Identification of bona fide B2 SINE retrotransposon transcription through single-nucleus RNA-seq of the mouse hippocampus

Sara B. Linker,[1] Lynne Randolph-Moore,[1] Kalyani Kottilil,[1] Fan Qiu,[1] Baptiste N. Jaeger,[2] Jerika Barron,[3] and Fred H. Gage[1]

[1]*Laboratory of Genetics, The Salk Institute for Biological Studies, La Jolla, California 92037, USA;* [2]*Laboratory of Neural Plasticity, Faculty of Medicine and Science, Brain Research Institute, University of Zurich, 8057 Zurich, Switzerland;* [3]*Biomedical Sciences Graduate Program, University of California, San Francisco, San Francisco, California 94143, USA*

Currently, researchers rely on generalized methods to quantify transposable element (TE) RNA expression, such as RT-qPCR and RNA-seq, that do not distinguish between TEs expressed from their own promoter (bona fide) and TEs that are transcribed from a neighboring gene promoter such as within an intron or exon. This distinction is important owing to the differing functional roles of TEs depending on whether they are independently transcribed. Here we report a simple strategy to examine bona fide TE expression, termed BonaFide-TEseq. This approach can be used with any template-switch based library such as Smart-seq2 or the single-cell 5′ gene expression kit from 10x, extending its utility to single-cell RNA-sequencing. This approach does not require TE-specific enrichment, enabling the simultaneous examination of TEs and protein-coding genes. We show that TEs identified through BonaFide-TEseq are expressed from their own promoter, rather than captured as internal products of genes. We reveal the utility of BonaFide-TEseq in the analysis of single-cell data and show that short-interspersed nuclear elements (SINEs) show cell type–specific expression profiles in the mouse hippocampus. We further show that, in response to a brief exposure of home-cage mice to a novel stimulus, SINEs are activated in dentate granule neurons in a time course that is similar to that of protein-coding immediate early genes. This work provides a simple alternative approach to assess bona fide TE transcription at single-cell resolution and provides a proof-of-concept using this method to identify SINE activation in a context that is relevant for normal learning and memory.

[Supplemental material is available for this article.]

Retrotransposons are a class of transposable elements (TEs) that integrate into the host genome through an RNA intermediate (Craig et al. 2015). It is important to be able to examine the transcription of TE RNA in order to understand the dynamics of TE regulation across physiological and pathological conditions. TEs exist throughout the genome, including within the introns and exons of genes (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002; Zhang et al. 2011), which imparts methodological constraints when attempting to distinguish the expression level of a true, bona fide TE that is expressing from its own promoter, as opposed to a TE sequence that is transcribed by an upstream gene promoter (passenger). Standard techniques such as qRT-PCR and RNA-sequencing (RNA-seq) do not distinguish bona fide TEs from passenger TEs; however, these methods are used readily in the field to assess TE biology in combination with computational tools that have been developed to count TE expression levels (Criscione et al. 2014; Jin et al. 2015; Lerat et al. 2017; Yang et al. 2019).

Analysis of TE expression is computationally challenging owing to the repetitive nature of TEs (Treangen and Salzberg 2012; Teissandier et al. 2019). Multiple methods exist that offer researchers a way to quantify TE abundance from high-throughput sequencing data such as TEtools (Lerat et al. 2017), RepEnrich (Criscione et al. 2014), SQuIRE (Yang et al. 2019), and TEtranscripts (Jin et al. 2015). Each of these methods shows relatively similar high true-positive rates when detecting repeat subfamilies, supporting the high accuracy in identifying a repeat subfamily from sequencing data (Teissandier et al. 2019). However, although these methods are advantageous for calling subfamily abundance from sequencing data, they fail to distinguish whether a TE-aligned read is derived from a bona fide or a passenger repeat element. For example, RepEnrich was one of the first packages that enabled users to count estimates for both genes and TEs simultaneously. This method is a useful technique for the community. However, RepEnrich does not distinguish bona fide from passenger elements; therefore, all TE-containing RNA sequences, whether they are present as a by-product of gene expression or from TE-directed transcription, are conflated together into one count estimate. The same is true of other TE transcript counting methods such as TEtranscripts and TEtools. SQuIRE can identify bona fide TE transcripts when those transcripts are uniquely mapped to the genome. However, there are lower unique alignment rates for young TEs, which are the more active elements (Brouha et al. 2003; Teissandier et al. 2019).

Given the evidence that bona fide TE transcription can have a range of physiological and pathological impacts on a system (Allen et al. 2004; Hasler and Strub 2006; Ahl et al. 2015; Zovoilis et al.

Corresponding authors: gage@salk.edu, sara.linker@gmail.com
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.262196.120.

2016; Percharde et al. 2018; Hernandez et al. 2020), future studies need to have an approach to estimate the level of TE expression from bona fide elements. Importantly, passenger and bona fide TEs have vastly different functional roles. A potent example is the case of LINE-1 (L1), in which retrotransposition is supported by bona fide expression from active L1 elements as opposed to passenger elements. Conversely, short interspersed elements (SINEs) and L1 elements within genes can be exonized or can aid in the recruitment of RNA-binding proteins to genes, thereby regulating local gene expression (Kaer et al. 2011; Zarnack et al. 2013; Attig et al. 2018). In the SINE family of TEs, bona fide elements also play roles in transcriptional and translational inhibition (Allen et al. 2004; Hasler and Strub 2006; Ahl et al. 2015; Zovoilis et al. 2016; Hernandez et al. 2020), whereas passenger SINEs can play a functional role in RNA trafficking and RNA editing (Athanasiadis et al. 2004; Buckley et al. 2011). These functional distinctions underscore the importance of accurately quantifying bona fide, versus passenger, TE transcription.

