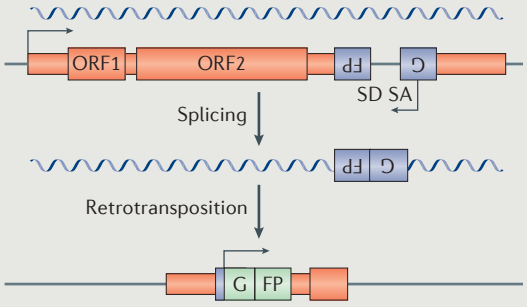
Antibodies to detect ORF2p have recently been reported[133–135], and immunostaining shows that it is expressed in most intestinal adenomas and colon adenocarcinomas, prostate intraepithelial neoplasias (PINs) and prostate adenocarcinomas, lung adenocarcinomas, and breast invasive ductal carcinomas[135]. Although more studies remain to be done to validate these findings and test their clinical relevance, these data suggest that LINE-1 ORF1p and ORF2p could have utility as cancer biomarkers for non-invasive screening or predicting clinical outcomes.

Functional effects of LINE-1 proteins in human cancers remain to be characterized in most of these disease contexts. As with other aspects of this exciting field, the repetitive and variable numbers of genomic sequences have posed unique challenges, for example, complicating approaches to loss-of-function experiments. Characterizing cellular RNA that may be sequestered by ORF1p[126] and addressing potential consequences of ORF2p endonuclease and reverse transcriptase activities apart from canonical retrotransposition are likely to emerge as important areas of investigation. Finally, although LTR retrotransposons are not competent for transposition, their persisting protein-coding potential may promote neoplastic properties during tumour progression. Oncogenic effects have been attributed to the splice products of the HERV-K(HML2) *env* gene, *rec* and *np9* (REF. 136), and immunosuppressive functions are associated with the HERV-H envelope[137]. Further studies to characterize the expression and functional significance of these virus-like proteins in cancers may be valuable.

Retrotransposition in cancer

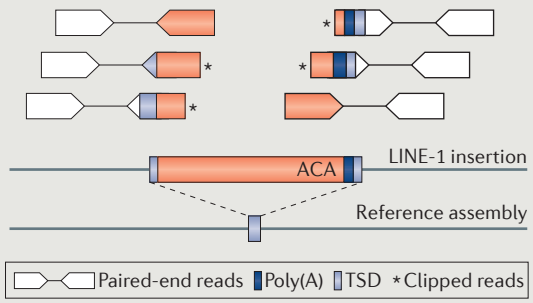
Somatic LINE-1 insertions occur in many types of human cancer, and they can be clinically significant mutagens. The first example of this was recognized in 1992 in a patient with colorectal cancer. In this patient, a 750-bp LINE-1 insertion interrupted a coding exon of adenomatous polyposis coli (*APC*) and crucially compromised *APC* function[138]. The *APC* gene is well-established to have tumour suppressor activities in the colon[139]. Inherited mutations in *APC* cause familial adenomatous polyposis (FAP) and 'second hit', as well as *de novo*, acquired mutations are commonplace early events in sporadic colorectal tumours[140–144]. Methods that harnessed next-generation sequencing for genome-wide profiling of LINE-1 insertions in tumour DNA for comparison with matched normal DNA took nearly 20 years to establish[39] (BOX 2). By 2016, these methods had matured and a second somatic LINE-1 insertion in the same *APC* exon was reported in an independent patient with colon cancer[72]. This insertion occurred at a different position in the exon and spanned 1.4 kb. Ultimately, the LINE-1 insertion in *APC* in this case was one of 27 somatically acquired LINE-1 insertions identified in the same tumour. The other insertions were distributed in intronic and intergenic spaces genome wide without obvious functional effects.

Box 1 | **Retrotransposition assays**



Retrotransposition assays have informed much of our understanding of long interspersed element-1 (LINE-1) activity. These assays provide a measurement of the processing of a LINE-1 RNA intermediate and the reverse transcription and integration of that processed sequence into the genome. Native LINE-1 does not have introns, and splicing of LINE-1 RNA sequences is presumed to typically preclude retrotransposition. These assays introduce a reporter gene (GFP) into the 3′ untranslated region (UTR) of an ectopic LINE-1 sequence in an antisense orientation (that is, opposing the direction of the LINE-1). The reporter is interrupted by an intronic sequence in the same direction as the LINE-1. It is operational only after the sequence has been retrotransposed[7–9]. The LINE-1 RNA must be spliced to remove this intron, and the reconstituted reporter must be reverse transcribed, integrated into the genome and expressed in sense. These assays have been used to demonstrate which LINE-1 loci are hot or highly retrotransposition competent[37,114], as well as to identify crucial LINE-1 sequences and cellular proteins important for retrotransposition ORF, open reading frame; SA, splice acceptor or 3′ splice site; SD, splice donor or 5′ splice site.

Box 2 | **Finding germline and somatic insertions**



Most polymorphic, inherited long interspersed element-1 (LINE-1) variants and all somatically acquired LINE-1 will not be incorporated in the human reference genome assembly. Instead, there are hundreds of thousands of look-alike sequences. Thus, identifying new insertions requires approaches tailored to analysing targeted sequencing (amplification typing of L1 active subfamilies (ATLAS)[167], LINE-1 sequencing (L1-seq)[39,148], transposon insertion profiling (TIP)-seq[132], retrotransposon capturing (RC)-seq[70], sequencing, identification and mapping of primed LINE-1 elements (SIMPLE)[168]) or whole-genome sequencing (WGS)[71,119,152]. Algorithms use paired-end and clipped reads that span the junction between the insertion sequence and the surrounding genomic sequence to pinpoint insertions. Several of these algorithms use the polyadenylation (poly(A)) tail at the 3′ end of the element, and one has been developed specifically to recognize 3′ transduction events. For WGS data and targeted sequencing approaches that capture both sides of the element (that is, the 5′ and 3′ ends), algorithms may recognize target site duplications (TSDs). Some assays use a one-sided PCR to amplify insertions and may have useful information only from the 3′ aspect (bold outlined paired-end reads, right side). The 3′ACA trinucleotide sequence distinguishes L1PA1 insertions from older LINE-1 subfamilies.

