RESEARCH ARTICLE

TeXP: Deconvolving the effects of pervasive and autonomous transcription of transposable elements

Fabio CP Navarro[1,2], **Jacob Hoops**[1,2], **Lauren Bellfy**[3], **Eliza Cerveira**[3], **Qihui Zhu**[3], **Chengsheng Zhang**[3], **Charles Lee**[3,4], **Mark B. Gerstein**[1,2,5]*

1 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut, United States of America, **2** Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America, **3** The Jackson Laboratory for Genomic Medicine, Farmington, Michigan, United States of America, **4** Department of Life Sciences, Ewha Womans University, Seoul, Korea, **5** Department of Computer Science, Yale University, New Haven, Connecticut, United States of America

* mark@gersteinlab.org

Citation: Navarro FCP, Hoops J, Bellfy L, Cerveira E, Zhu Q, Zhang C, et al. (2019) TeXP: Deconvolving the effects of pervasive and autonomous transcription of transposable elements. PLoS Comput Biol 15(8): e1007293. https://doi.org/10.1371/journal.pcbi.1007293

Editor: Isidore Rigoutsos, Thomas Jefferson University, UNITED STATES

Received: September 5, 2018

Accepted: July 26, 2019

Published: August 19, 2019

Data Availability Statement: 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. 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 S1 Table.

Abstract

The Long interspersed nuclear element 1 (LINE-1) is a primary source of genetic variation in humans and other mammals. Despite its importance, LINE-1 activity remains difficult to study because of its highly repetitive nature. Here, we developed and validated a method called TeXP to gauge LINE-1 activity accurately. TeXP builds mappability signatures from LINE-1 subfamilies to deconvolve the effect of pervasive transcription from autonomous LINE-1 activity. In particular, it apportions the multiple reads aligned to the many LINE-1 instances in the genome into these two categories. Using our method, we evaluated well-established cell lines, cell-line compartments and healthy tissues and found that the vast majority (91.7%) of transcriptome reads overlapping LINE-1 derive from pervasive transcription. We validated TeXP by independently estimating the levels of LINE-1 autonomous transcription using ddPCR, finding high concordance. Next, we applied our method to comprehensively measure LINE-1 activity across healthy somatic cells, while backing out the effect of pervasive transcription. Unexpectedly, we found that LINE-1 activity is present in many normal somatic cells. This finding contrasts with earlier studies showing that LINE-1 has limited activity in healthy somatic tissues, except for neuroprogenitor cells. Interestingly, we found that the amount of LINE-1 activity was associated with the with the amount of cell turnover, with tissues with low cell turnover rates (e.g. the adult central nervous system) showing lower LINE-1 activity. Altogether, our results show how accounting for pervasive transcription is critical to accurately quantify the activity of highly repetitive regions of the human genome.

Author summary

Repetitive sequences, such as LINEs, comprise more than half of the human genome. Due to their repetitive nature, LINEs are hard to grasp. In particular, we find that pervasive

Funding: This work was supported in part by a grant from the National Human Genome Research Institute (grants U24HG009446 and U41HG007497) and the National Cancer Institute of the NIH (grant P30CA034196). C.L. is a distinguished Ewha Womans University Professor, supported in part by the Ewha Womans University Research grant of 2016. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

transcription is a major confounding factor in transcriptome data. We observe that, on average, more than 90% of LINE signal derives from pervasive transcription. To investigate this issue, we developed and validated a new method called TeXP. TeXP accounts and removes the effects of pervasive transcription when quantifying LINE activity. Our method uses the broad distribution of LINEs to estimate the effects of pervasive transcription. Using TeXP, we processed thousands of transcriptome datasets to uniformly, and unbiasedly measure LINE-1 activity across healthy somatic cells. By removing the pervasive transcription component, we find that (1) LINE-1 is broadly expressed in healthy somatic tissues; (2) Adult brain show small levels of LINE transcription and; (3) LINE-1 transcription level is correlated with tissue cell turnover. Our method thus offers insights into how repetitive sequences and influenced by pervasive transcription. Moreover, we uncover the activity of LINE-1 in somatic tissues at an unmatched scale.

Introduction

Long interspersed nuclear element 1 (LINE-1) has attracted much attention in the last decade due to its capacity to promote genetic plasticity of 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]. There are multiple molecular mechanisms to deactivate LINE-1 instances, most prominently, the truncation of 5'UTR due to partial retrotransposition has resulted in mostly inactive and truncated copies of LINE-1 across the human genome[3–6]. Truncated copies of LINE-1 lack their internal promoter sequence and therefore, are expected to be dead-on-arrival. Although full-length LINE-1 activity has been described in both healthy and pathogenic tissues [3,7,8], quantifying its activity is remarkably difficult due to its repetitive nature. Until recently, LINE-1 retrotransposition was believed to occur in germ cells [9–11] and tumors [12–14], but not in somatic tissues. However, growing evidence suggests that LINE-1 is active in the neuroprogenitor cells and in other healthy somatic tissue at low levels [15–18].

As opposed to healthy tissues, tumor and tumor derived cell lines show higher levels of LINE-1 activity [13]. LINE-1 instances are likely to be activated due to broad demethylation of LINE-1 promoter [19]. The literature describes many other factors contributing to the constraints of LINE-1 activity pre- and post-transcriptionally [20]; however, little is known about its activation and impact in tumors [21]. A major challenge to asses LINE-1 activity is the requirement of either specialized assays [22,23] or multiple and complementary datasets [24], hindering estimation of autonomous LINE-1 transcription in a large number of samples. Moreover, affordable methods to quantify LINE-1 activity, such as those based on RNA [17,25,26], are largely confounded by the high copy number nature of LINE-1 and pervasive transcription [23], which refers to the idea that the majority of the genome is transcribed, beyond just the boundaries of known genes [27].

How much pervasive transcription influences the human transcriptome is still unclear [27–29]. Some researchers suggest that pervasive transcription is mostly derived from technical and biological noise and, therefore, might not be relevant in RNA sequencing experiments [30]. 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 autonomous transcription of LINE-1, on the other hand, derives from LINE-1 transcripts being fully transcribed from its internal promoter. Thus, by definition, since LINE-1 promoters are at the 5' extremity of LINE-1 elements, autonomous transcription is more likely to derive from full length LINE-1 instances. These transcripts could derive from both intronic or intergenic full-length LINE-1 instances.

