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 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 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 nature and the effect of pervasive transcription. Here, we develop a new method to estimate the activity of LINE-1 subfamilies, 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.

Introduction

LINE-1 is a DNA sequence capable of duplicating itself in a host genome and mobilizing its messenger RNA (mRNA) copies to new genomic locations via retrotransposition {Cost:2002ti, Kulpa:2006js, Ostertag:2001jl}; this process results in thousands of mostly inactive and truncated copies of LINE-1 across the genome, contributing to the genomic variability across individuals and between neurons. The major enzyme coded by LINE-1

SEN 4/17/2017 2:24 PM
Comment [1]: It should just be "LINE-1" because elements is part of the acronym.

SEN 4/18/2017 9:40 AM
Comment [2]: If you want to shorten LINE-1 to L1, you need to introduce that at first mention, and then use L1 every single time. For now, I kept everything as LINE-1.

SEN 4/17/2017 2:23 PM
Deleted: elements are

SEN 4/17/2017 2:25 PM
Deleted: sequences

SEN 4/17/2017 2:19 PM
Deleted: are specialized

SEN 4/17/2017 2:29 PM
Deleted: in

SEN 4/17/2017 2:28 PM
Deleted: themselves

SEN 4/17/2017 2:31 PM
Deleted: {Ostertag:2001jl}. LINE-1 mobilization mechanism is based on retrotransposition of mRNA

SEN 4/17/2017 2:32 PM
Deleted: thus, it creates

SEN 4/17/2017 2:32 PM
Deleted: -

SEN 4/17/2017 2:32 PM
Deleted: ,

SEN 4/18/2017 8:43 AM
Comment [3]: I felt this concept should be brought up sooner.

is comprised of a reverse transcriptase and an endonuclease domain {Piskareva:2006do}. The endonuclease domain has been shown to create double-strand breaks on DNA molecules {Gasior:2006dp}, which are then posteriorly corrected by endogenous DNA repair mechanisms. LINE-1 is known to be active in the mammalian germline {Wang:2006hr, Ewing:2010da, Schrider:2013di}, but until recently, was believed to be static in somatic cells. However, building evidence suggest that LINE-1 is active in some somatic tissues and in neural precursor cells during early development of the human brain {Evrony:2015it, Muotri:2005go, Kano:2009dt} {Belancio:2010ie}.

In addition, the somatic activity of LINE-1 has been extensively investigated in human tumors {Skowronski:1985te, Belancio:2010df, Tubio:2014gm}. As opposed to healthy tissues and cells, most human tumors and cell lines show a higher activity of LINE-1, likely due to broad demethylation of LINE-1 {Gainetdinov:2016fw, Igarashi:2010dz, Ogino:2008ey} {Patnala:2014dy, Philippe:2016cx}. Furthermore, LINE-1 hypomethylation has been linked to poor prognosis in several types of cancers. Recently, researchers have leveraged large-scale sequencing projects to search for evidence of LINE-1 mobilizations in cancer samples. Although LINE-1 has been shown to activate oncogenes and disrupt tumor suppressor genes {Scott:2016jq, Lee:2012cv, Shukla:2013bl, Helman:2014if, Tubio:2014gm}.

The assessment of LINE-1 activity requires elaborate essays {Doucet:2016ke} or multiple and complementary datasets {Philippe:2016cx}, hindering estimation of LINE-1

SEN 4/17/2017 2:32 PM
Deleted: elements …s comprised of a [1]

SEN 4/17/2017 3:19 PM
Comment [4]: Is this what you mean?

Fabio Navarro 4/18/2017 10:58 AM
Deleted: fixed

SEN 4/17/2017 3:19 PM
Deleted: These elements…INE-1 is [2]

SEN 4/18/2017 8:45 AM
Comment [5]: The way this was written suggested that you were saying "LINE-1 is active in neural precursor cells in early development of the human brain and in neural precursor cells in early development of other tissues. "

I think you meant "other tissues" to be independent of the neural precursor cells. If not, please change back

SEN 4/18/2017 8:45 AM
Comment [6]: Combine references.

SEN 4/18/2017 8:44 AM
Deleted: and a other tissues

SEN 4/17/2017 3:30 PM
Deleted: Thus, the LINE-1 elements when active, create double strand breaks and new LINE-1 insertions contributing to the genomic variability across individuals and between neurons.

SEN 4/17/2017 3:29 PM
Formatted: Highlight

SEN 4/17/2017 3:31 PM
Deleted: Aside from…n addition, the [3]

SEN 4/17/2017 3:38 PM
Comment [7]: Combine these references

SEN 4/17/2017 3:34 PM
Deleted: In addition…urthermore, the [4]

SEN 4/17/2017 3:36 PM
Comment [8]: Reference?

SEN 4/17/2017 3:38 PM
Deleted: Cancer cell lines were also shown to have high activity of LINE-1 eleme [5]

SEN 4/17/2017 3:43 PM
Comment [9]: I felt this sentence was [7]

Fabio Navarro 4/19/2017 10:56 AM
Deleted: , however, only a minority of [6]

SEN 4/17/2017 3:42 PM
Deleted: These studies also suggest [8]

SEN 4/17/2017 3:41 PM
Formatted: Highlight

SEN 4/17/2017 3:44 PM
Deleted: frequently …equires elabora [9]

activity on large scale datasets. More 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. In this study, we developed a new method to remove the effect of pervasive transcription on the quantification of LINE-1 transcription. Moreover, we analyzed the landscape of LINE-1 transcription in cell lines and tissues (healthy and tumorous), and demonstrated that human-specific LINE-1 (L1Hs) is autonomously transcribed in many tissues and correlates with the origin of small insertions and deletions (indels) in cancer cells.

Results

SEN 4/17/2017 3:45 PM
Deleted: their activity …n large scale … [10]

SEN 4/17/2017 3:55 PM
Comment [10]: I felt a more simple definition made sense to start.