Genomic LINE-1 copies that result from canonical retrotransposition reactions (FIG. 2) have five sequence features. First, these are almost universally members of 'young', currently active LINE-1 families (L1PA1 and L1PA2)[6,25]; old LINE-1 sequences are not substrates for retrotransposition. Second, LINE-1 and sequences dependent on LINE-1 ORF2p for retrotransposition have 3′ poly(A) sequences incorporated into their genomic copies, reflecting their RNA precursors[109,118]. Third, these insertions tend to occur at ORF2p endonuclease recognition sites, 5′ TTTT-AA 3′, where the gap is the nicked strand scissile bond[112,145]. Fourth, the copies are flanked by variable length target site duplications (TSDs)[146]. Fifth, LINE-1 sequences that are retrotransposed are typically 5′ truncated, and inversions of the 5′ end are also seen[147].

Tumour types prone to retrotransposition. Tumours of the gastrointestinal tract, particularly colon cancers, are prone to somatic LINE-1 activity. A study using targeted LINE-1 insertion site sequencing in 16 colorectal tumours and matched normal DNA samples identified 69 PCR-validated, tumour-specific insertions[148]. A survey of 43 tumours that had undergone whole-genome sequencing in The Cancer Genome Atlas (TCGA) showed more LINE-1 retrotransposition in colorectal cancers than in prostate and ovarian cancers, with one colorectal cancer case harbouring 106 somatic insertions[119]. In other cases the number of insertions ranged from a few to a few dozen[119]. Other luminal malignancies, namely oesophageal squamous carcinomas and adenocarcinomas[149,150], gastric and small bowel tumours[132], hepatocellular carcinomas[70] and PDACs[132,151], also show somatic LINE-1 retrotransposition.

Outside the gastrointestinal tract, NSCLCs[39,71] and head and neck cancers[71,152] frequently show somatically acquired LINE-1 insertions, as do prostate and ovarian cancers as mentioned above[71,119,153]. Meanwhile, haematolymphoid malignancies, namely plasma cell myelomas[119] and acute myeloid leukaemias (AMLs)[152], and high-grade central nervous system gliomas[39,119,130,154] have little evidence of somatic retrotransposition. Breast cancers also have relatively low levels[71].

Mechanisms underlying this predilection for LINE-1 retrotransposition in certain malignancies are not well understood. Overall, the types of tumour that acquire somatic LINE-1 insertions correspond to those types that show LINE-1 protein expression[127], and controls on expression are probably important. However, there is no clear correlation between the amount of ORF1p across tumour types and the numbers of retrotransposition events seen. Ovarian epithelial cancers as a group have the highest levels of ORF1p expression[127], but more modest numbers of LINE-1 insertions per tumour than gastrointestinal tract cancers[119]. Either non-functional LINE-1 loci are expressed in these ovarian tumours, or the LINE-1 proteins are necessary but not sufficient for retrotransposition.

Hot LINE-1 loci in human cancers. A proportion of LINE-1 insertions exhibit 3′ transduction. These result from a read-through of the poly(A) sequence of the source element such that the RNA intermediate and ultimately the newly acquired genomic insertion incorporate this downstream sequence — a signature of the source element locus[87,88,155–157]. A recent study of genomes from 244 patients with cancer found that approximately half of cases had somatically acquired retrotranspositions and a significant portion, almost one-quarter of these events (655 of 2,756 total insertions), had 3′ transduced sequences[71]. Nearly all these insertions (95%) were attributable to 72 germline LINE-1 loci. Two exceptionally hot elements at 22q12 and 6p24.1 accounted for more than one-third of somatic transductions in this study and frequently sourced numerous insertions per tumour. Although it is unclear to what extent this representation reflects their expression levels, inherent retrotransposition potential or their tendency to read through their poly(A) sequence, the emerging model is that, as in the germline, limited numbers of hot LINE-1 loci dominate retrotransposition in cancers.

Internal sequence variations (that is, within the LINE-1 sequence) can also be used to infer the source element of an insertion[72]. Depending on the complement of full-length L1Hs elements in an individual and the length of a somatically acquired insertion, it is possible to either unambiguously identify the source or narrow the possibilities. This approach was used to trace the source element of the most recently discovered somatic LINE-1 insertion at *APC*[72]. The source LINE-1 is a full-length, intact element at 17p11.2.