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 show that the vast majority of reads overlapping LINE-1 elements are derived from pervasive transcription and propose a method to address this issue. We validated the LINE-1 transcription landscape in well-established human cell lines and their cell compartments. Finally, we 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 transcriptional activity in most brain regions from adults. Instead, we found LINE-1 transcriptional activity in other somatic tissues consistent with an overall trend of LINE-1 activity in cell with higher turnover.

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 [31,32]. Fig 1A shows the correlation between the average number of reads mapping to LINE-1 subfamilies in healthy tissues and the number of bases in the reference genome annotated as the respective LINE-1 subfamily (Spearman's rank correlation rho = 0.94, $p < 2.2e\text{-}16$). We find that even ancient LINE-1 subfamilies can have thousands of reads from RNA-seq; for example, reads mapped ten times more frequently to ancient LINE-1 subfamilies, such as L1ME1 and L1M5, than some recently active LINE-1 subfamilies. In fact, most of the LINE-1 reads appeared to derive from subfamilies that are thought to be autonomously inactive (genomic fossils) for millions of years. 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 whole genome.

We model pervasive transcription as sufficiently broad transcription component that creates a bias in RNA-seq reads overlapping LINE-1 subfamilies. This bias tends to be correlated to the number of instances (or bases) in the reference genome because pervasive transcription is predicted to create small quantities of reads across large portions of the genome.

As opposed to the broad pervasive transcription model, under the narrow LINE-1 transcription model, we expect that a small number of LINE-1 instances to produce a much higher signal (number of reads) than the lowly expressed instances. Therefore, in order to investigate this issue, we first calculated the number of reads mapping to LINE-1 instances. We observed that, the vast majority of LINE-1 instances have a very small number of uniquely mapped reads (S1A Fig and S1B Fig) suggesting that indeed, pervasive transcription is capable of generating small number of transcripts overlapping many LINE-1 instances across the genome. A typical RNA-seq experiment has approximately 35.000 LINE-1 instance with a single read. In fact, we also observed that, on average, 99.94% of LINE-1 instances with at least one read have very small expression levels (RPKM < = 1).

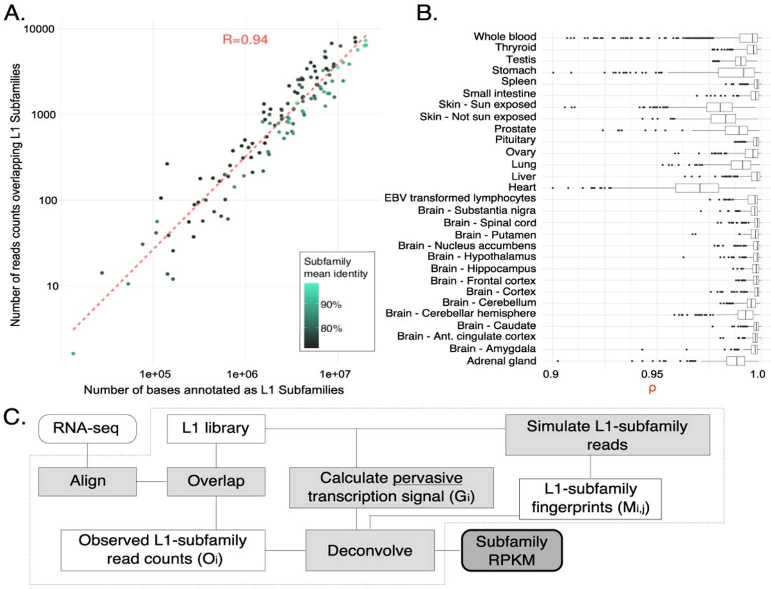

Fig 1. As pervasive transcription is a major factor leading to reads mapping to L1 instances, TeXP functions as an approach to decouple pervasive transcription from autonomous transcription. (A) The number of reads mapped to LINE-1 subfamilies is proportional to the number of bases annotated as the subfamily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

https://doi.org/10.1371/journal.pcbi.1007293.g001

To test the amount of signal generated by the top expressed instances and the lowly expressed instances we calculated the ratio of reads overlapping the most 10 expressed instances in each RNA-seq experiment and the sum of reads overlapping all instances with less than 10 reads. We find that the summation of the least expressed instances is, on average, one order of magnitude higher than the 10 mostly expressed instances. Together these results suggest that broad pervasive transcription is an important factor when quantifying LINE-1 subfamily transcription level.

We then divided samples by their tissues of origin (Fig 1B) and noticed that some tissues had smaller genomic-transcriptomic correlations, hinting at another confounding signal other than pervasive transcription 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 pipeline, TeXP, that uses mappability signatures from pervasive and simulated LINE-1 subfamilies autonomous transcription to deconvolve reads overlapping LINE-1 elements. TeXP counts the number of reads overlapping recently expanded LINE-1 subfamilies and calculates the best combination of signatures that explains the observed read counts. Specifically, TeXP regresses the proportion of reads derived from each signal, ensuring sparsity to estimate autonomous transcription of recent LINE-1 subfamilies (L1Hs, L1P1, L1PA2, L1PA3, L1PA4) and remove the effect of pervasive transcription (Fig 1C).

LINE-1 transcriptional activity in human cell lines

We benchmarked TeXP by estimating the autonomous transcription of LINE-1 subfamilies in RNA sequencing experiments of well-established human cell lines [31]. Fig 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). TeXP estimations were also