Fabio Navarro 4/19/2017 10:58 AM
Deleted: and at varied transcription levelsis defined as the transcription of regions well beyond the boundaries of known genes at varied transcription levels

SEN 4/17/2017 3:58 PM
Deleted: There are uncertainties raised regarding h…owever, how much pe … [11]

SEN 4/17/2017 5:07 PM
Comment [11]: Is this correct?

SEN 4/17/2017 4:03 PM
Deleted: and …n many tissues and i … [12]

Fabio Navarro 4/19/2017 11:00 AM
Deleted: /

SEN 4/17/2017 9:03 PM
Comment [12]: Please confirm with the Journal guidelines that there shouldn't be a separate Discussion section after Results.

Recently amplified LINE-1 subfamilies, such as L1Hs, are discarded from traditional transcript quantification essays due to insufficient mapping specificity to LINE-1. We analyzed RNA sequencing experiments on thousands of cell lines, healthy primary tissue, and tumors (Table 1) {GTExConsortium:2015fb}. In the majority of cases, the average number of reads mapping to LINE-1 subfamilies correlated with the number of bases annotated as the respective LINE-1 subfamily in the reference genome (Figure 1A; Spearman's rank correlation c=0.94, p < 2.2e-16). This correlation occurred in most LINE-1 subfamilies, regardless of the retrotransposon age or evidence of recent mobilization.

We hypothesized that this genomic-transcriptomic correlation might be indicative of pervasive transcription, with RNA fragments being transcribed proportionally to the number of copies of LINE-1 subfamilies in the genome. We noticed that many of the healthy samples had a smaller genomic-transcriptomic correlation, hinting at another signal confounding the genomic-transcriptome correlation (Figure 1B). We further hypothesized that deviations from a high genomic-transcriptome correlation could be due to autonomous transcription of the LINE-1 subfamilies (see Methods for details). We modeled the number of reads mapping to LINE-1 elements as the sum of signals emanating from pervasive transcription and the autonomous transcription of LINE-1 subfamilies. We estimated the signal derived from pervasive transcription by the number of bases annotated as each subfamily, and the signal from autonomous transcription by simulated reads from LINE-1 subfamilies' transcripts. We developed a software platform, TeXP (available at [GITHUB]), that creates signatures for pervasive

SEN 4/17/2017 5:08 PM
Deleted: mapping …o LINE-1L1 inst … [13]

Fabio Navarro 4/19/2017 11:02 AM
Deleted: performed

SEN 4/17/2017 5:09 PM
Deleted: analyzed …n thousands of … [14]

SEN 4/17/2017 5:13 PM
Comment [13]: Still correct?

SEN 4/17/2017 5:13 PM
Deleted: that …he average number … [15]

SEN 4/18/2017 8:58 AM
Comment [14]: Is this the exact p value? If so it should be "="

SEN 4/17/2017 5:14 PM
Deleted: The …his correlation holds … [16]

Fabio Navarro 4/19/2017 11:03 AM
Deleted: evolution

SEN 4/17/2017 5:16 PM
Deleted: and

Fabio Navarro 4/19/2017 11:04 AM
Deleted: any

SEN 4/17/2017 5:16 PM
Comment [15]: Correct?

SEN 4/17/2017 5:16 PM
Deleted: correlation, the …enomic- … [17]

SEN 4/18/2017 9:07 AM
Comment [16]: Is this what you mean?

SEN 4/17/2017 5:21 PM
Deleted: The

Fabio Navarro 4/19/2017 11:07 AM
Deleted: calculated

SEN 4/17/2017 5:21 PM
Deleted: is calculated …y the numb … [18]

SEN 4/18/2017 9:00 AM
Comment [17]: Is this correct?

SEN 4/17/2017 5:24 PM
Deleted: is a software platform …ha … [19]

and autonomous transcription and deconvolves the proportion of reads deriving from these signals (Figure 1C).

Activity of LINE-1 elements in human cell lines

We used TeXP to estimate the autonomous transcription of LINE-1 subfamilies in well-established cell lines from RNA sequencing experiments performed by ENCODE {ENCODEProjectConsortium:2012gc} (Table S1). We included fingerprints for five distinct LINE-1 subfamilies (L1Hs, L1P1, L1PA2, L1PA3, and L1PA4; Figure 2A), and detected signals from pervasive and L1Hs autonomous transcription, and at a lower frequency, L1PA2 autonomous transcription (Figure 2A and Figure S2). L1Hs and L1PA2 are the only LINE-1 subfamilies known to be capable of mobilization in germinative tissues {Ovchinnikov:2002in, Sudmant:2015kz} and, therefore active in the human genome. This result suggests that TeXP can avoid the erroneous quantification of inactive LINE-1 subfamilies.

We next sought to characterize the transcription of LINE-1 subfamilies in different cell compartments and different RNA fractions. We used four different cell lines (MCF-7, SK-MEL5, K562, and GM12878) for which the transcriptome has been carefully sequenced through the ENCODE project {ENCODEProjectConsortium:2012gc}. We analyzed total, polyadenylated (polyA+), and non-polyadenylated (polyA-) libraries from whole cells and cellular compartments such as the nucleus and cytoplasm.

SEN 4/18/2017 9:39 AM
Deleted: L1

SEN 4/18/2017 9:00 AM
Deleted: of

SEN 4/18/2017 9:01 AM
Deleted: with

SEN 4/18/2017 9:02 AM
Deleted: Even though we

SEN 4/17/2017 5:25 PM
Deleted: 5

SEN 4/17/2017 5:26 PM
Deleted: –

SEN 4/18/2017 9:02 AM
Deleted: we almost only

SEN 4/17/2017 5:27 PM
Deleted: transcription

SEN 4/18/2017 9:03 AM
Deleted: .

