Abstract

The broad distribution and high copy number of Long Interspaced Nuclear Element (LINE-1 or L1) elements across the human genome hinders the quantification of LINE-1 autonomous transcription. RNA sequencing analysis are specially confounded by RNA fragments emanating from pervasive transcription. We modeled and implemented a new approach that discern pervasive transcription from autonomous transcription of L1 subfamilies and estimate their autonomous transcription level. We processed thousands of RNA sequencing experiments to evaluate the autonomous activity of LINE-1 subfamilies in human cell lines, healthy organs and tumor tissue. We demonstrate that most of LINE-1 signal emanates from pervasive transcription, however recent and potentially active, LINE-1 subfamilies are transcribed in healthy tissues and tumors. Their transcription is found in cytoplasmic poly-adenylated transcripts. Moreover, we found that basal ganglia harbor higher activity of L1Hs than other adult brain regions, but, have relatively small autonomous transcription of L1 compared to tissues such as tibial nerve, testes, skin. Transcription of L1 is upregulated in most tumors when compared to their counterpart healthy tissue and is directly correlated to the number of indels in tumor samples, suggesting that emergence INDELs could be facilitated by the activity of L1 endonuclease and error prone DNA repair mediated by NHEJ.

Main

Long Interspersed Nuclear Element-1 (LINE-1 or L1) is a DNA sequence specialized in duplicating itself in a host genome {Ostertag:2001jl}. L1 elements are highly active in mammals germline {Wang:2006hr, Ewing:2010da, Schrider:2013di} and, due to its

replicative mechanism, it colonized approximately 17% of the human genome {Lander:2001hk}. L1 mobilization mechanism is based on retrotransposition of mRNA{Cost:2002ti, Kulpa:2006js, Ostertag:2001jl} creating thousands of - mostly inactive and truncated, copies across the genome, however, little is known about its activity in somatic tissues {Criscione:2014dp, Philippe:2016cx, Belancio:2010ie, Muotri:2005go, Kano:2009dt, Rangwala:2009bg, Hashimoto:2015ic}.

The phenomena known as pervasive transcription is defined as the transcription of regions well beyond the boundaries of known genes at varied transcription levels {BUZFClark:2011cc}. Pervasive transcription does not affect transcription quantification of protein coding genes since protein coding genes are present either as a single copy or low copy numbers in the genome. On the other hand, the transcription level quantification of transposable elements, including L1 elements, is specially affected by pervasive transcription due to its multi-copy nature hindering usage of short read technologies to assess the autonomous transcription of L1 elements.

Here we analyzed more than 9,000[N] RNA-Seq experiments from multiple datasets (Table 1). Twenty-five human tissues were investigated and the autonomous transcription of L1 subfamilies was estimated. We and others {GTExConsortium:2015fb} find that most of highly expressed and ancient repetitive elements – such as DNA transposons and LINE-2s; have their expression correlated to the expression of proximal genes. Implying that their transcription is mostly passive, and reliant on genomic vicinities to be transcribed (Supplementary Figure 1[N]). We find that most of L1 RNA sequencing reads

emanates from pervasive transcription, however, most of autonomous transcription signal derives from poly-adenylated and cytoplasmic transcripts. We further describe the human healthy somatic tissues that support high levels of L1 autonomous transcription and show the tumor samples consistently have up-regulation of L1 elements. The higher activity of L1 elements also correlates to the higher number of INDELs in tumors. Indicating that activation of L1 elements could create double strand breaks close to TTTAA motifs and increase genomic instability.

Due to their repetitive nature, recently amplified subfamilies such as L1Hs, are usually ignored transcript quantification essays due to insufficient read specificity mapping to L1 instances. We observe that, for most analyzed RNA sequencing experiments, the number of reads mapped to young L1 subfamilies elements correlates to the number of bases annotated as the respective L1 subfamily in the reference genome regardless of their recent evolutionary activity (Figure 1A – spearman ranked correlation c=0.94, p-value < 2.2e-16). We hypothesize that this correlation, the genomic-transcriptomic correlation, is consistent with pervasive transcription - a process that is known to transcribe random regions of the genome at low levels and, therefore, emanates RNA fragments proportionally to the number of copies of L1 subfamilies in the genome. We analyzed thousands of healthy tissue RNA sequencing experiments {GTExConsortium:2015fb} and calculated their genomic-transcriptomic correlation. We find a substantial number of samples from distinct tissues displays lower genomic-transcriptomic correlation (Figure 1B). We further hypothesized that samples with smaller genomic-transcriptomic correlation are the outcome of autonomous transcription of L1 coupled with pervasive