The most promising methods to distinguish bona fide TE elements directly assess RNA through modifications of 5′- or 3′-RACE or by combination with ChIP-sequencing approaches (Faulkner et al. 2009; Oler et al. 2012; Deininger et al. 2017; Karijolich et al. 2017). However, these methods often preclude simultaneous quantification of TEs and protein-coding genes, or they require multiple methods such as RNA- and ChIP-seq. Here, we sought to develop an approach that combines the separate advantages of these methods into one strategy that can (1) identify *bone fide* TE transcripts, (2) retain quantification of standard protein-coding genes, and (3) be applied to single-cell RNA-seq protocols.

Here, we couple the 5′-tagging performed in the Smart-seq-based approach with a dual TE- and protein-coding alignment strategy to directly examine TE dynamics in conjunction with protein-coding gene expression in data derived from the mouse hippocampus. We term this approach BonaFide-TEseq, and we performed a proof-of concept analysis by exploring SINE expression in response to behaviorally relevant neuronal activity.

The SINE family of TEs, which includes *Alu* and SVA in primates and B1 and B2 in rodents, is involved in biological processes such as cellular stress in response to heat shock, viral infection, or DNA damage (Jang and Latchman 1989; Liu et al. 1995; Rudin and Thompson 2001; Zovoilis et al. 2016). In 2004, using in situ hybridization, Kalkkila et al. identified that a subclass of rodent SINEs, B2, was up-regulated in response to global seizure-like neuronal activity, opening up an important question regarding whether B2 SINEs play a role in physiological contexts such as learning and memory (Kalkkila et al. 2004).

We previously showed that exposing mice that have been raised in a standard home-cage (HC) environment to 15 min in a novel environment (NE)—which included a larger exploration area, huts, tunnels, and a running wheel—was sufficient to elicit neuronal activity within the mouse hippocampus (Lacar et al. 2016). This exposure to a NE increased the number of neurons labeled with FOS protein, a marker of neuronal activity, and was associated with nascent transcription of activity-dependent genes (Lacar et al. 2016). Furthermore, this brief exposure was sufficient to habituate mice to the novel context, indicating that it drove the downstream signals required for learning and memory (Jaeger et al. 2018). Here, we aimed to identify bona fide TE sequences using single-cell RNA-seq data. We used BonaFide-TEseq to examine SINE expression in the mouse hippocampus as a function of a brief exposure to a NE at single-neuron resolution. Together, these results show the utility of single-cell data for examining bona fide retrotransposon expression.

Results

Detection of bona fide SINE transcripts in single-nucleus RNA-seq data

Our first goal was to develop an analytical method that was capable of distinguishing TEs that were transcribed from their own promoter, or "bona fide" transcripts (Fig. 1A), from TEs that were transcribed as a by-product of a protein-coding gene promoter, whether intronic or exonic, which we will hereafter refer to as "passenger" transcripts (Fig. 1B). To accomplish this task, we took advantage of the process of template switching during cDNA synthesis. Template switching is a procedure in which an oligo (template-switch oligo [TSO]) is added when the reverse transcriptase reaches the end of the template RNA molecule, thereby tagging the 5′-end of a transcript (Picelli et al. 2014). We previously used Smart-seq2, with TSO, to generate single-nucleus RNA-seq (snRNA-seq) libraries from individual neurons in the mouse hippocampus (Lacar et al. 2016; Jaeger et al. 2018). RNA was sequenced and subsequently separated into files based on the presence of the TSO sequence at the 5′-end (+TSO, −TSO). +TSO and −TSO files were then aligned to the rodent RepeatMasker reference that contains TEs and other non-TE-derived repetitive sequences. All reads were aligned to the mm10 transcriptome as is standard for RNA-seq analysis of genes. After normalization based on total counts, we determined the expression of all TE elements within hippocampal nuclei (Fig. 1C). Although passenger transcripts detected expression of long-terminal repeat containing elements (LTR), LINEs, satellite RNA, and SINEs, bona fide transcripts primarily detected expression of SINEs and small cytoplasmic RNA (Fig. 1D). Further analysis based on L1 or SINE subfamilies showed increased detection of younger elements in the bona fide transcripts compared with the passenger elements, supporting the ability of BonaFide-TEseq to detect the younger, more transcriptionally active elements (Supplemental Fig. S1).

It has been speculated that evolutionarily older elements are more likely to accumulate mutations in their promoter, thereby rendering ancient elements incapable of transcription. We determined the overall expression patterns of SINEs as well as other short repeats such as transfer RNAs (tRNAs) and signal recognition particle RNAs (srpRNAs) from single hippocampal nuclei. We identified that B2 elements were the most highly expressed, followed by B1 and then ID elements (Supplemental Fig. S2A). Evolutionary time was calculated as the average Smith–Waterman distance across all elements within the respective subfamily in the mouse mm10 reference genome. In general, younger subfamilies of B2 elements were detected at higher levels than older SINEs (Pearson's correlation coefficient = −0.57, $P < 0.002$) (Supplemental Fig. S2B). However, one ancient SINE, Proto-B1 (PB1), was expressed at a higher level than expected given the trendline. Upon inspection of the TSO-PB1 promoter, we noted that the A box sequence that was present in the active mouse B1 element was also present in the expressed PB1 sequences (5′-TGGCGCACGC) (Supplemental Fig. S3A,B), confirming that the TSO-bound PB1 sequences were indeed from bona fide expression of Pol III elements. The ancient PB1 element originated from the 7SL gene, which is still ubiquitously expressed in mammals. Therefore, we compared the PB1 sequence to the 7SL promoter region. Indeed, the consensus TSO-bound PB1 promoter was