SEN 4/17/2017 5:26 PM
Deleted: is also detected, but mostly at low frequency when compared to L1 ... [20]

SEN 4/17/2017 5:27 PM
Deleted: -

SEN 4/17/2017 5:28 PM
Deleted: Reassuringly, these

SEN 4/17/2017 5:28 PM
Deleted: two LINE-1 subfamilies

SEN 4/17/2017 5:28 PM
Deleted: previously described as

Fabio Navarro 4/19/2017 11:10 AM
Deleted: , have the presence/absenc ... [21]

SEN 4/17/2017 5:28 PM
Deleted: Suggesting

Fabio Navarro 4/19/2017 11:14 AM
Deleted:

Fabio Navarro 4/19/2017 11:12 AM
Deleted: is specific and

Fabio Navarro 4/19/2017 11:15 AM
Deleted: s

Fabio Navarro 4/19/2017 11:11 AM
Deleted: erroneous

SEN 4/17/2017 6:31 PM
Deleted: In order to

SEN 4/18/2017 9:04 AM
Comment [18]: Correct reference?

SEN 4/17/2017 6:31 PM
Deleted: we estimated the autonom ... [22]

Fabio Navarro 4/19/2017 11:15 AM
Deleted: prepared

We first focused on MCF-7, a cell line derived from breast cancer and previously described as having remarkable levels of L1Hs transcription {Philippe:2016cx, Belancio:2010ie}. We found that polyA+ libraries from the whole cell yielded extremely high levels of expression of L1Hs transcript (180.7 reads per kilobase of transcript per million mapped reads [RPKM]). In order to understand the translational competency of LINE-1 transcripts, we compared MCF-7 whole cell polyA+ and polyA- fractions (Figure 2A). We found that, despite a similar number of reads mapping to LINE-1 elements (Figure S3), the polyA+ transcripts yielded 3.7x more autonomous L1Hs transcript (RPKM). This result suggests that most of the autonomous L1Hs signal is derived from mature polyA+ transcripts; conversely, the polyA- transcripts are enriched for pervasive transcription. We further analyzed the cytoplasmic subset of polyA+ transcripts (Figure 2A). We found an autonomous/pervasive ratio of approximately 0.45, in agreement with the whole cell polyA+ fraction (0.51). By contrast, when we evaluated the nuclear polyA+ fraction, we found a small autonomous/pervasive ratio (0.02), with transcription levels 30x smaller than the whole cell polyA+ fraction. Together, these data 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).

When we analyzed other cancer-derived cell lines (SK-MEL-5 and K-562), we found no evidence of L1Hs autonomous transcription in most cell compartments or RNA fractions (Figure 2B). Although at smaller levels, L1Hs autonomous transcription occurred in whole-cell polyA+ samples (2.4 and 8.8 RPKM, respectively). These findings suggest that cancer cell lines can have a wide range of L1Hs autonomous transcription. We

SEN 4/17/2017 6:33 PM
Deleted: has carefully sequenced the transcriptome of MCF-7, SK-MEL5, K562, and GM12878 cell lines. Total, polyadenylated (polyA+) and non-polyadenylated (polyA-) libraries were prepared from cell whole cell and compartments such as the nuclear and cytoplasm. …e first focused on MCF … [23]

SEN 4/18/2017 9:06 AM
Comment [19]: Correct?

SEN 4/17/2017 6:43 PM
Deleted: …ell polyA+ and polyA- tra … [24]

SEN 4/18/2017 9:08 AM
Comment [20]: Correct?

SEN 4/18/2017 9:07 AM
Deleted: 1…elements being similar … [25]

SEN 4/17/2017 6:41 PM
Comment [21]: Should there be a number here?

SEN 4/17/2017 6:42 PM
Deleted: ,…suggesting …uggests th … [26]

SEN 4/17/2017 6:48 PM
Comment [22]: I switched the order to be consistent with above.

SEN 4/17/2017 6:48 PM
Deleted: Other …hen we analyzed c … [27]

SEN 4/17/2017 6:48 PM
Deleted: have …e found no evidenc … [28]

SEN 4/18/2017 9:21 AM
Comment [23]: Please confirm that this is the right order.

SEN 4/17/2017 6:50 PM
Deleted: Suggesting …hese findings … [29]

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

used ddPCR to detect autonomous and pervasive transcriptions of L1Hs in these four lines, MCF-7, K562, SK-MEL-5, GM12878. We quantified the autonomous and pervasive transcription levels based on the assumption that most expression on the 5' end of the L1Hs transcript was 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 numbers of the truncated L1H 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).

In order to confirm our *in-silico* quantification of autonomous transcription of L1Hs, we used *HPRT1* 5' end as reference, and our ddPCR analysis detected 12,600 copies of the full-length transcripts/ng in MCF-7. For K562 and SK-MEL-5, our ddPCR assays detected 1,512 copies and 1,708 copies of the full-length transcript/ng in K562 and SK-MEL-5, respectively. For GM12878, HeLa, and HepG2 cell lines, we expected no autonomous expression of L1Hs however, our ddPCR assays detected low levels of autonomous transcription of L1Hs. GM12878 had 655 copies of full-length transcript/ng,

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/21/2017 12:08 PM
Comment [24]: Move to methods?

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Line spacing: double

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

HeLa had 696 copies of full-length transcript/ng, and HepG2 had 964 copies of full-length transcript/ng (Fig. 2B., Table 2).

We next aimed to investigate the differences between cancer and healthy cell lines, using ENCODE and GTEx RNA sequencing datasets from cell lines derived from primary tissue. We found that Epstein-Barr virus (EBV)-transformed cell lines (lymphoblastic and fibroblastic) had very distinct patterns of L1Hs autonomous transcription: lymphoblast (blood-derived) cell lines showed no or very little transcription of L1Hs (Figure S6) with approximately 84% of samples having an RPKM of zero, whereas fibroblastic (skin-derived) cell lines showed a consistently higher level of L1Hs autonomous transcription (median 1.5 RPKM) with 58.7% of samples with an RPKM higher than 1. We further confirmed this result using ENCODE's GM12878, a lymphoblastic cell line derived from a healthy individual's blood. Using GM12878, we found no autonomous L1Hs regardless of the cell component (Figure 2B). These results might explain why fibroblast-derived induced pluripotent stem cells (iPSCs) support retrotransposition of L1Hs, and suggest that lymphoblastic-derived iPSCs should be more stable (retrotranspositionally) than fibroblasts {Klawitter:2016ff}.