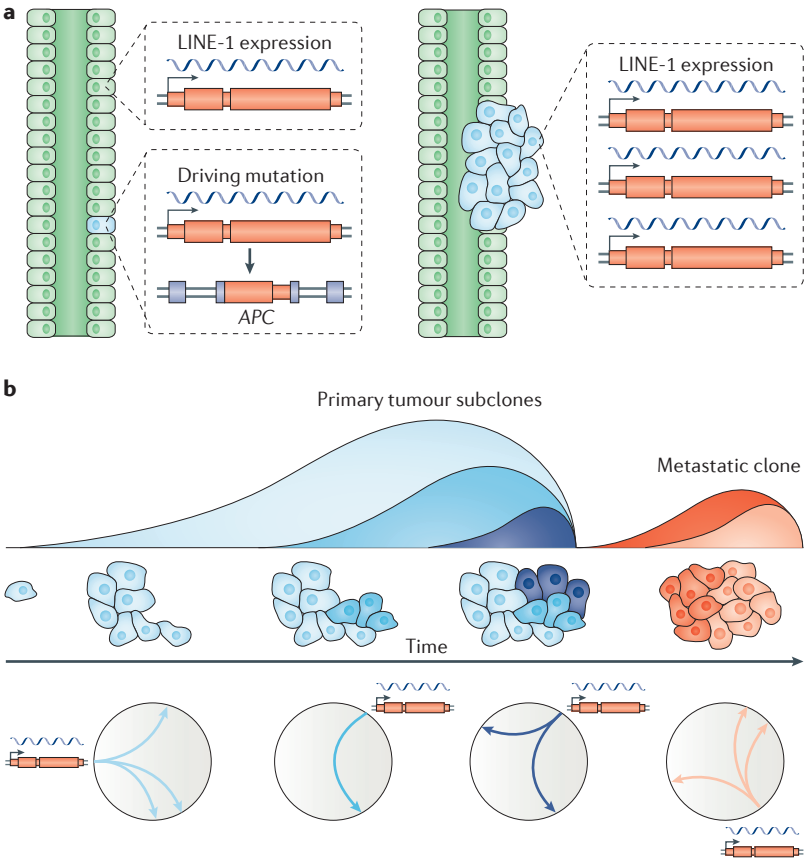


Figure 3 | **LINE-1 insertions accrue in cancer cell clones. a** | Precancerous and normal cells can allow long interspersed element-1 (LINE-1) expression, although it is unclear how pervasive this is. Left: the diagram shows a LINE-1 source element that escaped somatic silencing in normal epithelial tissue lining the gastrointestinal tract (indicated by green cells). In most cells, this has no effect. In a rare cell, this will generate an insertion that promotes tumour development, such as one identified in the adenomatous polyposis coli (*APC*) locus. Right: cells that have undergone transformation may become more permissive for LINE-1 expression. **b** | A model of how transiently active source LINE-1 can contribute to the genetic evolution of cancers. Clonal evolution of a primary tumour and its metastatic offshoot is shown with the quantity of tumour cells denoted by the size of the evolving clones. Time is depicted on the *x* axis. Four successive LINE-1 source elements are illustrated as being active over the course of tumour development; these are colour-coded to match the subclone in which each is active. The circular plots depict genomic locations arrayed clockwise as the circumference, with arrows drawn between source LINE-1 positions and their corresponding acquired insertions (arrowheads). The genetic heterogeneity of the malignant population that results from acquired insertion events is indicated by the clusters of tumour cells.

Two additional source elements on chromosomes 12 and 14 were responsible for most of the other somatic insertions in that case.

Retrotransposition of other sequences can also occur in cancers, although these have received less attention to date. Somatically acquired *Alu* insertions appear to be outnumbered by LINE-1 retrotransposition events, comprising only about 5.5% of instances (10 insertions) reported in the TCGA survey study[119]. Somatically acquired pseudogene insertions have also been described (42 in 660 cancer cases) with greatest prevalence in NSCLCs and colorectal cancers[158,159].

Timing of retrotransposition. When do somatic LINE-1 insertions occur? The *APC* insertion described in 2016 likely served as an early, driver mutation in the development of the tumour. Consistent with this, its source element was found to have escaped silencing in adjacent, normal tissue by both RNA-seq and LINE-1 promoter methylation analysis[72]. Other somatically acquired LINE-1 insertions in intronic and intergenic intervals — and likely passenger mutations — can also occasionally be detected in adjacent, normal tissue and in cancer precursor lesions[149,151].

That said, fully transformed cells are likely to provide a more optimal environment for LINE-1 expression and ongoing retrotransposition. Established malignancies show LINE-1 promoter hypomethylation and ORF1p expression as discussed in the preceding sections. Interestingly, though, the phenomenon of LINE-1 expression and retrotransposition appears to occur with spatial and temporal restrictions in tumour development. Studies using 3′ transduced sequences to identify source elements show that these hot LINE-1 loci are not simultaneously active throughout the tumour. Instead, different phases of cancer growth will be punctuated by the transient activities of different source elements[71] (FIG. 3). Consistent with this, in the evolution of pancreatic cancers, a primary tumour may show a high number of somatically acquired LINE-1, whereas its metastatic outgrowth would accrue no additional LINE-1, or the converse[132]. These data suggest an episodic or heterogeneous quality to LINE-1 activity in cancers.

Functions of new insertions. Retrotransposed sequences can compromise gene function. Insertions of LINE-1 and *Alu* can directly disrupt coding exons (FIG. 4). They can also interfere with mRNA splicing when they land close to exons. Both mechanisms are well-described ways for insertions to produce loss-of-function alleles in human genetic diseases[6]. The somatic LINE-1 insertions discovered at the *APC* locus in colon cancer[72,138] fit this model. These occur in a crucial coding exon of a crucial tumour suppressor gene. There is no doubt that they are causative, or driver, mutations in the development of these cancers.

Newly retrotransposed sequences most commonly introduce small fragments of LINE-1 to non-coding regions of the genome already replete with repetitive DNAs. These somatically acquired insertions are widely distributed without strong hot spots near genes or enrichments in tumour suppressor or oncogene loci[160]. My interpretation of these findings is that these mutations are epiphenomenal to oncogenesis — as a rule, neutral, passenger mutations.

Rules can have important exceptions. The large majority of retroelement insertions described to date in cancers are in intronic and intergenic intervals. Whether these are consequential has now become a central question in the field. Like other mutations in non-coding DNA, insertion of a mobile DNA has the potential to disrupt protein binding sites and regulatory sequences[70]. Beyond this potential for disruption, transposable

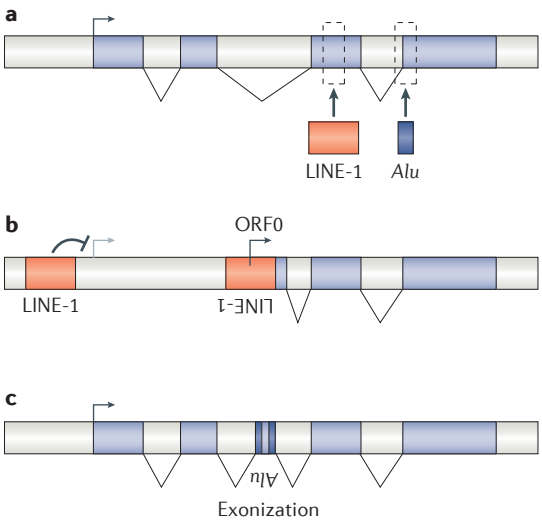


Figure 4 | **Functional impacts of transposable elements.** The schematic shows a hypothetical gene locus with exons marked in grey. The rightward pointing arrow depicts a promoter, and thin black lines denote mRNA splicing patterns. **a** | Long interspersed element-1 (LINE-1) and *Alu* insertions within or adjacent to exons can interrupt protein-coding sequences or disrupt splicing, resulting in loss-of-function alleles, such as the insertion in the coding exon of adenomatous polyposis coli (*APC*), or alleles producing altered mRNAs. **b** | Transposable element insertions in non-coding DNA could disrupt regulatory sequences (left). Full-length LINE-1 sequences also have the potential to introduce promoter activity and produce chimeric transcripts and protein coding sequences. For example, a LINE-1-antisense promoter (L1-ASP) transcript juxtaposing the LINE-1 open reading frame 0 (ORF0) in-frame with downstream exons is shown (right)[98]. **c** | An intronic, antisense *Alu* insertion provides a new exon sequence in a process known as exonization.

