Abstract

Long Interspaced Nuclear Element 1 (LINE-1 or L1) is one of the most important elements in the human genome. They comprise approximately 17% of our genome and mounting evidence suggest S that LINE-1 elements are not stable junk DNA but are highly active in the human germline and developing tissue. Their activity contributes to the creation of changes and variation in WD the host genome through the insertion of new LINE-1 and the creation of double strand breaks in terms of inter-individual variation and somatic variation. The estimation of their transcriptional activity remains poorly understood because of their highly duplicate d nature and the effect of pervasive transcription. Here, we develop a new method to estimate the activity of LINE-1 subfamilies, TRANS our method can deconvolve autonomous transcription from the overall pervasive transcription signal showing that a small number of LINE-1 subfamilies are active in human tissues, in particular, L1Hs. Furthermore, we show that there is a great variability of L1Hs activity between different tissues. Surprisingly, there is less LINE-1 activity in the human adult brain while there is high activity in tissue such as the nerve, ? testis and skin. We finally show that LINE-1 is more active in certain cancer cells and, in particular, its activity is coupled to the creation of indels in cancer tissue.

Comment [FN1]: rewrite

MORE

Introduction

LINE-1 has attracted much attention in the last decade due to its capacity to create variations in the human genome. LINE-1 is a DNA sequence capable of duplicating itself by mobilizing its messenger RNA (mRNA) to new genomic locations via retrotransposition {Cost:2002ti, Kulpa:2006js, Ostertag:2001jl}; this process resulted in thousands of mostly inactive and truncated copies of LINE-1 across the human genome

{Lander:2001hk}. Although LINE-1 activity has been described in both healthy and pathogenic tissues {Ostertag:2001jl, Hancks:2012ij, Burns:2017jv}, LINE-1 activity is remarkably difficult to study due to their repetitive nature. LINE-1 activity was believed to occur only in germ cells {Wang:2006hr, Ewing:2010da, Sudmant:2015kz}, and cancer tissues {Skowronski:1985te, Belancio:2010df, Tubio:2014gm}. However, building evidence suggest that LINE-1 is active in somatic tissue {Muotri:2005go, Kano:2009dt, Belancio:2010ie, Evrony:2015it}

CONTRA

As opposed to healthy tissues, most human tumors and cell lines show a higher activity of LINE-1, likely due to broad demethylation of LINE-1 promoter {Ogino:2008ey, Igarashi:2010dz, Patnala:2014dy, Philippe:2016cx, Gainetdinov:2016fw}. LINE-1 is activity is constrained at many levels, however, little is known about their activation in somatic, tumors, and germinative cells. Therefore, the assessment of LINE-1 activity requires elaborate essays {Doucet:2016ke} or multiple and complementary datasets {Philippe:2016cx}, hindering estimation of LINE-1 activity on large scale datasets. Moreover, affordable methods to quantify LINE-1 activity, such as those based on RNA sequencing {Belancio:2010ie, Rangwala:2009bg, Criscione:2014dp}, are confounded by pervasive transcription and the highly duplicated nature of LINE-1.

Pervasive transcription refers to the idea that the majority of the genome is transcribed, beyond just the known genes {BUZFClark:2011cc}. However, how much pervasive transcription influences the human transcriptome remains uncertain {Jacquier:2009hz, Clark:2011cc, Lee:2015cw}. Some scientists suggest that pervasive transcription is mostly derived from technical and biological noise and, therefore, might not be relevant in RNA sequencing experiments {vanBakel:2010bt}. Others suggest that pervasive transcription has a stochastic nature, and if sequenced at enough depth, the majority of the genome may be transcribed. With this theory, pervasive transcription should not affect quantification of the transcription of protein-coding genes, which are present either as a single copy or low copy numbers in the genome. However, the quantification of the transcription of transposable elements, including LINE-1, would be especially affected by pervasive transcription due to the multi-copy nature of these genes.

The activation of LINE-1 can lead to the expression of its major enzyme – ORF2p. ORF2p is a reverse transcriptase and also contains an endonuclease domain {Piskareva:2006do}. The endonuclease domain in ORF2p has been shown to create double-strand breaks on DNA molecules {Gasior:2006dp}, which are then corrected by endogenous DNA repair mechanisms. Furthermore, LINE-1 activation has been linked to poor prognosis in several types of cancers {Ogino:2008ey}. Recently, researchers have leveraged large-scale sequencing projects to search for evidence of LINE-1 mobilizations in cancer samples. However, LINE-1 has been shown to rarely activate oncogenes or disrupt tumor suppressor genes {Lee:2012cv, Shukla:2013bl, Helman:2014if, Tubio:2014gm, Scott:2016jq} .

The present paper presents a new method to remove the effect of pervasive transcription on RNA sequencing datasets and reliably quantify LINE-1 subfamily transcription. We describe and validate the landscape of LINE-1 transcription in well

REP.
NEURO
+
CANCER

established human cell lines. Furthemore, we find the L1Hs transcription in healthy adult tissues is correlated with cell turnover. Morever, L1Hs transcription correlates with the number of small insertions and deletions (indels) in cancer cells.

Results

Recently amplified LINE-1 subfamilies, such as L1Hs, are discarded from traditional transcript quantification essays due to the insufficient mapping specificity of LINE-1 instances. Before addressing the LINE-1 multi-mappability issue, we quantified the number of reads overlapping LINE-1 subfamilies in thousands of RNA sequencing experiments from human cell lines and healthy primary tissue (Table 1) {GTExConsortium:2015fb}. Figure 1A shows that the average number of reads mapping to LINE-1 subfamilies is correlated with the number of bases annotated as the respective LINE-1 subfamily in the majority of RNA sequencing experiments (Spearman's rank correlation c=0.94, $p < 2.2e-16$). We observed that the correlation is driven by recently retrotransposed LINE-1 subfamilies and by ancient LINE-1 subfamilies. Particularly, reads map ten times more frequently to ancient LINE-1 subfamilies, such as L1ME1 and L1PA3, than recently expanded LINE-1 subfamilies. Since ancient L1 subfamilies have no evidence of recent activity in the human lineage, we hypothesized that this "genomic-transcriptomic" correlation should be indicative of pervasive transcription. In this model, the stochastic nature of RNA polymerase II transcription drives the creation of RNA fragments proportionally to the number of copies of LINE-1 subfamilies in the genome.

We then divided samples per tissues of origin (Figure 1B) and noticed that many tissues had a smaller genomic-transcriptomic correlation, hinting at another confounding signal creating reads overlapping LINE-1 subfamilies. We hypothesize that deviations from a high genomic-transcriptome correlation could be derived from autonomous transcription of the LINE-1 subfamilies (see Methods for details). Thus, we modeled the number of reads mapping to LINE-1 elements as the sum of signals collectively emanating from pervasive transcription and the autonomous transcription of LINE-1 subfamilies. We estimated the signal derived from pervasive transcription as previously described and the signal derived from autonomous transcription was calculated by simulating reads from LINE-1 subfamilies' transcripts. We developed a software platform, TeXP (available at https://github.com/gersteinlab/TeXP), that creates signatures for pervasive and autonomous transcription and deconvolves reads overlapping L1 elements into pervasive and autonomous transcription (Figure 1C).