transcription, and, the deviation from the genomic-transcriptomic correlation would be an effect of RNA sequencing reads emanating from the autonomous transcription of L1 subfamilies (Methods for details). To distinguish signals originating from **autonomous** transcription of L1 subfamilies and **pervasive** transcription of L1 subfamilies we developed TeXP (Figure 1C), (available at [GITHUB]), a comprehensive suite that creates subfamily **fingerprints** and process RNA-seq experiments to estimate the proportion of reads emanating from pervasive and autonomous L1 transcription (Figure 1C).

We used TeXP to estimate the autonomous transcription of L1 subfamilies in well-established cell-lines with RNA sequencing experiments performed by ENCODE {ENCODEProjectConsortium:2012gc} (Table S1). We first noticed that despite adding fingerprints for 5 distinct L1 subfamilies (L1Hs, L1P1, L1PA2, L1PA3 and L1PA4 – Figure 2A) we mostly detect signals from pervasive transcription and L1Hs autonomous transcription. L1PA2 is also detectable, but at low levels when compared to L1Hs and pervasive transcription (Figure 2B - Figure S2). We find that MCF-7, a cell line derived from breast cancer, shows a remarkable high level of L1Hs transcription (180.7 RPKM) as in agreement with previous works {Philippe:2016cx}. We further investigated the transcription level of L1 subfamilies in different MCF-7 cell compartments and transcripts fractions. Using ENCODE datasets, we first compared MCF-7 WholeCell polyA+ and polyA- transcript fractions (Figure 2C). We find that, despite the number of reads overlapping L1 elements being similar (Figure S3) the selection of polyA+ transcripts yields 3.7x more autonomous L1Hs transcripts (RPKM) then PolyA- libraries, suggesting that most of the autonomous L1Hs signal is being generated by mature polyadenylated

transcripts. We further analyzed Cytoplasmic PolA+ fraction - a subset of WholeCell PolyA+ transcripts (Figure 2C). We found that the autonomous/pervasive ratio is approximately 0.45, in agreement to WholeCell PolyA+ fraction (0.51); therefore, suggesting that WholeCell PolyA+ signal is predominated by mature transcripts in the cytoplasm. As a control, we evaluated Nuclear PolyA+ fraction (Figure 2C). We found that nuclear transcripts show the smallest autonomous/pervasive ratio (0.02) and transcription level 30 times smaller than WholeCell PolyA+ fractions, therefore suggesting that a very small fraction of WholeCell PolyA+ L1Hs autonomous transcription is derived from RNA fragments from the nucleous (Figure S4).

We further analyzed ENCODE RNA-Seq datasets for other cell lines and found that GM12878, a lymphoblastic cell line derived from a healthy individual's blood, has no autonomous L1Hs regardless of the cell component. As well as the most of the compartments from K-562 and SKMEL5 (Figure 2D). However, in contrast to GM12878, SK-MEL-5 and K-562 are cancer derived cell lines and show transcription level of respectively 8.8 and 2.4 RPKM in whole-cell polyA+ datasets. Suggesting that cancer cell lines can have a large range of L1Hs autonomous transcription. With these results in mind we processed GTEx RNA sequencing experiments from K562 and Epstein-Barr Virus transformed cell-lines derived from healthy tissues. K-562 samples showed remarkable consistency across different GTEx batches, with median RPKM at 12.14 (1.47 RPKM standard deviation – Figure S6) and similar levels to ENCODE K-562.

