LINE-1 Retrotransposon expression in cancerous, epithelial and neuronal cells revealed by 5'-single cell RNA-Seq

Wilson McKerrow[1,2], Shane A. Evans[3], Azucena Rocha[4], John Sedivy[4,5], Nicola Neretti[4], Jef D. Boeke[1,2,6], and David Fenyö[1,2,*]
[1]Institute for Systems Genetics, NYU Langone Health, New York, NY, USA
[2]Department of Biochemistry and Molecular Pharmacology, NYU Langone Health, New York, NY, USA
[3]Center for Computational Molecular Biology, Brown University, Providence, RI, USA
[4]Molecular Biology, Cell Biology, and Biochemistry, Brown University, Providence, RI, USA
[5]Center on the Biology of Aging, Brown University, Providence, RI, USA
[6] Department of Biomedical Engineering, NYU Tandon School of Engineering, Brooklyn 11201, NY, USA

*Correspondence: David@Fenyolab.org

Abstract

LINE-1 retrotransposons are known to be expressed in early development, in tumors and in the germline. Less is known about LINE-1 expression at the single cell level, especially outside the context of cancer. Because LINE-1 elements are present at a high copy number, many transcripts that are not driven by the LINE-1 promoter nevertheless terminate at the LINE-1 3' UTR. Thus, 3' targeted single cell RNA-seq datasets are not appropriate for studying LINE-1. However, 5' targeted single cell datasets provide an opportunity to analyze LINE-1 expression at the single cell level. Most LINE-1 copies are 5' truncated, and a transcript that contains the LINE-1 5' UTR as its 5' end is likely to have been transcribed from its promoter. We developed a method, L1-sc (LINE-1 expression for single cells), to quantify LINE-1 expression in 5' targeted 10x genomics single cell RNA-seq datasets. Our method confirms that LINE-1 expression is high in cancer cells, but low or absent from immune cells. We also find that LINE-1 expression is elevated in epithelial compared to immune cells outside of the context of cancer and that it is also elevated in neurons compared to glia in the mouse hippocampus.

Introduction

In the human genome, LINE-1 is the only family of retrotransposons – sequences capable of copying themselves to new genomic loci via an RNA intermediate – that remains active and autonomous[1]. LINE-1 is expressed during early embryonic development, but is (at least mostly) repressed in healthy somatic tissues. In about 50%

of tumors, LINE-1 is de-repressed, leading to high expression of its ORF1 protein and frequent retrotransposition[2,3]. LINE-1 can insert into and disrupt key cancer suppressor genes[4,5] or mediate large scale structural variation[3]. While the evidence is clear that LINE-1 expression is far higher in cancer compared to normal, several studies indicate that a small amount of LINE-1 expression can be detected in normal tissues[6–8].

Single cell RNA-seq (scRNA-seq) data presents an opportunity to identify LINE-1 expressing cells both in cancer and in healthy tissue. Unfortunately, the most popular scRNA-seq method (10x Genomics 3' chemistry) is unsuitable for detection of LINE-1 and other retrotransposon transcripts, many of which are 6kbp or longer. In humans, only the human-specific L1Hs family of LINE-1 remains transcriptionally and transpositionally active. However, the sequences of the primate specific L1PA2-4 families are highly similar to L1Hs and as a result, alignment algorithms will distribute some L1Hs reads among these families[8]. From L1Hs and L1PA2-4 alone, the 3' end of LINE-1 is present in 25,000 copies in the human reference genome (hg38). As a result, any noise from DNA contamination or readthrough transcription into 5' truncated element copies will be massively amplified. Since most of the element copies are defective and 5' truncated, detecting the signal of transcription initiating from the LINE-1 promoter should be much easier to detect by using sequencing focused on transcript 5' ends; indeed, the 5' end of L1Hs/L1PA2-4 is only present in 2,000 copies. This 5' vs 3' imbalance means that 5' targeted scRNA-seq is much more promising for the analysis of LINE-1 expression. Knowing that a transcript starts at the 5' UTR strongly supports its being a LINE-1 RNA, whereas a transcript containing the LINE-1 3' UTR may simply coincidentally terminate at one of the many LINE-1 3' UTR polyadenylation sites.

In this study, we present L1-sc (LINE-1 expression for single cells) a simple method based on our previous algorithm, MapRRCon[9], to quantify LINE-1 RNA expression in single cells sequenced on the 10x genomics platform with a 5' targeted library preparation. We validated this method on tumor data, confirming that LINE-1 expression in cancer cells is readily detected by this method. Then we applied this method to non-tumor data from 15 healthy tissues, where it revealed that LINE-1 expression can be detected, and is elevated in epithelial cells over other cell types. We further confirmed LINE-1 expression in skin epithelial cells (keratinocytes) by analyzing primary cell culture data from the ENCODE project. Finally, we looked for LINE-1 expression in the mouse hippocampus where we found it to be higher in neurons compared to glia.

Results

3' targeted scRNA-seq is not appropriate for LINE-1 quantitation

To count LINE-1 transcripts in 10x Genomics data, we built a custom LINE-1 CellRanger transcriptome. Beginning with the hg38 human reference genome, we masked all LINE-1 sequences from the L1Hs and L1PAx (i.e. L1PA2, L1PA3, etc.)

subfamilies. We then added back the L1Hs consensus sequence as a decoy chromosome. We also included the dfam[10] sequences that are available from L1PA families as additional decoys. Most of these cover only the 3’ end of LINE-1, but 5’ end sequences for L1PA10 and older elements are also included. We truncated the consensus LINE-1 sequences to omit the polyA tract as this could lead to misalignments. Rather than spreading LINE-1 aligned reads throughout the genome, this method forces the aligner to collect them in one place, making them much easier to quantify. We then appended an “L1Hs” transcript covering the full-length L1Hs consensus sequence (excepting polyA). We used CellRanger’s count function to generate the raw expression matrix (including LINE-1). Then we used the Seurat package for visualization and analysis[11].