Older 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

ToDisc

Fabio Navarro 4/20/2017 4:48 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/21/2017 1:55 PM
Deleted: -

SEN 4/17/2017 6:51 PM
Deleted: We further analyzed …NCC… [30]

SEN 4/18/2017 9:37 AM
Comment [25]: This sentence felt out of place here, so I moved it to later in the text.

SEN 4/18/2017 9:32 AM
Comment [26]: Or "more ancient"?

SEN 4/17/2017 7:47 PM
Comment [27]: I thought this made sense as the rationale for the experiment. Feel free to change it back if you prefer.

SEN 4/17/2017 7:47 PM
Deleted: We …herefore, we also …[31]

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.

Since cell lines have to undergo the major transformation process of immortalization, we evaluated the transcription of LINE1 subfamilies in primary tissue to assess their transcription under normal conditions. We analyzed the levels of autonomous transcription of LINE-1 subfamilies of 7,429 GTEx primary tissue samples (Table S2). We removed 129 samples from further analysis as they lacked sufficient reads of overlapping LINE-1 elements. Similar to the immortalized cell lines, we found that only L1Hs was autonomously transcribed, with L1P1, L1PA2, L1AP3, and L1PA4 having only 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 2B; L1Hs RPKM histogram), the cancer cell lines reached 180 RPKM. Conversely, 2,520 (34.3%) GTEx RNA sequencing experiments from primary tissues had no or very little (<1 RPKM) evidence of L1Hs autonomous transcription. Interestingly, when we compared immortalized cell lines and primary tissue derived from the skin and blood we found a similar pattern of L1Hs autonomous transcription levels. Only one sample from skin had an L1Hs autonomous transcription level of smaller than 1 RPKM, and most (74.6%) of the whole blood samples had no transcriptional activity of L1Hs. This finding suggests that the

SEN 4/17/2017 7:43 PM
Comment [28]: True?

SEN 4/18/2017 9:32 AM
Deleted: or …nd genes with L1Hs in … [32]

SEN 4/17/2017 7:45 PM
Deleted: Suggesting …his finding in … [33]

SEN 4/18/2017 9:37 AM
Comment [29]: I pulled this from above. Is this a good place for it, as rationale for looking in primary tissue?

SEN 4/17/2017 7:48 PM
Deleted: O…29ne…hundred twenty … [34]

Fabio Navarro 4/21/2017 2:03 PM
Deleted: many

SEN 4/17/2017 7:51 PM
Deleted: small

SEN 4/17/2017 7:52 PM
Comment [30]: I'm not sure what these numbers refer to.

Fabio Navarro 4/21/2017 2:03 PM
Deleted: , 2,520 (34.3%)

SEN 4/17/2017 7:52 PM
Deleted: find …ound a similar patter … [35]

EBV-transformed cell lines partially preserve the L1Hs transcription level from their tissue of origin.

Activity of LINE-1 elements in primary tissue

The human brain is thought to support high levels of somatic LINE-1 retrotransposition during early development {Thomas:2012km, Muotri:2010go, Muotri:2005go, Coufal:2009kb}. We were therefore surprised to find that adult tissue samples from almost all brain regions were amongst the tissues with the lowest levels of L1Hs autonomous transcription (Figure 3). The brain regions that showed the highest autonomous transcription of L1Hs were those related to the striatum, a portion of the basal ganglia. The putamen and caudate regions showed consistently 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 (described below).

We next evaluated LINE-1 subfamily transcription in the other tissues sequenced by GTEx. Like for the brain, liver, pancreas, and spleen tissue also had very little or no autonomous transcription of L1Hs (91.2%, 82.9%, 88.9% of samples, respectively, had an RPKM < 1). The tissues that showed the highest activity of L1Hs were nerve (tibia), skin (both exposed and not exposed to the sun), prostate, lung, vagina, and testis (Figure 3). Interestingly, testis previously was reported to support genomic mobilization of L1Hs. 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.

SEN 4/18/2017 9:41 AM
Comment [31]: Is there a reason the previous paragraph isn't included in this section?

SEN 4/18/2017 9:39 AM
Deleted: L1

SEN 4/18/2017 9:42 AM
Deleted: its …arly development ... [36]

Fabio Navarro 4/21/2017 2:04 PM
Deleted: healthiesty

SEN 4/17/2017 7:57 PM
Deleted: with

Fabio Navarro 4/21/2017 2:05 PM
Deleted: and had

SEN 4/17/2017 7:57 PM
Deleted: lesser …f L1Hs autonomou ... [37]

SEN 4/17/2017 8:07 PM
Comment [32]: I moved this sentence down because it broke up the flow on the brain.

SEN 4/17/2017 8:02 PM
Deleted: …rain regions that showed ... [38]

SEN 4/18/2017 9:46 AM
Comment [33]: Correct that you are referring to those in the next paragraph?

SEN 4/17/2017 8:06 PM
Deleted: Surprised by the fact that brain samples have little or no transcription of L1Hs we looked for evidence of…e ... [39]

SEN 4/17/2017 8:12 PM
Comment [34]: Is this what you mean? Is there a reference?

SEN 4/17/2017 8:12 PM
Deleted: (Figure 3)… As in…n agre ... [40]

SEN 4/17/2017 8:13 PM
Comment [35]: Reference a figure?

