Comprehensive survey of LINE-1 transcriptional activity in human cell lines, healthy somatic tissue, and tumors

Fabio CP Navarro 1,2; Jacob Hoops 1,2; Lauren Bellfy 4; Eliza Cerveira 4; Qihui Zhu 4; Chengsheng Zhang 4; Charles Lee 4; Mark B. Gerstein 1,2,3

1 Program in Computational Biology and Bioinformatics, 2 Department of Molecular Biophysics and Biochemistry, and 3 Department of Computer Science, Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520 (Mark.Gerstein (at) Yale.edu); 4 The Jackson Laboratory for Genomic Medicine, Farmington, CT.

Abstract

Long interspersed nuclear element 1 (LINE-1) is a main source of variation in humans and other mammals. However, LINE-1 activity is difficult to study because of its highly repetitive nature and the effects of pervasive transcription. We developed and validated a method to gauge LINE-1 transcriptional activity accurately by taking into account pervasive transcription. We evaluated the individual cell-line compartments and found that most L1 transcription signal derives from the cytoplasm. This method also allowed us to perform comprehensive, uniform, and unbiased measurements of LINE-1 activity across healthy somatic cells, and tumor cells. Previously, LINE-1 was shown to be active in human germline and tumor cells but not in healthy somatic tissue, with the exception of some activity in the human brain. In contrast, we found that LINE-1 activity was limited in the central nervous system, but present in some normal somatic cells and tumor cells. Interestingly, the amount of LINE-1 activity was associated with the amount of cell turnover and, in tumor cells, with the amount of genomic instability. Our results suggest a mechanism in which LINE-1 activity gives rise to insertions and deletions overlapping LINE-1 target sites, potentially contributing to the mutagenic landscape in tumors.

Introduction

Long interspersed nuclear element 1 (LINE-1) has attracted much attention in the last decade due to its capacity to create variation in the human genome. LINE-1 is a DNA sequence capable of duplicating itself and other DNA sequences by mobilizing messenger RNAs (mRNAs) to new genomic locations via retrotransposition (*1*-*3*); this process has resulted in thousands of mostly inactive and truncated copies of LINE-1 across the human genome (*4*). Although LINE-1 activity has been described in both healthy and pathogenic tissues (*3*, *5*, *6*), quantifying its activity is remarkably difficult due to its repetitive nature. Until recently, LINE-1 retrotransposition was believed to occur in germ cells (*7*-*9*) and tumors (*10*-*12*), but not in somatic tissues. However, growing evidence suggests that LINE-1 is active in the human brain and in other healthy somatic tissue at low levels (*13*-*16*).

As opposed to healthy tissues, tumors and cancer cell lines show higher levels of LINE-1 activity (*11*). LINE-1 instances are likely to be activated due to broad demethylation of LINE-1 promoter (*17*). The current literature describes many other factors contributing to the constraints of LINE-1 activity pre- and post-transcriptionally (*18*); however, little is known about its activation and impact in tumors (*19*). A major challenge is that the assessment of LINE-1 activity requires either elaborate assays (*20*, *21*) or multiple and complementary datasets (*22*), hindering estimation of LINE-1 activity in a large number of samples. Moreover, affordable methods to quantify LINE-1 activity, such as those

based on RNA (*15*, *23*, *24*), are confounded by the highly duplicated nature of LINE-1 and pervasive transcription (*21*), which refers to the idea that the majority of the genome is transcribed, beyond just the boundaries of known genes (*25*).

How much pervasive transcription influences the human transcriptome is still unclear (*25-27*). Some researchers suggest that pervasive transcription is mostly derived from technical and biological noise and, therefore, might not be relevant in RNA sequencing experiments (*28*). Others suggest that pervasive transcription has a stochastic nature, and if sequenced at enough depth the majority of the genome may be transcribed. With either theory, pervasive transcription should not affect quantification of the transcription of protein coding genes, which are present either in single copy or low copy numbers in the genome. However, the quantification of the transcriptional activity of transposable elements, including LINE-1, would be greatly affected by pervasive transcription due to their multi-copy nature.

The activation of LINE-1 can lead to the expression of its major enzyme, ORF2 protein (ORF2p). ORF2p is comprised of a reverse transcriptase and an endonuclease domain (*29*). The endonuclease domain of ORF2p has been shown to create double-strand breaks on DNA molecules (*30*), which are then corrected by endogenous DNA repair mechanisms. In addition, LINE-1 activation, and consequent activation of ORF2p, has been linked to poor prognosis in colorectal cancers (*31*). Recently, researchers have leveraged large-scale genome sequencing projects to search for evidence of LINE-1

mobilizations in cancer samples. However, little evidence supports that LINE-1 directly activates oncogenes or disrupts tumor suppressor genes (*12*, *32*-*35*).