We applied this method to 3’ targeted lung adenocarcinoma (LUAD) scRNA-seq pooled from 8 primary tumor samples[12]. We identified T cells (CD3E+), B cells (MS4A1+), plasma cells (MZB1+), macrophages (LYZ+), mast cells (KIT+), fibroblasts (COL1A2+), endothelial cells (PLVAP+) and epithelial/cancer cells (KRT8+). Cell types are shown in figure 1A and marker gene expression in figure S1. Looking at LINE-1 aligning reads, we did see a strong peak in read coverage at the 3’ end of the L1Hs decoy sequence, indicating that we are able to identify transcripts terminating at the LINE-1 3’ UTR (figure S2A). However, we also saw a high level of inferred “LINE-1” expression across all cell types (figure 1B). This is an unreasonable result given that previous IHC experiments show that LINE-1 is specifically overexpressed in cancer cells[2], and our own analyses of tumor transcriptomes and proteomes indicate that LINE-1 expression is highly enriched in tumors compared to adjacent normal tissue[13]. Thus, we concluded that, in this analysis, any signal from LINE-1 has been almost completely overwhelmed by the presence of other transcripts that terminate at or near the 3’ end of LINE-1. These are likely the result of readthrough/pervasive transcription and/or cryptic splicing into or just upstream of LINE-1 loci.

5’ targeted scRNA-seq from a tumor shows that LINE-1 expression is cancer/epithelial cell specific

Given our lack of success in seeing cancer cell specificity with 3’ targeted data, we sought alternatives and turned our attention to 5’ targeted single cell RNA-seq data. We started by looking at the example lung squamous cell carcinoma (LSCC) 5’ scRNA-seq data provided by 10x Genomics[14]. Read alignments to the L1Hs consensus sequence showed a jagged peak at the 5’ end of L1Hs that we believe to be a meaningful signal, but also a smaller peak near the 3’ end which we believe reflects a form of background related to the excessively high copy number of LINE-1 3’ sequences in the genome/transcriptome (figure S2B). We observed a similar profile for the non-tumor 5’ targeted dataset (figure S2C, next subsection). Given the high copy number of LINE-1 copies in the genome, we reasoned that some reads aligning to

LINE-1 may not actually be from RNA molecules that initiate at the LINE-1 5' UTR. To filter out such noise we tried truncating the L1Hs transcript to various lengths: 100 bases, 150 bases, 200, 250, 300, 350, 400, 450, 500, 1kb, 2kb and full length. In each case we found higher LINE-1 expression in two KRT7+ clusters that we determined to be cancer/epithelial cells, and sporadic expression in the other (KRT7-) clusters. We chose to proceed with the 150 base L1Hs annotation as it yielded the greatest enrichment in LINE-1 expression for KRT7+ (cancer) over KRT7- (microenvironment) cells (10.9x).

In this dataset we identified T cells, B cells, NK cells, plasma cells, macrophages, mast cells, fibroblasts, epithelial/cancer cells and mitotic cells (figure 2A). As expected, we found consistent expression of LINE-1 in epithelial/cancer cells (73%), but only sporadic expression in immune cells, ranging from 4.4% in macrophages to 10.5% in plasma cells (figure 2B,C). This is in stark contrast to the results from 3' targeted sc-RNAseq, indicating that LINE-1 can only be measured using a 5' targeted library (figure 2D).

LINE-1 expression is also detected in non-tumor epithelial cells

The elevated LINE-1 expression in cancer cells and the barely detectable expression in immune cells provides strong evidence that LINE-1 expression can be accurately identified from 5' targeted scRNA-seq. We therefore wanted to know whether this method could shed light on the longstanding question of whether LINE-1 is expressed in non-cancerous tissues and cell types, and if so, to what extent. A recent study performed 5' targeted scRNA-seq on 15 human tissues from a single donor[15]. We first wanted to know whether LINE-1 is expressed in these samples and if so, in which cell types. To that end, we pooled cells across the 15 tissues and reanalyzed this data for LINE-1 expression using our L1sc tool. The most prevalent cell types in this dataset were: T cells (CD3E+), B cells (MS4A1+), plasma cells (MZB1+), macrophages (LYZ+), muscle cells (ACTA2+), fibroblasts (COL1A2+), endothelial cells (PECAM+) and epithelial cells (KRT 5/8/10+) (figures 3A, S4). Cell types clustered together across tissue of origin, excepting epithelial cells, which clustered separately (figure 3B). This reflects the tissue specificity of epithelial cells (e.g. keratinocytes in the skin, enterocytes in the intestine or hepatocytes in the liver.) We found that while LINE-1 expression is indeed lower (34%) in these non-tumor tissues compared to the LSCC tumor, LINE-1 expression is still prevalent in epithelial cells, with about 30% being positive for at least one LINE-1 read (figure 3C). In contrast, LINE-1 expression is only detected in 5% of non-epithelial cells. Overall LINE-1 expression is 6.8x higher in epithelial compared to non-epithelial cells. Epithelial cell LINE-1 expression is fairly consistent across tissue of origin, but is lower in skin and small intestine where only 20% and 18% of epithelial cells are positive for LINE-1 expression, respectively (figure 3D).

Additional evidence for LINE-1 expression in keratinocytes (skin epithelia)

While one study did show ORF1p staining in skin and esophagus by immunohistochemistry (IHC)[7], and several studies report evidence for LINE-1 RNA in healthy tissues using less specific methods[6,8], LINE-1 expression in epithelial cells has not been widely described. Therefore, we used primary cell culture data for epithelial and non-epithelial cells from the ENCODE project to perform an independent validation. Publicly available ENCODE human primary culture RNA-seq data includes 3 keratinocyte (skin epithelial cell) samples along with 2 from B cells, 2 from CD14+ monocytes, 6 from endothelial cells, 1 from mammary epithelia, 2 from lung fibroblasts and 2 from myoblasts. All are poly A enriched and strand specific. We used L1EM[16] to identify the expression of full-length mRNAs from the active LINE-1 subfamily (L1Hs) in these samples. Primary keratinocyte (KC) cultures had the highest LINE-1 expression at 4.7 fragments per million (FPM) on average in this dataset. We also identified full length L1Hs transcripts in both B cell samples (2.0 FPM on average) and in one of the two myoblast samples (1.0 FPM). LINE-1 mRNAs were not detected above noise in the other samples (figure 4A). The fact that we found the highest LINE-1 expression in KCs and that LINE-1 expression was low or absent in non-epithelial subtypes supports the results of our single cell analysis. A lack of LINE-1 expression in mammary epithelia was surprising, but 5' single cell RNA-seq from breast tissue was not among those considered.