SEN 4/17/2017 8:13 PM
Deleted: {Belancio:2010ie}… With t ... [41]

Based on the signal emanating from pervasive transcription, we next estimated a pervasive transcription index (PI) 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 PI, 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 PIs have been shown to have low transcriptional diversity {GTExConsortium:2015fb}, suggesting that the PI might be a good proxy for tissue transcription diversity.

SEN 4/17/2017 8:16 PM
Deleted: can also…ext estimated a … [42]

Previous research has suggested that LINE-1 activity could be associated with an individual's age {Cho:2015bx, VanMeter:2014gs} and body mass {MarquesRocha:2016iw}; this suggests that changes in the methylation state of LINE-1 elements could lead to transcription of LINE-1 subfamilies. Having estimated the transcription level of L1Hs and having access to the phenotypes of the samples, we tested whether the autonomous transcription of L1Hs correlates with sample age or body mass index (BMI). In most of the tissues, we did not observe significant correlations with subject age, most likely due to low levels of L1Hs autonomous transcription (Figure 3). We did observe significant positive correlations with the samples' age in lung, skeletal muscle, fibroblast cell lines, adipose tissue, skin, breast,

SEN 4/17/2017 8:20 PM
Deleted: works …esearch have …as … [43]

and testis, ranging from 0.17 to 0.28 (Figure 3, red triangles; Table S3). Intriguingly, we found that prostate and whole blood samples had 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). BMI was recently reported to be inversely correlated with the methylation of LINE-1 elements; 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.

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

SEN 4/18/2017 9:54 AM
Comment [36]: I thought it was more logical to put this result first.

SEN 4/17/2017 8:24 PM
Deleted: find …ound that pP…ostate … [44]

SEN 4/18/2017 9:51 AM
Comment [37]: What could this mean? What makes it intriguing?

SEN 4/18/2017 9:53 AM
Deleted: All other tissues with a significant correlation between L1Hs autonomous activity and age, show a positive correlation. Lung, Skeletal Muscle, Fibroblast cell lines, Adipose tissue, Skin, Breast, and Testis show significant positive correlation to the sample age ranging from 0.17 to 0.28 (Figure 3, red triangles - Table S3). …ther tissues … [45]

SEN 4/18/2017 9:56 AM
Comment [38]: Reference?

SEN 4/18/2017 9:58 AM
Comment [39]: Do you want to comment on what this could mean (here or in a discussion section, if there is to be one)?

SEN 4/17/2017 8:27 PM
Deleted: B

SEN 4/17/2017 8:27 PM
Deleted: - …able S4). We …inally, f … [46]

SEN 4/18/2017 9:59 AM
Comment [40]: Can you include some rationale as to why you thought there could be a difference?

SEN 4/18/2017 9:51 AM
Formatted: Font:Not Italic

SEN 4/17/2017 8:28 PM
Deleted: **…ctivity of LINE-1 L1** … [47]

SEN 4/18/2017 9:59 AM
Deleted: We further

SEN 4/18/2017 10:34 AM
Formatted: Tabs: 4.38", Left

SEN 4/18/2017 10:18 AM
Deleted: c…uld have a…e more hig… [48]

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. To test this hypothesis, we assessed the frequency of indels in the genome 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. Namely, we investigated the occurrence of indels close to the sequences 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 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 investigate if LINE-1 endonuclease could be driving double-strand breaks resulting in the enrichment of indels, we tested whether the LINE-1 endonuclease target motif (TTTAA) was enriched in sequences flanking indels. We found that regardless of the tissue of origin, there was an enrichment of the motif TTTAA in the 50 nucleotides (nts)

SEN 4/17/2017 8:40 PM
Deleted: levels like…evels slightly lo ... [49]

SEN 4/17/2017 8:44 PM
Deleted: suggest …ypothesized that ... [50]

SEN 4/17/2017 8:54 PM
Deleted: -

SEN 4/17/2017 8:54 PM
Formatted: Font:12 pt

SEN 4/17/2017 8:54 PM
Formatted: Font:12 pt

SEN 4/17/2017 8:55 PM
Deleted: …trand breaks resulting in ... [51]

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 across neighbor regions collectively. 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 suggest 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, forming 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, for example, and therefore create 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).

SEN 4/17/2017 8:55 PM
Deleted: INDEL

SEN 4/17/2017 8:55 PM
Deleted: INDEL

SEN 4/17/2017 8:57 PM
Deleted: is

SEN 4/17/2017 8:57 PM
Deleted:

SEN 4/17/2017 8:58 PM
Deleted: are

SEN 4/17/2017 8:59 PM
Deleted: INDELs

SEN 4/17/2017 8:59 PM
Deleted: INDELs

SEN 4/17/2017 8:59 PM
Deleted: –

SEN 4/18/2017 10:37 AM
Deleted: to

SEN 4/18/2017 10:37 AM
Deleted: its

SEN 4/18/2017 10:37 AM
Deleted: messenger RNA

SEN 4/17/2017 9:00 PM
Deleted: RNP

SEN 4/17/2017 9:00 PM
Deleted:

SEN 4/18/2017 10:38 AM
Deleted: messenger

SEN 4/18/2017 10:39 AM
Deleted: it

SEN 4/17/2017 9:00 PM
Deleted:

SEN 4/18/2017 10:39 AM
Deleted: , for example,

SEN 4/17/2017 9:01 PM
Deleted:

SEN 4/17/2017 8:43 PM
Deleted: Non Homologous End Joining (

SEN 4/17/2017 9:01 PM
Deleted:)

SEN 4/18/2017 10:40 AM
Deleted: ,

SEN 4/18/2017 10:39 AM
Deleted: creating

SEN 4/17/2017 9:01 PM
Deleted: insertions and deletions (INDELs)

SEN 4/17/2017 8:43 PM
Deleted: –

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:

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

Where O_{L1Hs} is the observed number of reads mapping to L1Hs, T is the total number of reads mapped to L1 instances, P_{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 ε and ϵ

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 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 ({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.