This paper presents a new method to remove the effect of pervasive transcription on RNA sequencing datasets and reliably quantify LINE-1 subfamily transcriptional activity. We first validated the LINE-1 transcription landscape in well-established human cell lines and their cell compartments. We then surveyed LINE-1 activity in a variety of healthy somatic tissues. Although somatic retrotransposition has been mainly studied in the human brain, we found surprisingly little transcription activity in most brain regions. Instead, we found LINE-1 transcription activity in other somatic tissues and tumors, consistent with an overall trend of LINE-1 activity in frequently dividing cells. Moreover, we found that instances of the LINE-1 subfamily L1Hs are highly active in tumors. We also demonstrated that LINE-1 transcription drives the creation of small insertions and deletions (indels) in the tumoral genome, indicating a strong correlation between LINE1 transcription and genomic instability.

Results

Recently amplified LINE-1 subfamilies, such as L1Hs, are frequently discarded from traditional transcript quantification assays 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 tissues (*36*, *37*). Figure 1A shows the high correlation between the average number of reads mapping to LINE-1

subfamilies and the number of bases annotated as the respective LINE-1 subfamily (Spearman's rank correlation rho=0.94, $p < 2.2e-16$). The correlation was mostly driven by ancient LINE-1 subfamilies; specifically, reads mapped ten times more frequently to ancient LINE-1 subfamilies, such as L1ME1 and L1M5, than recently active LINE-1 subfamilies. In fact, most of the LINE-1 reads appeared to derive from subfamilies that are thought to be inactive (genomic fossils) and not autonomously transcribed. As an explanation for this counterintuitive result, we hypothesized that this "genomic-transcriptomic" correlation might be indicative of pervasive transcription. In this model, the stochastic nature of RNA polymerase II transcription would drive the creation of RNA fragments proportionally to the number of copies of LINE-1 subfamilies in the genome.

We then divided samples by their tissues of origin (Figure 1B) and noticed that some tissues had smaller genomic-transcriptomic correlations, hinting at another confounding signal creating reads overlapping LINE-1 subfamilies. We hypothesized that deviations from a high genomic-transcriptome correlation could be derived from autonomous transcription of the LINE-1 subfamilies (see Methods for details). We then developed a software platform, TeXP (available at https://github.com/gersteinlab/texp), that uses mappability signatures from pervasive and simulated LINE-1 subfamilies autonomous transcription to deconvolve reads overlapping LINE-1 elements (Figure 1C). TeXP counts the number of reads overlapping recently expanded LINE-1 subfamilies, and calculates the best signature fit that explains the observed read counts. Specifically,

TeXP regresses the proportion of reads derived from each signal, ensuring sparsity (see Methods for details).

LINE-1 transcriptional activity in human cell lines

We benchmarked TeXP by estimating the autonomous transcription of LINE-1 subfamilies in an RNA sequencing experiment of well-established human cell lines (*36*). Figure 2A shows the proportion of reads mapped to LINE-1 subfamilies using a naïve method (left panel) and proportions of reads from each signature using TeXP (right panel). In the naïve method (Figure 2A; left panel), cytoplasmic and whole-cell polyadenylated (polyA)+ samples had an enrichment of reads mapping to L1Hs and L1PA2 when compared to whole-cell transcripts without a polyadenylated tail (whole-cell polyA-) and nuclear RNA samples. The enrichment of L1Hs reads was consistent with increased transcription of full-length L1Hs (Figure S1). The estimates after applying TeXP (Figure 2A; right panel) revealed two major signals in MCF-7 RNA sequencing experiments: pervasive transcription and L1Hs autonomous transcription. This analysis suggests that reads mapped to ancient LINE-1 subfamilies, such as L1PA3 and L1PA4, are mostly derived from pervasive transcription. TeXP also detected L1PA2 transcription but at lower intensity and frequency (Figure 2A and Figure S2). This result is consistent with L1Hs and L1PA2 being the only two LINE-1 subfamilies capable of mobilizing in the human germline and tumors (*9*, *38*).

MCF-7, a cell line derived from breast cancer, was previously described as having remarkably high levels of L1Hs autonomous transcription (*15*, *22*). To investigate the

source of this L1Hs transcription, we analyzed RNA sequencing experiments from MCF-7. The transcriptome of MCF-7, and many other cell lines, has been carefully and consistently sequenced through the Encyclopedia of DNA elements (ENCODE) project. Leveraging these ENCODE cell line datasets, we assessed L1Hs autonomous transcription in distinct cell compartments (*36*). First, we found that MCF-7 whole-cell polyA+ samples had extremely high levels of L1Hs transcription (180.7 RPKM), in agreement with the literature. Selecting whole-cell polyA- samples reduced the signal of L1Hs autonomous transcription by 73% (Figure 2A), suggesting that most of the signal was derived from mature polyA+ transcripts. Furthermore, we tested whether L1Hs transcripts are derived from cytoplasmic (mature) or nuclear (pre-mRNA) portions of the cell. We found that nuclear transcripts were enriched for pervasive transcription (autonomous/pervasive ratio 0.02), whereas cytoplasmic transcripts had an autonomous/pervasive ratio similar to transcripts derived from whole-cell polyA+ samples (0.45 and 0.51, respectively – Figure 2A). 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. Analyzing other lymphoblastic and cancer-derived cell lines such as GM12878, SK-MEL-5 and K-562 yielded no evidence of L1Hs autonomous transcription in most cell compartments or RNA fractions, despite low levels of L1Hs autonomous transcription in whole-cell polyA+ samples (0, 8.8 and 8.4 RPKM, respectively. Figure 2B and Table S1).