As a further validation to ascertain whether we could also identify KC LINE-1 expression in an independent dataset, we considered a second publicly available 5' targeted single cell RNA-seq dataset from skin (without cancer)[17]. Excluding samples that were sorted for a specific cell type or analyzed with 3' targeted library, this study included 2 healthy skin samples and one from a patient with drug induced eosinophilia and systemic symptoms (DRESS). However, overlapping cell types in the three samples did not always co-cluster and only the DRESS sample had a large number of KCs.

This sample includes KCs (KRT5+ or KRTDAP+), T cells (CD3E+), macrophages (LYZ+), fibroblasts (COL1A2+), endothelial cells (PLVAP+), smooth muscle (ACT2A+) and mitotic cells (MKI67+). The KC clusters can be further subdivided into proliferating (MKI67+), basal (KRT5+), squamous (KRTDAP+), and inflamed (IL1B+). See figure 4B for cell types and figure S5 for marker gene expression. Consistent with the results above, 33% of basal and squamous KCs were positive for LINE-1 expression, but only 8% of other cells were. Excepting the mitotic cells, LINE-1 expression is 2.2x higher in KC compared to other cells

LINE-1 expressing rectal epithelial cells also express genes associated with intronless mRNA export, heat shock and DNA damage

We next performed a more in-depth investigation of LINE-1 expression in rectal epithelial cells, as these had the highest LINE-1 detection rate (figure 3D). Two types of epithelial cells were identified in the rectal sample: goblet cells (SPINK4+) and enterocytes (CA2+) (figures 5A, S6). LINE-1 expression is higher in enterocytes (figures 5B,C), with 54% of cells being positive for LINE-1. We then wanted to know what other genes are expressed in LINE-1 positive rectal enterocytes. To that end, we performed differential expression between enterocytes in which LINE-1 was detected and those where it was not. We then performed GSEA[18] using Reactome[19], PID[20], WikiPathways[21], KEGG[22] and BioCarta[23] gene sets. At FDR < 5%, this yielded 28 significantly enriched gene sets (full list in table S1). Many of these are related to nuclear export generally and the export of intronless mRNAs (LINE-1 is such a mRNA). Specifically, the first and fourth most strongly enriched gene sets were “Interactions of Rev with host cellular proteins” (Reactome) and “Transport of mature mRNAs derived from intronless transcripts” (Reactome). Rev is an HIV protein required for export of unspliced HIV RNAs. Highly enriched gene sets also included two related to heat stress (“cellular response to heat stress” (Reactome) and “regulation of hsf1 mediated heat shock response” (Reactome)) and two related to DNA damage (“sumoylation of dna damage response and repair proteins” (Reactome) and “ATR pathway” (PID)).

Antisense LINE-1 transcription

In addition to the sense promoter that can yield retrotransposon competent LINE-1 RNAs, the LINE-1 5’ UTR also includes an antisense promoter[24]. LINE-1 antisense transcripts encode an ORF0 protein that is proposed to aid retrotransposition and can splice into flanking coding regions[25]. Because the 10x genomics data we looked at is strand specific, it is also possible to quantify antisense LINE-1 expression in these cells. To that end, we added an antisense LINE-1 transcript that also covers the first 500 bases of L1Hs, but is oriented in the reverse direction, to our database. Figure 4 shows the antisense LINE-1 expression in the tumor and non-tumor datasets analyzed. Overall, we observe that antisense LINE-1 transcripts are most prevalent in epithelial cells, but at a lower level. In the LSCC sample, antisense LINE-1 expression is about 70% of sense LINE-1 expression and is 5.1 times higher in epithelial/cancer cells compared to other cells (figure 6A). In the non-tumor samples, antisense LINE-1 expression is about 60% of sense LINE-1 is expression and is 2.1 times higher in epithelial compared to normal cells (figure 6B). Antisense LINE-1 expression is most prevalent in the bladder where it is detected in 32% of epithelial cells, and is 3.3 times more abundant compared to other epithelial tissues.

We then wanted to know whether sense and antisense LINE-1 positive cells overlap more than would be predicted by cell type and the number of genes detected in each cell. To that end, each epithelial cell was given two 0 or 1 (binary) labels depending on whether they are positive for sense and antisense reads. We then calculated the partial correlation between the sense and antisense labels with respect to the number of features quantified in each cell. We found significant and positive correlations in both the tumor and the non-tumor epithelial cells, but the partial correlation is much higher in the LSCC tumor (0.22) than in the non-tumor epithelial cells (0.05).

LINE-1 expression in mouse hippocampal neurons

It has been proposed that neuronal lineages in the hippocampus are an exception to the rule that LINE-1 expression and retrotransposition is (at least mostly) repressed in somatic tissues[26–28]. To explore this, we adapted L1sc for the analysis of 5' targeted single nucleus (sn) RNA-seq from the mouse hippocampus. As with the human version of L1-sc, we masked the mouse specific L1Md elements from the mm39 mouse reference genome and then added back the L1Md subfamily consensus sequences available from dfam[10]. Unlike in humans, mice have multiple active LINE-1 subfamilies, from the L1MdTf, L1MdGf and L1MdA lineages[29]. We therefore appended 8 LINE-1 transcripts to the mouse transcriptome, each including the first 150 bases of: L1MdTf_I, L1MdTf_II, L1MdTf_III, L1MdGf_I, L1MdGf_II, L1MdA_I, L1MdA_II and L1MdA_III.

Astrocytes, microglia, oligodendrocytes, oligodendrocyte progenitor cells (OPCs), neurons and interneurons were identified in the mouse hippocampus (figure 7A). Overall the L1MdTf family LINE-1 was the most readily detected. We therefore focused on the two youngest Tf subfamilies (I and II) and combined their signal as they are ~99% identical. 11% of neurons and interneurons were positive for L1MdTf I/II compared to only 3% of glia (astrocytes, microglia, oligodendrocytes and OPCs). Overall L1MdTf I/II are expressed 2.5x higher in neurons and interneurons compared to glia (figure 7B,C).

Discussion

Single cell RNA sequencing (scRNA-seq) is a powerful tool that has generated considerable excitement and a range of new insights. However, this technology has by and large not been applied to questions in retrotransposon biology. Here we have shown that while it is difficult, if not impossible, to identify LINE-1 expression in the much more commonly performed 3' targeted scRNA-seq procedure, we can readily quantify LINE-1 in single cells from 5' targeted data using our L1-sc method. This makes it possible to identify LINE-1 expressing cells and analyze the heterogeneity in LINE-1 expression both in cancer and non-cancer tissues.