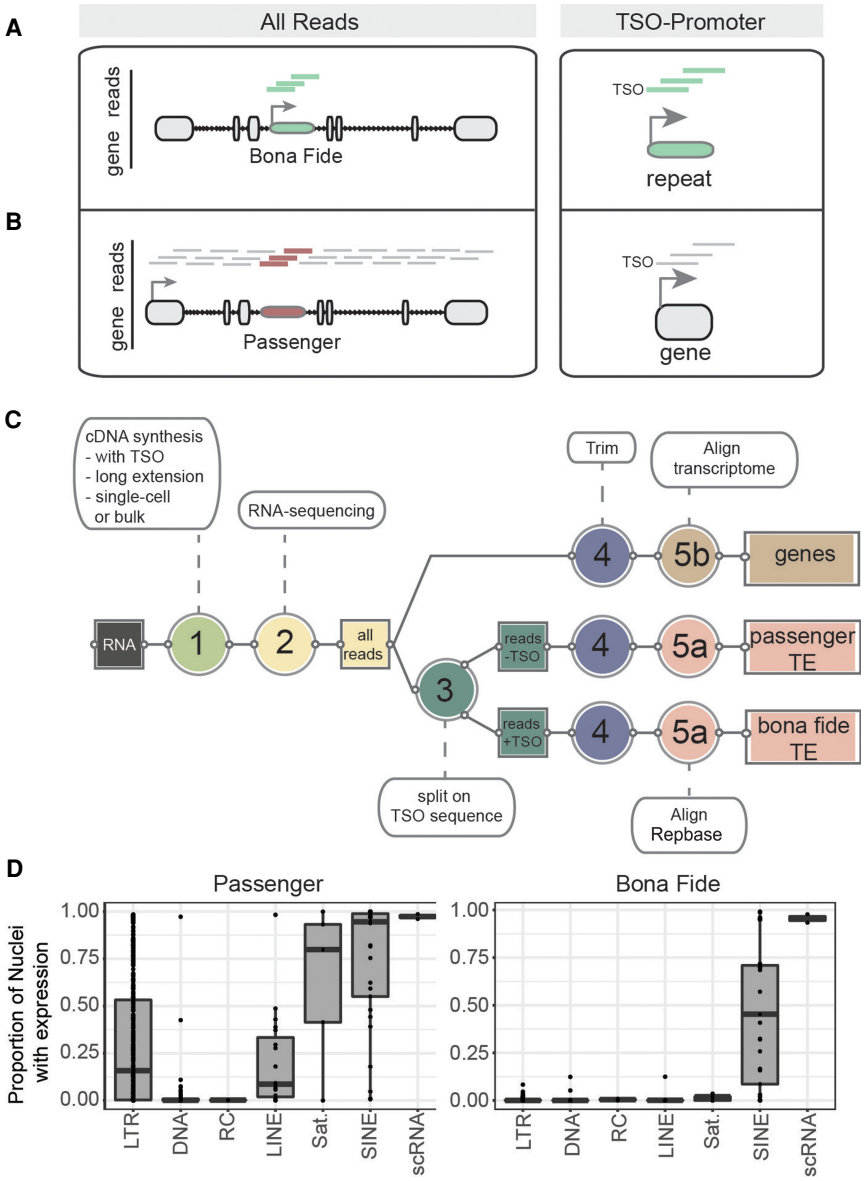

Figure 1. Separation of bona fide and passenger transcripts TE elements can be detected via template-switch oligo (TSO)–assisted cDNA synthesis. (*A*, *left*) Diagram of short reads generated by sequencing bona fide TE elements (green). (*Right*) Filtering for reads that are linked to the TSO will identify reads that capture the 5′-end of the TE. (*B*, *left*) Sequencing cDNA from a Pol II–derived gene with a passenger TE element in the intron (red). (*Right*) The TSO is bound to the 5′-end of the gene, not the TE. Note that both bona fide and passenger elements generate identical short read information and cannot be distinguished without additional information about the promoter sequence at the 5′-end. (C) BonaFide-TEseq workflow begins with (1) first-strand cDNA synthesis with a TSO, ensuring long first-strand synthesis times to extend to the end of long transcripts. This synthesis step is compatible with both single-cell and bulk RNA-sequencing (RNA-seq) libraries. (2) Sequencing the libraries, in this case with Illumina short reads. (3) All reads are split into separate files based on the capture of the TSO sequence at the 5′-end. (4) Each file: +TSO reads, −TSO reads, and all reads are trimmed of low-quality bases and of the TSO adapter sequence. (5a) Aligning the +TSO and −TSO read files against the Repbase reference for the corresponding lineage, in this case, rodent. (5b) Aligning all trimmed reads to the transcriptome reference using standard procedures such as RSEM. The outputs at the end of this procedure are three separate counts files: +TSO (bona fide TEs), −TSO (passenger TEs), and a standard gene expression matrix. (*D*) Proportion of nuclei with detectable expression classified as passenger (*left*) or bona fide (*right*) TEs within each annotated family. (LTR) Long terminal repeat; (RC) rolling circle; (Sat.) satellite; (scRNA) small cytoplasmic RNA.

identical to both the 7SL and PB1 promoter sequences (Supplemental Fig. S3C). Therefore, although it was possible that we detected expression of the PB1 element, a more parsimonious explanation was that the transcripts labeled as PB1 were detecting expression of the canonical 7SL element. Together these results suggested that TSO-bound transcription decreased linearly with SINE evolutionary age. Together with the Pol III promoter and terminator analyses, these results indicated that the TSO filtering strategy captured bona fide repeat transcripts, enabling discovery of SINE expression dynamics across the mammalian brain.

To determine if TSO-bound elements were detecting younger TEs than other previously published TE counting algorithms, we compared our approach to two previously developed methods that map to the genome (TEtranscripts and SQuIRE). We identified higher overall expression when using TEtranscripts or SQuIRE compared with bona fide elements detected using BonaFide-TEseq (Supplemental Fig. S4A), indicating that the filtering procedure to identify bona fide elements reduced the overall count estimates. Analysis of expression as a function of age identified that elements identified as bona fide by BonaFide-TEseq were enriched for younger TEs compared with both TEtranscripts and SQuIRE (Supplemental Fig. S4B,C).