Validation of LINE-1 autonomous transcription

To validate the quantification of L1Hs autonomous transcription, we used a reference panel of cell lines: MCF-7, K-562, HeLa, HepG2, SK-MEL-5, and GM12878. We used droplet digital PCR (ddPCR) to quantify autonomous and pervasive transcription levels. For these experiments, we assumed that expression on the 5' end of the L1Hs transcript was mostly derived from autonomous transcription, and expression on the 3' end was derived from a combination of autonomous and pervasive transcription. We initially designed and tested multiple assays targeting different regions of the L1Hs locus, and proceeded with the two best performing assays (Table S2). The first assay targeted ORF1, directly adjacent to the 5'UTR, representing the 5' end of the transcript. The second one targeted ORF2 about 1.5 kb upstream of the 3' UTR, representing the 3' end of the transcript. We completed the same design process for ORF2 to find the copy numbers of the truncated L1Hs transcripts (i.e., the transcripts missing the 5' end of L1H) (Figure 2C, Table S3). Since autonomous transcription results in the full-length transcript of L1Hs, we quantified the level of pervasive transcription by subtracting expression of the 5' end (ORF1) from the 3' end (ORF2) (shown in terms of transcript levels, percentage, and fold change compared to autonomous transcription, Figure 2D-F).

Comment [S 1]: Figure 2C hasn't yet been referenced. Is this the right place for it?

Comment [S 2]: Correct? You haven't yet referenced Figure 2E or 2F

Figure 2D shows the relative quantification of L1Hs transcripts in these four cell lines using the *HPRT1* 5' end as a reference. The ddPCR analysis detected 12,600 copies of full-length transcripts/ng in MCF-7 cells. In agreement with our *in-silico* result, K562 and SK-MEL-5 had 1,512 and 1,708 copies of full-length transcript/ng, respectively. For the GM12878 cell line, we expected to find no autonomous expression of L1Hs; however,

our ddPCR assays detected low levels of autonomous transcription of L1Hs (655 copies of full-length transcript/ng; Figure 2D, Table 2). Overall, the quantification of L1Hs autonomous transcription using ddPCR was highly correlated with the quantification using TeXP (Spearman correlation, rho=0.99, p-value=3.803e-06). This suggests that TeXP can remove most of the noise derived from pervasive transcription, although it is insensitive to samples with little LINE-1 autonomous transcription.

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

Researchers have long thought that LINE-1 instances are completely silenced in most somatic cells. LINE-1 is silenced by the methylation of its promoter (*17*), which should preclude the transcription of mature LINE-1 mRNAs in healthy somatic tissue. To test whether LINE-1 subfamilies are completely silenced in somatic tissue, we analyzed LINE-1 transcription in 7,429 primary tissue samples from the Genotype-Tissue Expression (GTEx) project (*37*) (Table S4). Similar to the cell lines, we found that L1Hs was autonomously transcribed; L1P1, L1PA2, L1AP3, and L1PA4 only had residual or spurious autonomous transcription in healthy tissues (Figure S5). Furthermore, we found that pervasive transcription was the major signal in the RNA sequencing datasets, accounting for 91.7%, on average, of the reads overlapping LINE-1 instances (Figure S12). Overall, healthy tissues had a narrower range of L1Hs autonomous transcription levels than cell lines, with the peak transcription level of 47 RPKM (Figure 3; L1Hs RPKM histogram) versus 180 RPKM in the cell lines (Table S1). We found no or very little (<1 RPKM) evidence of L1Hs autonomous transcription in 2,520 (34.3%) of the

GTEx RNA sequencing experiments from primary tissues. Together, these results indicate that L1Hs is broadly transcribed in some healthy somatic tissues, polyadenylated, and present in the cytoplasm. Therefore, if post-transcriptional regulatory constraints do not completely shut down LINE-1 activity, we expect that LINE-1 should 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 corresponding primary tissue to understand the changes induced by cell line immortalization. EBV immortalization causes drastic changes in the expression of cell cycle, apoptosis, and alternative splicing pathways (*39-41*). Overall, we found that EBV-transformed cell lines derived from different tissues (lymphoblastic and fibroblastic) had distinct patterns of L1Hs autonomous transcription; lymphoblast (blood-derived) cell lines had 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 had higher levels of L1Hs autonomous transcription (median 1.5 RPKM) with 58.7% of samples having an RPKM higher than 1. In general, EBV-immortalized cell lines reflected 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 below 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 found that both skin cells and lymphocytes had a drastic down-regulation of L1Hs

autonomous transcription (Figure S11). This finding suggests that EBV-transformed cell lines partially preserve the L1Hs transcription level from their tissue of origin, potentially explaining why fibroblast-derived induced pluripotent stem cells support higher levels of LINE-1 retrotransposition (*42*).