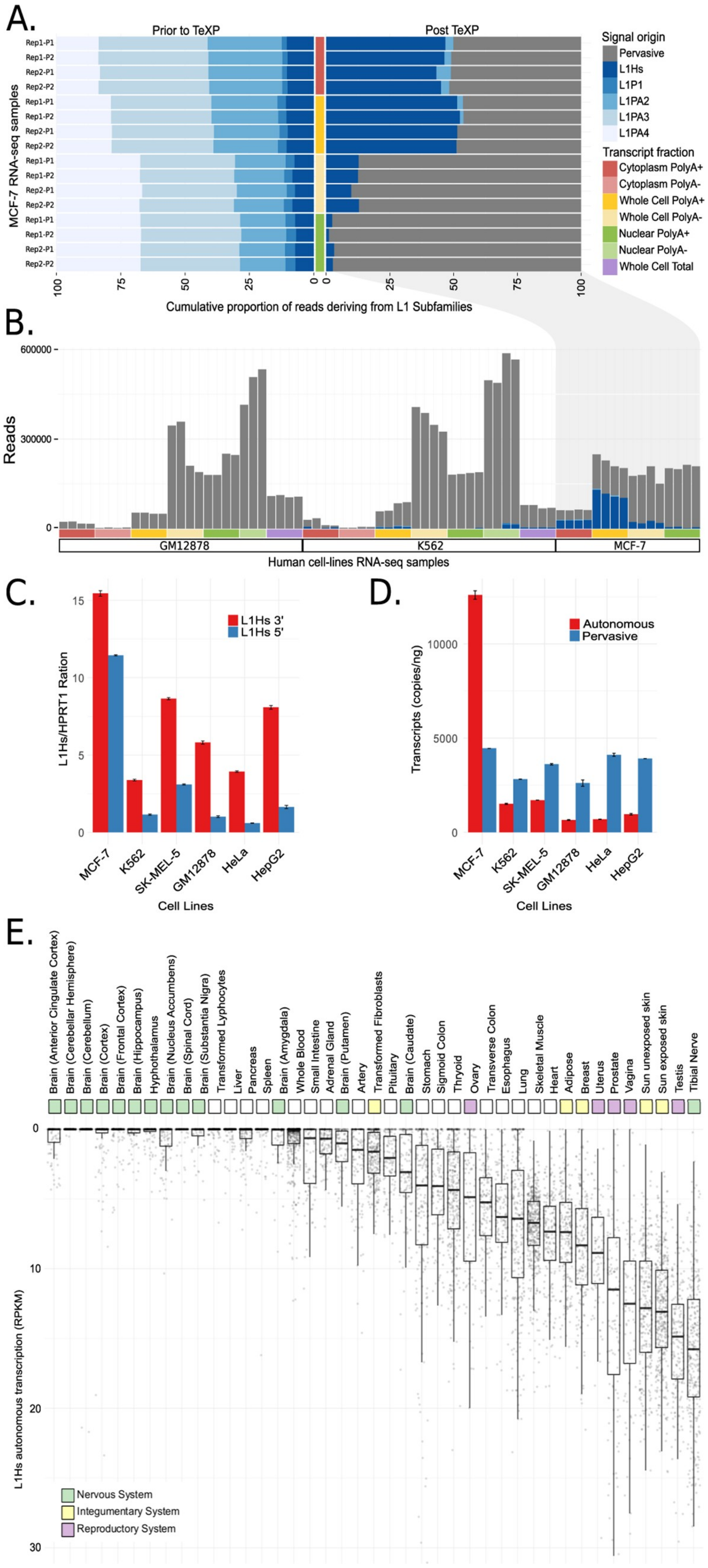
A.
Prior to TeXP
Post TeXP
MCF-7 RNA-seq samples
Rep1-P1
Rep1-P2
Rep2-P1
Rep2-P2
100
75
50
25
0
Cumulative proportion of reads deriving from L1 Subfamilies
Signal origin
Pervasive
L1Hs
L1P1
L1PA2
L1PA3
L1PA4
Transcript fraction
Cytoplasm PolyA+
Cytoplasm PolyA-
Whole Cell PolyA+
Whole Cell PolyA-
Nuclear PolyA+
Nuclear PolyA-
Whole Cell Total
B.
Reads
600000
300000
0
GM12878
K562
MCF-7
Human cell-lines RNA-seq samples
C.
L1Hs/HPRT1 Ration
15
10
5
0
L1Hs 3'
L1Hs 5'
MCF-7
K562
SK-MEL-5
GM12878
HeLa
HepG2
Cell Lines
D.
Transcripts (copies/ng)
10000
5000
0
Autonomous
Pervasive
MCF-7
K562
SK-MEL-5
GM12878
HeLa
HepG2
Cell Lines
E.
Brain (Anterior Cingulate Cortex)
Brain (Cerebellar Hemisphere)
Brain (Cerebellum)
Brain (Cortex)
Brain (Frontal Cortex)
Brain (Hippocampus)
Hyphothalamus
Brain (Nucleus Accumbens)
Brain (Spinal Cord)
Brain (Substantia Nigra)
Transformed Lyphocytes
Liver
Pancreas
Spleen
Brain (Amygdala)
Whole Blood
Small Intestine
Adrenal Gland
Brain (Putamen)
Artery
Transformed Fibroblasts
Pituitary
Brain (Caudate)
Stomach
Sigmoid Colon
Thryoid
Ovary
Transverse Colon
Esophagus
Lung
Skeletal Muscle
Heart
Adipose
Breast
Uterus
Prostate
Vagina
Sun unexposed skin
Sun exposed skin
Testis
Tibial Nerve
L1Hs autonomous transcription (RPKM)
0
10
20
30
Nervous System
Integumentary System
Reproductory System

Fig 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 L1Hs in the cell lines is shown using ddPCR. (C) The ratio of L1Hs 5' and 3' transcripts shows the enrichment of the 3' end of L1Hs 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. All data were run in duplicate. All errors bars are mean ± SEM. These data represent two independent experiments. (E) L1Hs autonomous transcription landscape of human healthy primary tissues. Each point is a RNA sequencing experiment, separated by tissue of origin.

https://doi.org/10.1371/journal.pcbi.1007293.g002

compared to other transposable element quantification pipelines (S2 Fig). In the naïve method (Fig 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 (S3 Fig). The estimates after applying TeXP (Fig 2A; right panel) revealed two major signals in MCF-7 RNA sequencing experiments: pervasive transcription and L1Hs autonomous transcription. The difference between the naïve method and TeXP suggests that reads mapped to ancient LINE-1 subfamilies, such as L1PA3 and L1PA4, are mostly derived from pervasive transcription. TeXP also detected residual L1PA2 transcription in a small number of samples (Fig 2A and S4 Fig). This result is consistent with L1Hs and L1PA2 being the only two LINE-1 subfamilies capable of autonomous transcription and autonomous mobilization in the human germline and tumors [11,33].

MCF-7, a cell line derived from breast cancer, was previously described as having remarkably high levels of L1Hs autonomous transcription [17,24]. The transcriptome of MCF-7 and many other cell lines were 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 (S5 Fig and S6 Fig) [31]. 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% (Fig 2A), suggesting that most of the signal was derived from mature polyA+ LINE-1 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 highly 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–Fig 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. Fig 2B and S1 Table).