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,

Fabio Navarro 4/20/2017 5:34 PM
Formatted: Line spacing: double

Fabio Navarro 4/20/2017 5:34 PM
Formatted: Font:(Default) Arial

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.

Fabio Navarro 4/20/2017 5:34 PM
Formatted: Font:(Default) Arial

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.

Fabio Navarro 4/20/2017 5:34 PM
Formatted: Font:(Default) Arial

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.

Fabio Navarro 4/20/2017 5:34 PM
Formatted: Font:(Default) Arial

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.

Fabio Navarro 4/20/2017 5:35 PM
Formatted: Font:Bold, Italic

Fabio Navarro 4/20/2017 5:34 PM
Formatted: Font:(Default) Arial

Fabio Navarro 4/20/2017 5:34 PM
Formatted: Line spacing: double

References

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

	Assay Name	Sequence (5' → 3')
FAM Labelled	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

	Reference	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

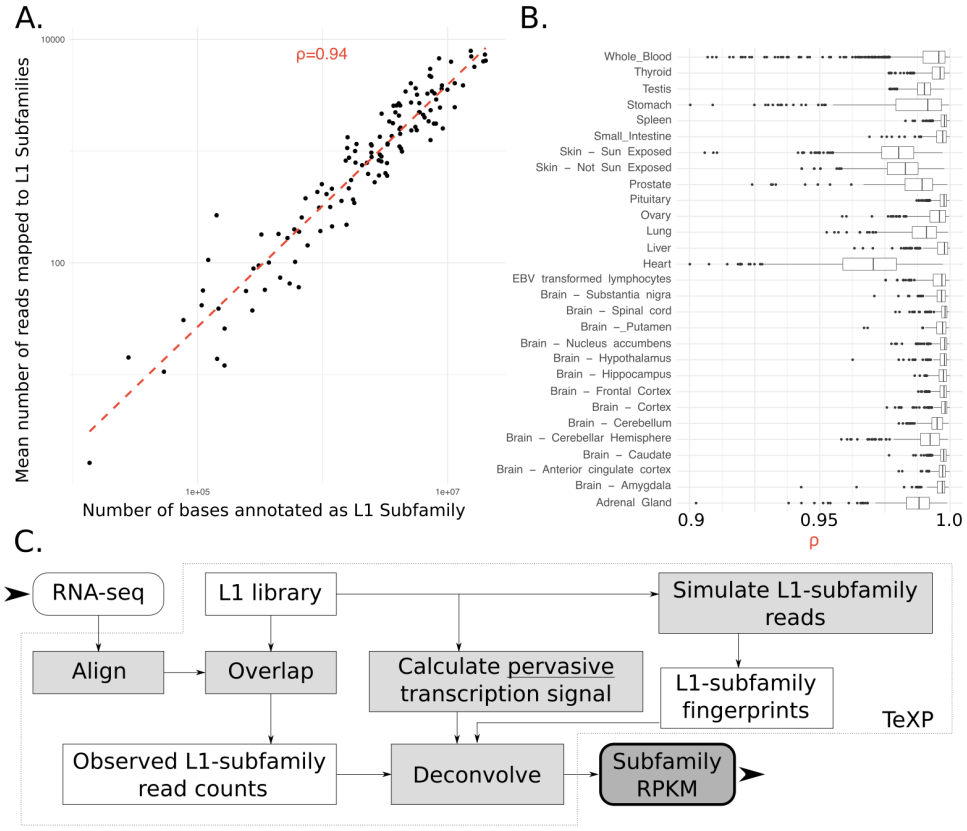
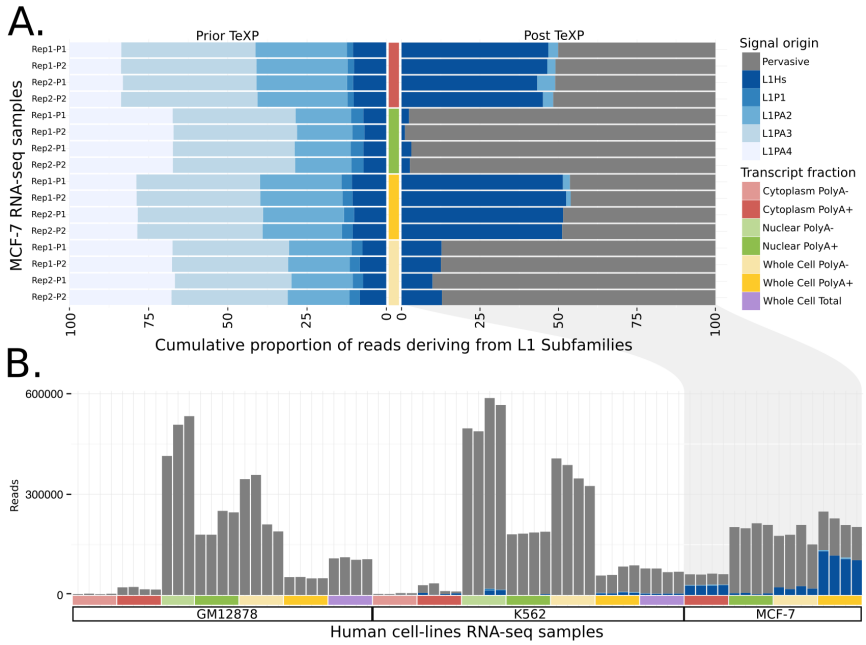


Figure 1. (A) The number of reads mapped to LINE-1 subfamilies is proportional to the number of bases annotated as the subfamily for most RNA sequencing experiments. (B) Healthy human tissues show varied distributions of the genomic-transcriptomic correlation. (C) TeXP pipeline description.

SEN 4/17/2017 9:05 PM
Comment [41]: Can you give an overall summary sentence for Figure 1 before getting into the subparts?