Human tissues show remarkable variability of L1Hs autonomous transcription. We found that L1Hs autonomous transcription is inversely correlated to the time it takes 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 (*43*), are amongst the tissues with the lowest levels of L1Hs autonomous transcription (Figure 3). In particular, the human cerebellum, which has no transcription of L1Hs, is likely to have strong repression of L1Hs autonomous transcription. This result contradicts the literature that suggests that the human brain supports high levels of somatic LINE-1 retrotransposition; however, most of these studies were based on neural precursors that correspond to the early development stage of the human brain (*13*, *44-46*). Conversely, brain samples extracted from the striatum, putamen, and caudate, all regions associated with the basal ganglia, had higher levels of L1Hs autonomous transcription compared to other brain regions (T-test basal ganglia vs. all other brain tissues, t = -7.0943; p value = 9.867e-12 – Figure 3); importantly, these levels were still low compared to other tissues. Other tissues with low cell turnover rates, such as liver, pancreas, and spleen, also showed very little or no autonomous transcription of L1Hs (91.2%, 82.9%, 88.9% of samples, respectively, had a L1HS RPKM < 1 – Figure 3). Conversely, germinative tissues have been proposed to support

somatic activity of L1Hs elements (*47*). Our results suggest that this trend is more general, and most tissues associated with the reproductive system sustain higher levels of L1Hs autonomous transcription (Figure 3). In addition, we found that the tissues with the highest levels of L1Hs autonomous transcription were enriched for high cell turnover; these included the nerve (tibia), skin (both exposed and not exposed to the sun), prostate, lung, and vagina (Figure 3).

Previous research have suggested that LINE-1 activity could be correlated with an individual's age (*48-50*); specifically, as individuals age LINE-1 may lose methylation marks in its promoter and be derepressed. Having uniformly 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. In most tissues we did not observe significant correlations, 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 S5). Intriguingly, contrary to our expectation of higher L1Hs transcriptional activity in older individuals, we found that prostate and whole blood samples showed 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., tibia nerve and ovary).

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 higher basal 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 levels 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 matched tissue, reaching up to 13x higher expression in LUSC samples (Figure S8).

We hypothesized that these genomes would have consistently higher genomic instability due to the activity of L1Hs endonuclease. Ideally, we would use somatic LINE-1 insertions or chromosomal rearrangements in order to assess the activity of LINE-1; however, these analyses demand large-scale structural variation calling on whole genome sequencing datasets. Therefore, to test this hypothesis we assessed the frequency of indels in the exome in the same samples we estimated L1Hs activity as a proxy for the overall level of genomic instability. 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 showed a significant correlation between autonomous LINE-1 transcription and the number of indels (Figure 4A), all samples combined had a significantly high correlation (0.49, p value < 2.2e-16). To further assess the association between these two variables, we focused on 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 at TTT|AA loci (*30*, *51*). 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 (*52*). The NHEJ pathway is known to be error-prone and particularly active in the cancer context, creating small indels as well as large duplications, deletions, and translocations (*53*). We tested whether the LINE-1 endonuclease target motif (TTT|AA) was enriched in sequences flanking indels and found that regardless of the tissue of origin, the motif TTT|AA was enriched in the 50 nucleotides (nts) flanking the indel. We further selected 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 (Figure 5) 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 TTT|AA motifs on nuclear DNA and creates double-strand breaks. Instead of initiating the reverse transcription of the LINE-1 mRNA, the ORF2p 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).

Conclusion

Previous to this study, LINE-1 was thought to be active in germline and tumor cells but not normal somatic cells, with the exception of hints of activity in brain cells (*19*). Here we performed a comprehensive and unbiased analysis of LINE-1 transcriptional activity across different cell types and somatic tissues. Surprisingly, we found that LINE-1 was active in normal cells, especially epithelial cells, but not in brain cells. This result is in agreement with LINE-1 activity being correlated with cell proliferation rate. We also found high activity of LINE-1 in tumor cells, which appeared to be associated with a particular type of mutation mechanism in the tumoral genome, contributing to the creation of indels.

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 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.

Our model proposes the number of reads overlapping L1Hs instances as 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 LINE-1 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 (*25*). 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 LINE-1 bases annotated as the subfamily i in the **Equation 3**:

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

On the other had mappability fingerprints, which represents how reads deriving from LINE-1 transcripts would be mapped to the genome, are created by aligning simulated reads deriving from putative L1 transcripts from each L1 subfamily. 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 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 based on fastq read length.