Validation of LINE-1 autonomous transcription

To validate the quantification of L1Hs autonomous transcription, we performed droplet digital PCR (ddPCR) to estimate autonomous and pervasive transcription levels on a reference panel of six cell lines: MCF-7, K-562, HeLa, HepG2, SK-MEL-5, and GM12878. For these experiments, we assumed that expression on the 5' end of the L1Hs transcript can be used as an approximation to autonomous transcription due to the large imbalance of 5' truncated and full-length copies. The expression on the 3' end, on the other hand, is an approximation to the 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 (S2 Table). The first assay targeted ORF1, directly adjacent to the 5'UTR, representing the 5' end of the transcript. The second assay 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 L1Hs) (Fig 2C, S3 Table). Since autonomous transcription results in an enrichment full-length transcript of L1Hs, we estimated an approximation to the level of autonomous transcription as the pervasive transcription created by full-length transcription subtracted from the expression of the 5' end (ORF1).

Fig 2D shows the relative quantification of L1Hs transcripts in these four cell lines using the *HPRT1* 5' end as a reference. We estimate that the ddPCR analysis detected approximately 2,412.8 copies of autonomously transcribed transcripts/ng in MCF-7 cells. In agreement with our *in-silico* result, K562 and SK-MEL-5 had 990.0 and 1,075.6 copies of autonomously transcribed 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 (233.0 copies of autonomously transcribed transcript/ng; Fig 2D, S3 Table). Overall, the quantification of L1Hs autonomous transcription using ddPCR was highly correlated with the quantification using TeXP (Spearman correlation, rho = 0.9, p-value = 0.01394, S7 Fig). 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 (S8 Fig). To address TeXP sensitivity in relation to noise we first tested TeXP under an ideal experimental setup. In this simulation, the number of observed reads overlapping L1 subfamilies is a simple combination of known proportions of these signals (i.e. pervasive and autonomous transcription). For example, we simulate read counts where 30% of the reads derive from LIHs autonomous transcription and the remaining 70% derive from pervasive transcription. We use simulated read counts as input to TeXP and calculate the root mean square error (rmse) between the known proportions and estimated proportions. As observed in S15 Fig (solid line), under this condition, the TeXP estimations of read counts is nearly identical to the simulated read counts (median(rmse) = 0.0003). We also tested the effect of noise in the TeXP estimations. To that end, we modeled the reads that cross-map across LINE-1 subfamilies as a noise component as a Poisson process deriving from random sequencing of cDNA fragments. For these simulations, the Poisson noise was added to read counts deriving from of known proportions of signals as described above. We tested two scenarios, one with noise equivalent to 10% of all observed reads (dashed line) and 20% of the of all observed reads (long dashed lines). In these conditions, the median TeXP RMSEs were respectively 0.014 and 0.028. Overall our observations suggest that, under low noise conditions, TeXP should be able to detect autonomous L1Hs transcription even when L1Hs has low transcription levels. However, we also describe the TeXP capacity to detect low levels of L1Hs autonomous transcription degrades with higher levels of noise in RNA-se data. In extreme situations, such as when 20% of the read counts are derived for a Poisson process, the RMSE can be up to 0.083.

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 healthy somatic cells. LINE-1 is silenced by the methylation of its promoter [19], 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 [32] (S4 Table). 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 (S9 Fig). Furthermore, we found that pervasive transcription was the major signal in most RNA sequencing datasets, accounting for 91.7%, on average, of the reads overlapping LINE-1 instances (S10 Fig and S14 Fig). Overall, healthy tissues had a narrower range of L1Hs autonomous transcription levels than cell lines, with the peak transcription level of 47 RPKM (Fig 2E) versus 180 RPKM in the cell lines (S1 Table). 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. Therefore, if post-transcriptional regulatory constraints do not completely silence LINE-1 activity, one could expect that LINE-1 to play an important role in creating genetic diversity across somatic cells within an individual.

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 [34–36]. 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 (S11 Fig) 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 (S12 Fig). 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 [37].

Human solid tumors display increased levels of LINE-1 activity [38–40]. In order to assess if this increase is a result from pervasive transcription or an increase in autonomous transcription in LINE-1 we run TeXP in RNA-seq from healthy thyroid samples and solid thyroid tumors. We than calculate the distribution of number of reads deriving from pervasive and autonomous transcription. We first observed that there indeed a significant difference in the number of reads from LINE-1 elements (S13 Fig). We also found that compared to healthy thyroid samples, tumor samples display higher levels of autonomous transcription and lower levels of pervasive transcription suggesting that autonomous transcription LINE-1 is driving most of the increase in LINE-1 expression in thyroid tumor samples.

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; Pearson correlation: cor = -0.6668968; p-value = 0.04). No correlation was found between cell turnover and pervasive transcription (Pearson correlation: cor = 0.3983474; p-value = 0.2883). Tissues suggested to have low cell turnover, such as the human brain [41], are amongst the tissues with the lowest levels of L1Hs autonomous transcription

(Fig 2E). In particular, the human cerebellum, which has no transcription of L1Hs, is likely to have strong repression of L1Hs autonomous transcription. This result seems to contradict the literature that suggests that the human brain supports high levels of somatic LINE-1 retrotransposition; however, most of these studies were based on neuro precursors that correspond to the early development stage of the human brain [15,42–44]. 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 –Fig 2E); 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 –Fig 2E). Conversely, germinative tissues have been proposed to support somatic activity of L1Hs elements [45]. Our results suggest that this trend is more general, and most tissues associated with the reproductive system sustain higher levels of L1Hs autonomous transcription (Fig 2E). 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 (Fig 2E).

Discussion

Prior to this study, the effects of pervasive transcription on the estimates of transposable elements activity were largely ignored. Here, we showed that most of the RNA-seq reads matching to LINE-1 instances derive from pervasive transcription, highlighting the importance of these effects. In order to account for the effect of pervasive transcription on the quantification of LINE-1 activity we developed TeXP, a method that uses the widespread nature and the mappability signatures of LINE-1 subfamilies to account for and remove the effects of pervasive transcription from L1Hs, L1P1, L1PA2, L1PA3, and L1PA4 subfamilies. We compared TeXP estimates to other strategies such as naïve counts and others established methods such as SalmonTE [46] and TETranscript [47] (S2 Fig). Our estimations suggest that the pervasive transcription component is frequently missed and can be a confounding factor in some quantifications, for example, we observed that over 75% of reads mapping to ancient LINE-1 subfamilies derive from inactive subfamilies (Fig 2A and S14 Fig).