Transcriptional activity of LINE-1 in human cell lines

As a first step, we benchmarked TeXP by estimating the autonomous transcription of LINE-1 subfamilies in well-established cell lines RNA sequencing experiments {ENCODEProjectConsortium:2012gc} (Table S1). Figure 2A shows the proportion of reads mapped to LINE-1 subfamilies using a naïve method (left panel) and signal proportions post TeXP processing (right panel) in three MCF-7 cell compartments (Cytoplasm, Nuclear and Whole Cell). As it can be seen in the naïve method panel (Figure 2A - left panel), Cytoplasm polyA+ and Whole Cell polyA+ have an enrichment of reads mapping to L1Hs and L1PA2 when compared to Whole Cell polyA- (light yellow) and Nuclear (green) RNA sequencing experiments. This enrichment of L1Hs reads is consistent with the transcription of full-length L1Hs transcripts (Figure S1). Accordingly, the panel displaying estimates after applying TeXP (Figure 2A - right panel) shows two major signals in MCF-7 RNA-seq experiments: pervasive transcription (gray) and L1Hs autonomous transcription (dark blue). This analysis suggests that reads mapped to ancient L1 subfamilies, such as L1PA3 and L1PA4, are mostly derived from pervasive transcription. Furthermore, TeXP only finds significant evidence of L1Hs autonomous transcription, despite L1PA2 also being detected but at a lower frequency (Figure 2A and Figure S2). Noticeably, L1Hs and L1PA2 are the only LINE-1 subfamilies known to be capable of mobilization in germinative tissues {Ovchinnikov:2002in, Sudmant:2015kz}.

MCF-7 is a cell line derived from breast cancer and was previously described as having remarkable high levels of L1Hs transcription {Philippe:2016cx, Belancio:2010ie}. In order to investigate the source of the L1Hs autonomous transcription, we analyzed RNA sequencing experiments from MCF-7 cell compartments and RNA fractions. First, we observed that, in agreement with the literature, whole cell polyA+ experiments yield extremely high levels of L1Hs transcriptions (180.7 RPKM). Selecting whole cell transcripts without polyadenilated tail (whole cell polyA-) reduces the signal of L1Hs autonomous transcription by 73%, suggesting that most of the signal is derived from mature poly-adenylated transcripts. Furthermore, we tested whether L1Hs transcripts are derived from cytoplasmic (mature) or nucleolar (pre-mRNA) portions of the cell; We find that nucleolar transcripts were vastly enriched for pervasive transcription (autonomous/pervasive ratio 0.02), while cytoplasmic transcripts have an autonomous/pervasive ratio similar to transcripts derived from whole cell polyA+ (0.45 and 0.51 respectively). Together, these results suggest that most of the LINE-1 autonomous transcription signal is derived from mature transcripts in the cytoplasm and only a small fraction of signal is derived from fragmented LINE-1 transcripts in the nucleus (Figure S4). Analyzing other cancer-derived cell lines such as SK-MEL-5 and K-562, yielded no evidence of L1Hs autonomous transcription in most cell compartments or RNA fractions (Figure 2B). However, we found smaller levels of L1Hs autonomous transcription in whole cell polyA+ samples (2.4 and 8.8 RPKM, respectively). This reference panel of cell lines were then used to validate TeXP L1Hs autonomous transcription estimations.

Validation of LINE-1 autonomous transcription

We used ddPCR to detect autonomous and pervasive transcriptions of L1Hs in four cell lines, MCF-7, K562, SK-MEL-5, and GM12878. We quantified the autonomous and pervasive transcription levels based on the assumption that expression on the 5' end of the L1Hs transcript was mostly derived from autonomous transcription, while expression on the 3' end is from pervasive transcription. We had initially designed and tested multiple assays targeting different regions of the L1Hs locus, but subsequently proceeded with two of the best performing assays (Table 2). The first one was located in ORF1 directly adjacent to the 5'UTR and represented the 5' end of the transcript. The second one was located in ORF2 about 1.5 kb upstream of the 3' UTR and represented the 3' end of the transcript. The same process was completed for ORF2 to find the copy

JAP

numbers of the truncated L1Hs transcripts (i.e., the transcripts missing the 5’ end of L1H) (Table 3). Since the autonomous transcription results in a full-length transcript of L1Hs, we quantified the pervasive transcription level by subtracting the 5’ end expression (ORF1) from the 3’ end expression (ORF2).

Figure 2D shows the relative quantification of L1Hs transcripts in these four cell lines using *HPRT1* 5’ end as reference. The ddPCR analysis detected 12,600 copies of the full-length transcripts/ng in MCF-7. In agreement with our *in-silico* result, K562 and SK-MEL-5, had 1,512 copies and 1,708 copies of the full-length transcript/ng respectively. For GM12878 cell line, we expected no autonomous expression of L1Hs however, our ddPCR assays detected low levels of autonomous transcription of L1Hs (Fig. 2B., Table 2) (GM12878 had 655 copies of full-length transcript/ng). Overall, the quantification of L1Hs autonomous transcription using ddPCR is highly correlated to the TeXP quantification (spearman correlation, 0.99, p value = 3.803e-06); suggesting that TeXP is able to remove most of the noise derived from pervasive transcription, however, it can be insensitive to samples with little LINE-1 autonomous transcription.

Landscape of LINE-1 subfamily transcription in healthy primary tissue and cells lines.

It has been long thought that LINE-1 instances are completely silenced in most somatic cells. The major mechanism responsible for this repression is thought to the methylation of LINE-1 promoter which would preclude the transcription of mature LINE-1 mRNAs. To test whether LINE-1 subfamilies are completely silenced in somatic tissue we analyzed 7,429 GTEx primary tissue samples (Table S2) and removed [N] samples from further analysis as they lacked sufficient reads of overlapping LINE-1 elements. Similar to the cancer derived cell lines we found that only L1Hs is autonomously transcribed, and conversely, L1P1, L1PA2, L1AP3, and L1PA4 only have residual or spurious autonomous transcription in healthy tissues (Figure S5). Overall, healthy tissues had a narrower range of L1Hs autonomous transcription levels when compared to cancer cell lines. Whereas the highest L1Hs autonomous transcription in healthy tissues was 46.66 RPKM (Figure 3; L1Hs RPKM histogram), the cancer cell lines reached 180 RPKM. By contrast, 2,520 (34.3%) GTEx RNA sequencing experiments from primary tissues had no or very little (<1 RPKM) evidence of L1Hs autonomous transcription. All together these results suggest that differently from expected, L1Hs is broadly transcribed in healthy somatic tissues, are polyadenylated and present in the cytoplasm. Other major post-transcriptional mechanism could play a central role in constraining L1Hs retrotransposition. For example, during its translation, transportation to the nucleous or in later steps during integration. Otherwise, we expect that LINE-1 could play a major role in creating diversity across intra-individual somatic cells.