In the tumor sample we analyzed, LINE-1 expression was nearly universal across cancer cells, but nearly absent from immune cells. This reinforces the effectiveness of our method and reflects what we already know to be true from bulk tissue RNA and proteomics: that LINE-1 is dramatically de-repressed in many tumors. In particular, LINE-1 is most often de-repressed in epithelial derived tumors (carcinomas)[30]. More surprisingly, we also found that in non-cancer samples, many epithelial cells also express LINE-1. This raises the question of whether high LINE-1 expression in tumors is due to de-repression during tumor development or to the expansion of a pre-existing population of LINE-1 expressing cells. We also found that, in the mouse hippocampus, LINE-1 expression is higher in neurons compared to glia. This is consistent with reports that LINE-1 remains active during neuronal development[26–28]. It may help resolve a puzzling previous observation: LINE-1 expression and retrotransposition seem to be largely absent in glioblastoma (a brain cancer derived from astrocytes)[31,32].

Two caveats should be noted when considering our results. Firstly, in the cancer datasets we still measure some LINE-1 expression in the microenvironment. While there is no highly sensitive LINE-1 assay that can completely rule out low-level expression in these cells, these quantifications likely indicate that, while we have enabled identification of LINE-1 expressing cell types, there is still some background noise. Secondly, our identification of LINE-1 expression in mouse hippocampal neurons comes from single nucleus sequencing, whereas the human datasets are all single cell derived. Because LINE-1 RNA/ORF1/ORF2 RNPs are primarily cytoplasmic[33], it remains possible that enrichment in neuronal nuclei reflects a distinct population of LINE-1 RNAs.

5’ targeted scRNA-seq methods remain less popular than 3’ methods, but our results show that only 5’ targeted methods are able to clearly identify the cell types that express LINE-1. We would encourage researchers working with samples that may express LINE-1 to consider a 5’ targeted library prep method.

Methods

Construction of the L1-sc cell ranger index

For the reference genome L1Hs and L1PAx family repeats were masked in hg38 using bedtools maskfasta. L1Hs and L1PAx annotations were downloaded in bed format from UCSC genome table browser (https://genome.ucsc.edu/cgi-bin/hgTables). Then this masked reference was concatenated with the human L1Hs consensus sequence from repbase[34] and all available L1PA consensus sequences from dfam (https://www.dfam.org/browse?name_accession=L1PA&clade_descendants=true). For transcript annotation, we started with RefSeq GRCh38 annotation in gtf format and then added an L1Hs transcript. For the 3’ analysis the L1Hs transcript covered the entire element. For the 5’ analysis it only covered the first 150 bases. Finally a CellRanger index was built with cellranger mkref.

The mouse version of L1-sc was generated in the same manner as human. We first used bedtools maskfasta to mask L1Md sequences (downloaded from the UCSC genome table browser in bed format) from the mm39 mouse genome. We then added all the available L1Md consensus sequences in dfam (https://dfam.org/browse?name_accession=L1Md&classification=root;Interspersed_Repeat;Transposable_Element;Class_I_Retrotransposition;LINE;Group-II;Group-1;L1-like;L1-group;L1&clade=10088&clade_descendants=true) as decoy chromosomes. LINE-1 transcripts were appended to the mouse transcriptome covering the first 150 bases of L1MdTf_I, L1MdTf_II, L1MdTf_III, L1MdGf_I, L1MdGf_II, L1MdA_I, L1MdA_II and L1MdA_III.

Human single cell RNA-seq analysis

A count matrix was built using cellranger count and the index described above (default parameters). Clustering and umap embedding was performed in Seurat according to the PBMC 3k tutorial (https://satijalab.org/seurat/v3.2/pbmc3k_tutorial.html). Filtering of cells was performed according to the original analysis of each dataset. No filtering was performed on the LSCC sample. For the non-cancerous tissues, cells with more than 25% MT-RNA, less than 500 genes or less than 1000 RNA molecules detected were removed. In both datasets, cell types were then identified using marker genes: CD3E for T cells, MS4A1 for B cells, MZB1 for plasma cells, GNLY for NK cells, LYZ for macrophages, KIT for mast cells, COL1A2 for fibroblasts, ACTA2 for muscle, PECAM1 for endothelial cells, MLANA for melanocytes, and keratin (KRT) genes for epithelial cells. UMAP plots were made in Seurat, and violin plots were made using ggplot2.

Mouse hippocampus snRNA-seq sequencing and analysis

For snRNA-seq the hippocampus was dissected from the brains of 4-month old C57BL/6 mice. There were a total of 4 animals used. The hippocampi from 2 mice were pooled together into 1 sample, and the other 2 mice were pooled into another sample. Nuclei were isolated from minced hippocampi tissue using the Nuclei PURE Prep Nuclei Isolation Kit with a Dounce B homogenizer. Samples were subjected to a sucrose gradient, and nuclei were further purified and counted. We targeted 5,000 nuclei per sampled to load onto a 10x Chromium chip using VD(J) chemistry. We targeted 50,000 sequence reads per nuclei on an illumina HiSeq device.

Fastq files from snRNA-seq were aligned using CellRanger Version 3.0.2 to a custom-made pre-mRNA reference for the mm10 genome that was created according to 10x Genomics instructions. snRNA-seq data was analyzed in the Seurat package. Data was normalized using the SC-Transform function in Seurat, and samples for the snRNA-seq were integrated together into one Seurat object using the Integrate Data function. UMAP projections were generated using the “Integrated” data assay on the resulting Seurat objects. Cells were filtered to include those with less than 30% mitochondrial RNA and at least 200 genes detected. Cell types were identified using a manually curated gene marker list.

Competing interests

David Fenyö is a Founder and President of The Informatics Factory, and serves or served on the Scientific Advisory Board or consults for: Spectragen Informatics, Protein Metrics, Preverna. Jef Boeke is a Founder and Director of CDI Labs, Inc., a Founder of Neochromosome, Inc, a Founder and SAB member of ReOpen Diagnostics, and serves or served on the Scientific Advisory Board of the following: Sangamo, Inc., Modern Meadow, Inc., Sample6, Inc. and the Wyss Institute. John Sedivy is a cofounder of Transposon Therapeutics, Inc., serves as Chair of its Scientific Advisory Board, and consults for Astellas Innovation Management LLC, Atropos Therapeutics, Inc. and Gilead Sciences, Inc.