We used TeXP to perform a comprehensive analysis of LINE-1 transcriptional activity across different cell types and somatic tissues. Previous studies suggest that LINE-1 is active in germline and tumor cells, but not in most normal somatic cells with the exception of hints of activity in neuro-precursor cells [21]. Somatic mosaicism of transposable elements, in particular LINE-1, has been carefully characterized in the human brain and, despite some disagreement on the exact rate that LINE-1 retrotranspose in the human brain, it is clear that LINE-1 frequently create somatic copies in the human brain, in particular, in neuroprecursor, cortex and caudate nucleus cells [15,48,49]. The somatic mobilization in other tissues, nonetheless, remain obscure and lack a systematic investigation. A first step is to comprehensive characterize other tissues in terms of their LINE-1 transcriptome activity. Surprisingly, we found that LINE-1 was active in many healthy human tissues, particularly in epithelial cells. As we only detected a limited amount of LINE-1 activity in adult brain cells, our findings are in agreement with LINE-1 activity correlating with cell proliferation rate.

We validated many aspects of TeXP using ddPCR probes designed to quantify pervasive and autonomous transcription of L1Hs across human cell lines. These results show that our method lacks the sensitivity to measure autonomous transcription. In particular, TeXP underestimates L1Hs transcription levels when the signal-to-noise ratio is small. One could imagine using unique regions of LINE-1 instances after removing the pervasive transcription signal to improve the transcriptional quantification. However, removing pervasive transcription from individual instances is not trivial and should be carefully investigated.

As a tool, TeXP could be useful in several scopes beyond this study. Pervasive transcription should also affect the quantification of other transposable elements and repetitive regions accounting for large portions of the human genome. Our method could be further used to estimate the autonomous transcription levels of pseudogenes, SINEs (ALUs) and HERVs. Furthermore, TeXP could be used in model organisms to distinguish the effects of pervasive transcription beyond humans. The mouse genome, for example, has evidence of higher rates of retrotransposition but little is known about the activity of LINE-1 in somatic cells. Moreover, some of the results we describe here could be extended and uncover important biological insights. For example, assays such as induced pluripotent stem cells clones [50], RC-seq [51] and L1-seq [52], among others, could be used to carefully characterize of the rate of somatic retrotransposition in tissues with higher rates of autonomous transcription of L1Hs. Additionally, TeXP could be used to further investigate the biases found in individuals from different ancestral backgrounds [53]. The TeXP approach to quantify transcriptional activity by removing pervasive transcription could also be expanded to investigate the activity of LINE-1 during embryonic development or in pathological tissues, such as tumors.

Materials and methods

Modeling pervasive and autonomous LINE-1 transcription

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

Our model proposes that the number of reads overlapping L1Hs instances as described by the **Eq 1**:

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

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, $\varepsilon_{pervasive}$ 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 autonomous transcription L1 Subfamily *j*. This model can be further generalized as the **Eq 2**:

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

We selected these five LINE-1 subfamilies based on the rates of cross-mappability of simulated data (S3 Fig). In particular, we are interested in removing the effect of pervasive transcription from the estimates of L1Hs autonomous transcription. We simulated reads from L1Hs transcripts we observed that most (>90%) of the reads emanating from L1Hs autonomous transcription map to the L1Hs, L1PA2, L1PA3, L1PA4 and L1P1 subfamilies. Older subfamilies such as L1PA5 and L1PA6, for example, correspond to less than approximately 5% of L1Hs cross-mappable reads.

In contrast to L1Hs which only a fourth of the reads map back to L1Hs, older elements such as L1PA5 and L1PA6 have a much higher self-mapping rates, 60% and 70% respectively.

Therefore, these subfamilies should be less affected by confounding factors deriving from pervasive transcription.

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 [27]. 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 **Eq 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 **Eq 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 (S15 Fig). 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 (S16 Fig). 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 (S3 Fig). 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%—S3 Fig).

The hidden variables $\boldsymbol{\varepsilon}$ 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 $\boldsymbol{\varepsilon}$ and ϵ in **Eq 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 availability

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

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 S1 Table.

Cell turnover rate

Mutation load and cell turn-over rate were extracted from the compilation of somatic mutation rate in Tomasetti et al [55].

Pervasive versus Autonomous transcription of L1Hs transcripts

More ancient elements such as DNA transposons and LINE-2 have been shown to be primarily transcribed pervasively, hitchhiking the transcription of nearby autonomously transcribed regions [32]. 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. Furthermore, we tested if and enrichment of pervasive transcription deriving from intronic regions would create a background signal distinct from the pervasive transcription derived from a whole genome model. We correlated the number of LINE-1 instances from each subfamily in intergenic and intronic regions based on GENCODE v29 and found a statistically significant correlation between the number of instances in both regions (Spearman corr = 0.979057, p-value $<$ 2.2e-16—S17 Fig).

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 L1Hs transcript expression in the cell lines described above. Since L1Hs 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 L1Hs 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 L1Hs transcriptions. We also designed two reference assays on the housekeeping gene *HPRT1*, which targeted the 5' and 3' ends of the transcript, respectively (S2 Table). 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.

In order to estimate the levels of pervasive and autonomous transcription using ddPCR we use the following formulation:

$$X3' = a + p$$

$$X5' = a + fp$$

Where $X3'$ and $X5'$ are the transcript/μL measurements of the 3' and 5' L1Hs primers; *a* and *p* are autonomous and pervasive transcription estimates, respectively; and *f* is the fraction of L1Hs instances that could potentially generate full-length transcripts. We estimate that *f* is approximately 12%. Using this model, the estimated level of L1Hs autonomous transcription using two primers and ddPCR is:

$$a = 0.88 * X5' - 0.12 * X3'$$

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 (S3 Table). 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.