We then compared the landscape of LINE-1 subfamily transcription in Epstein-Barr Virus (EBV)-immortalized cell lines and their primary tissue to understand the changes induced by immortalization. EBV immortalization induces drastic changes in the expression of cell cycle, apoptosis and alternative splicing pathways {Bolotin:2014kt, Caliskan:2011kx, Min:2010in}. Overall, we found that EBV-transformed cell lines derived from different tissues (lymphoblastic and fibroblastic) have distinct patterns of

L1Hs autonomous transcription; lymphoblast (blood-derived) cell lines have no or little autonomous transcription of L1Hs (Figure S6) with approximately 84% of samples having an estimated RPKM equal to zero, whereas fibroblastic (skin-derived) cell lines consistently have higher levels of L1Hs autonomous transcription (median 1.5 RPKM) with 58.7% of samples with an RPKM higher than 1. In general, EBV-immortalized cell lines reflect their tissue of origin. While most (74.6%) of the whole blood samples had no transcriptional activity of L1Hs, only one sample from skin had an L1Hs autonomous transcription level of smaller than 1 RPKM. We further selected patients with both primary and EBV-transformed cell lines to assess whether the EBV transformation could change L1Hs autonomous transcription. We find that both skin and lymphocytes have a drastic down-regulation of L1Hs autonomous transcription (Figure S11). This finding suggests that the EBV-transformed cell lines partially preserve the L1Hs transcription level from their tissue of origin and might also explain why fibroblast-derived induced pluripotent stem cells (iPSCs) support higher levels of LINE-1 retrotransposition {Klawitter:2016ff}.

Human tissues show remarkable variability of L1Hs autonomous transcription. We found that L1Hs autonomous transcription is inversely correlated to time cells to divide (cell turnover rate - spearman correlation: rho=-0.7551126; p-value = 0.01865). Tissues suggested to have low cell turnover, such as the human brain {Spalding:2005fa}, are amongst the tissues with the lowest levels of L1Hs autonomous transcription (Figure 3). In particular, the human cerebellum, which samples are likely to have strong repression of L1Hs autonomous transcription. This result is in apparent opposition to the literature that suggests that the human brain supports high levels of somatic LINE-1 retrotransposition, however, most of these studies are based on neuro-precursors which corresponds to the early development stage of the human brain {Thomas:2012km, Muotri:2010go, Muotri:2005go, Coufal:2009kb}. Conversely, brain samples extracted from the striatum, putamen and caudate – all regions associated with the basal ganglia; have higher levels of L1Hs autonomous transcription compared to the other brain regions (T-test basal ganglia vs. all other brain tissues, t = -7.0943; p value = 9.867e-12); importantly, these levels were still low compared to other tissues. Other tissues with low cell turnover rate such as liver, pancreas, and spleen samples also show very little or no autonomous transcription of L1Hs (91.2%, 82.9%, 88.9% of samples, respectively, had an RPKM < 1). On the other hand, germinative tissues have been proposed to support somatic activity of L1Hs elements {Iskow:2010gh}. In fact, our results (Figure 3) suggest that this trend is more general, and most tissues associated with the reproductive system sustain high levels of L1Hs autonomous transcription. In addition to the reproductive system association, we found that the tissues with highest L1Hs autonomous transcriptions are also enriched for high cell turnover. The nerve (tibia), skin (both exposed and not exposed to the sun), prostate, lung, vagina (Figure 3) are the five tissues with highest level of L1Hs transcription.

Comment [FN6]: Many of these tissues have never been investigated regarding the activity of L1Hs and none of the analyzed tissues were investigated in the scale shown here. In agreement with previous works that used Northern blot to quantify full-length LINE-1 transcripts {Belancio:2010ie}, we found that esophagus, prostate, stomach, and heart muscle supported high-transcriptional activity (Figure 3).

Previous research have suggested that LINE-1 activity could be correlated with an individual’s age {Cho:2015bx, VanMeter:2014gs} {Bjornsson:2008fs}; Specifically, as individuals age, methylation marks in LINE-1 promoters might be lost and LINE-1 are derrepressed. Having estimated the transcription level of L1Hs and having access to the phenotypes of the GTEx samples, we tested whether the autonomous transcription of L1Hs correlates with sample age or body mass index (BMI). In most tissues, we did not observe significant correlations with subject age, most likely due to low levels of L1Hs autonomous transcription (Figure 3). However, we did observe significant positive correlations ranging from 0.17 to 0.28 with the samples' age in lung, skeletal muscle, fibroblast cell lines, adipose tissue, skin, breast, and testis, (Figure 3, red triangles; Table S3). Intriguingly, contrary to the expectation of higher L1Hs transcriptional activity in older individuals, we found that prostate and whole blood samples show an inverse correlation with age; prostate samples had the highest L1Hs transcriptional activity in 20-30 years old individuals. Other tissues with relatively high autonomous transcription of LINE-1 showed no correlation (e.g., tibial nerve and ovary).

Comment [FN9]: BMI was recently reported to be inversely correlated with the methylation of LINE-1 elements{MarquesRocha:2016iw}; however, we only found a correlation between L1Hs transcriptional activity and BMI in breast tissue (corr=0.23, FDR=0.046; Figure 3, blue circles; Table S4). Finally, we tested if samples of skin exposed to the sun showed any significant enrichment of L1Hs autonomous transcription compared to skin not exposed to sun. We found that both groups of samples (exposed and not exposed) had similarly high levels of L1Hs autonomous transcription, with slightly (but not significantly) higher L1Hs activity in samples of tissue exposed to the sun.

Based on the signal emanating from pervasive transcription, we next estimated a pervasive transcription index for each RNA sequencing experiment. We defined the PI as the number of reads with overlapping LINE-1 subfamilies and emanating from pervasive transcription, normalized by the total number of aligned reads in an RNA sequencing experiment. Overall, we found that testis and cerebellum were amongst the tissues with the highest pervasive transcription level (median 1,056 and 906.3 pervasive transcription index, respectively). Conversely, whole blood and skeletal muscle were amongst the tissues with the lowest levels of pervasive transcription (134.9 and 223.8 PI, respectively) (Figure S7). Interestingly, tissues with smaller pervasive transcription index have been shown to have low transcriptional diversity {GTExConsortium:2015fb}, suggesting that the pervasive transcription index might be a good proxy for tissue transcription diversity.