References

1. Burns KH, Boeke JD. Human transposon tectonics. *Cell*. 2012;149(4):740-752. doi:10.1016/j.cell.2012.04.019
2. Rodić N, Sharma R, Sharma R, et al. Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am J Pathol*. 2014;184(5):1280-1286. doi:10.1016/j.ajpath.2014.01.007
3. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, et al. Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat Genet*. 2020;52(3):306-319. doi:10.1038/s41588-019-0562-0
4. Miki Y, Nishisho I, Horii A, et al. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res*. 1992;52(3):643-645.
5. 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*. 2016;26(6):745-755. doi:10.1101/gr.201814.115
6. Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P. Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res*. 2010;38(12):3909-3922. doi:10.1093/nar/gkq132
7. Doucet-O'Hare TT, Rodić N, Sharma R, et al. LINE-1 expression and retrotransposition in Barrett's esophagus and esophageal carcinoma. *Proc Natl Acad Sci U S A*. 2015;112(35):E4894-E4900. doi:10.1073/pnas.1502474112
8. Navarro FC, Hoops J, Bellfy L, et al. TeXP: Deconvolving the effects of pervasive and autonomous transcription of transposable elements. *PLOS Comput Biol*. 2019;15(8):e1007293. doi:10.1371/journal.pcbi.1007293
9. Sun X, Wang X, Tang Z, et al. Transcription factor profiling reveals molecular choreography and key regulators of human retrotransposon expression. *Proc Natl Acad Sci U S A*. 2018;115(24):E5526-E5535. doi:10.1073/pnas.1722565115
10. Wheeler TJ, Clements J, Eddy SR, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. *Nucleic Acids Res*. 2013;41(Database issue):D70-D82. doi:10.1093/nar/gks1265
11. Stuart T, Butler A, Hoffman P, et al. Comprehensive Integration of Single-Cell Data. *Cell*. 2019;177(7):1888-1902.e21. doi:10.1016/j.cell.2019.05.031
12. Laughney AM, Hu J, Campbell NR, et al. Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nat Med*. 2020;26(2):259-269. doi:10.1038/s41591-019-0750-6
13. McKerrow W, Wang X, Mita P, et al. LINE-1 expression in cancer correlates with DNA damage response, copy number variation, and cell cycle progression. *bioRxiv*. Published online June 28, 2020:2020.06.26.174052. doi:10.1101/2020.06.26.174052

14. vdj_v1_hs_nsclc_5gex -Datasets -Single Cell Immune Profiling -Official 10x Genomics Support. Accessed October 22, 2020. https://support.10xgenomics.com/single-cell-vdj/datasets/2.2.0/vdj_v1_hs_nsclc_5gex
15. He S, Wang L-H, Liu Y, et al. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome Biol*. 2020;21(1):294. doi:10.1186/s13059-020-02210-0
16. McKerrow WH, Fenyö D. L1EM: A tool for accurate locus specific LINE-1 RNA quantification. *Submitt Bioinforma*.
17. Kim D, Kobayashi T, Voisin B, et al. Targeted therapy guided by single-cell transcriptomic analysis in drug-induced hypersensitivity syndrome: a case report. *Nat Med*. 2020;26(2):236-243. doi:10.1038/s41591-019-0733-7
18. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102
19. Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498-D503. doi:10.1093/nar/gkz1031
20. Schaefer CF, Anthony K, Krupa S, et al. PID: the Pathway Interaction Database. *Nucleic Acids Res*. 2009;37(Database issue):D674-D679. doi:10.1093/nar/gkn653
21. Martens M, Ammar A, Riutta A, et al. WikiPathways: connecting communities. *Nucleic Acids Res*. 2021;49(D1):D613-D621. doi:10.1093/nar/gkaa1024
22. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27
23. Nishimura D. BioCarta. *Biotech Softw Internet Rep*. 2001;2(3):117-120. doi:10.1089/152791601750294344
24. Speek M. Antisense promoter of human L1 retrotransposon drives transcription of adjacent cellular genes. *Mol Cell Biol*. 2001;21(6):1973-1985. doi:10.1128/MCB.21.6.1973-1985.2001
25. Denli AM, Narvaiza I, Kerman BE, et al. Primate-specific ORF0 contributes to retrotransposon-mediated diversity. *Cell*. 2015;163(3):583-593. doi:10.1016/j.cell.2015.09.025
26. 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(7044):903-910. doi:10.1038/nature03663
27. Muotri AR, Marchetto MCN, Coufal NG, et al. L1 retrotransposition in neurons is modulated by MeCP2. *Nature*. 2010;468(7322):443-446. doi:10.1038/nature09544
28. Evrony GD, Cai X, Lee E, et al. Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*. 2012;151(3):483-496. doi:10.1016/j.cell.2012.09.035
29. Sookdeo A, Hepp CM, McClure MA, Boissinot S. Revisiting the evolution of mouse LINE-1 in the genomic era. *Mob DNA*. 2013;4(1):3. doi:10.1186/1759-8753-4-3
30. Ardeljan D, Taylor MS, Ting DT, Burns KH. The human LINE-1 retrotransposon: an emerging biomarker of neoplasia. *Clin Chem*. 2017;63(4):816-822. doi:10.1373/clinchem.2016.257444
31. Achanta P, Steranka JP, Tang Z, et al. Somatic retrotransposition is infrequent in glioblastomas. *Mob DNA*. 2016;7:22. doi:10.1186/s13100-016-0077-5
32. Carreira PE, Ewing AD, Li G, et al. Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. *Mob DNA*. 2016;7:21. doi:10.1186/s13100-016-0076-6
33. Mita P, Wudzinska A, Sun X, et al. LINE-1 protein localization and functional dynamics during the cell cycle. eLife. doi:10.7554/eLife.30058
34. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6:11. doi:10.1186/s13100-015-0041-9