Supporting information

S1 Fig. Broad expression vs Narrow expression of L1Hs instances. Broad model of pervasive transcription is evidenced by the number of uniquely mapped reads overlapping LINE-1 (L1Hs) instances. **A.** Most "expressed" LINE-1 (L1Hs) instances have small amount of uniquely mapped reads (1–5) suggesting low levels of transcription throughout the genome. **B.** Distribution of the read count ratio. Ratio between the Top10 mostly expressed instances and all instances having 10 reads or less across all GTEx RNA-seq assays–On average only 10% of the reads derive from top expressed L1Hs instances.
(PDF)

S2 Fig. Comparison of L1 transcription level estimation methods on MCF-7 datasets. Fraction of reads deriving from 4 subfamilies and pervasive transcription from four methods.
(PDF)

S3 Fig. L1 Subfamily mappability fingerprint. Simulated transcripts from putative L1Hs, L1P1, L1PA2, L1PA3, L1PA4, L1PA5 and L1PA6 were aligned to the reference genome using the same parameters as the TeXP pipeline. The proportion of reads mapped to each subfamily was calculated.
(PDF)

S4 Fig. L1 Subfamily RPKM in ENCODE RNA-seq samples. Most samples have zero RPKM (denser bar at the bottom). L1Hs has more samples with higher RPKM than any other subfamily followed by L1PA2 the second most recent L1 Subfamily.
(PDF)

S5 Fig. MCF7 L1 subfamilies absolute read count. Every four bars are experiments respectively from whole cell polyA-; whole cell polyA+; cytoplasm polyA+ and nuclear polyA+.
(PDF)

S6 Fig. MCF7 L1 subfamilies absolute read deconvolution. Every four bars are experiments respectively from whole cell polyA-; whole cell polyA+; cytoplasm polyA+ and nuclear polyA +. Gray and Dark blue bars refer to pervasive transcription and L1Hs autonomous transcription signal respectively.
(PDF)

S7 Fig. Correlation between TeXP L1Hs estimations and ddPCR. Estimates of L1Hs autonomous transcription from ddPCR (Y-axis) and TeXP (X-axis). (Spearman correlation, rho = 0.99, p-value = 3.803e-06)
(PDF)

S8 Fig. Estimation of TeXP sensibility under multiple simulation scenarios. Input to TeXP were simulated as a combination of known signals (i.e. X pervasive transcription + (1-X) L1Hs autonomous transcription). Y-axis represents the root means square error for TeXP estimations at 3 different scenarios (0%, 10% and 20% error rate). Ribbons represent 25%-75% of TeXP rmse.
(PDF)

S9 Fig. L1 Subfamily RPKM in GTex samples. Most samples have zero RPKM (more dense bar at the bottom). L1Hs has more samples with higher RPKM than any other subfamily followed by L1PA2 the second most recent L1 Subfamily.
(PDF)

S10 Fig. Percentage of pervasive transcription on processed GTex samples. Most signal mapping to LINE-1 samples is derived from pervasive transcription.
(PDF)

S11 Fig. Estimation of L1Hs autonomous transcription in GTEx cell lines. Most of EBV transformed cell lines have no autonomous transcription of L1Hs (bottom box). Transformed fibroblasts, derived from skin, have intermediate autonomous transcription of L1Hs (at lower levels than Skin samples). And K-562, derived from Leukemia tumor, has consistently high autonomous transcription of L1Hs across distinct batches.
(PDF)

S12 Fig. EBV transformation silences L1Hs autonomous transcription. When comparing primary tissue and EBV transformed cell-lines from the same individuals we noticed a consistent decrease in the autonomous transcription of (A) Skin samples and EBV-Transformed fibroblasts ($t = 22.5743$, $df = 153.878$, p-value $< 2.2e-16$) and (B) Whole-blood and EBV--Transformed lymphocytes ($t = 4.8937$, $df = 182.036$, p-value = 2.171e-06).
(PDF)

S13 Fig. Estimation of tumoral pervasive and autonomous transcription. Distribution of normalized number of reads (RPM) derived from pervasive transcription and autonomous transcription of LINE-1 from tumor and healthy samples.
(PDF)

S14 Fig. Pervasive transcription index per GTEx tissue. Pervasive transcription index was estimated for each GTEx sample and ordered by pervasive transcription median.
(PDF)

S15 Fig. L1Hs transcript simulated as RNA sequencing reads. Different aligners were used to assess the construction of mappability fingerprints in the human reference genome. One hundred independent simulations of L1Hs transcript reads were independently mapped to the reference genome using bowtie2 (red), bwa (green) and star (blue). The box plot represents the distribution of the number of reads mapped to each L1 subfamily in the reference genome.
(PDF)

S16 Fig. Read length effect on mappability of L1 subfamilies. Simulating L1Hs reads with different length yield distinct proportions of reads mapped to each subfamily. As expected, the longer the read, the higher the proportion of reads correctly mapped originating subfamily.
(PDF)

S17 Fig. Similarity of intergenic vs intronic pervasive signal. Correlation between the number of LINE-1 elements in intergenic (x-axis) and intronic (y-axis) regions (Spearman corr = 0.979057, p-value < 2.2e-16).
(PDF)

S1 Table. ENCODE RNA-seq experiments and L1Hs RPKM estimations.
(PDF)

S2 Table. Primer and probe sequences for L1Hs target regions and *HPRT1* reference regions.
(PDF)

S3 Table. Unnormalized quantification of L1Hs transcripts. Comparison of the absolute expression of copies of full-length transcript/ng of L1Hs autonomous transcript (ORF1) and L1Hs pervasive transcript (ORF2) when run with both references.
(PDF)

S4 Table. Number of samples per tissue type. Bladder, Kidney Cortex and minor salivary gland were eliminated from further analysis.
(PDF)

S5 Table. L1 autonomous transcription level and age correlation coefficient and significance for each GTEx tissue.
(PDF)

Author Contributions

Conceptualization: Fabio CP Navarro, Mark B. Gerstein.

Data curation: Fabio CP Navarro.

Formal analysis: Fabio CP Navarro.

Funding acquisition: Mark B. Gerstein.

Investigation: Fabio CP Navarro, Jacob Hoops.

Methodology: Fabio CP Navarro.

Software: Fabio CP Navarro.

Supervision: Mark B. Gerstein.

Validation: Lauren Bellfy, Eliza Cerveira, Qihui Zhu, Chengsheng Zhang, Charles Lee.

Visualization: Fabio CP Navarro.

Writing – original draft: Fabio CP Navarro.

Writing – review & editing: Fabio CP Navarro, Mark B. Gerstein.