Our findings of bona fide TE expression indicated that SINEs were the primary TEs detected in the adult mouse hippocampus. L1 elements are known to be transcribed in early neural stem and progenitor cells but not in adult neurons (Ostertag et al. 2002; Muotri et al. 2005; Belancio et al. 2010). We next wanted to determine if the Smart-seq2 pipeline was capable of detecting L1 elements or if the lack of expression in adult neurons was because of a technical artifact. We examined Smart-seq2 data generated from quiescent and activated adult neural stem cells (NSCs) within the mouse subventricular zone (Llorens-Bobadilla et al. 2015), with the expectation of observing elevated L1 transcription in this system. UMAP dimensionality reduction on the gene expression data set revealed four distinct clusters that matched the cell types identified by Llorens-Bobadilla et al. (2015), including activated NSCs (aNSCs: *Mki67*, *Slc1a3*, and *Sox9*), quiescent NSCs (qNSCs: *Slc1a3* and *Sox9*), oligodendrocytes (*Sox10*, *Mbp*), and neuroblasts

(*Dcx* and *Cd24a*) (Supplemental Fig. S5A, B). As expected, we detected high expression of L1 across all NSCs, particularly the young L1 element L1MdA_IV. Llorens-Bobadilla et al. (2015) activated NSCs through ischemia and identified an associated increase in qNSC populations. We observed that NSCs that were either from the control group or from the ischemia group had heightened levels of L1MdA_IV expression in comparison to the interferon-gamma knockdown (Supplemental Fig. S5C,D). Furthermore, although clustering based on genes clearly separated out cell types, clustering based on repeat elements separated out cell states (i.e., ischemia, control, and ischemia + interferon-gamma knockdown) (Supplemental Fig. S5B). This finding is intriguing in light of work linking L1 transcription and inflammation (Thomas et al. 2017; De Cecco et al. 2019). Together, these analyses indicate that LINEs are detectable in Smart-seq2 RNA-seq data.

To further investigate the capability of our approach to identify bona fide TEs, we examined TSO-TE hybrid reads, which were used to quantify expression of the SINEs. The TSO-SINE hybrids were likely to be generated from true bona fide SINEs; however, an alternative hypothesis was that the 5′-end of the RNA molecule was degraded and contained a SINE sequence at random, thereby creating a false TSO-SINE sequence. To distinguish between these two possibilities, we first examined the distribution of TSO-mobile element positions. As expected, the reads containing TSO sequences were significantly enriched for the 5′-end of the repeat element, whereas reads without TSO sequences were uniformly distributed across the entire repeat, with a significantly increased distance from the start position ($P < 2.2 \times 10^{-16}$) (Supplemental Fig. S6A,B), indicating that TSO-bound reads were non-randomly detecting the promoter region of repeat elements. This finding was unique to the TSO-bound reads, as the total repeat-aligned reads were uniformly covered across the length of the elements, indicating a lack of 3′-bias during cDNA synthesis (Supplemental Fig. S6C,D).