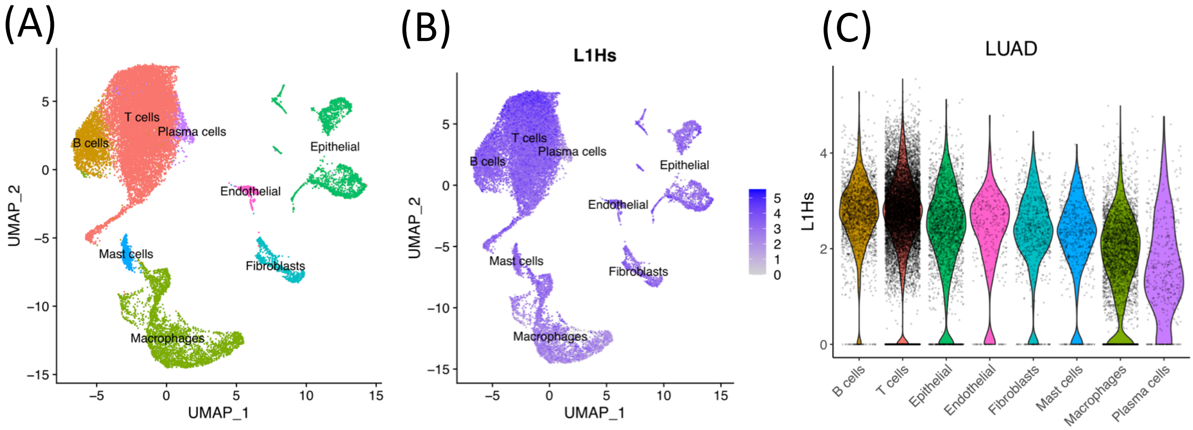

Figure 1. “LINE-1” expression estimated from 3’ targeted data from lung adenocarcinoma (LUAD) tumors. (A) UMAP and clustering. (B) UMAP of “LINE-1” expression. (C) “LINE-1” expression in each cluster sorted from highest to lowest “LINE-1” expression. “LINE-1” is in quotes as these quantifications include all transcripts terminating at LINE-1 3’ UTR and likely do not reflect actual LINE-1 de-repression. Expression level is normalized using the NormalizeData function in Seurat.

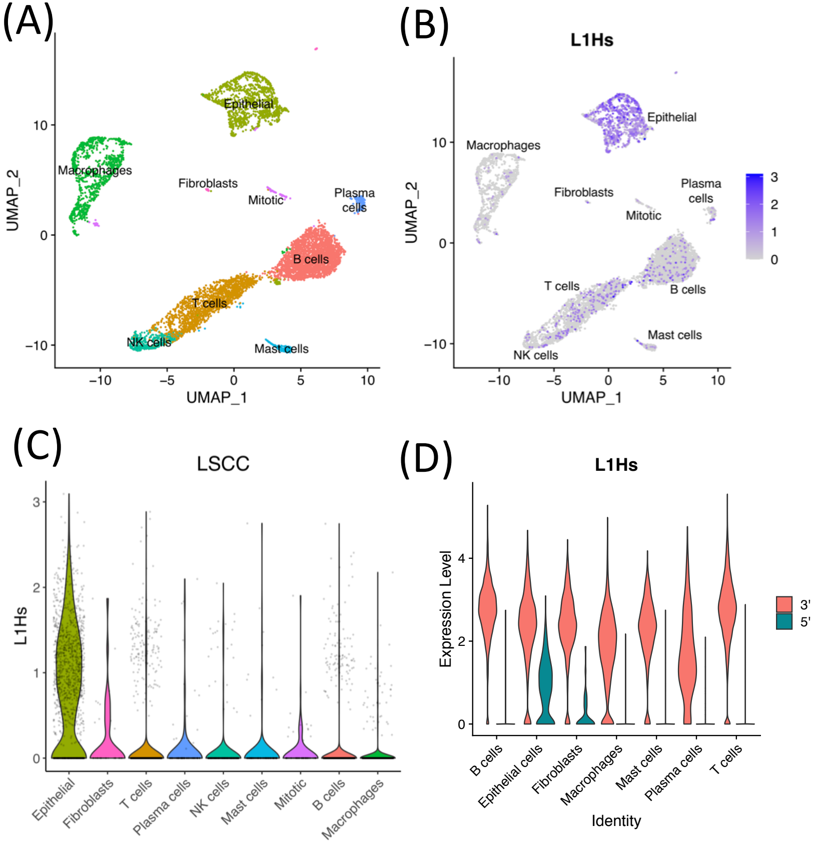

Figure 2. LINE-1 expression in lung squamous cell carcinoma (LSCC) tumor cells. (A) UMAP and clustering of cells. (B) UMAP of LINE-1 expression in cells. (C) LINE-1 expression in each cluster, ordered from highest to lowest LINE-1 expression. (D) Direct comparison of LINE-1 quantifications in the 3’ (LUAD) vs 5’ (LSCC) data sets for cell types identified in both.