References

1. Cost GJ, Feng Q, Jacquier A, Boeke JD. Human L1 element target-primed reverse transcription in vitro. EMBO J. 2002; 21: 5899–5910. https://doi.org/10.1093/emboj/cdf592 PMID: 12411507
2. Kulpa DA, Moran JV. Cis-preferential LINE-1 reverse transcriptase activity in ribonucleoprotein particles. Nat Struct Mol Biol. 2006; 13: 655–660. https://doi.org/10.1038/nsmb1107 PMID: 16783376
3. Ostertag EM, Kazazian HH. Biology of mammalian L1 retrotransposons. Annu Rev Genet. 2001; 35: 501–538. https://doi.org/10.1146/annurev.genet.35.102401.091032 PMID: 11700292
4. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. Nature Publishing Group; 2001; 409: 860–921. https://doi.org/10.1038/35057062 PMID: 11237011
5. Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, et al. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. 2000; 10: 1496–1508. https://doi.org/10.1101/gr.149400 PMID: 11042149
6. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet. 2011; 12: 615–627. https://doi.org/10.1038/nrg3030 PMID: 21850042
7. Hancks DC, Kazazian HH. Active human retrotransposons: variation and disease. Curr Opin Genet Dev. 2012; 22: 191–203. https://doi.org/10.1016/j.gde.2012.02.006 PMID: 22406018
8. Burns KH. Transposable elements in cancer. Nat Rev Cancer. Nature Publishing Group; 2017;: 1–10. https://doi.org/10.1038/nrc.2017.35 PMID: 28642606
9. Wang J, Song L, Grover D, Azrak S, Batzer MA, Liang P. dbRIP: a highly integrated database of retrotransposon insertion polymorphisms in humans. Hum Mutat. 2006; 27: 323–329. https://doi.org/10.1002/humu.20307 PMID: 16511833
10. Ewing AD, Kazazian HH. High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. 2010; 20: 1262–1270. https://doi.org/10.1101/gr.106419.110 PMID: 20488934
11. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature Publishing Group. 2015; 526: 75–81. https://doi.org/10.1038/nature15394 PMID: 26432246
12. Skowronski J, Singer MF. Expression of a cytoplasmic LINE-1 transcript is regulated in a human teratocarcinoma cell line. Proc Natl Acad Sci USA. National Academy of Sciences; 1985; 82: 6050–6054. https://doi.org/10.1073/pnas.82.18.6050 PMID: 2412228
13. Belancio VP, Roy-Engel AM, Deininger PL. All y'all need to know 'bout retroelements in cancer. Seminars in Cancer Biology. Elsevier Ltd; 2010; 20: 200–210. https://doi.org/10.1016/j.semcancer.2010.06.001 PMID: 20600922
14. Tubio JMC, Li Y, Ju YS, Martincorena I, Cooke SL, Tojo M, et al. Mobile DNA in cancer. Extensive transduction of nonrepetitive DNA mediated by L1 retrotransposition in cancer genomes. Science. 2014; 345: 1251343. https://doi.org/10.1126/science.1251343 PMID: 25082706
15. Muotri AR, Chu VT, Marchetto MCN, Deng W, Moran JV, Gage FH. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. Nature. 2005; 435: 903–910. https://doi.org/10.1038/nature03663 PMID: 15959507
16. Kano H, Godoy I, Courtney C, Vetter MR, Gerton GL, Ostertag EM, et al. L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev. 2009; 23: 1303–1312. https://doi.org/10.1101/gad.1803909 PMID: 19487571
17. Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P. Somatic expression of LINE-1 elements in human tissues. Nucleic Acids Res. 2010; 38: 3909–3922. https://doi.org/10.1093/nar/gkq132 PMID: 20215437
18. Evrony GD, Lee E, Mehta BK, Benjamini Y, Johnson RM, Cai X, et al. Cell lineage analysis in human brain using endogenous retroelements. Neuron. 2015; 85: 49–59. https://doi.org/10.1016/j.neuron.2014.12.028 PMID: 25569347
19. Hata K, Sakaki Y. Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene. 1997; 189: 227–234. https://doi.org/10.1016/s0378-1119(96)00856-6 PMID: 9168132
20. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009; 10: 691–703. https://doi.org/10.1038/nrg2640 PMID: 19763152

21. Erwin JA, Marchetto MC, Gage FH. Mobile DNA elements in the generation of diversity and complexity in the brain. Nature Publishing Group. 2014; 15: 497–506. https://doi.org/10.1038/nrn3730 PMID: 25005482

22. Doucet TT, Kazazian HH. Long Interspersed Element Sequencing (L1-Seq): A Method to Identify Somatic LINE-1 Insertions in the Human Genome. Methods Mol Biol. New York, NY: Springer New York; 2016; 1400: 79–93. https://doi.org/10.1007/978-1-4939-3372-3_5 PMID: 26895047

23. Deininger P, Morales ME, White TB, Baddoo M, Hedges DJ, Servant G, et al. A comprehensive approach to expression of L1 loci. Nucleic Acids Res. 2017; 45: e31–e31. https://doi.org/10.1093/nar/gkw1067 PMID: 27899577

24. Philippe C, Vargas-Landin DB, Doucet AJ, van Essen D, Vera-Otarola J, Kuciak M, et al. Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. Elife. eLife Sciences Publications Limited; 2016; 5: 166. https://doi.org/10.7554/eLife.13926 PMID: 27016617

25. Rangwala SH, Zhang L, Kazazian HH. Many LINE1 elements contribute to the transcriptome of human somatic cells. Genome Biol. BioMed Central Ltd; 2009; 10: R100. https://doi.org/10.1186/gb-2009-10-9-r100 PMID: 19772661

26. Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. Transcriptional landscape of repetitive elements in normal and cancer human cells. BMC Genomics. BioMed Central; 2014; 15: 583–17. https://doi.org/10.1186/1471-2164-15-583 PMID: 25012247

27. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, Rinn JL, et al. The reality of pervasive transcription. PLoS Biol. 2011; 9: e1000625–discussion e1001102. https://doi.org/10.1371/journal.pbio.1000625 PMID: 21765801

28. Jacquier A. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. Nat Rev Genet. 2009; 10: 833–844. https://doi.org/10.1038/nrg2683 PMID: 19920851

29. Lee H-G, Kahn TG, Simcox A, Schwartz YB, Pirrotta V. Genome-wide activities of Polycomb complexes control pervasive transcription. Genome Res. 2015. https://doi.org/10.1101/gr.188920.114 PMID: 25986499

30. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. Most "dark matter" transcripts are associated with known genes. Eddy SR, editor. PLoS Biol. Public Library of Science; 2010; 8: e1000371. https://doi.org/10.1371/journal.pbio.1000371 PMID: 20502517

31. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature Publishing Group. 2012; 489: 57–74. https://doi.org/10.1038/nature11247 PMID: 22955616

32. Consortium GTEx. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science. 2015; 348: 648–660. https://doi.org/10.1126/science.1262110 PMID: 25954001

33. Ovchinnikov I, Rubin A, Swergold GD. Tracing the LINEs of human evolution. Proc Natl Acad Sci USA. National Acad Sciences; 2002; 99: 10522–10527. https://doi.org/10.1073/pnas.152346799 PMID: 12138175

34. Bolotin E, Armendariz A, Kim K, Heo S-J, Boffelli D, Tantisira K, et al. Statin-induced changes in gene expression in EBV-transformed and native B-cells. Human Molecular Genetics. 2014; 23: 1202–1210. https://doi.org/10.1093/hmg/ddt512 PMID: 24179175

35. Caliskan M, Cusanovich DA, Ober C, Gilad Y. The effects of EBV transformation on gene expression levels and methylation profiles. Human Molecular Genetics. 2011; 20: 1643–1652. https://doi.org/10.1093/hmg/ddr041 PMID: 21289059

36. Min JL, Barrett A, Watts T, Pettersson FH, Lockstone HE, Lindgren CM, et al. Variability of gene expression profiles in human blood and lymphoblastoid cell lines. BMC Genomics. BioMed Central; 2010; 11: 96. https://doi.org/10.1186/1471-2164-11-96 PMID: 20141636

37. Klawitter S, Fuchs NV, Upton KR, Muñoz-Lopez M, Shukla R, Wang J, et al. Reprogramming triggers endogenous L1 and Alu retrotransposition in human induced pluripotent stem cells. Nat Commun. 2016; 7: 10286. https://doi.org/10.1038/ncomms10286 PMID: 26743714

38. Ogino S, Nosho K, Kirkner GJ, Kawasaki T, Chan AT, Schernhammer ES, et al. A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. J Natl Cancer Inst. 2008; 100: 1734–1738. https://doi.org/10.1093/jnci/djn359 PMID: 19033568

39. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. Cold Spring Harbor Lab; 2016; 26: 745–755. https://doi.org/10.1101/gr.201814.115 PMID: 27197217

40. Carreira PE, Richardson SR, Faulkner GJ. L1 retrotransposons, cancer stem cells and oncogenesis. FEBS J. 2014; 281: 63–73. https://doi.org/10.1111/febs.12601 PMID: 24286172

41. Spalding KL, Bhardwaj RD, Buchholz BA, Druid H, Frisén J. Retrospective birth dating of cells in humans. Cell. 2005; 122: 133–143. https://doi.org/10.1016/j.cell.2005.04.028 PMID: 16009139

42. Thomas CA, Paquola ACM, Muotri AR. LINE-1 retrotransposition in the nervous system. Annu Rev Cell Dev Biol. 2012; 28: 555–573. https://doi.org/10.1146/annurev-cellbio-101011-155822 PMID: 23057747

43. Muotri AR, Marchetto MCN, Coufal NG, Oefner R, Yeo G, Nakashima K, et al. L1 retrotransposition in neurons is modulated by MeCP2. Nature. 2010; 468: 443–446. https://doi.org/10.1038/nature09544 PMID: 21085180

44. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, Lovci MT, et al. L1 retrotransposition in human neural progenitor cells. Nature. 2009; 460: 1127–1131. https://doi.org/10.1038/nature08248 PMID: 19657334

45. Iskow RC, McCabe MT, Mills RE, Torene S, Pittard WS, Neuwald AF, et al. Natural mutagenesis of human genomes by endogenous retrotransposons. Cell. 2010; 141: 1253–1261. https://doi.org/10.1016/j.cell.2010.05.020 PMID: 20603005

46. Jeong H-H, Yalamanchili HK, Guo C, Shulman JM, Liu Z. An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. Pac Symp Biocomput. WORLD SCIENTIFIC; 2018; 23: 168–179. https://doi.org/10.1142/9789813235533_0016 PMID: 29218879

47. Jin Y, Tam OH, Paniagua E, Hammell M. TEtranscripts: A package for including transposable elements in differential expression analysis of RNA-seq datasets. Bioinformatics. Oxford University Press; 2015;: btv422. https://doi.org/10.1093/bioinformatics/btv422 PMID: 26206304

48. Upton KR, Gerhardt DJ, Jesuadian JS, Richardson SR, Sánchez-Luque FJ, Bodea GO, et al. Ubiquitous L1 Mosaicism in Hippocampal Neurons. Cell. The Authors; 2015; 161: 228–239. https://doi.org/10.1016/j.cell.2015.03.026 PMID: 25860606

49. Faulkner GJ, Garcia-Perez JL. L1 Mosaicism in Mammals: Extent, Effects, and Evolution. Trends Genet. 2017. https://doi.org/10.1016/j.tig.2017.07.004 PMID: 28797643

50. Abyzov A, Mariani J, Palejev D, Zhang Y, Haney MS, Tomasini L, et al. Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. Nature. 2012; 492: 438–442. https://doi.org/10.1038/nature11629 PMID: 23160490

51. Baillie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, et al. Somatic retrotransposition alters the genetic landscape of the human brain. Nature. 2011; 479: 534–537. https://doi.org/10.1038/nature10531 PMID: 22037309

52. Ewing AD, Gacita A, Wood LD, Ma F, Xing D, Kim M-S, et al. Widespread somatic L1 retrotransposition occurs early during gastrointestinal cancer evolution. Genome Res. Cold Spring Harbor Lab; 2015;: gr.196238.115. https://doi.org/10.1101/gr.196238.115 PMID: 26260970

53. Wang L, Rishishwar L, Mariño-Ramírez L, Jordan IK. Human population-specific gene expression and transcriptional network modification with polymorphic transposable elements. Nucleic Acids Res. 2017; 45: 2318–2328. https://doi.org/10.1093/nar/gkw1286 PMID: 27998931

54. Goeman JJ. L1 penalized estimation in the Cox proportional hazards model. Biom J. WILEY-VCH Verlag; 2010; 52: 70–84. https://doi.org/10.1002/bimj.200900028 PMID: 19937997

55. Tomasetti C, Vogelstein B. Cancer etiology. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science. American Association for the Advancement of Science; 2015; 347: 78–81. https://doi.org/10.1126/science.1260825 PMID: 25554788