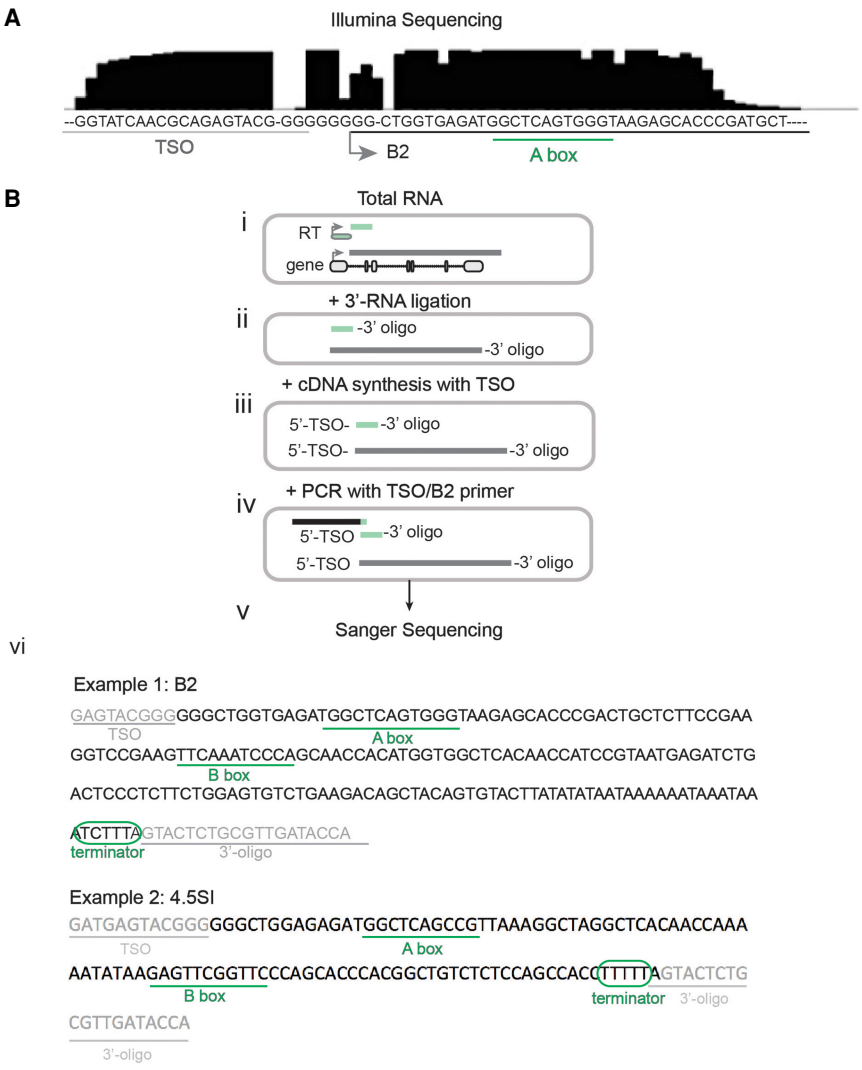

Figure 2. SINE-promoted transcripts detected via 5′-capture. (*A*) Consensus sequence of all reads that match the reference B2 sequence and contain the TSO sequence. Green line indicates Pol III A box sequence. (*B*) Diagram of 5′- and 3′-end tagging to identify if TSO-bound B2 elements contain a Pol III termination sequence. (*i*) Total RNA is isolated from tissue; (*ii*) oligos are directly ligated onto the 3′-end of RNA; (*iii*) cDNA synthesis with template switching begins with a primer at the 3′-oligo and extends to attach the TSO sequence on the 5′-end. (*iv*) PCR with an oligo that spans the junction between the 5′-TSO sequence and the B2 promoter and a second primer that captures the 3′-end. (*v*) The whole PCR product is cloned for Sanger sequencing, with no size selection. (*vi*) Two examples of Sanger-sequenced clones with hallmarks of Pol III transcription. (Cf. Supplemental Data for sequences.)

We next directly examined the junction between the TSO and SINE sequences. Consensus sequences were generated for the highly transcribed SINE family, B2, by accumulating all TSO-B2 reads and aligning them with Clustal Omega (Higgins and Sharp 1988). The consensus B2 5′-end was directly downstream from the TSO sequence, separated by Gs that were added on by the reverse transcriptase during synthesis to prime template switching (Fig. 2A). The consensus sequence derived from TSO-B2 reads contained the canonical A box promoter, further confirming that the TSO-bound SINEs were enriched for bona fide elements that were driven by Pol III promoters.

To determine whether TSO-bound SINEs were indeed Pol III transcribed, we examined the 3′-end of TSO-tagged SINEs for Pol III termination sequences by direct 3′-end ligation of an RNA adapter to total RNA followed by PCR amplification with a hybrid primer that complemented the TSO-B2 junction (Fig. 2B; Supplemental Data). The total PCR product was then cloned and Sanger sequenced. Similar to the observations with high-throughput RNA-seq, the 5′-end of the B2 element was directly downstream from the TSO, with the addition of three Gs that were added on by the reverse transcriptase during synthesis. This B2 sequence mapped with 100% identity to three loci in the mouse mm10 genome (Chr 15: 93,110,810–93,110,991; Chr 4:

155,830,944–155,831,125; Chr 5: 65,439,599–65,439,781). Pol III termination sequences are usually composed of a string of four or more T nucleotides, with flexibility for an extra interjected nucleotide (Orioli et al. 2011; Gao et al. 2018). At the 3′-end, we observed a clean boundary between the 3′-oligo and the noncanonical Pol III termination sequence, "TCTTT" (Fig. 2, vi, example 1). It is important to note that this finding may indicate Pol III pausing rather than termination. Furthermore, given the longer length of the Sanger-sequenced B2 element, we were able to detect both the A box (13 bp+) and the presence of the B box at position 59 (Geiduschek and Tocchini-Valentini 1988); both are required for Pol III transcription.

We also identified a clone with high similarity to B2 at the 5′-end but that was expressed by a separate Pol III–transcribed RNA, 4.5SI RNA (Fig. 2, vi, example 1; Gogolevskaya and Kramerov 2010; Koval et al. 2012). This element maps to four loci in the mm10 genome (Chr 6: 128,839,049–128,839,148; Chr 6: 128,798,861–128,798,960; Chr 6: 128,760,077–128,760,176; Chr 16: 33,047,591–33,047,690), contains the 4.5SI A and B box sequences, and terminates with the canonical "TTTT" (Fig. 2B, vi, example 2). Because 4.5SI is not included in the RepeatMasker reference database, we analyzed our TSO-containing reads for sequences that matched the 5′-portion of the 4.5SI sequence (AGAGATGGCTCAGCCGTTA) and identified that this sequence was present and enriched within reads classified as B2L_S (85.8% of reads annotated as B2L_S), indicating that the Pol III–transcribed 4.5SI RNA was detected in the single-nucleus data.

Together, termination at the Pol III terminator sequence and the presence of an A box and B box indicate that the TSO capture approach was able to identify bona fide Pol III–transcribed repeat sequences.

We next wanted to determine whether filtering on TSO sequence was compatible with TE identification from 10x data as well as Smart-seq2 data. The 10x 5′ Gene Expression kit captures the 5′-end of transcripts using a TSO in a method that is similar to Smart-seq2. However, given the lower overall transcript counts per cell in 10x data sets, it is feasible that there is not enough information to generate meaningful results. We used a publicly available data set from human CD45$^+$ cells isolated from a fresh kidney tumor sample (SRR6798781) (Neal et al. 2018). A total of 3870 cells were identified with a mean sequencing depth of 13,983 and average detection of 462 genes per cell. This finding is in contrast to the mouse hippocampal Smart-seq2 data, which aligned an average of 1.17 million reads per cell and had a mean gene detection of 5637 genes per nucleus. Despite the low coverage, RepeatMasker mapping of passenger reads identified the presence of 190 element subfamilies in 703 cells, 16 of which were also identified in bona fide reads from 27 cells. In contrast to TE expression in the mouse hippocampus, the highest TE element expression in the CD45$^+$ kidney tumor cells was L1HS (Supplemental Fig. S7), which is consistent with previous studies that have identified heightened L1HS expression across multiple tumor types (Ardeljan et al. 2017). We further explored the validity of L1HS detection by assessing the position of the TSO sequence within the reads. As expected, TSO sequences were enriched near the promoter of the L1HS elements, in contrast to the start position of passenger L1HS elements, which occurred uniformly across the 6-kb consensus sequence (Supplemental Fig. S7). We further performed UMAP dimensionality reduction based on TE sequences and identified that L1HS expression drove clustering within this plot, indicating that L1HS expression was a large driver of TE variability within the CD45$^+$ data set (Supplemental Fig. S7). Together, these results indicated that 10x 5′ gene expression data can be used to assess bona fide repeat expression. However, users should ensure sequencing to a depth of saturation to maximize bona fide TE detection.