Both EBV transformed cell-lines – lymphoblastic and fibroblastic cell lines; show very distinct patterns of L1Hs autonomous transcription. While lymphoblast (blood derived)

cell-lines have no or very little transcription of L1Hs (Figure S6) with approximately 84% of samples with RPKM equal to zero, fibroblastic cell lines (skin derived) show consistent higher L1Hs autonomous transcriptions (median 1.5 RPKM) and 58.7% of samples with L1Hs RPKM higher than 1. We also tested whether our estimation of L1Hs transcription level is correlated to genes containing or adjacent to L1Hs instances. We found no significant difference between the correlation distribution of a random set of genes or genes with L1Hs in exons, introns, 3kb upstream or 3kb downstream. Suggesting that our estimation of L1Hs autonomous transcription level is not significantly influenced by non-autonomous L1Hs transcription adjacent or contained by protein coding genes loci.

We analyzed the levels of autonomous transcription of L1 subfamilies of 7,429 GTEx samples (Table S2). One hundred twenty nine samples were removed from further analysis because there was not enough reads overlapping L1 elements. We find that only L1Hs is autonomously transcribed while L1P1, L1PA2, L1AP3 and L1PA4 have only residual or spurious autonomous transcription in healthy tissues (Figure S5). Overall, we found that healthy tissues have a narrower range of L1Hs autonomous transcription level than cancer cell lines. While the highest L1Hs autonomous transcription in healthy tissues was 46.66 RPKM (Figure 2B – L1Hs RPKM histogram) cancer cell lines reached 180 RPKM. Conversely, many GTEx RNA sequencing experiments had no or very small (<1 RPKM) evidence of L1Hs autonomous transcription, 2,520 (34.3%). Interestingly, when we compared skin and blood samples to the cell lines derived from these tissues we find a similar pattern of L1Hs autonomous transcription level. While only one sample from Skin have L1Hs autonomous transcription level smaller than 1 RPKM, most (74.6%) of

the Whole blood samples have no transcriptional activity of L1Hs. Suggesting that the EBV transformed cell-lines partially preserve the L1Hs transcription level from the tissue of origin.

Surprisingly, adult brain tissue samples are, regardless of brain region, amongst the healthy tissues with lesser L1Hs autonomous transcription (Figure 2B). Other tissues such as, Liver, Pancreas and Spleen, also have very little or no autonomous transcription of L1Hs (respectively 91.2%, 82.9%, 88.9% of samples have RPKM < 1). Most of the brains regions show a notably low autonomous transcription of L1Hs except for regions related to the Basal Ganglia. The Putamen and Caudate regions show consistently higher levels – but still low compared to other tissues; 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). Among the tissues with higher activity of L1Hs we highlight Nerve(Tibia), Skin (both exposed and not exposed to sun), Prostate, Lung, Vagina and Testis - that was previously reported as one of the tissues supporting genomic mobilization of L1Hs (Figure 2).

Based on the signal emanating from pervasive transcription we could estimate a pervasive transcription index for each RNA sequencing experiment. We define the pervasive transcription index as the number reads overlapping L1 subfamilies and emanating from pervasive transcription, normalized it by the total number of aligned reads in a RNA sequencing experiment. Overall we find that testis and cerebellum are amongst the tissues with highest pervasive transcription level (median 1,056 and 906.3 PI respectively). Conversely, Whole blood and Skeletal Muscle are amongst the tissues with

the lowest levels of pervasive transcription (134.9 and 223.8 PI respectively) (Figure S7). Interestingly, tissues with smaller Pervasive Index are also described as tissues with low transcriptional diversity {GTExConsortium:2015fb}, suggesting that pervasive transcription index might be a good proxy for tissue transcription diversity.

Having access to samples' phenotype we asked if the autonomous transcription of L1Hs is correlated to sample Age or Body Mass Index (BMI). We find that most of tissues, most likely due to their low levels of L1Hs autonomous transcription (Figure 2) do not significantly correlate with age. Intriguingly, we find that Prostate and Whole Blood samples have an inverse correlation with age, and prostate samples having the higher L1Hs transcriptional activity at 20-30 years old individuals. All other tissues with 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. Other tissues with relatively high autonomous transcription of L1 show no correlation at all (Tibial Nerve and Ovary). (Figure 2B – Table S3). Despite BMI being recently reported as inversely correlated with the methylation of L1 elements we only found a correlation between L1Hs transcriptional activity and BMI in Breast tissue (corr=0.23 FDR=0.046; Table S4). We finally tested if samples of skin exposed to sun does show any significant enrichment of L1Hs autonomous transcription compared to Skin not exposed to Sun. We found that both group of samples (exposed and not exposed) have similar high levels of L1Hs autonomous transcription, with slightly (and not significantly) higher L1Hs activity in samples of tissue exposed to the sun.