element insertions are unique in that they have the ability to introduce to a locus a complex sequence module (FIG. 4). Transpositions of full-length LINE-1 may position its promoter to produce new RNA isoforms and thus break genes[161]. Insertions with 3′ transduction events encompassing a downstream exon can introduce this into a new RNA, in a phenomenon known as exon shuffling[71,156]. *Alu* repeats, which are under-studied in cancers, have internal sequences that have a tendency to be 'exonized'[162,163]. Finally, gene rearrangements associated with retrotransposition[27,164,165] or post-insertion recombinations that delete intervening sequence[166] may be highly significant. Moving forward, we should seek out those retrotransposed sequences with content likely to have functional effects, those associated with selectively advantaged tumour subclones and those recurring independently in different patients.

Conclusions

Much of our DNA is a historical record of retroelement invasions — fixed and fragmented repeats that have gone cold. However, hot LINE-1 sequences persist, reverse transcribe themselves and trap likenesses of cellular RNAs in our genome in the process. In recent years, it has come to light that ancient mobile element promoters and modern LINE-1 loci are expressed as RNA and protein in the context of cancer, and that LINE-1 is interspersing new sequences in cancer genomes. There are many implications of these findings. The selective expression of repetitive sequences in cancers may provide opportunities for cancer detection or cancer cell targeting, and their RNAs and proteins may have roles in transformation. Moreover, acquired insertions have demonstrated potential to act as driver mutations, and can both destroy and create functions at loci in which they fall.

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
3. Boissinot, S., Chevret, P. & Furano, A. V. L1 (LINE-1) retrotransposon evolution and amplification in recent human history. *Mol. Biol. Evol.* **17**, 915–928 (2000).
4. Huang, C. R., Burns, K. H. & Boeke, J. D. Active transposition in genomes. *Annu. Rev. Genet.* **46**, 651–675 (2012).
5. Kazazian, H. H. Jr. *et al.* Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature* **332**, 164–166 (1988).
6. Hancks, D. C. & Kazazian, H. H. Jr. Roles for retrotransposon insertions in human disease. *Mob. DNA* **7**, 9 (2016).
7. Moran, J. V. *et al.* High frequency retrotransposition in cultured mammalian cells. *Cell* **87**, 917–927 (1996).
8. Ostertag, E. M., Prak, E. T., DeBerardinis, R. J., Moran, J. V. & Kazazian, H. H. Jr. Determination of L1 retrotransposition kinetics in cultured cells. *Nucleic Acids Res.* **28**, 1418–1423 (2000).
9. Kopera, H. C. *et al.* LINE-1 cultured cell retrotransposition assay. *Methods Mol. Biol.* **1400**, 139–156 (2016).
10. Ostertag, E. M. *et al.* A mouse model of human L1 retrotransposition. *Nat. Genet.* **32**, 655–660 (2002).
11. An, W. *et al.* Active retrotransposition by a synthetic L1 element in mice. *Proc. Natl Acad. Sci. USA* **103**, 18662–18667 (2006).
12. Batzer, M. A. & Deininger, P. L. Alu repeats and human genomic diversity. *Nat. Rev. Genet.* **3**, 370–379 (2002).
13. Xing, J. *et al.* Mobile elements create structural variation: analysis of a complete human genome. *Genome Res.* **19**, 1516–1526 (2009).
14. Ewing, A. D. & Kazazian, H. H. Jr. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* **20**, 1262–1270 (2010).
15. Ewing, A. D. & Kazazian, H. H. Jr. Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. *Genome Res.* **21**, 985–990 (2011).
16. Hormozdiari, F. *et al.* Alu repeat discovery and characterization within human genomes. *Genome Res.* **21**, 840–849 (2011).
17. Stewart, C. *et al.* A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.* **7**, e1002236 (2011).
18. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
19. Richardson, S. R. *et al.* The influence of LINE-1 and SINE retrotransposons on mammalian genomes. *Microbiol. Spectr.* **3**, MDNA3-0061-2014 (2015).
20. Boeke, J. D., Garfinkel, D. J., Styles, C. A. & Fink, G. R. Ty elements transpose through an RNA intermediate. *Cell* **40**, 491–500 (1985).
21. Dewannieux, M., Esnault, C. & Heidmann, T. LINE-mediated retrotransposition of marked Alu sequences. *Nat. Genet.* **35**, 41–48 (2003).
22. Hancks, D. C., Goodier, J. L., Mandal, P. K., Cheung, L. E. & Kazazian, H. H. Jr. Retrotransposition of marked SVA elements by human L1s in cultured cells. *Hum. Mol. Genet.* **20**, 3386–3400 (2011).
23. Raiz, J. *et al.* The non-autonomous retrotransposon SVA is *trans*-mobilized by the human LINE-1 protein machinery. *Nucleic Acids Res.* **40**, 1666–1683 (2012).
24. Esnault, C., Maestre, J. & Heidmann, T. Human LINE retrotransposons generate processed pseudogenes. *Nat. Genet.* **24**, 363–367 (2000).
25. Wei, W. *et al.* Human L1 retrotransposition: *cis* preference versus *trans* complementation. *Mol. Cell. Biol.* **21**, 1429–1439 (2001).
26. Buzdin, A. *et al.* A new family of chimeric retrotranscripts formed by a full copy of U6 small nuclear RNA fused to the 3′ terminus of L1. *Genomics* **80**, 402–406 (2002).
27. Gilbert, N., Lutz, S., Morrish, T. A. & Moran, J. V. Multiple fates of L1 retrotransposition intermediates in cultured human cells. *Mol. Cell. Biol.* **25**, 7780–7795 (2005).
28. Garcia-Perez, J. L., Doucet, A. J., Bucheton, A., Moran, J. V. & Gilbert, N. Distinct mechanisms for *trans*-mediated mobilization of cellular RNAs by the LINE-1 reverse transcriptase. *Genome Res.* **17**, 602–611 (2007).
29. Doucet, A. J., Droc, G., Siol, O., Audoux, J. & Gilbert, N. U6 snRNA pseudogenes: markers of retrotransposition dynamics in mammals. *Mol. Biol. Evol.* **32**, 1815–1832 (2015).
30. Schueler, M. G. & Sullivan, B. A. Structural and functional dynamics of human centromeric chromatin. *Annu. Rev. Genom. Hum. Genet.* **7**, 301–313 (2006).
31. Mefford, H. C. & Trask, B. J. The complex structure and dynamic evolution of human subtelomeres. *Nat. Rev. Genet.* **3**, 91–102 (2002).