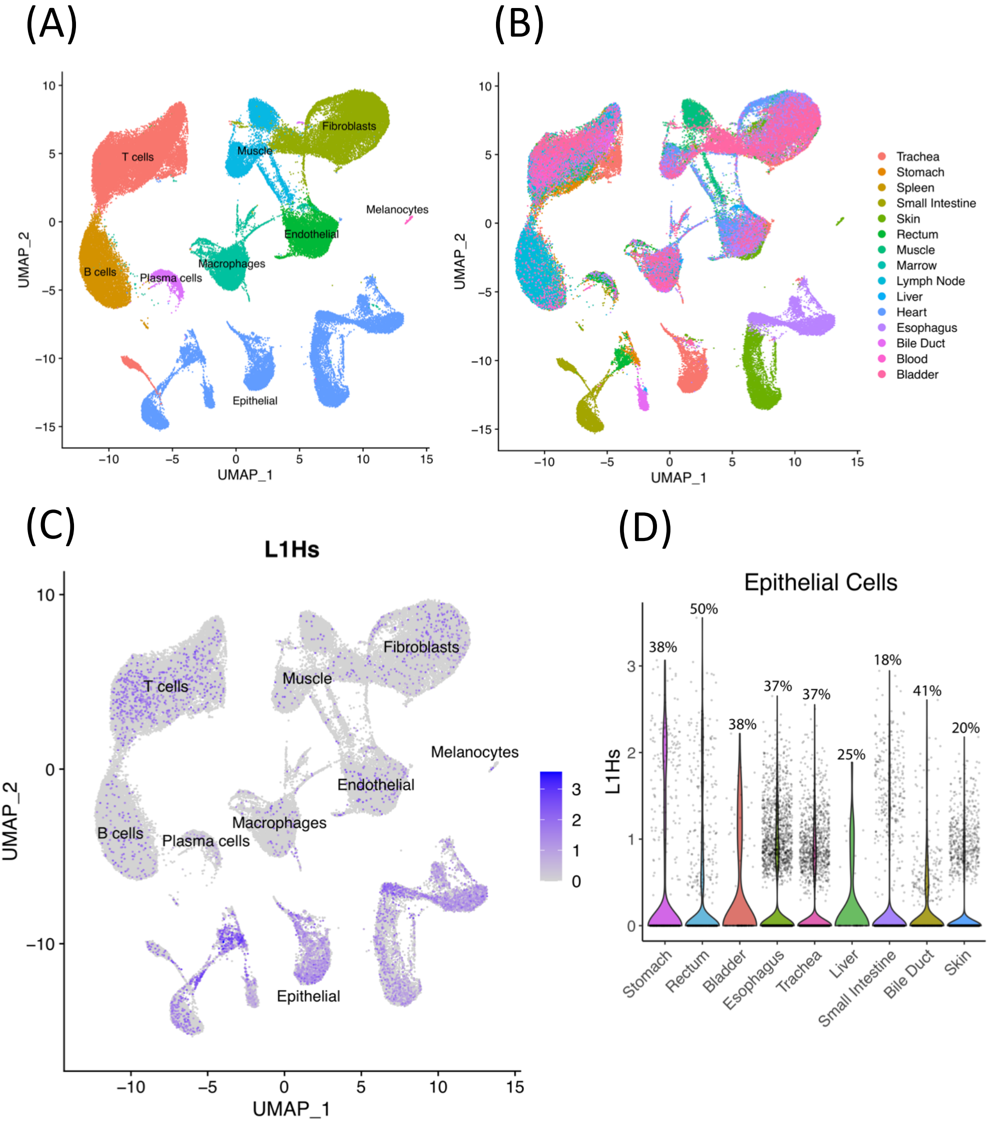

Figure 3. LINE-1 expression estimated in non-tumor cells. UMAP embeddings labeled by (A) cell type, (B) tissue of origin, and (C) LINE-1 expression. (D) Violin plot of LINE-1 expression in epithelial cells from each tissue. Sorted from highest to lowest LINE-1 expression. Percents atop the plots are the fraction of cells that are positive for at least one LINE-1 read.

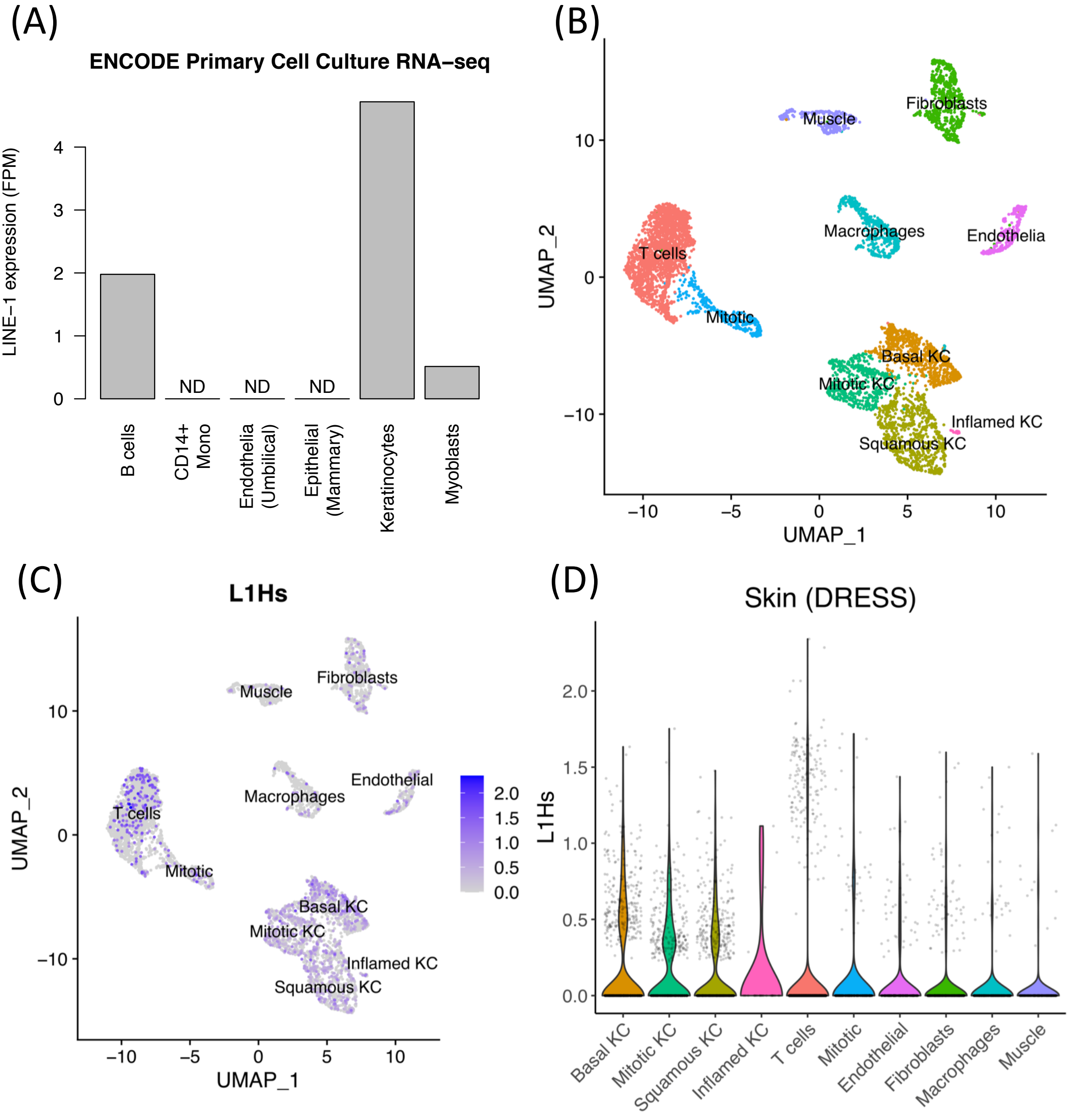

Figure 4. Additional evidence for LINE-1 expression in keratinocytes (epithelial skin cells, KC). (A) LINE-1 expression in primary cell culture from ENCODE. ND = full length LINE-1 mRNA not detected. UMAP embeddings for cells in the DRESS skin dataset labelled by cell type (A) and LINE-1 expression (B). (D) Violin plot of LINE-1 expression in each cell type.