We further investigated the impact of L1 autonomous transcription in cancer samples. We hypothesized that tissues with higher transcription of L1 elements in a healthy context could have higher susceptibility to activation L1Hs activity and consequent genomic instability mediated by L1 reverse transcriptase. We found that recent literature supports our hypothesis. For example, Tubio et. al. describe Lung Cancer as the tumor type with have highest rates of L1 mobilization {Tubio:2014gm}. We investigated the autonomous transcription level of L1Hs of over 2,500 cancer samples originating from 6 tumor types. Namely we evaluated the autonomous transcription level of L1Hs in Lung (LUAD and LUSC), Prostate Adenocarcinoma (PRAD), Brain – Lower Grade Glioma (LGG), Thyroid Carcinoma and Skin Cancer Melanoma (SKCM). We found that SKCM (Skin cancer) supports autonomous L1Hs transcription at levels similar to healthy tissue (2.38 times lower). On the other hand, tumor derived from Lung consistently have higher levels of L1Hs autonomous transcription in the cancer counterparts reaching up to 13 times higher expression in Lung Squamous Cell Carcinoma (LUSC - Figure S8). We predict that these genomes would consistently have higher genomic instability due to the activity of L1Hs endonuclease. To test this hypothesis, we assessed the frequency INDELs in the genome of samples that we analyzed L1Hs autonomous transcription. In total, we analyzed somatic INDELs from 2,504 tumors. We selected samples from TCGA with tumor origin on Lung, Skin, Thyroid and Prostate to search for signatures originating from L1Hs endonuclease activity. Namely we investigated the occurrence of INDELs close to the sequences recognized L1 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 the Non-Homologous End Joining

(NHEJ) pathway{ODriscoll:2006cz}. The NHEJ pathway is known to be error-prone, specially in the tumoral context, creating small insertions and deletions 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 have a significant correlation between autonomous L1 transcription (Figure 3B) and number of INDELs (Figure 3C), all samples have a significantly high correlation (0.49, p-value < 2.2e-16). To further investigate if L1 endonuclease could be driving an enrichment of INDELs we tested whether there is an enrichment of the L1 endonuclease target motif (TTTAA) in sequences flanking INDELs. We found that regardless the tissue of origin, there is an enrichment o the motif TTTAA in the 50nt flanking the INDEL. We further select motifs closer to the INDEL coordinate (-3;+3nt) and found that the effect is even more pronounced (Figure 3D). Finally, we evaluated the distribution of the endonuclease target motif across neighbor regions collectively. We found that most TTTAA motifs are concentrated around position 0 or 1, meaning that they perfectly overlap the breakpoint of INDELs for both insertions (Figure 3E) and deletions (Figure 3F).

Methods

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

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

GTEx raw RNA sequencing data.

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

ENCODE raw RNA sequencing data.

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

TeXP model.

TeXP models the number of reads overlapping L1 elements as 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:

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

Where T is the total number of reads mapped to L1 instances, P_{L1Hs} defines the proportion of bases 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 to the **Equation1**:

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

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 informations from fastq raw files such as read length and quality encoding. Fastq filies 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. 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 $P_{i,pervasive}$ as the proportion of bases annotated as the subfamily i:

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

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 full length instances based on RepeatMasker annotation and the reference genome (hg38). Read from putative full length 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 median percentage of counts across all simulations:

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

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 tends 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 $M_{i.j}$ generated for each RNA sequencing experiment we estimate the signal proportion ε and ϵ in **Equation 1** 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. Source code is available at https://github.com/fabiocpn/TeXP. A docker image is also available for users at dockerhub under fnavarro/texp.