SINEs show cell-type specificity in the mouse hippocampus

Although there is evidence to suggest that SINEs are transcribed in response to potent pathological cellular stress, it is unclear whether SINE expression in basal physiological conditions is a random event or if it is controlled by a predictable cell state. Identifying whether SINEs are expressed according to a predictable set of rules in the healthy mammalian brain would be informative about their putative functional impact.

We therefore used the BonaFide-TEseq approach to first examine cell-type dependence of SINE expression within the mouse hippocampus using snRNA-seq data from a study by Jaeger et al. (2018), in which nuclei were extracted from the hippocampi of mice that were housed under standard laboratory conditions until sacrifice. This HC environment ensured that a majority of the neurons were relatively inactive upon RNA analysis. Nuclei were sorted by FACS on staining for RBFOX3 (also known as NEUN), PROX1, CTIP2, and FOS. RBFOX3$^+$ PROX1$^-$CTIP2$^-$ nuclei corresponded to a mixture of CA3 pyramidal neurons as well as GABAergic interneurons (CA3$^+$), RBFOX3$^+$ PROX1$^-$CTIP2$^+$ nuclei corresponded to CA1 pyramidal neurons, RBFOX3$^+$PROX1$^+$CTIP2$^-$ were vasoactive intestinal polypeptide-expression (VIP) interneurons, and RBFOX3$^+$PROX1$^+$CTIP2$^+$ corresponded to dentate granule (DG) neurons. No FOS$^+$ neurons were isolated from the HC mice. FOS$^-$ staining, which marked neurons that were not recently activated, were sorted by fluorescence-activated cell sorting (FACS) for each cell type (Jaeger et al. 2018).

In the current study, we performed t-SNE analysis using either bona fide or passenger SINE expression from this data set. Bona fide SINE expression from FOS$^-$ neurons was sufficient to cluster cell types in the absence of gene information (Fig. 3A, right). Conversely, when using identical t-SNE parameters (initial dimensions = 15, perplexity = 22), passenger SINE expression displayed no segregation based on cell type (Fig. 3A, left), indicating that bona fide SINE expression was associated with nonrandom cell type–specific dynamics.

The separation in t-SNE space was driven by expression of a few SINEs, in particular B2_Mm1t, which showed high expression in CA1 and CA3$^+$ and lower expression in DG and VIP neurons (Fig. 3B). We next used linear regression analysis to determine whether B2_Mm1t expression was directly correlated with other repetitive elements or gene expression dynamics in FOS$^-$ CA1, CA3$^+$, DG, and VIP neurons. Sixty elements and genes were significantly associated with B2_Mm1t expression after multiple-testing correction: 53 and seven were positively and negatively correlated, respectively ($P_{adj} < 0.05$) (Fig. 3C,D; Supplemental Table S1). B2_Mm1a, PB1/7SL, and 4.5SRNA were the top correlated repetitive elements. The Pol III transcripts PB1/7SL and 4.5SRNA (4.5SH) were inversely correlated with B2_Mm1t, indicating that B2 transcription was not merely a response to a global increase in the levels of Pol III transcription. In addition to repetitive elements, many genes were associated with B2_Mm1t expression, including *H2-T23* and *Malat1* (Fig. 3C,D). To further determine if B2 expression was associated with cell-type identity, we built a random forest classifier that was trained to discriminate broad cell types based only on repetitive elements and genes that had been