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 ε and ϵ

The known variables O_i, T, the vector P_i, the mappability fingerprint matrix $M_{i.j}$ are used to estimate the signal proportion ε and ϵ in **Equation 2** by solving a linear regression. We used lasso regression (L1 regression) to maintain sparsity. We used the R package penalized ((*54*) - 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 transcripts.

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 (*37*). 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 S2). 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 S3). 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

1. G. J. Cost, Q. Feng, A. Jacquier, J. D. Boeke, Human L1 element target-primed reverse transcription in vitro. *EMBO J.* **21**, 5899–5910 (2002).
2. D. A. Kulpa, J. V. Moran, Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. *Nat. Struct. Mol. Biol.* **13**, 655–660 (2006).
3. E. M. Ostertag, H. H. Kazazian, Biology of mammalian L1 retrotransposons. *Annu. Rev. Genet.* **35**, 501–538 (2001).
4. E. S. Lander *et al.*, Initial sequencing and analysis of the human genome. *Nature*. **409**, 860–921 (2001).
5. D. C. Hancks, H. H. Kazazian, Active human retrotransposons: variation and disease. *Curr. Opin. Genet. Dev.* **22**, 191–203 (2012).
6. K. H. Burns, Transposable elements in cancer. *Nat. Rev. Cancer*, 1–10 (2017).
7. J. Wang *et al.*, dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. *Hum. Mutat.* **27**, 323–329 (2006).
8. A. D. Ewing, H. H. Kazazian, High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. *Genome Res.* **20**, 1262–1270 (2010).
9. P. H. Sudmant *et al.*, An integrated map of structural variation in 2,504 human genomes. *Nature Publishing Group*. **526**, 75–81 (2015).
10. J. Skowronski, M. F. Singer, Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 6050–6054 (1985).
11. V. P. Belancio, A. M. Roy-Engel, P. L. Deininger, All y’all need to know ‘bout retroelements in cancer. *Seminars in Cancer Biology*. **20**, 200–210 (2010).

12. J. M. C. Tubio *et al.*, Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. *Science*. **345**, 1251343 (2014).

13. A. R. Muotri *et al.*, Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*. **435**, 903–910 (2005).

14. H. Kano *et al.*, L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. *Genes Dev.* **23**, 1303–1312 (2009).

15. V. P. Belancio, A. M. Roy-Engel, R. R. Pochampally, P. Deininger, Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res*. **38**, 3909–3922 (2010).

16. G. D. Evrony *et al.*, Cell lineage analysis in human brain using endogenous retroelements. *Neuron*. **85**, 49–59 (2015).

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

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

19. J. A. Erwin, M. C. Marchetto, F. H. Gage, Mobile DNA elements in the generation of diversity and complexity in the brain. *Nature Publishing Group*. **15**, 497–506 (2014).

20. T. T. Doucet, H. H. Kazazian, Long Interspersed Element Sequencing (L1-Seq): A Method to Identify Somatic LINE-1 Insertions in the Human Genome. *Methods Mol. Biol.* **1400**, 79–93 (2016).

21. P. Deininger *et al.*, A comprehensive approach to expression of L1 loci. *Nucleic Acids Res.* **45**, e31–e31 (2017).

22. C. Philippe *et al.*, Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. *Elife*. **5**, 166 (2016).

23. S. H. Rangwala, L. Zhang, H. H. Kazazian, Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol.* **10**, R100 (2009).

24. S. W. Criscione, Y. Zhang, W. Thompson, J. M. Sedivy, N. Neretti, Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics.* **15**, 583–17 (2014).

25. M. B. Clark *et al.*, The reality of pervasive transcription. *PLoS Biol.* **9**, e1000625–discussion e1001102 (2011).

26. A. Jacquier, The complex eukaryotic transcriptome: unexpected pervasive

transcription and novel small RNAs. *Nat. Rev. Genet.* **10**, 833–844 (2009).

27. H.-G. Lee, T. G. Kahn, A. Simcox, Y. B. Schwartz, V. Pirrotta, Genome-wide activities of Polycomb complexes control pervasive transcription. *Genome Res.* (2015), doi:10.1101/gr.188920.114.

28. H. van Bakel, C. Nislow, B. J. Blencowe, T. R. Hughes, Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* **8**, e1000371 (2010).

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

30. S. L. Gasior, T. P. Wakeman, B. Xu, P. L. Deininger, The human LINE-1 retrotransposon creates DNA double-strand breaks. *J. Mol. Biol.* **357**, 1383–1393 (2006).

31. S. Ogino *et al.*, A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. *J. Natl. Cancer Inst.* **100**, 1734–1738 (2008).

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

33. R. Shukla *et al.*, Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell*. **153**, 101–111 (2013).

34. E. Helman *et al.*, Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome Res.* **24**, 1053–1063 (2014).

35. E. C. Scott *et al.*, A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. *Genome Res.* **26**, 745–755 (2016).

36. ENCODE Project Consortium, An integrated encyclopedia of DNA elements in the human genome. *Nature Publishing Group*. **489**, 57–74 (2012).