L1 endonuclease motif enrichment analysis

Number of 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 5nt is shown here. Using Agilent capture was used to define exomic region. The same number of indels for each cancer type was simulated across the exomic (as defined above) and we estimated the expected number INDELs close to the indel breakpoing by counting the number of simulated indels close to the TTTAA motif. The statistical significance of the enrichment of TTTAA motif was calculated using chi-squared test.

References

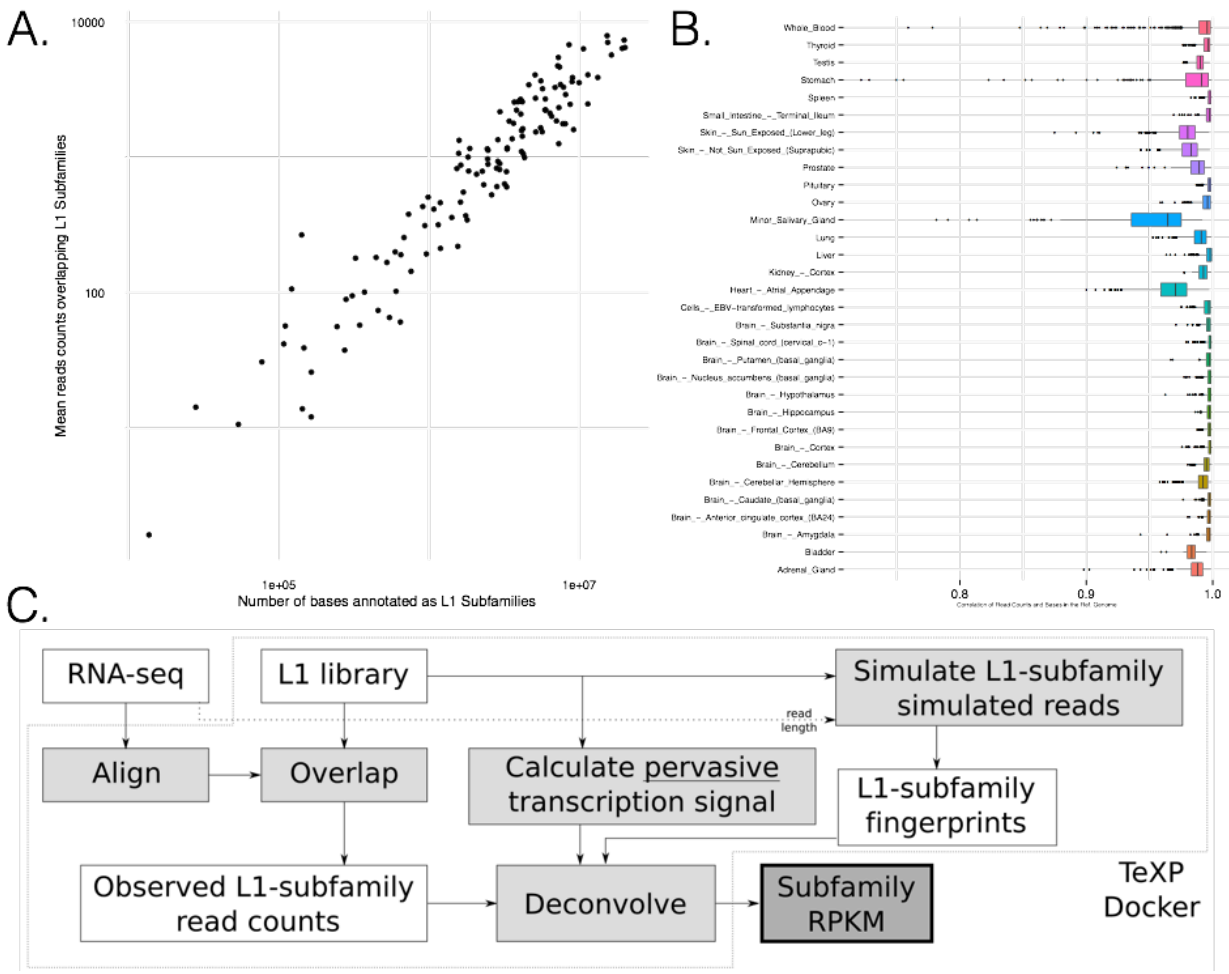

Figure 1. RNA sequencing experiments have a high genomic-transcriptomic correlation where the number of reads overlapping L1 subfamilies is proportional to the number of bases annotated as the subfamily (A). (B) Human healthy tissues show varied distributions of genomic-transcriptomic correlation. (C) TeXP pipeline description.