Activity of LINE-1 elements in human cancer

Finally, we investigated the impact of LINE-1 autonomous transcription in cancer samples. We hypothesized that tissues with a higher transcription of LINE-1 elements in a healthy context would be more susceptible to L1Hs activity and consequent genomic instability mediated by LINE-1 reverse transcriptase. We investigated the autonomous transcription level of L1Hs from over 2,500 cancer samples originating from six tumor types: lung adenocarcinoma, lung squamous cell carcinoma (LUSC), prostate adenocarcinoma, brain lower grade glioma, thyroid carcinoma, and skin cutaneous melanoma (SKCM). We found that SKCM tissue supported autonomous L1Hs transcription at levels slightly lower (2.38x) than healthy tissue. By contrast, tumors derived from lung consistently had higher levels of L1Hs autonomous transcription in their cancer counterparts, reaching up to 13x higher expression in LUSC (Figure S8).

We hypothesized that these genomes would have consistently higher genomic instability due to the activity of L1Hs endonuclease. Ideally, one would use somatic LINE-1 insertions or chromosomal rearrangements in order to assess the activity of LINE-1, however, these analysis demand large scale Whole Genome Sequencing structural variation calling. Therefore, to test this hypothesis, we assessed the frequency of indels in the exome of our samples. In total, we analyzed somatic indels from 2,504 tumors. We selected lung, skin, thyroid, and prostate samples from the

Cancer Genome Atlas to search for signatures originating from L1Hs endonuclease activity. We first compared the correlation between exonic indels and the autonomous transcription of L1Hs. While not all tissues had a significant correlation between autonomous LINE-1 transcription and the number of indels (Figure 4A), all samples had a significantly high correlation (0.49, p value < 2.2e-16). To further assess the causation of these two variables we focused signatures created by LINE-1 endonuclease. Namely, we investigated the occurrence of indels close to the motif recognized by LINE-1 endonuclease. L1Hs endonuclease creates double-strand break points in TTT|AA loci {Feng:1996we, Gasior:2006dp}. We hypothesized that the double-strand breaks created by L1Hs are corrected by endogenous double-strand break correction mechanisms such as the non-homologous end joining (NHEJ) pathway {ODriscoll:2006cz}. The NHEJ pathway is known to be error-prone, especially in the tumoral context, creating small indels as well as large duplications, deletions, and transversions {Onozawa:2014cv}. We tested whether the LINE-1 endonuclease target motif (TTTAA) was enriched in sequences flanking indels and found that regardless of the tissue of origin, there was an enrichment of the motif TTTAA in the 50 nucleotides (nts) flanking the indel. We further select motifs closer to the indel coordinate (-3;+3 nt) and found that the effect was even more pronounced (Figure 4B). Finally, we evaluated the distribution of the endonuclease target motif relative to the position of the detected indel. We found that most TTTAA motifs were concentrated around position 0 or 1, meaning that they perfectly overlapped the break point of indels for both insertions (Figure 4C) and deletions (Figure 4D). Together, these results suggest that LINE-1 could lead to the creation of indels in somatic cells. We propose a model in which autonomously active

LINE-1 instances are transcribed in somatic cells. These polyadenylated transcripts follow the expected life cycle of LINE-1. ORF1p and ORF2p proteins are translated and associate with their mRNA, creating a ribonucleoprotein particle complex that is imported back to the nucleus. In the nucleus, the endonuclease domain targets TTTAA motifs on nuclear DNA and creates double-strand breaks. Instead of initiating the reserve transcription of the LINE-1 mRNA, the endonuclease aborts the insertion and dissociates from the DNA molecule. Endogenous mechanisms detect and correct double-strand breaks using error-prone NHEJ creating small indels close to the target site (Figure 5).

Methods

Tumor and Normal exon sequencing, INDEL and RNA sequencing data.

Exonic data and INDEL calling were obtained from the Genomic Data Center data portal (https://gdc-portal.nci.nih.gov). RNA-seq raw files were downloaded from the legacy archive (https://gdc-portal.nci.nih.gov/legacy-archive).

GTEx raw RNA sequencing data.

Raw RNA sequencing datasets from healthy tissues were obtained from Database of Genotypes and Phenotypes (DB-Gap - https://dbgap.ncbi.nlm.nih.gov) accession number phs000424.v6.p1.

ENCODE raw RNA sequencing data.

Raw RNA sequencing data from cancer cell lines were obtained from the ENCODE data portal (https://www.encodeproject.org/search). We selected RNA-seq experiments from immortalized cell lines with multiple cellular fractions and transcripts selection experiments. Accessions and cell lines are available in TableS1.

TeXP model.

TeXP models the number of reads overlapping L1 elements as the composition of signals deriving from pervasive transcription and full-length L1 autonomous transcripts from distinct L1 subfamilies.

For example, the number of reads overlapping is L1Hs instances is described by the Equation 1:

$$O_{L1Hs} = T{*}G_{L1Hs}{*}\epsilon_{pervasive} + T{*}M_{L1Hs,L1Hs}{*}\varepsilon_{L1Hs} + T{*}M_{L1Hs,L1PA2}{*}\varepsilon_{L1PA2} + \cdots + T{*}M_{L1Hs,j}{*}\varepsilon_{j}$$

Where O_{L1Hs} is the observed number of reads mapping to L1Hs, T is the total number of reads mapped to L1 instances, G_{L1Hs} defines the proportion of L1 bases in the genome annotated as L1Hs, $\epsilon_{pervasive}$ is the percentage of reads emanating from pervasive transcription, M is the mappability fingerprint (defined bellow) that describes what is the proportion of reads emanating from the signal $j \in \{L1Hs, L1P1, L1PA2, L1PA3, L1PA4\}$ that maps to L1 subfamily $i \in \{L1Hs, L1P1, L1PA2, L1PA3, L1PA4\}$ and ε is the percentage of reads emanating from the L1 Subfamily j. This model can be further generalized as the **Equation 2**:

$$O_i = T(G_i\epsilon_{pervasive} + M_{i,j}\varepsilon_j)$$

The number of reads mapped to each subfamily O_i is measured by analyzing paired-end or single-end RNA sequencing experiments independently. TeXP extracts basic information from fastq raw files such as read length and quality encoding. Fastq files are filtered to remove homopolymer reads and low quality reads using in-house scripts and FASTX suite (http://hannonlab.cshl.edu/fastx_toolkit/). Reads are mapped to the reference genome (hg38) using bowtie2 (parameters: --sensitive-local -N1 --no-unal). Multiple mapping reads are assigned to one of the best alignments. Reads overlapping L1 elements from Repeat Masker annotation of hg38 are extracted and counted per subfamily. The total number of reads T is defined as $T = \sum_i O_i$.

Pervasive transcription and mappability fingerprints of L1 subfamily transcripts.