37. GTEx Consortium, Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. **348**, 648–660 (2015).

38. I. Ovchinnikov, A. Rubin, G. D. Swergold, Tracing the LINEs of human evolution. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 10522–10527 (2002).

39. E. Bolotin *et al.*, Statin-induced changes in gene expression in EBV-transformed and native B-cells. *Human Molecular Genetics*. **23**, 1202–1210 (2014).

40. M. Caliskan, D. A. Cusanovich, C. Ober, Y. Gilad, The effects of EBV transformation on gene expression levels and methylation profiles. *Human Molecular Genetics*. **20**, 1643–1652 (2011).

41. J. L. Min *et al.*, Variability of gene expression profiles in human blood and lymphoblastoid cell lines. *BMC Genomics*. **11**, 96 (2010).

42. S. Klawitter *et al.*, Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. *Nat Commun*. **7**, 10286 (2016).

43. K. L. Spalding, R. D. Bhardwaj, B. A. Buchholz, H. Druid, J. Frisén, Retrospective birth dating of cells in humans. *Cell*. **122**, 133–143 (2005).

44. C. A. Thomas, A. C. M. Paquola, A. R. Muotri, LINE-1 retrotransposition in the nervous system. *Annu. Rev. Cell Dev. Biol.* **28**, 555–573 (2012).

45. A. R. Muotri *et al.*, L1 retrotransposition in neurons is modulated by MeCP2. *Nature*. **468**, 443–446 (2010).

46. N. G. Coufal *et al.*, L1 retrotransposition in human neural progenitor cells. *Nature*. **460**, 1127–1131 (2009).

47. R. C. Iskow *et al.*, Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*. **141**, 1253–1261 (2010).

48. H. T. Bjornsson *et al.*, Intra-individual change over time in DNA methylation with familial clustering. *JAMA*. **299**, 2877–2883 (2008).

49. M. Van Meter *et al.*, SIRT6 represses LINE1 retrotransposons by ribosylating KAP1 but this repression fails with stress and age. *Nat Commun*. **5**, 5011 (2014).

50. Y. H. Cho *et al.*, The Association of LINE-1 Hypomethylation with Age and Centromere Positive Micronuclei in Human Lymphocytes. *PLoS ONE*. **10**, e0133909 (2015).

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

52. M. O'Driscoll, P. A. Jeggo, The role of double-strand break repair — insights from human genetics. *Nat. Rev. Genet.* **7**, 45–54 (2006).

53. M. Onozawa *et al.*, Repair of DNA double-strand breaks by templated nucleotide sequence insertions derived from distant regions of the genome. *Proceedings of the National Academy of Sciences*. **111**, 7729–7734 (2014).