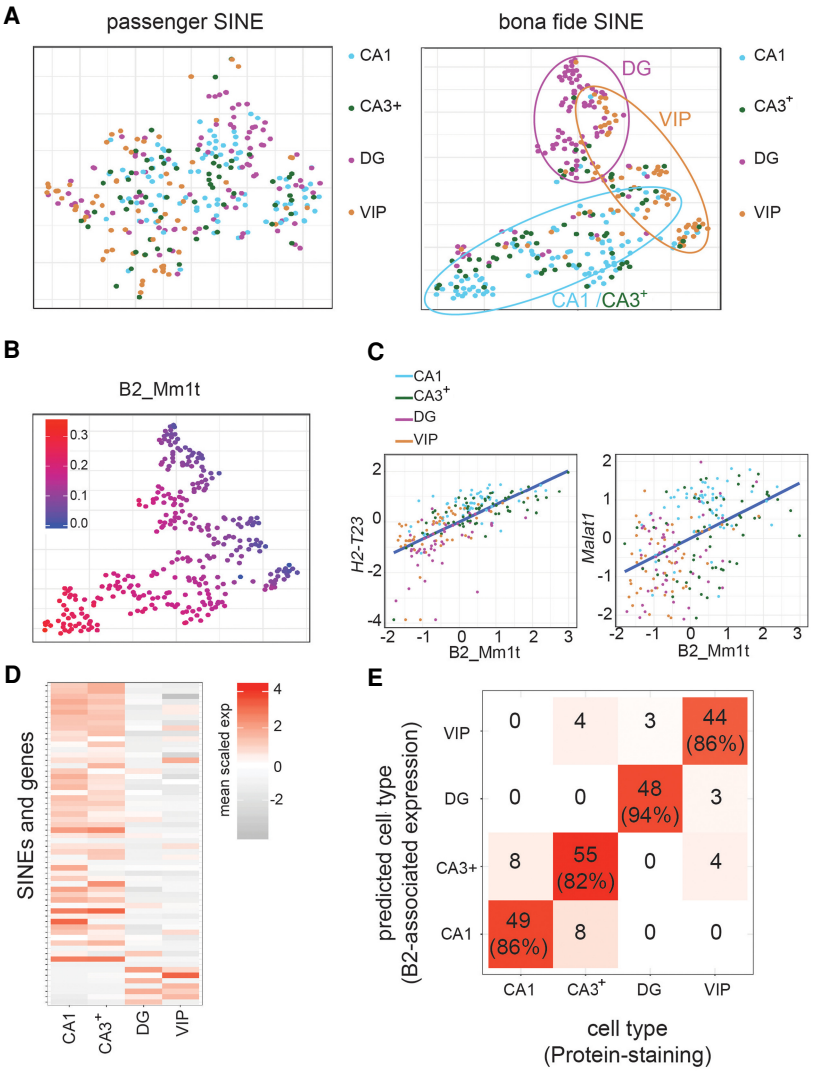

Figure 3. Cell type–specific expression of SINEs (*A*) t-SNE plot using passenger SINEs (*left*) or bona fide SINEs (*right*). Each dot represents a single nucleus colored by the staining pattern: PROX1$^-$CTIP2$^+$ = CA1, PROX1$^-$CTIP2$^-$ CA3$^+$, PROX1$^+$CTIP2$^+$ = DG, PROX1$^+$CTIP2$^-$ = VIP. (*B*) The same bona fide SINE t-SNE plot as in *A*, *right*, colored by B2-Mm1t expression. (*C*) Correlation between top B2-Mm1t–associated genes with B2-Mm1t expression in FOS$^-$ hippocampal neurons. (*D*) Heatmap of the average expression of all genes associated with B2_Mm1t as a function of cell type. (*E*) Confusion matrix from random forest model trained only on B2-associated SINEs transcripts to predict cell type–associated protein staining. Numbers indicate the number of nuclei predicted as the corresponding cell type.

previously identified to have a significant association with B2_Mm1t expression. Indeed, the classifier distinguished CA1 and CA3$^+$ neurons with a true positive rate of 86%, VIP neurons at a rate of 90%, and DG neurons at a rate of 94% (Fig. 3E). These rates were only slightly lower than our original estimates using all genes [CA1 = 91%, CA3 = 73%, VIP = 99%, DG = 100%) (Jaeger et al. 2018), indicating that B2 expression and B2 correlates shown cell-type specificity.

B2 retrotransposons are elevated in response to behaviorally relevant neural activity

To identify whether specific elements responded to neuronal activity, we analyzed FOS$^+$ neurons from mice that were exposed to a NE for 15 min, followed by 1 h in the HC. We have previously shown that this NE exposure is sufficient to elicit neural activity in the hippocampus, which can be detected by FOS staining (Fig. 4A; Lacar et al. 2016). When using only bona fide SINE expression, DG FOS$^+$ and FOS$^-$ nuclei clustered separately from one another, whereas CA1 and VIP FOS$^+$ and FOS$^-$ neurons were indistinguishable (Fig. 4B). Conversely, t-SNEs based on passenger elements showed no separation between FOS$^+$ and FOS$^-$ in any of the three cell types (Supplemental Fig. S8A), indicating a cell type–specific up-regulation of SINEs in response to neural activity. The top SINEs associated with FOS status in DG neurons were the B2 elements B2_Mm1a (logFC = 1.03, $P_{adj} < 1.07 \times 10^{-10}$) and B2_Mm1t (logFC = 0.96, $P_{adj} < 1.78 \times 10^{-9}$ (Fig. 4C,D).

To validate that B2 expression was indeed elevated in response to neural activity, we further examined SINE expression via northern blot. To ensure activation of a majority of hippocampal neurons, mice were injected peritoneally with PTZ, a Gamma-aminobutyric acid type A receptor (GABR) antagonist that inhibits inhibitory neurons, thereby activating the glutamatergic population throughout the brain. Hippocampi were collected and showed elevated expression of the IEGs *Arc* (Student's *t*-test *P*-value $< 5.6 \times 10^{-5}$) and *Fos* (Student's *t*-test *P*-value $< 2.8 \times 10^{-8}$) via qPCR (Supplemental Fig. S8B). B2 expression was not detected as being up-regulated via qPCR, likely owing to the inability to discriminate bona fide from passenger elements. Indeed, northern blot analysis with a probe to B2 showed elevated expression of the bona fide 180-bp band in PTZ-treated mice compared with saline-treated mice (Supplemental Fig. S8C). The high-molecular-weight smear is indicative of passenger B2 elements contained within poly(A)-containing transcripts. The presence of this high-molecular-weight smear underscores the necessity of discriminating bona fide from passenger elements when quantifying TE expression from RNA-seq data. The BC1 noncoding RNA did not show changes in expression in association with activity (Supplemental Fig. S8D). Together, these results indicated that bone fide B2 expression was up-regulated in response to neural activity in the mouse hippocampus, as indicated by BonaFide-TEseq analysis of the RNA-seq data.

Similar to results based on overall gene expression differences, DG neurons were the most transcriptionally active. In our previous study, we identified that these transcriptional changes extended to hours past the initial activating event (Jaeger et al. 2018). Therefore, we next wanted to determine the late expression dynamics of SINEs in DG neurons in response to neural activity.