Pervasive transcription is defined as the transcription of regions well beyond the boundaries of known genes {BUZFClark:2011cc}. We rationalized that the signal emanating from pervasive transcription would correlate to the number of bases annotated as each subfamily in the reference genome (hg38). We used Repeat Masker to count the number of instances and number of bases in hg38 annotated as the subfamily $i \in \{L1Hs, L1PA2, L1PA3, L1PA4, L1P1\}$. We define P_i as the proportion of bases annotated as the subfamily i in the Equation 3:

$$P_i = \frac{B_i}{\sum_j B_j}, j \in \{L1Hs, L1PA2, L1PA3, L1PA4, L1P1\}$$

Mappability fingerprints are created by aligning simulated reads deriving from putative L1 transcripts from each L1 subfamily and the expected signal from pervasive transcription. For each L1 subfamily, we extract the sequences of instances based on RepeatMasker annotation and the reference genome (hg38). Read from putative transcripts are generated using wgsim (https://github.com/lh3/wgsim - parameters: -1 [RNA-seq mean read length] –N 100000 -d0 –r0.1 -e 0). One hundred simulations are performed and reads are aligned to the human reference genome (hg38) using the same parameters described in the model session. The three-dimensional count matrix C is defined as the number of reads mapped to the subfamily $i \in \{L1Hs, L1PA2, L1PA3, L1PA4, L1P1\}$ emanating from the set of full-length transcripts $j \in \{L1Hs, L1PA2, L1PA3, L1PA4, L1P1\}$ in the simulation k. The matrix M is defined as the median percentage of counts across all simulations as in Equation 4:

$$M_{i.j} = median_{k \in \{1,2,..,100\}} \left(\frac{C_{i,j,k}}{\sum_{f \in \{L1Hs,L1PA2,L1PA3,L1PA4,L1P1\}} C_{i,f,k}} \right)$$

We tested whether different aligners yield different mappability fingerprints. BWA, STAR, and bowtie2 yielded very similar results (Figure S9). As L1 transcripts are not spliced, we decided to integrate bowtie2 as the main TeXP aligner. We further tested the effect of read length on L1Hs subfamily mappability fingerprints (Figure S10). To counter the effects of distinct read lengths TeXP constructs L1 mappability fingerprints libraries. If the read length used by the user is not available, TeXP creates it on the fly and include it to the L1 mappability fingerprint library.

We simulated reads emanating from their respective L1 subfamily transcripts and aligned these reads to the human reference genome creating a mappability fingerprint for each L1 subfamily (Figure S1). When we analyzed the L1 subfamily mappability fingerprints we observed that younger L1 subfamilies tend to have more reads mapped to other L1 subfamilies. For example, we find that only approximately 25% of reads from L1Hs (the most recent – and supposedly active L1) maps back to loci annotated as L1Hs. While older subfamilies such as L1PA4, have a higher proportion of reads mapping back to its instances (~70% - Figure S1).

The hidden variables $\boldsymbol{\varepsilon}$ and $\boldsymbol{\epsilon}$

By using O_i, T, the vector P_i, the mappability fingerprint matrix $M_{i.j}$ is generated for each RNA sequencing experiment we estimate the signal proportion $\boldsymbol{\varepsilon}$ and $\boldsymbol{\epsilon}$ in **Equation 2** by solving a linear regression. We used lasso regression (L1 regression) to maintain sparsity. We used the R package penalized ({Goeman:2010db} - parameters: unpenalized=~0, lambda2=0, positive=TRUE, standardize=TRUE, plot=FALSE, minsteps=10000, maxiter=1000).

TeXP

TeXP was developed as a combination of bash, R and python scripts. The source code is available at https://github.com/fabiocpn/TeXP. A docker image is also available for users at dockerhub under fnavarro/texp. \

TeXP consistency

To test whether the TeXP LINE-1 subfamily quantification is consistent across distinct RNA sequencing experiments we used GTEx RNA sequencing of the K-562 transcriptome. GTEx resequenced K562 RNA sequencing libraries for 102 sequencing batches. K-562 samples showed remarkable consistency across different GTEx batches, with median RPKM at 12.14 (1.47 RPKM standard deviation – Figure S6).

L1 endonuclease motif enrichment analysis

The exonic indels were extracted from GDC. For small insertions, we extracted 50 nucleotides flanking the small insertion coordinate. For small deletions, we extracted 50 nucleotides flanking the small deletion and the deleted sequence. We counted the number L1-endonuclease recognition motif (TTTAA) close of indels. We used three different flanking regions threshold: 50nt (as extracted), 10nt and 3nt. All strategies yielded similar results and only the 5nt analysis is shown here. Using Agilent capture was used to define the exonic regions. The same number of indels for each cancer type was simulated across the exonic (as defined above) and we estimated the expected number INDELs close to the indel breakpoint by counting the number of simulated indels close to the TTTAA motif. The statistical significance of the enrichment of TTTAA motif was calculated using the chi-squared test.

Passive versus Autonomous transcription of L1Hs transcritps.

More ancient elements such as DNA transposons and LINE-2 have been shown to be primarily transcribed passively, hitchhiking the transcription of nearby autonomously transcribed regions {GTExConsortium:2015fb}. Therefore, we tested whether our estimation of L1Hs transcription level correlated with genes containing or adjacent to L1Hs instances. We found no significant difference between the correlation distribution of a random set of genes and genes with L1Hs in exons or introns or within 3kb upstream or 3kb downstream of L1Hs. This finding indicates that our estimation of L1Hs autonomous transcription is not significantly influenced by non-autonomous L1Hs transcription adjacent or contained by protein-coding genes’ loci.

Cell Culture and Culture Conditions

All the cell lines used in this study were obtained from the American Type Culture Collection (ATCC) (Manassas, VA, USA). MCF-7 cells were cultured in Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 (DMEM/F12; Gibco). HeLa, SK-MEL-5, and HepG2 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM; Gibco). K562 and GM12878 cells were cultured in RPMI 1640 (Gibco). All cell culture media were supplemented with 10% fetal bovine serum (FBS) (Atlanta Biologics) and 1% penicillin/streptomycin (Fisher Scientific). All cells were cultured and expanded using the standard methods.

RNA Extraction and cDNA Synthesis

RNA was extracted using the RNeasy PLUS Mini Kit and the QIAshredders (Qiagen) following the manufacturer's protocol. All samples were treated with DNase I (New England BioLabs Inc.) to remove any remaining genomic DNA. RNA concentration was determined by Qubit 2.0 Fluorometer (Invitrogen). RNA quality was determined by Nanodrop (Thermo Scientific) and 2100 BioAnalyzer with the Agilent RNA 6000 Nano kit (Agilent Technologies). Approximately 5 µg of RNA was used for synthesis of the cDNA using the iScript Advanced cDNA Synthesis Kit (Bio-Rad). The final cDNA product was quantified and a working solution of 10 ng/µL was prepared for the subsequent studies.