32. de Koning, A. P., Gu, W., Castoe, T. A., Batzer, M. A. & Pollock, D. D. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* **7**, e1002384 (2011).
33. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8**, 973–982 (2007).
34. Giordano, J. *et al.* Evolutionary history of mammalian transposons determined by genome-wide defragmentation. *PLoS Comput. Biol.* **3**, e137 (2007).
35. Boissinot, S., Entezam, A., Young, L., Munson, P. J. & Furano, A. V. The insertional history of an active family of L1 retrotransposons in humans. *Genome Res.* **14**, 1221–1231 (2004).
36. Boissinot, S. & Furano, A. V. The recent evolution of human L1 retrotransposons. *Cytogenet. Genome Res.* **110**, 402–406 (2005).
37. Beck, C. R. *et al.* LINE-1 retrotransposition activity in human genomes. *Cell* **141**, 1159–1170 (2010). **Description of full-length LINE-1 polymorphisms in human populations, and identification of a subset of highly active elements for retrotransposition.**
38. Huang, C. R. *et al.* Mobile interspersed repeats are major structural variants in the human genome. *Cell* **141**, 1171–1182 (2010).
39. Iskow, R. C. *et al.* Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell* **141**, 1253–1261 (2010). **Demonstration that massively parallel sequencing can be targeted to recover germline and somatically acquired retrotransposon insertions.**
40. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12**, 236 (2011).
41. Witherspoon, D. J. *et al.* Mobile element scanning (ME-Scan) by targeted high-throughput sequencing. *BMC Genomics* **11**, 410 (2010).
42. Witherspoon, D. J. *et al.* Mobile element scanning (ME-Scan) identifies thousands of novel Alu insertions in diverse human populations. *Genome Res.* **23**, 1170–1181 (2013).
43. Jurka, J., Zietkiewicz, E. & Labuda, D. Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Res.* **23**, 170–175 (1995).
44. Smit, A. F. & Riggs, A. D. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* **23**, 98–102 (1995).
45. Hancks, D. C., Mandal, P. K., Cheung, L. E. & Kazazian, H. H. Jr. The minimal active human SVA retrotransposon requires only the 5′-hexamer and Alu-like domains. *Mol. Cell. Biol.* **32**, 4718–4726 (2012).
46. Hancks, D. C. & Kazazian, H. H. Jr. SVA retrotransposons: evolution and genetic instability. *Semin. Cancer Biol.* **20**, 234–245 (2010).
47. Smit, A. F. Identification of a new, abundant superfamily of mammalian LTR-transposons. *Nucleic Acids Res.* **21**, 1863–1872 (1993).
48. Dewannieux, M. *et al.* Identification of an infectious progenitor for the multiple-copy HERV-K human endogenous retroelements. *Genome Res.* **16**, 1548–1556 (2006).
49. Belshaw, R. *et al.* Genomewide screening reveals high levels of insertional polymorphism in the human endogenous retrovirus family HERV-K(HML2): implications for present-day activity. *J. Virol.* **79**, 12507–12514 (2005).
50. Hughes, J. F. & Coffin, J. M. Human endogenous retrovirus K solo-LTR formation and insertional polymorphisms: implications for human and viral evolution. *Proc. Natl Acad. Sci. USA* **101**, 1668–1672 (2004).
51. Turner, G. *et al.* Insertional polymorphisms of full-length endogenous retroviruses in humans. *Curr. Biol.* **11**, 1531–1535 (2001).
52. Smit, A. F. & Riggs, A. D. Tiggers and DNA transposon fossils in the human genome. *Proc. Natl Acad. Sci. USA* **93**, 1443–1448 (1996).
53. Roth, D. B. & Craig, N. L. VDJ recombination: a transposase goes to work. *Cell* **94**, 411–414 (1998).
54. Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature* **394**, 744–751 (1998).
55. Hencken, C. G., Li, X. & Craig, N. L. Functional characterization of an active Rag-like transposase. *Nat. Struct. Mol. Biol.* **19**, 834–836 (2012).
56. Hiom, K., Melek, M. & Gellert, M. DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell* **94**, 463–470 (1998).
57. Kapitonov, V. V. & Jurka, J. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol.* **3**, e181 (2005).
58. Huang, S. *et al.* Discovery of an active RAG transposon illuminates the origins of V(D)J recombination. *Cell* **166**, 102–114 (2016).
59. Liu, D. *et al.* The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase. *Mol. Cell. Biol.* **27**, 1125–1132 (2007).
60. Majumdar, S., Singh, A. & Rio, D. C. The human *THAP9* gene encodes an active P-element DNA transposase. *Science* **339**, 446–448 (2013).
61. Henssen, A. G. *et al.* Genomic DNA transposition induced by human PGBD5. *Elife* **4,** e10565 (2015).
62. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
63. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nat. Rev. Mol. Cell Biol.* **12**, 246–258 (2011).
64. Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
65. Jacobs, F. M. *et al.* An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* **516**, 242–245 (2014).
66. Quenneville, S. *et al.* The KRAB–ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. *Cell Rep.* **2**, 766–773 (2012).
67. Rowe, H. M. *et al.* KAP1 controls endogenous retroviruses in embryonic stem cells. *Nature* **463**, 237–240 (2010).
68. Turelli, P. *et al.* Interplay of TRIM28 and DNA methylation in controlling human endogenous retroelements. *Genome Res.* **24**, 1260–1270 (2014).
69. Alves, G., Tatro, A. & Fanning, T. Differential methylation of human LINE-1 retrotransposons in malignant cells. *Gene* **176**, 39–44 (1996).