SEN 4/17/2017 9:03 PM
Deleted: overlapping

SEN 4/17/2017 9:04 PM
Deleted: healthy


A.
Prior TeXP
Post TeXP
MCF-7 RNA-seq samples
Rep1-P1
Rep1-P2
Rep2-P1
Rep2-P2
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
Cumulative proportion of reads deriving from L1 Subfamilies
B.
Reads
600000
300000
0
GM12878
K562
MCF-7
Human cell-lines RNA-seq samples

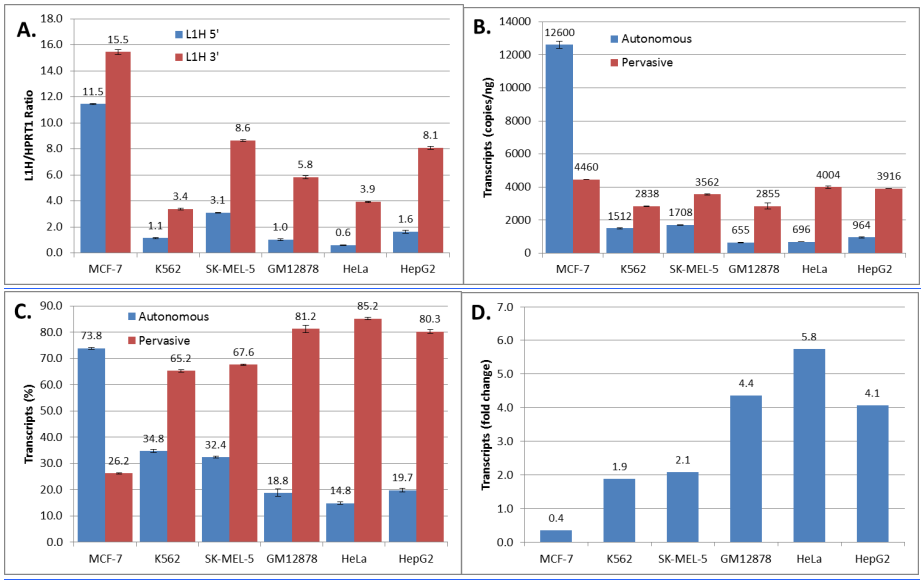

A.
L1H 5'
L1H 3'
L1H/HPRT1 Ratio
11.5
15.5
1.1
3.4
3.1
8.6
1.0
5.8
0.6
3.9
1.6
8.1
MCF-7
K562
SK-MEL-5
GM12878
HeLa
HepG2
B.
Autonomous
Pervasive
Transcripts (copies/ng)
12600
4460
1512
2838
1708
3562
655
2855
696
4004
964
3916
C.
Autonomous
Pervasive
Transcripts (%)
73.8
26.2
34.8
65.2
32.4
67.6
18.8
81.2
14.8
85.2
19.7
80.3
D.
Transcripts (fold change)
0.4
1.9
2.1
4.4
5.8
4.1

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.

SEN 4/17/2017 9:05 PM
Comment [42]: Can you give an overall summary sentence for Figure 2 before getting into the subparts?

SEN 4/17/2017 9:06 PM
Deleted: ,

SEN 4/17/2017 9:07 PM
Deleted: TeXP processing (left) and post

SEN 4/17/2017 9:07 PM
Deleted: (right)

SEN 4/17/2017 9:07 PM
Deleted:

SEN 4/17/2017 9:07 PM
Deleted: -

Fabio Navarro 4/20/2017 5:39 PM
Formatted: Font:(Default) Arial

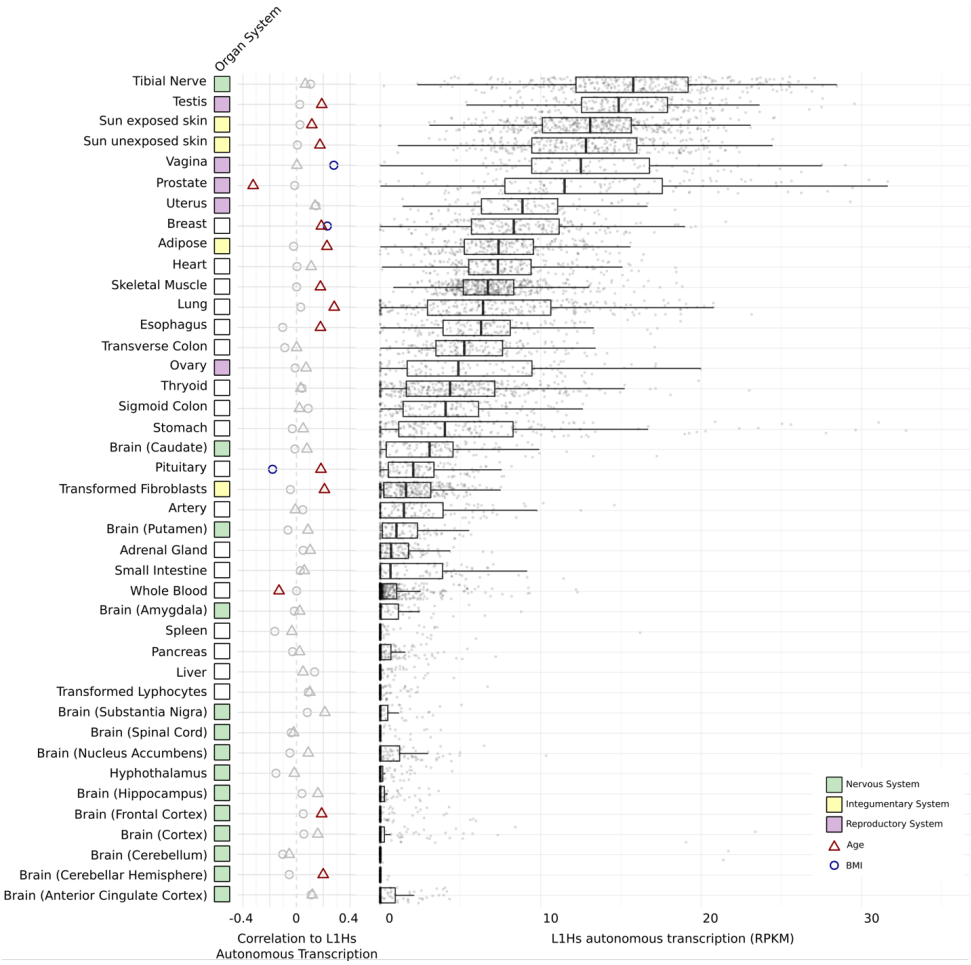


Figure 3. L1Hs autonomous transcription level on human healthy primary tissues. The left panel describes the correlation between L1Hs autonomous transcription and the subject's age (triangles) and BMI (circles). Significant correlations are colored. The right panel describes the panorama of L1Hs autonomous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