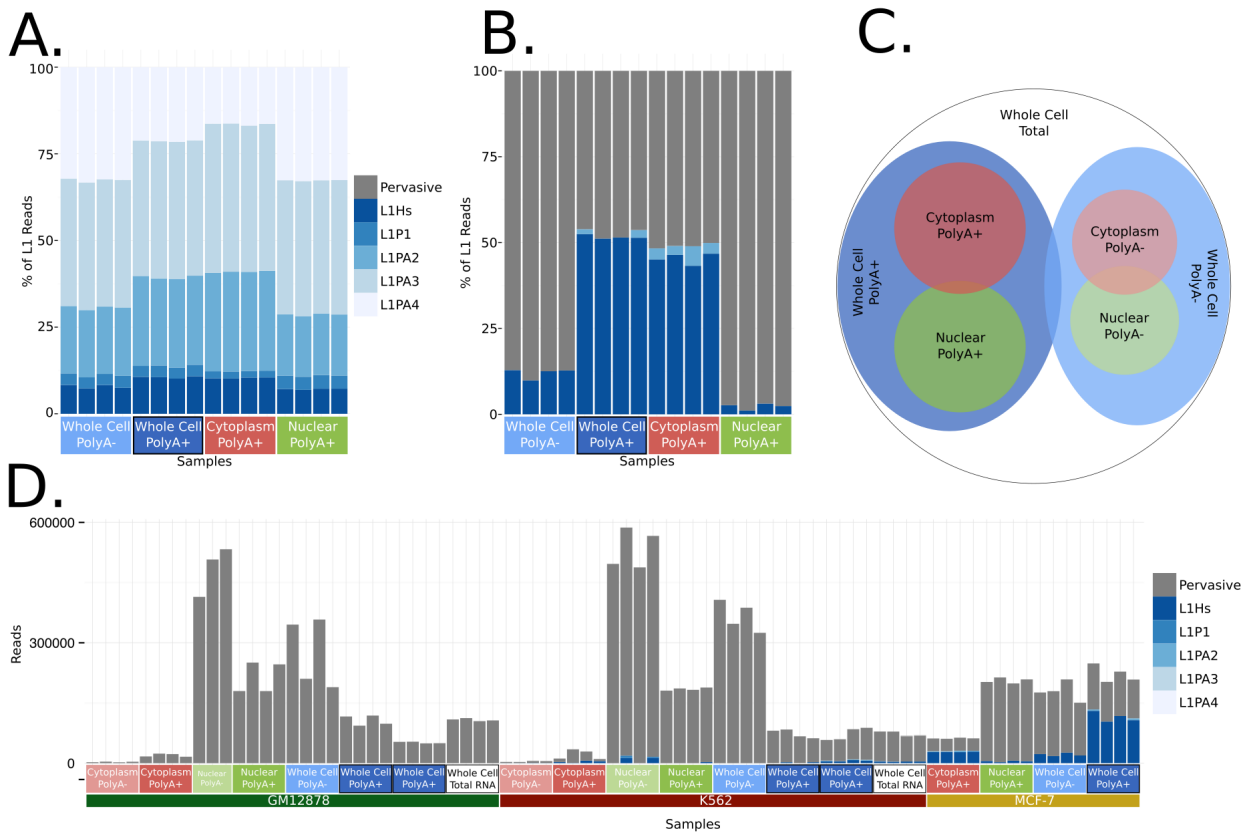

Figure 2. (A) Fraction of reads overlapping L1P1, L1PA2, L1PA3, L1PA4 and L1Hs Subfamilies in MCF-7 RNA sequencing experiments from different transcript partitions. (B) Fraction of reads assigned to pervasive transcription and L1 subfamilies autonomous transcription in MCF-7 RNA sequencing experiments from different transcript partitions. (C) Schematics of MCF-7 RNA fractions. (D) Absolute number of reads emanating from pervasive transcription and L1 subfamilies autonomous transcriptions from three ENCODE cell lines (GM12878, K562 and MCF7) for distinct RNA fractions.