70. Shukla, R. *et al.* Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153**, 101–111 (2013). **Mapping of LINE-1 insertions in hepatocellular carcinoma and tracking of inherited and acquired LINE-1 in non-coding DNA that relate to pathogenesis.**
71. Tubio, J. M. *et al.* Mobile DNA in cancer: extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science* **345**, 1251343 (2014). **Demonstration of a set of source elements responsible for LINE-1 retrotransposition with 3′ transduction in human cancers.**
72. Scott, E. C. *et al.* A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016). **Description of a second somatically acquired LINE-1 insertion in the *APC* gene, which could be traced to a source element that escaped silencing in the normal colon.**
73. Daskalos, A. *et al.* Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. *Int. J. Cancer* **124**, 81–87 (2009).
74. Saito, K. *et al.* Long interspersed nuclear element 1 hypomethylation is a marker of poor prognosis in stage IA non-small cell lung cancer. *Clin. Cancer Res.* **16**, 2418–2426 (2010).
75. Ogino, S. *et al.* A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. *J. Natl Cancer Inst.* **100**, 1734–1738 (2008).
76. Iwagami, S. *et al.* LINE-1 hypomethylation is associated with a poor prognosis among patients with curatively resected esophageal squamous cell carcinoma. *Ann. Surg.* **257**, 449–455 (2013).
77. van Hoesel, A. Q. *et al.* Hypomethylation of LINE-1 in primary tumor has poor prognosis in young breast cancer patients: a retrospective cohort study. *Breast Cancer Res. Treat.* **134**, 1103–1114 (2012).
78. Harada, K. *et al.* LINE-1 methylation level and patient prognosis in a database of 208 hepatocellular carcinomas. *Ann. Surg. Oncol.* **22**, 1280–1287 (2015).
79. Pattamadilok, J. *et al.* LINE-1 hypomethylation level as a potential prognostic factor for epithelial ovarian cancer. *Int. J. Gynecol. Cancer* **18**, 711–717 (2008).
80. Hall, L. L. *et al.* Stable C_0T-1 repeat RNA is abundant and is associated with euchromatic interphase chromosomes. *Cell* **156**, 907–919 (2014).
81. Skowronski, J., Fanning, T. G. & Singer, M. F. Unit-length line-1 transcripts in human teratocarcinoma cells. *Mol. Cell. Biol.* **8**, 1385–1397 (1988).
82. Belancio, V. P., Hedges, D. J. & Deininger, P. LINE-1 RNA splicing and influences on mammalian gene expression. *Nucleic Acids Res.* **34**, 1512–1521 (2006).
83. Belancio, V. P., Roy-Engel, A. M. & Deininger, P. The impact of multiple splice sites in human L1 elements. *Gene* **411**, 38–45 (2008).
84. Deininger, P. & Belancio, V. P. Detection of LINE-1 RNAs by northern blot. *Methods Mol. Biol.* **1400**, 223–236 (2016).
85. Philippe, C. *et al.* Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *Elife* **5,** e13926 (2016). **Description of a tailored approach to RNA-seq and ChIP-sequencing, which can be used to identify loci expressing LINE-1 unit transcripts.**
86. Deininger, P. *et al.* A comprehensive approach to expression of L1 loci. *Nucleic Acids Res.* **45**, e31 (2016).
87. Pickeral, O. K., Makalowski, W., Boguski, M. S. & Boeke, J. D. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res.* **10**, 411–415 (2000).
88. Goodier, J. L., Ostertag, E. M. & Kazazian, H. H. Jr. Transduction of 3′-flanking sequences is common in L1 retrotransposition. *Hum. Mol. Genet.* **9**, 653–657 (2000).
89. Matlik, K., Redik, K. & Speek, M. L1 antisense promoter drives tissue-specific transcription of human genes. *J. Biomed. Biotechnol.* **2006**, 71753 (2006).
90. Nigumann, P., Redik, K., Matlik, K. & Speek, M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics* **79**, 628–634 (2002).
91. Criscione, S. W. *et al.* Genome-wide characterization of human L1 antisense promoter-driven transcripts. *BMC Genomics* **17**, 463 (2016).
92. Roman-Gomez, J. *et al.* Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia. *Oncogene* **24**, 7213–7223 (2005).
93. Lin, L. *et al.* Multiple forms of genetic instability within a 2-Mb chromosomal segment of 3q26.3–q27 are associated with development of esophageal adenocarcinoma. *Genes Chromosomes Cancer* **45**, 319–331 (2006).
94. Cruickshanks, H. A. & Tufarelli, C. Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. *Genomics* **94**, 397–406 (2009).
95. Weber, B., Kimhi, S., Howard, G., Eden, A. & Lyko, F. Demethylation of a LINE-1 antisense promoter in the cMet locus impairs Met signalling through induction of illegitimate transcription. *Oncogene* **29**, 5775–5784 (2010).
96. Wolff, E. M. *et al.* Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. *PLoS Genet.* **6**, e1000917 (2010).
97. Hur, K. *et al.* Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. *Gut* **63**, 635–646 (2014).
98. Denli, A. M. *et al.* Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell* **163**, 583–593 (2015).
99. Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.* **9**, 397–405 (2008).
100. Lamprecht, B. *et al.* Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. *Nat. Med.* **16**, 571–579 (2010).
101. Babaian, A. *et al.* Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene* **35**, 2542–2546 (2016).
102. Scarfo, I. *et al.* Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of *ERBB4* transcripts. *Blood* **127**, 221–232 (2016).
103. Lock, F. E. *et al.* Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. *Proc. Natl Acad. Sci. USA* **111**, E3534–E3543 (2014).
104. Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA* **7**, 24 (2016).