SEN 4/17/2017 9:08 PM
Deleted: subject's

SEN 4/17/2017 9:08 PM
Deleted:

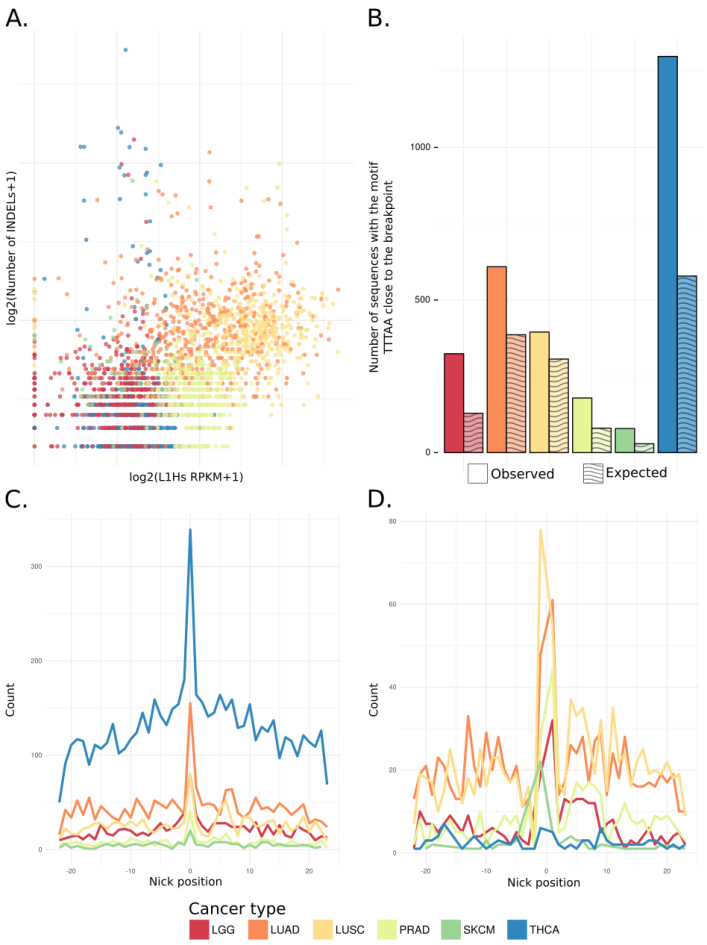


Figure 4. (A) The correlation between L1Hs autonomous expression and the number of indels in tumor samples is shown. (B) An overrepresentation of the TTTAA motif close to (-3|+3nt) indels (dark) is shown compared to null (light). (C) An overrepresentation of the TTT|AA in the indel break point on small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

SEN 4/17/2017 9:08 PM
Comment [43]: Can you give an overall summary sentence for Figure 4 before getting into the subparts?

SEN 4/17/2017 9:09 PM
Deleted: Correlation

SEN 4/17/2017 9:08 PM
Deleted: INDELs

SEN 4/17/2017 9:09 PM
Deleted: Overrepresentation

SEN 4/17/2017 9:08 PM
Deleted: INDELs

SEN 4/17/2017 9:09 PM
Deleted: O

SEN 4/17/2017 9:09 PM
Deleted: INDEL

SEN 4/17/2017 9:09 PM
Deleted: Overrepresentation

SEN 4/17/2017 9:09 PM
Deleted: INDEL

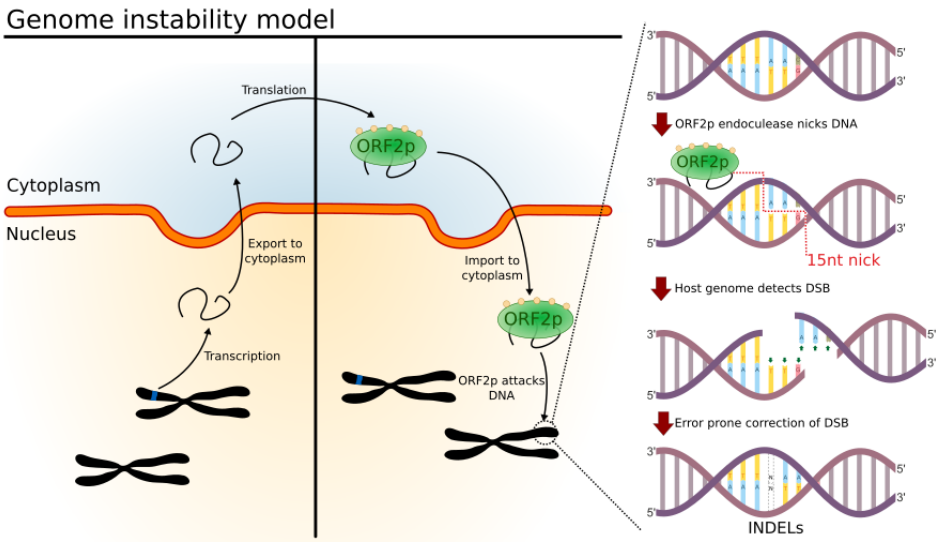


Figure 5. Model for LINE-1 favoring genome instability.