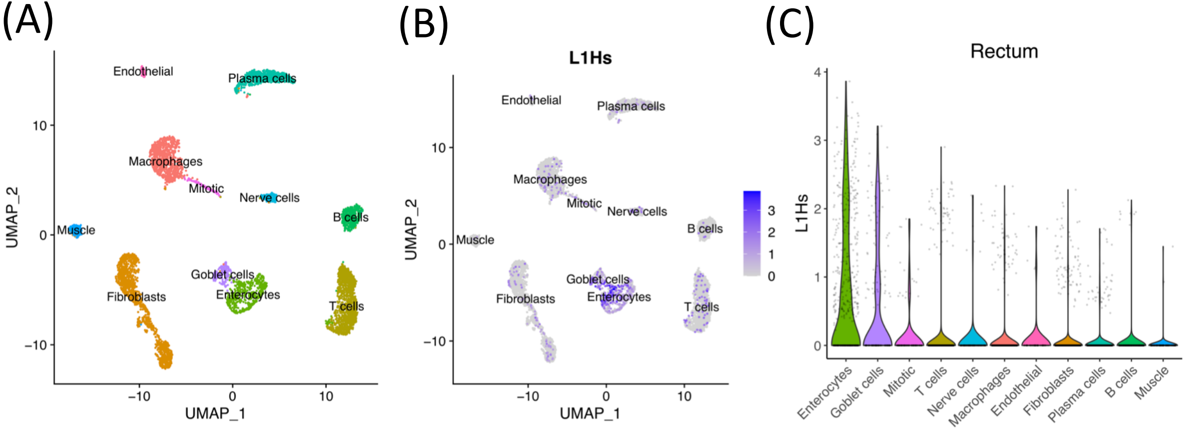

Figure 5. LINE-1 expression in rectal tissue. (A) UMAP and clustering of cells. (B) UMAP of LINE-1 expression in cells. (C) LINE-1 expression in each cluster, ordered from highest to lowest LINE-1 expression.

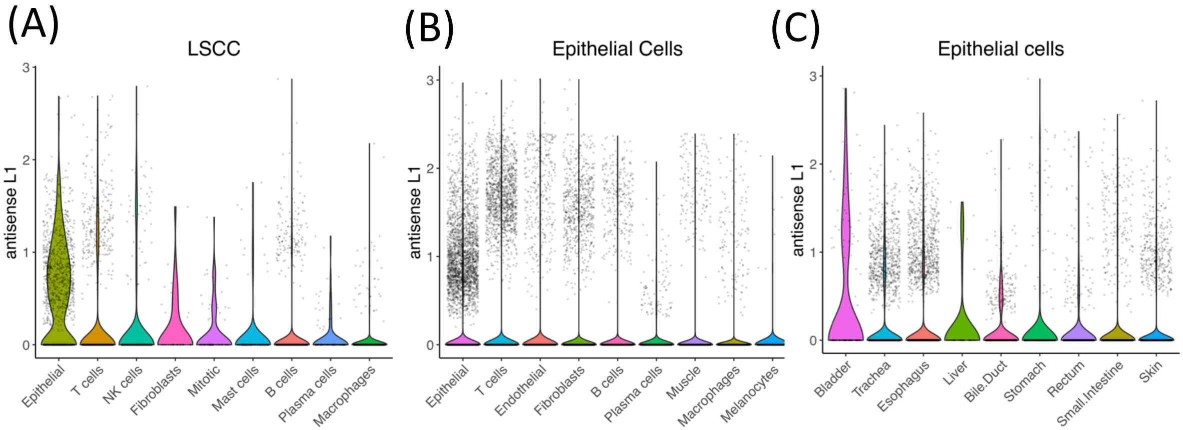

Figure 6. Antisense LINE-1 transcription by cell type in (A) Cancer (LSCC) and (B) 15 non-tumor tissues. (C) Antisense LINE-1 expression in non-tumor epithelial cells by tissue of origin.

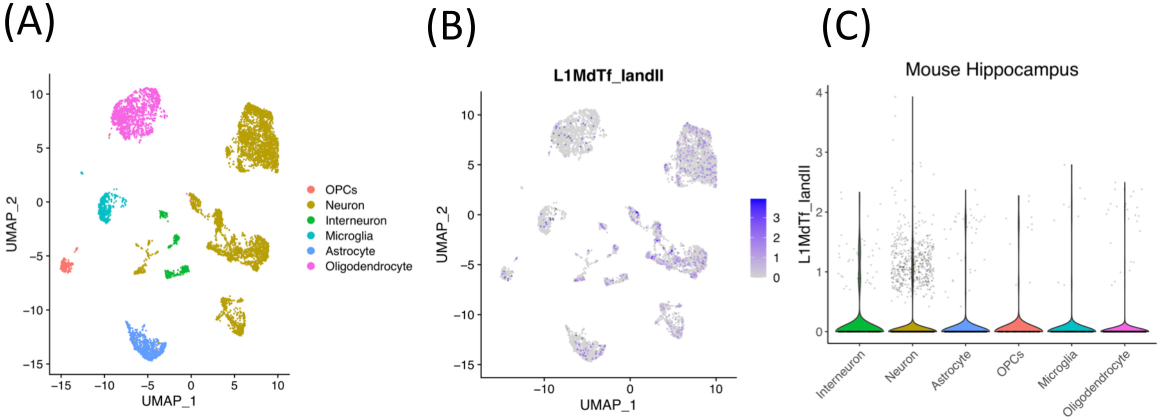

Figure 7. LINE-1 (L1MdTf I and II) expression in the mouse hippocampus. (A) UMAP embedding of mouse hippocampal cells, colored by cell type. (B) UMAP embedding of LINE-1 expression from the L1MdTf I and II subfamilies in mouse hippocampal cells. (C) Violin plot, showing LINE-1 expression in each of the mouse hippocampal cell types.