105. Cruickshanks, H. A. *et al.* Expression of a large LINE-1-driven antisense RNA is linked to epigenetic silencing of the metastasis suppressor gene *TFPI-2* in cancer. *Nucleic Acids Res.* **41**, 6857–6869 (2013).
106. Domansky, A. N. *et al.* Solitary HERV-K LTRs possess bi-directional promoter activity and contain a negative regulatory element in the U5 region. *FEBS Lett.* **472**, 191–195 (2000).
107. Chiappinelli, K. B. *et al.* Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **162**, 974–986 (2015).
108. Holmes, S. E., Singer, M. F. & Swergold, G. D. Studies on p40, the leucine zipper motif-containing protein encoded by the first open reading frame of an active human LINE-1 transposable element. *J. Biol. Chem.* **267**, 19765–19768 (1992).
109. Dombroski, B. A., Mathias, S. L., Nanthakumar, E., Scott, A. F. & Kazazian, H. H. Jr. Isolation of an active human transposable element. *Science* **254**, 1805–1808 (1991).
110. Luan, D. D., Korman, M. H., Jakubczak, J. L. & Eickbush, T. H. 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 (1993).
111. Christensen, S. M., Bibillo, A. & Eickbush, T. H. Role of the *Bombyx mori* R2 element N-terminal domain in the target-primed reverse transcription (TPRT) reaction. *Nucleic Acids Res.* **33**, 6461–6468 (2005).
112. Feng, Q., Moran, J. V., Kazazian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
113. Mathias, S. L. *et al.* Reverse transcriptase encoded by a human transposable element. *Science* **254**, 1808–1810 (1991).
114. Brouha, B. *et al.* Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. USA* **100**, 5280–5285 (2003). **Study that compared commonly occurring LINE-1 insertions for retrotransposition activity *in vitro* and established the term hot LINE-1.**
115. Ullu, E. & Tschudi, C. Alu sequences are processed 7SL RNA genes. *Nature* **312**, 171–172 (1984).
116. Zietkiewicz, E., Richer, C., Sinnett, D. & Labuda, D. Monophyletic origin of Alu elements in primates. *J. Mol. Evol.* **47**, 172–182 (1998).
117. Ahl, V., Keller, H., Schmidt, S. & Weichenrieder, O. Retrotransposition and crystal structure of an Alu RNP in the ribosome-stalling conformation. *Mol. Cell* **60**, 715–727 (2015).
118. Doucet, A. J., Wilusz, J. E., Miyoshi, T., Liu, Y. & Moran, J. V. A. 3′ Poly(A) tract is required for LINE-1 retrotransposition. *Mol. Cell* **60**, 728–741 (2015).
119. Lee, E. *et al.* Landscape of somatic retrotransposition in human cancers. *Science* **337**, 967–971 (2012). **Demonstration of the different types of human cancer that support LINE-1 retrotransposition.**
120. Hohjoh, H. & Singer, M. F. Ribonuclease and high salt sensitivity of the ribonucleoprotein complex formed by the human LINE-1 retrotransposon. *J. Mol. Biol.* **271**, 7–12 (1997).
121. Hohjoh, H. & Singer, M. F. Sequence-specific single-strand RNA binding protein encoded by the human LINE-1 retrotransposon. *EMBO J.* **16**, 6034–6043 (1997).
122. Martin, S. L., Branciforte, D., Keller, D. & Bain, D. L. Trimeric structure for an essential protein in L1 retrotransposition. *Proc. Natl Acad. Sci. USA* **100**, 13815–13820 (2003).
123. Khazina, E. *et al.* Trimeric structure and flexibility of the L1ORF1 protein in human L1 retrotransposition. *Nat. Struct. Mol. Biol.* **18**, 1006–1014 (2011).
124. Hohjoh, H. & Singer, M. F. Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.* **15**, 630–639 (1996).
125. Kulpa, D. A. & Moran, J. V. Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum. Mol. Genet.* **14**, 3237–3248 (2005).
126. Taylor, M. S. *et al.* Affinity proteomics reveals human host factors implicated in discrete stages of LINE-1 retrotransposition. *Cell* **155**, 1034–1048 (2013).
127. Rodic, N. *et al.* Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am. J. Pathol.* **184**, 1280–1286 (2014). **A survey of different types of human cancer that aberrantly express LINE-1-encoded ORF1p.**
128. Leibold, D. M. *et al.* Translation of LINE-1 DNA elements *in vitro* and in human cells. *Proc. Natl Acad. Sci. USA* **87**, 6990–6994 (1990).
129. Ardeljan, D., Taylor, M. S., Ting, D. T. & Burns, K. H. The human long interspersed element-1 retrotransposon: an emerging biomarker of neoplasia. *Clin. Chem.* http://dx.doi.org/10.1373/clinchem.2016.257444 (2017).
130. Achanta, P. *et al.* Somatic retrotransposition is infrequent in glioblastomas. *Mob. DNA* **7**, 22 (2016).
131. Wylie, A. *et al.* p53 genes function to restrain mobile elements. *Genes Dev.* **30**, 64–77 (2016).
132. Rodic, N. *et al.* Retrotransposon insertions in the clonal evolution of pancreatic ductal adenocarcinoma. *Nat. Med.* **21**, 1060–1064 (2015). **Profile of LINE-1 insertion sites in primary and metastatic PDAC samples, inferring when and where new insertions were acquired.**
133. Goodier, J. L., Zhang, L., Vetter, M. R. & Kazazian, H. H. Jr. 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 (2007).
134. Sokolowski, M. *et al.* Development of a monoclonal antibody specific to the endonuclease domain of the human LINE-1 ORF2 protein. *Mob. DNA* **5**, 29 (2014).
135. De Luca, C. *et al.* Enhanced expression of LINE-1-encoded ORF2 protein in early stages of colon and prostate transformation. *Oncotarget* **7**, 4048–4061 (2016).