Droplet Digital PCR (ddPCR)

Droplet Digital PCR (ddPCR) System (Bio-Rad Laboratories) was utilized to quantify the L1H transcript expression in the cell lines described above. Since L1H is a highly repetitive and heterogeneous target, we had initially designed and tested a panel of primers and probes that targeted the 5' untranslated region (5'UTR), the open reading frame 1 (ORF1), the open reading frame 2 (ORF2), and the 3' untranslated region (3'UTR) of the L1H locus, respectively. After a pilot screening study, we selected the two assays covering ORF1 and ORF2, which not only exhibited overall better performance, but also could help us to distinguish autonomous and pervasive L1H transcriptions. We also designed two reference assays on the housekeeping gene *HPRT1*, which targeted the 5' and 3' ends of the transcript, respectively (Table 1). All the ddPCR primers and probes were designed based on the human genome reference hg19 (GRCh37) and synthesized by IDT (Integrated DNA Technologies, Inc. Coralville, Iowa, USA).

The ddPCR reactions were performed according to the protocol provided by the manufacturer. Briefly, 10ng DNA template was mixed with the PCR Mastermix, primers, and probes to a final volume of 20 µL, followed by mixing with 60 µL of droplet generation oil to generate the droplet by the Bio-Rad QX200 Droplet Generator. After the droplets were generated, they were transferred into a 96-well PCR plate and then heat-sealed with a foil seal. PCR amplification was performed using a C1000 Touch thermal cycler and once completed, the 96-well PCR plate was loaded on the QX200 Droplet Reader. All ddPCR assays performed in this study included two normal human controls (NA12878 and NA10851) and two mouse controls (NSG and XFED/X3T3) as well as a no-template control (NTC, no DNA template). All samples and controls were run in duplicates. Data was analyzed utilizing the QuantaSoft™ analysis software provided by the manufacturer (Bio-Rad). Data were presented in copies of transcript/µL format which was mathematically normalized to copies of transcript/ng to allow for comparison between cell lines.

Reference house-keeping gene (**HPRT1**)

We designed two assays targeting the 5' and 3' ends of the *HPRT1* transcript, respectively, and used as the reference controls in this study (Table 3). The reference gene expression level was found to be constant within each cell line, but varied between cell lines. In addition, while 4 of the 6 cell lines had similar 5' and 3' end expression, K562 and GM12878 both had increased 3' end expression. This could be from different isoforms being expressed with different frequencies[3]. For the 5' end expression of *HPRT*, SK-MEL-5, GM12878, and HepG2 were all around 600 copies of transcript/ng. The remaining were all around 1200 copies of transcript/ng. When looking at the 3' end expression, we found that SK-MEL-5 and HepG2 were around 750 copies of transcript/ng, while MCF-7, GM12878, and HeLa were around 1350 copies of transcript/ng, and K562 was close to 1800 copies of transcript/ng. The slight difference between the 5' end and the 3' end expression levels in the same cell line could be explained by a potential 3' end bias in the cDNA synthesis. However, all the reference assays were consistent between experiments and did not affect the target expression.

References

Table 2. Primer and probe sequences for L1H target regions and *HPRT1* reference regions

	Assay Name	Sequence (5' → 3')
FAM Label	L1H ORF1 FWD	ACAAAGCTGGATGGAGAATG

	L1H ORF1 REV	GTTTGAATGTCCTCCCGTAG
	L1H ORF1 Probe	ACGAGCTGAGAGAAGAAGGCT
	L1H ORF2 FWD	AAATACCATTTGACCCAGCC
	L1H ORF2 REV	ATACGTGTGCATGTGTCTTT
	L1H ORF2 Probe	TCCCATTACTGGGTATATACCCA
HEX Labelled	*HPRT1* 5' End FWD	ACCAGGTTATGACCTTGATTT
	HPRT1 5' End REV	TCCATGAGGAATAAACACCC
	HPRT1 5' End Probe	TGCATACCTAATCATTATGCTGAGGA
	HPRT1 3' End FWD	CCAGACAAGTTTGTTGTAGGA
	HPRT1 3' End REV	CCAGTTTCACTAATGACACAAA
	HPRT1 3' End Probe	CCCTTGACTATAATGAATACTTCAGGG

Table 3. Quantification of L1H transcripts. Comparison of the expression of the copies of full-length transcript/ng of L1H autonomous transcript (ORF1) when run with both references and copies of truncated transcript/ng of L1H pervasive transcript (ORF2) when run with both references

	Referenc e	**MFC-7**	**K562**	**SK-MEL-5**	**GM12878**	**HeLa**	**HepG2**
ORF1-Autonomous Transcription (copies of full-length transcript/ng)	*HPRT1* 5' End	12600	1512	1708	655	696	964
	HPRT1 3' End	14050	1604	1810	735	709	1028
ORF2-Pervasive Transcription (copies of truncated transcript/ng)	*HPRT1* 5' End	4460	2838	3562	2855	4004	3916
	HPRT1 3' End	3370	3136	3720	2975	4381	4482

Figures

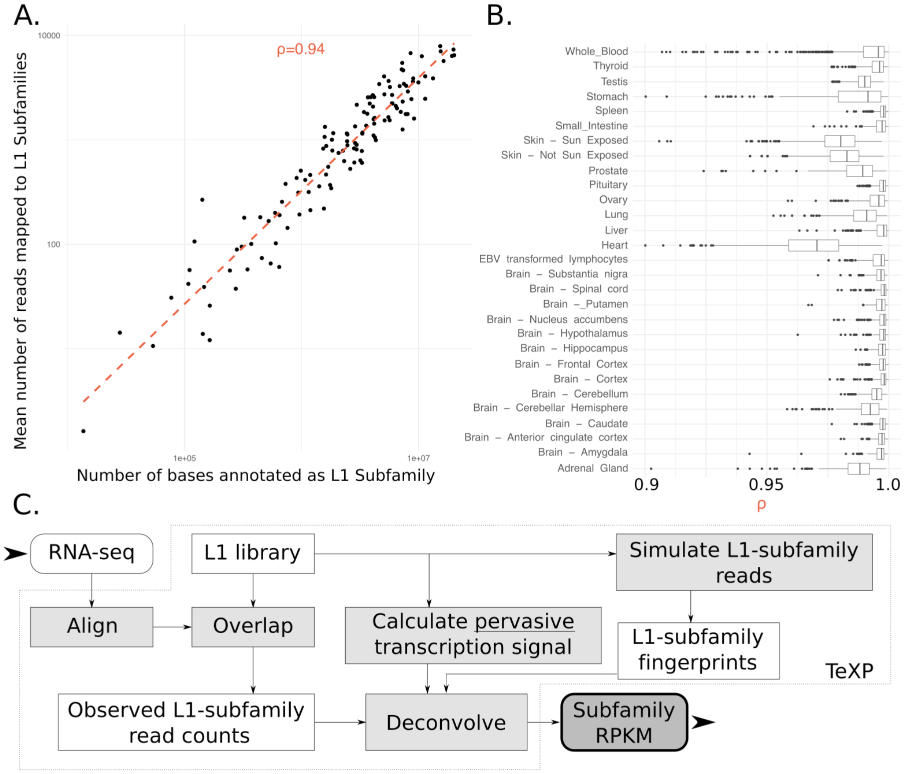
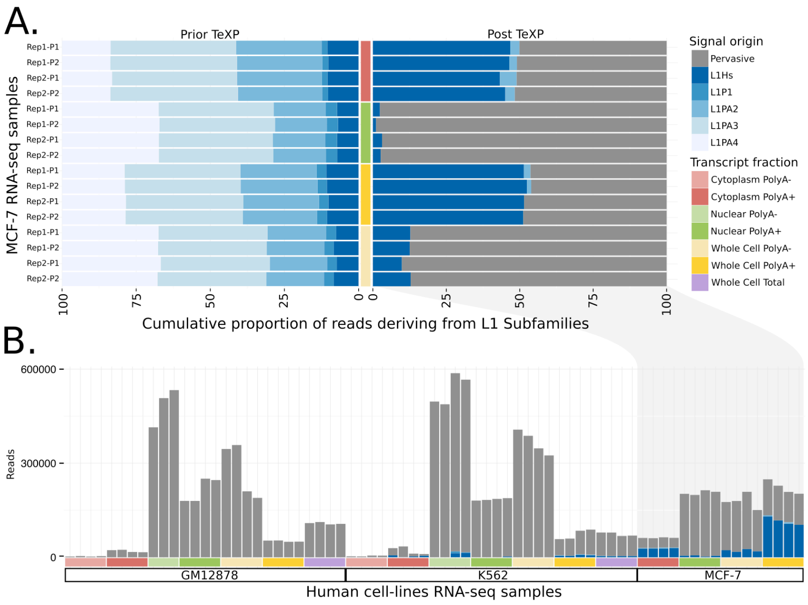