54. J. J. Goeman, L1 penalized estimation in the Cox proportional hazards model. *Biom J*. **52**, 70–84 (2010).

Figures

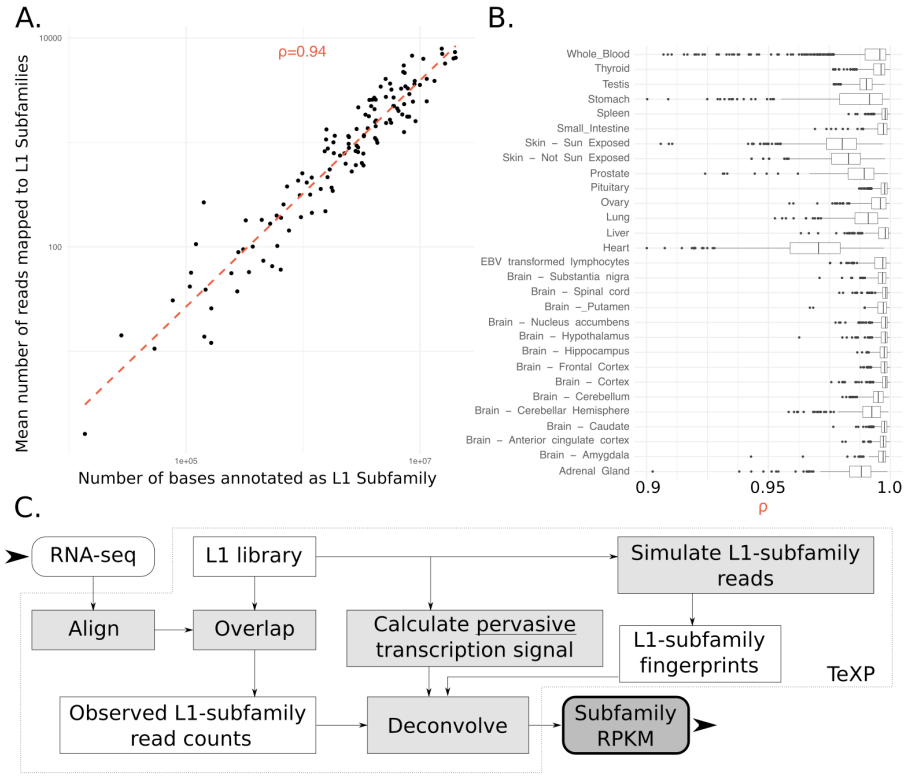
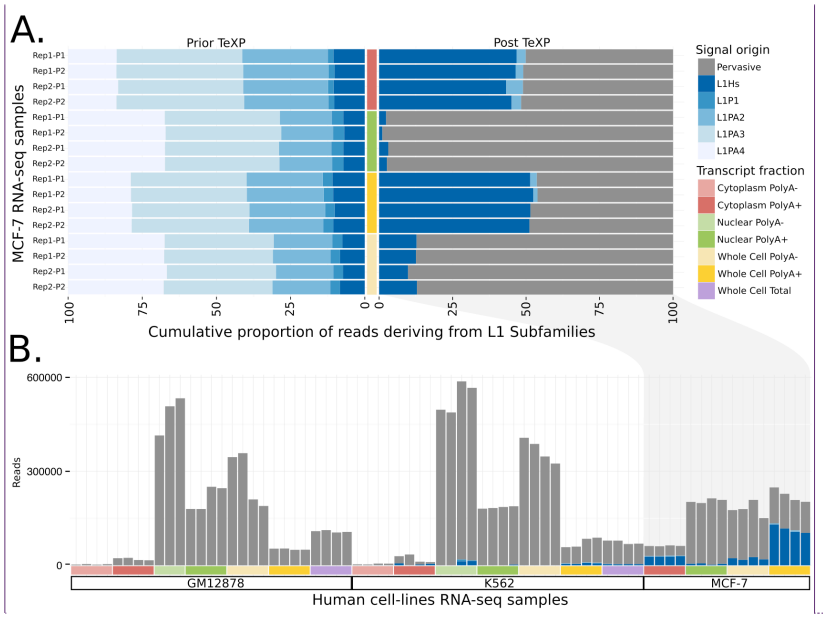
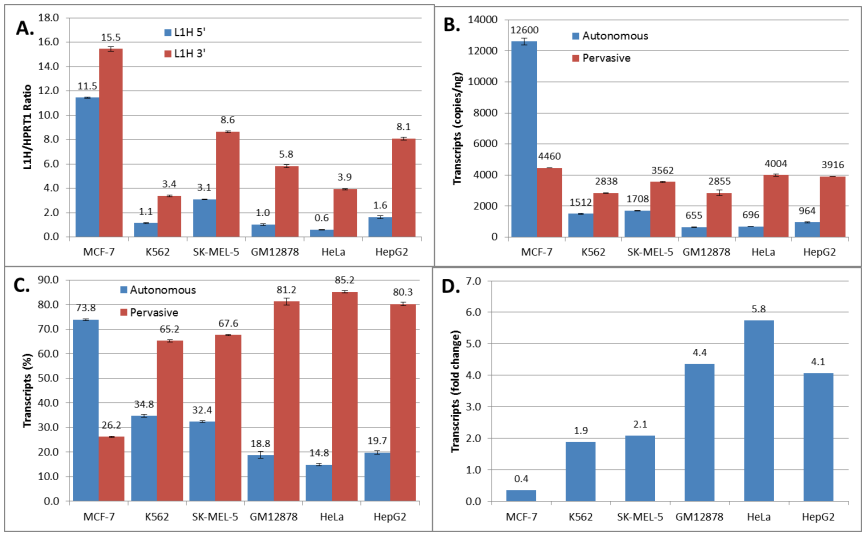


Figure 1. As pervasive transcription is a major factor leading to reads mapping to L1 instances, TeXP functions as an approach to decouple pervasive transcription from autonomous transcription. (A) The number of reads mapped to LINE-1 subfamilies is proportional to the number of bases annotated as the subfamily for most RNA sequencing experiments. (B) Healthy human tissues show varied distributions of the genomic-transcriptomic correlation. (C) Pipeline chart describes the TeXP approach.





Comment [S 3]: The figures here need to be relettered.

Figure 2. Quantification and validation of L1Hs autonomous transcription in human cell lines. (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. (C-D) The quantification of autonomous and pervasive transcripts of L1H in the cell lines is shown using ddPCR. (C) The ratio of L1H 5' and 3' transcripts shows the enrichment of the 3' end of L1H for all cell lines. (D) The absolute quantification of autonomous and pervasive transcripts reveals higher expression of pervasive compared to autonomous transcripts in all cell lines except MCF-7. (E) The percentage of autonomous and pervasive transcription shows higher expression of pervasive compared to autonomous transcripts in all cell lines except MCF-7. (F) The fold change between autonomous and pervasive transcription is shown. 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 as a reference. All data were run in duplicate. All errors bars are mean ± SEM. These data represent two independent experiments.

Comment [S 4]: In the figure itself, it should say either "prior to" or "pre". Right now it just says "prior"

Comment [S 5]: Change the letters within the actual figures.