136. Denne, M. *et al.* Physical and functional interactions of human endogenous retrovirus proteins Np9 and rec with the promyelocytic leukemia zinc finger protein. *J. Virol.* **81**, 5607–5616 (2007).
137. Mangeney, M., de Parseval, N., Thomas, G. & Heidmann, T. The full-length envelope of an HERV-H human endogenous retrovirus has immunosuppressive properties. *J. Gen. Virol.* **82**, 2515–2518 (2001).
138. Miki, Y. *et al.* Disruption of the *APC* gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res.* **52**, 643–645 (1992). **Description of the discovery of a somatically acquired LINE-1 insertion in the *APC* gene that causes colon cancer.**
139. Fodde, R., Smits, R. & Clevers, H. APC, signal transduction and genetic instability in colorectal cancer. *Nat. Rev. Cancer* **1**, 55–67 (2001).
140. Groden, J. *et al.* Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* **66**, 589–600 (1991).
141. Joslyn, G. *et al.* Identification of deletion mutations and three new genes at the familial polyposis locus. *Cell* **66**, 601–613 (1991).
142. Nishisho, I. *et al.* Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science* **253**, 665–669 (1991).
143. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
144. Powell, S. M. *et al.* APC mutations occur early during colorectal tumorigenesis. *Nature* **359**, 235–237 (1992).
145. Cost, G. J., Feng, Q., Jacquier, A. & Boeke, J. D. Human L1 element target-primed reverse transcription *in vitro*. *EMBO J.* **21**, 5899–5910 (2002).
146. Jurka, J. Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. *Proc. Natl Acad. Sci. USA* **94**, 1872–1877 (1997).
147. Ostertag, E. M. & Kazazian, H. H. Jr. Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. *Genome Res.* **11**, 2059–2065 (2001).
148. Solyom, S. *et al.* Extensive somatic L1 retrotransposition in colorectal tumors. *Genome Res.* **22**, 2328–2338 (2012).
149. Doucet-O'Hare, T. T. *et al.* LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc. Natl Acad. Sci. USA* **112**, E4894–E4900 (2015).
150. Doucet-O'Hare, T. T. *et al.* Somatically acquired LINE-1 insertions in normal esophagus undergo clonal expansion in esophageal squamous cell carcinoma. *Hum. Mutat.* **37**, 942–954 (2016).
151. Ewing, A. D. *et al.* Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. *Genome Res.* **25**, 1536–1545 (2015).
152. Helman, E. *et al.* Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063 (2014).
153. Tang, Z. *et al.* Human transposon insertion profiling: analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. *Proc. Natl Acad. Sci. USA* **114**, E733–E740 (2017).
154. Carreira, P. E. *et al.* Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. *Mob. DNA* **7**, 21 (2016).
155. Holmes, S. E., Dombroski, B. A., Krebs, C. M., Boehm, C. D. & Kazazian, H. H. Jr. A new retrotransposable human L1 element from the LRE2 locus on chromosome 1q produces a chimaeric insertion. *Nat. Genet.* **7**, 143–148 (1994).
156. Moran, J. V., DeBerardinis, R. J. & Kazazian, H. H. Jr. Exon shuffling by L1 retrotransposition. *Science* **283**, 1530–1534 (1999).
157. Macfarlane, C. M. *et al.* Transduction-specific ATLAS reveals a cohort of highly active L1 retrotransposons in human populations. *Hum. Mutat.* **34**, 974–985 (2013).
158. Ewing, A. D. *et al.* Retrotransposition of gene transcripts leads to structural variation in mammalian genomes. *Genome Biol.* **14**, R22 (2013).
159. Cooke, S. L. *et al.* Processed pseudogenes acquired somatically during cancer development. *Nat. Commun.* **5**, 3644 (2014).
160. Mir, A. A., Philippe, C. & Cristofari, G. euL1db: the European database of L1HS retrotransposon insertions in humans. *Nucleic Acids Res.* **43**, D43–D47 (2015).
161. Wheelan, S. J., Aizawa, Y., Han, J. S. & Boeke, J. D. Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. *Genome Res.* **15**, 1073–1078 (2005).
162. Sorek, R., Ast, G. & Graur, D. Alu-containing exons are alternatively spliced. *Genome Res.* **12**, 1060–1067 (2002).
163. Lev-Maor, G. *et al.* Intronic Alus influence alternative splicing. *PLoS Genet.* **4**, e1000204 (2008).
164. Symer, D. E. *et al.* Human l1 retrotransposition is associated with genetic instability *in vivo*. *Cell* **110**, 327–338 (2002).
165. Gilbert, N., Lutz-Prigge, S. & Moran, J. V. Genomic deletions created upon LINE-1 retrotransposition. *Cell* **110**, 315–325 (2002).
166. Lehrman, M. A. *et al.* Mutation in LDL receptor: Alu–Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. *Science* **227**, 140–146 (1985).
167. Rahbari, R. & Badge, R. M. Combining amplification typing of L1 active subfamilies (ATLAS) with high-throughput sequencing. *Methods Mol. Biol.* **1400**, 95–106 (2016).
168. Streva, V. A. *et al.* Sequencing, identification and mapping of primed L1 elements (SIMPLE) reveals significant variation in full length L1 elements betweenindividuals. *BMC Genomics* **16**, 220 (2015).
169. Quentin, Y. Fusion of a free left Alu monomer and a free right Alu monomer at the origin of the Alu family in the primate genomes. *Nucleic Acids Res.* **20**, 487–493 (1992).

Competing interests statement

The author declares competing interests: see Web version for details.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

FURTHER INFORMATION

The Cancer Genome Atlas: http://cancergenome.nih.gov

ALL LINKS ARE ACTIVE IN THE ONLINE PDF