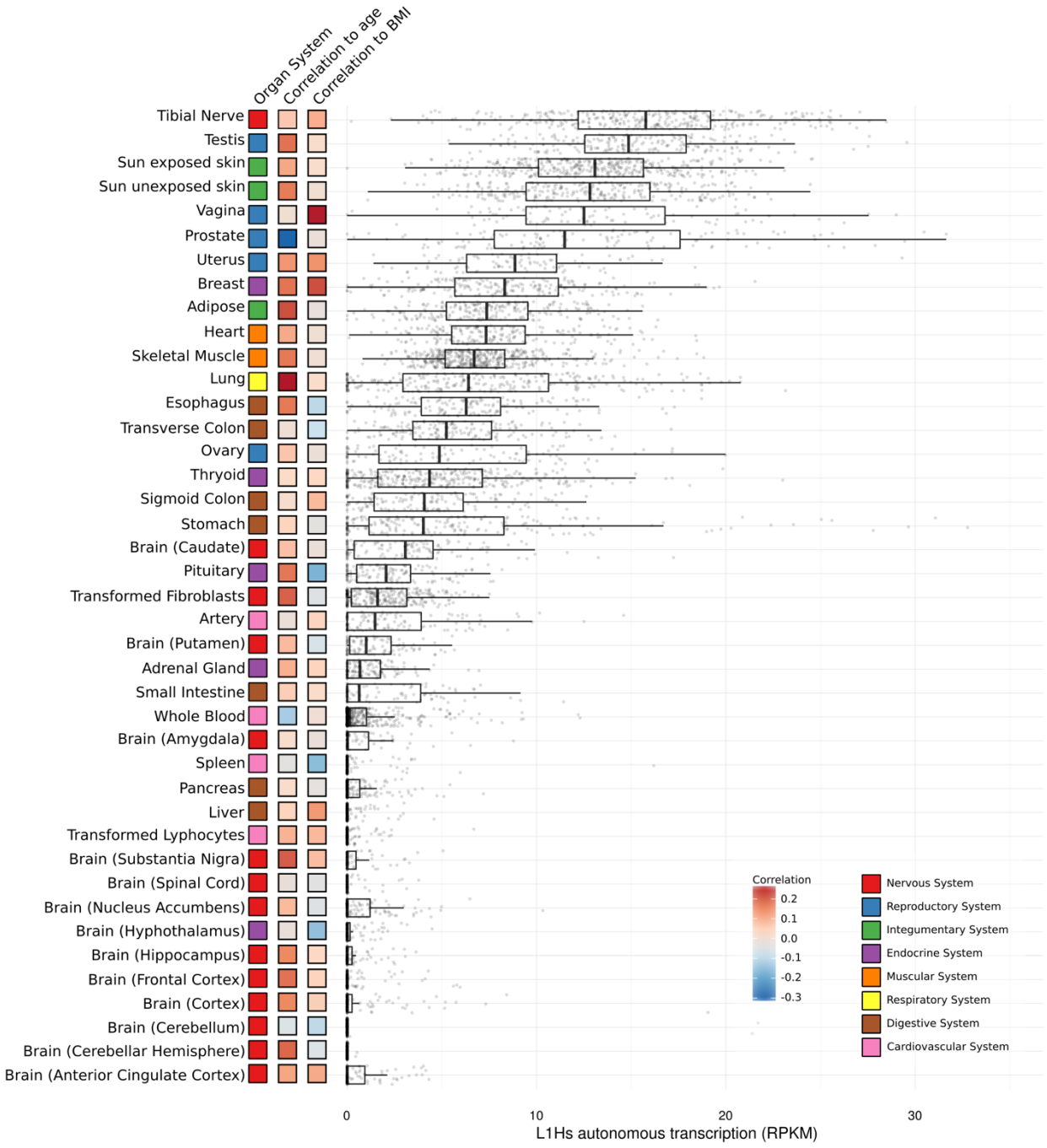

Figure 3. L1Hs autonomous transcription level on Human healthy somatic tissues. Each point is a RNA sequencing experiment, separated by tissue of origin. Coloring indicates the organ system of the tissue, correlation of L1Hs autonomous transcription and age, and correlation of L1Hs autonomous transcription and BMI respectively.