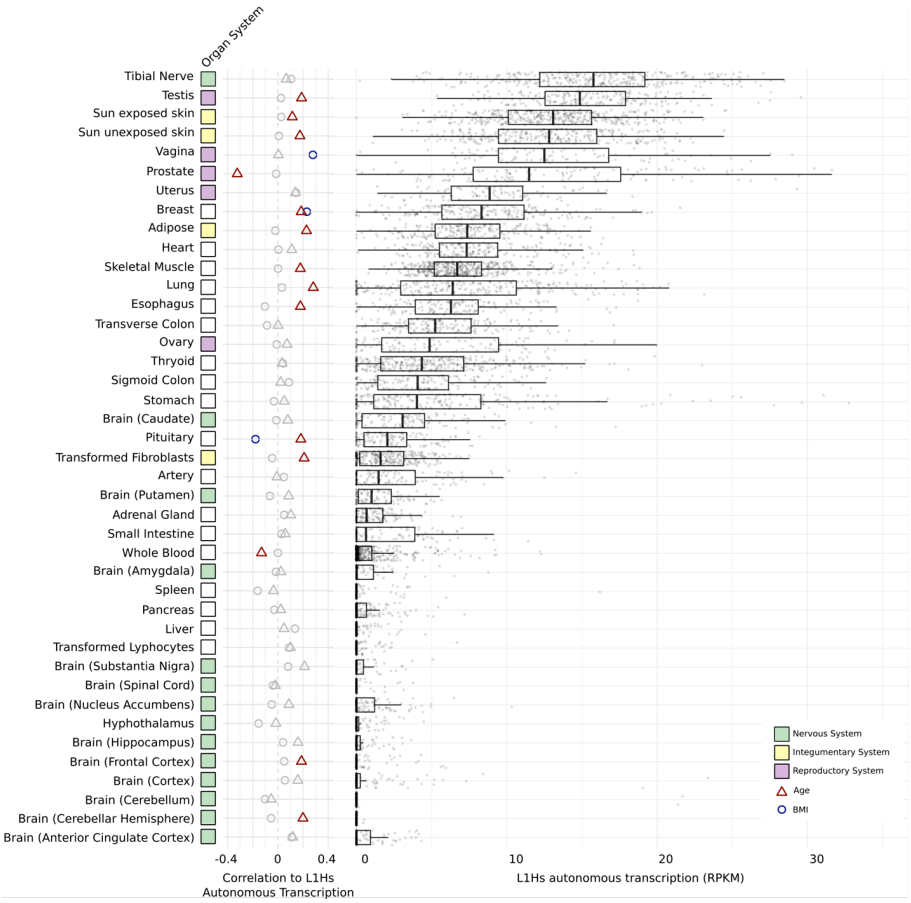


Figure 3. L1Hs autonomous transcription levels in human healthy primary tissues. The left panel describes the correlation between L1Hs autonomous transcription and the subject's age (triangles) and BMI (circles). Significant correlations are colored. The right panel describes the panorama of L1Hs autonomous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

Figure 4. L1 Endonuclease contributes to genomic instability and creation of indels. (A) The correlation between L1Hs autonomous expression and the number of indels in tumor samples is shown. (B) An overrepresentation of the TTT|AA motif close to (-3|+3nt) indels (dark) is shown compared to null (light). (C-D) An overrepresentation of the TTT|AA in the indel break point on small insertions (C) and small deletions (D) is shown.

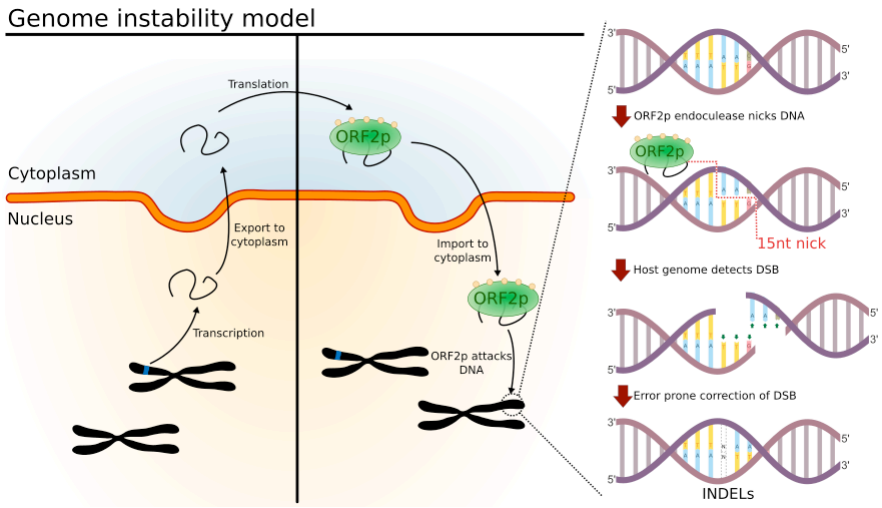


Figure 5. Model for LINE-1 favoring genome instability.