Figure 1. (A) The number of reads mapped to LINE-1 subfamilies is proportional to the number of bases annotated as the subfamily for most RNA sequencing experiments. (B) Healthy human tissues show varied distributions of the genomic-transcriptomic correlation. (C) TeXP pipeline description.

Comment [S 10]: Can you give an overall summary sentence for Figure 1 before getting into the subparts?


A.
Prior TeXP
Post TeXP
MCF-7 RNA-seq samples
Rep1-P1
Rep1-P2
Rep2-P1
Rep2-P2
100
75
50
25
0
Cumulative proportion of reads deriving from L1 Subfamilies
Signal origin
Pervasive
L1Hs
L1P1
L1PA2
L1PA3
L1PA4
Transcript fraction
Cytoplasm PolyA-
Cytoplasm PolyA+
Nuclear PolyA-
Nuclear PolyA+
Whole Cell PolyA-
Whole Cell PolyA+
Whole Cell Total
B.
600000
300000
0
Reads
GM12878
K562
MCF-7
Human cell-lines RNA-seq samples

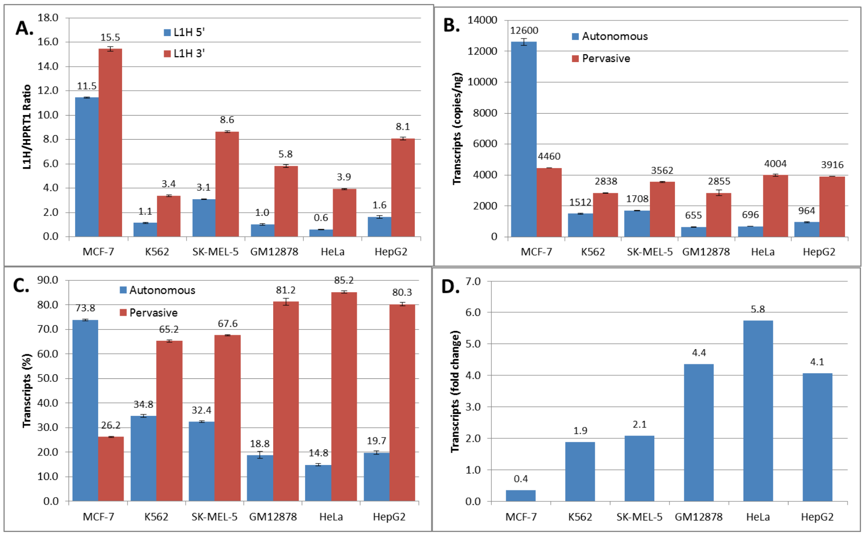

A.
L1H 5'
L1H 3'
L1H/HPRT1 Ratio
18.0
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.0
11.5
15.5
1.1
3.4
3.1
8.6
1.0
5.8
0.6
3.9
1.6
8.1
MCF-7
K562
SK-MEL-5
GM12878
HeLa
HepG2
B.
Autonomous
Pervasive
Transcripts (copies/ng)
14000
12000
10000
8000
6000
4000
2000
0
12600
4460
1512
2838
1708
3562
655
2855
696
4004
964
3916
MCF-7
K562
SK-MEL-5
GM12878
HeLa
HepG2
C.
Autonomous
Pervasive
Transcripts (%)
90.0
80.0
70.0
60.0
50.0
40.0
30.0
20.0
10.0
0.0
73.8
26.2
34.8
65.2
32.4
67.6
18.8
81.2
14.8
85.2
19.7
80.3
MCF-7
K562
SK-MEL-5
GM12878
HeLa
HepG2
D.
Transcripts (fold change)
7.0
6.0
5.0
4.0
3.0
2.0
1.0
0.0
0.4
1.9
2.1
4.4
5.8
4.1
MCF-7
K562
SK-MEL-5
GM12878
HeLa
HepG2

Figure 2. (A) The proportion of reads emanating from pervasive transcription and L1P1, L1PA2, L1PA3, L1PA4, and L1Hs subfamilies in MCF-7 RNA sequencing experiments are shown from the different cell compartments and transcript fractions prior to (left) and after (right) TeXP processing. (B) The absolute number of reads emanating from pervasive transcription and LINE-1 subfamilies are shown across the distinct cell and transcript fractions of the human-derived cell lines GM12878, K-562, and MCF7. Quantification of autonomous and pervasive transcripts of L1H in the cell lines using ddPCR. (**C)** Ratio of L1H 5' and 3' transcripts showing the enrichment of the 3' end of L1H for all cell lines. (**D)** Absolute quantification of autonomous and pervasive transcripts showing higher expression of pervasive transcripts compared to autonomous in all cell lines except MCF-7. (**E)** Percentage of autonomous and pervasive transcription showing a higher expression of pervasive transcripts compared to autonomous in all cell lines except MCF-7. (**F)** Fold change between autonomous and pervasive transcription. Fold changes above 1.0 indicates higher pervasive transcription. Fold changes below 1.0 indicates higher autonomous transcription. The data were run against *HPRT1* 5' end reference. All data were run in duplicate. All errors bars are mean ± SEM. These data represent two independent experiments.

Comment [S 11]: Can you give an overall summary sentence for Figure 2 before getting into the subparts?