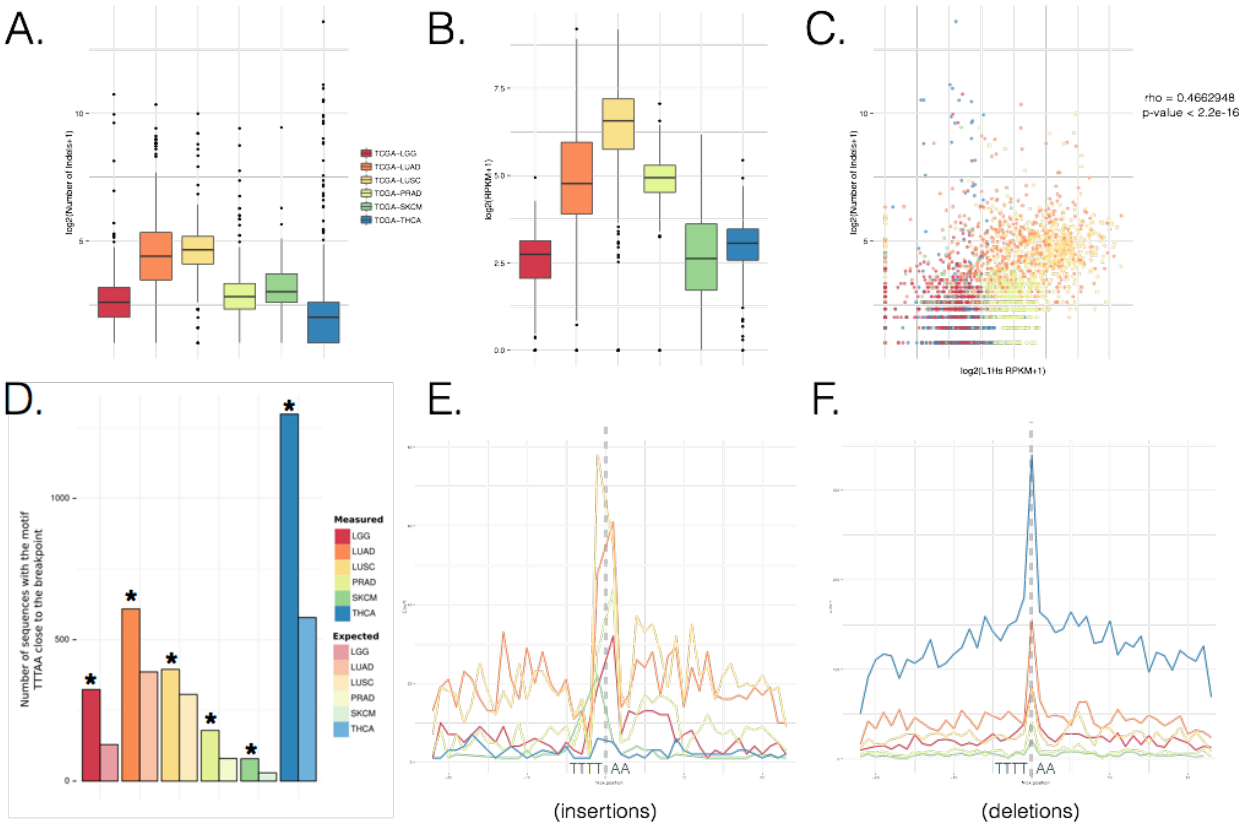

Figure 4. (A) Distribution of L1Hs autonomous transcription level in tumor samples from six cancer types. (B) Distribution of the number of INDELs in tumor samples from six cancer types. (C) Correlation between L1Hs autonomous expression and the number of INDELs in tumor samples. (D) Overrepresentation of TTTAA motif close to (-3|+3nt) INDELs (dark) compared to null (light). (E) Overrepresentation of the TTT|AA in the INDEL breakpoint on small insertions (F) Overrepresentation of the TTT|AA in the INDEL breakpoint on small deletions.