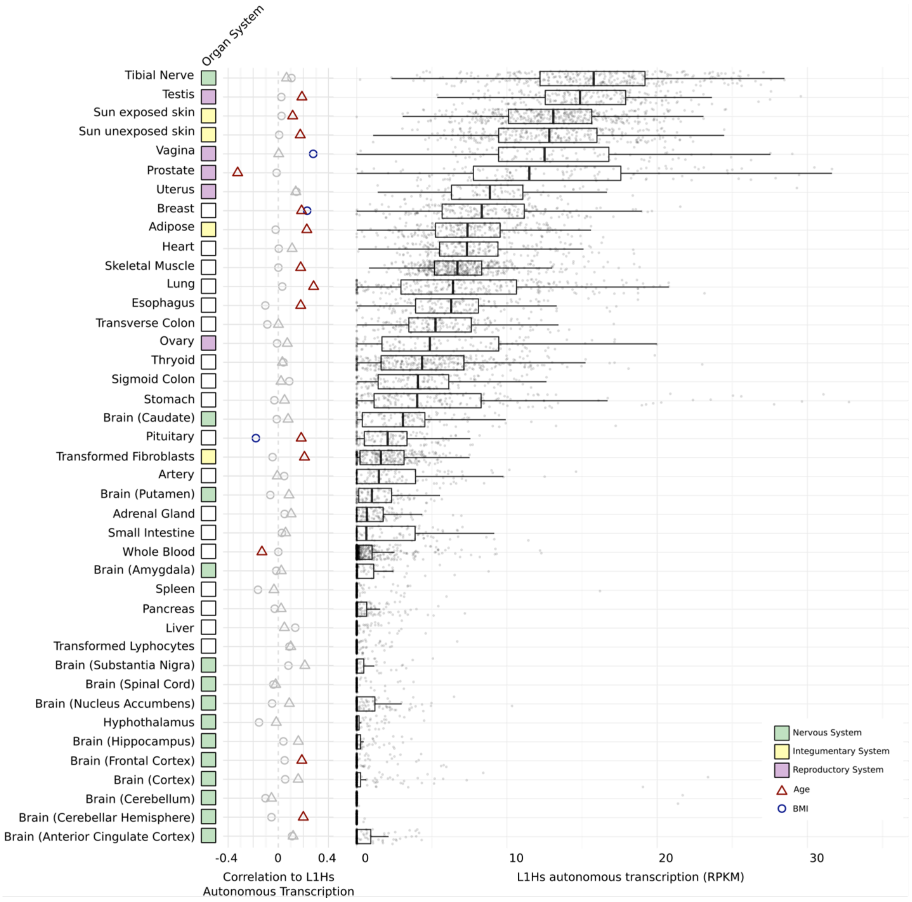


Figure 3. L1Hs autonomous transcription level on human healthy primary tissues. The left panel describes the correlation between L1Hs autonomous transcription and the subject's age (triangles) and BMI (circles). Significant correlations are colored. The right panel describes the panorama of L1Hs autonomous transcription on different tissues. Each point is an RNA sequencing experiment, separated by tissue of origin.

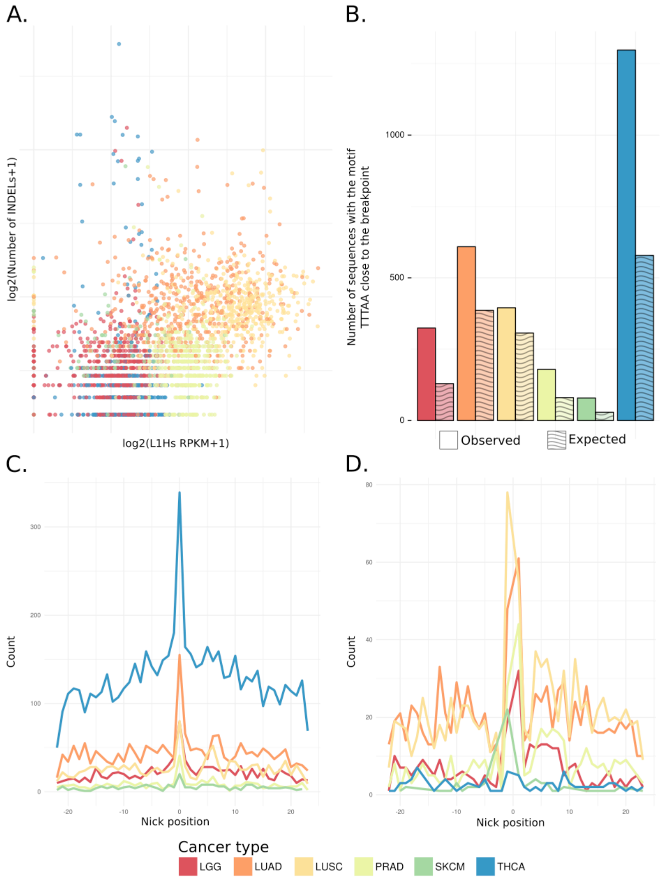


Figure 4. (A) The correlation between L1Hs autonomous expression and the number of indels in tumor samples is shown. (B) An overrepresentation of the TTTAA motif close to (-3|+3nt) indels (dark) is shown compared to null (light). (C) An overrepresentation of the TTT|AA in the indel break point on small insertions is shown. (D) An overrepresentation of the TTT|AA in the indel break point on small deletions is shown.

Comment [S 12]: Can you give an overall summary sentence for Figure 4 before getting into the subparts?

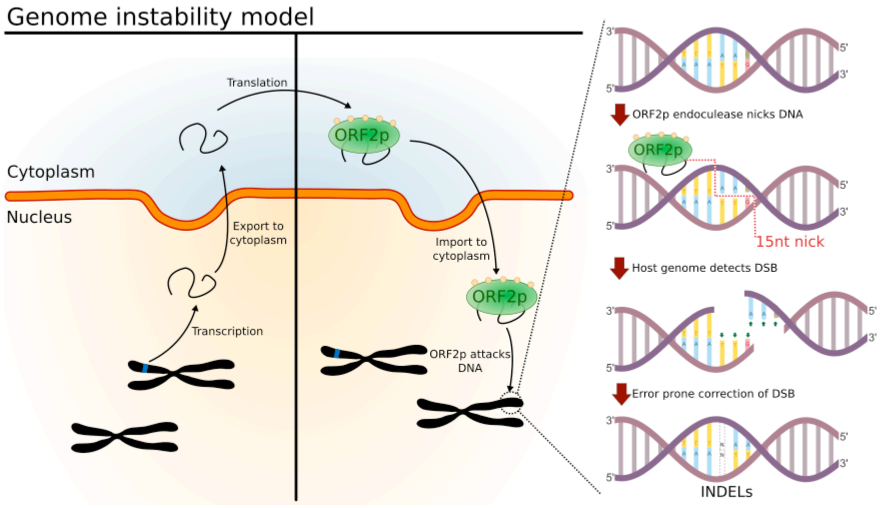


Figure 5. Model for LINE-1 favoring genome instability.