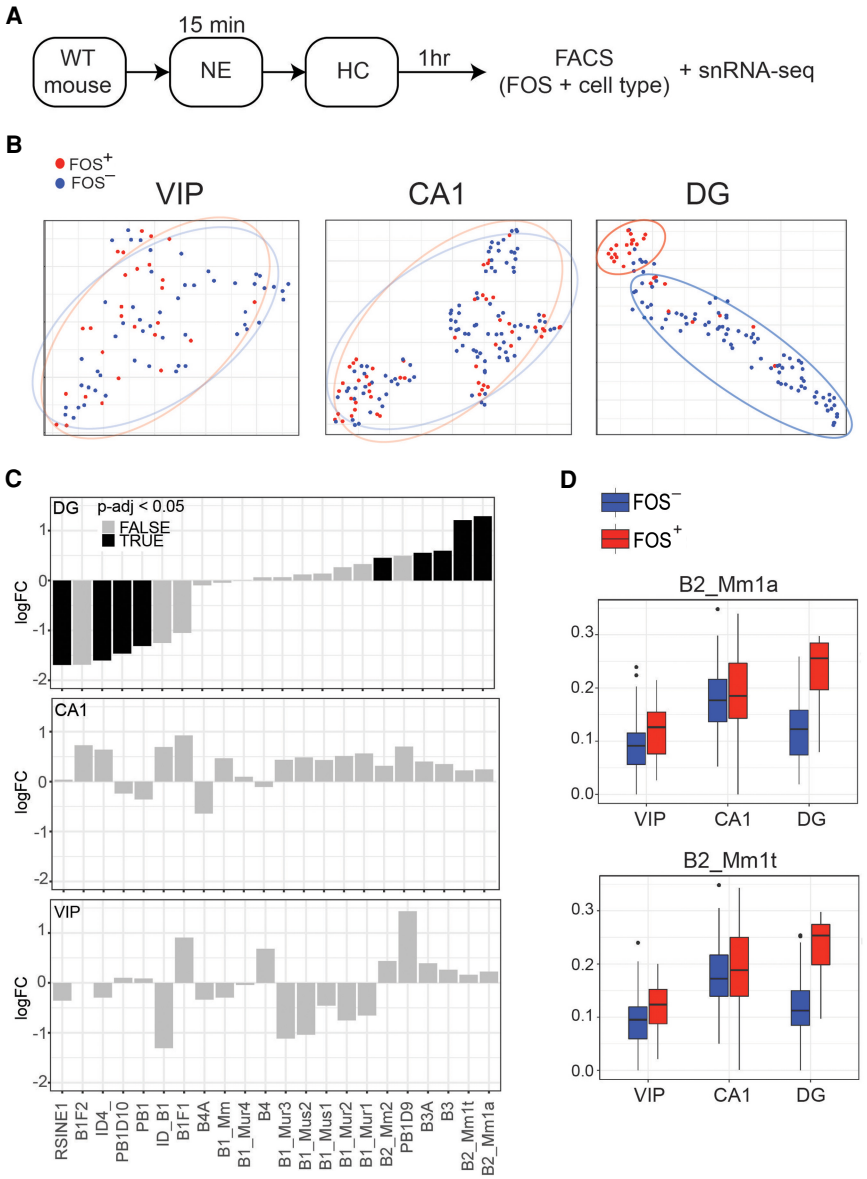

Figure 4. SINEs are up-regulated by activity in dentate granule (DG) cells. (*A*) Study design from Jaeger et al. (2018), in which mice were initially raised in a home-cage (HC) environment with minimal stimulation. They were then exposed to a novel environment (NE) for 15 min, which was sufficient to activate the IEG response. They were returned to the HC for 1 h to facilitate sufficient elevation of IEG expression. Nuclei were extracted, sorted on FOS, and prepared for snRNA-seq. (*B*) T-SNE of FOS^+ (red) and FOS^- (blue) nuclei from either VIP, CA1, or DG neurons. (*C*) log-Fold change (logFC) of all SINEs detected in DG, CA1, or VIP nuclei. Black = SINEs differentially expressed between FOS^+ versus FOS^- neurons after multiple-testing correction. (*D*) B2_Mm1a and B2_Mm1t expression in VIP, CA1, or DG neurons in FOS^- or FOS^+ nuclei. Boxes indicate upper, middle, and lower quartile range. Dots indicate measurements beyond 1.5 times the interquartile range.

The study of protein-coding genes identified three dynamics of activity-dependent genes: (1) short-term IEGs, elevated 1 h after exposure to novelty and returned to baseline by 4 h; (2) sustained IEGs, elevated at 1 h and sustained expression 4 h after exposure to novelty; and (3) late activity-dependent genes, elevated expression only 4 h after novelty (Fig. 5A,B; Jaeger et al. 2018). The identification of this last group of late activity-dependent genes indicated that a second wave of transcription occurred after the initial IEG response (Jaeger et al. 2018). Conversely, a t-SNE using only SINEs separated nuclei along a single axis (Fig. 5C). This axis was associated with FOS protein stain, indicating that it was linked to activity; however, unlike the separation based on protein-coding genes (i.e., *Sorcs3*), these "late" DG cells (DGCs), which were activated >4 h before sorting, did not separate out into their own cluster. Instead they clustered along with recently activated DGCs. This finding indicated that SINEs were up-regulated soon after activation and then maintained their expression level over the subsequent hours, a dynamic that was similar to other IEGs such as *Arc* (Fig. 5D,E). B2 elements most strongly illustrated this IEG-like response with elevated expression quickly, at 1 h, and sustained expression over the subsequent time points (Fig. 5E).

To determine which genes were associated with activity-dependent changes in B2 expression, we examined the association of repetitive elements and genes with B2_Mm1t after controlling for a continuous correlate of activity, *Arc* expression. This approach enabled the identification of genes that were associated with B2_Mm1t independently of the main effect of activity that might be induced by other confounding activity-dependent mechanisms in the cell (Fig. 6A). We identified 76 and 41 genes and repetitive elements that were positively and negatively associated with B2_Mm1t expression, respectively, with an FDR cutoff of 0.05 (Supplemental Table S2). As expected, the similarly expressed B2_Mm1a was the top correlated transcript in association with B2_Mm1t ($P_{adj} < 2.2 \times 10^{-16}$). Similar genes that were identified given only FOS^- neurons were again identified in the activated condition, indicating a consistent effect of B2 expression on the cell. Many of these genes were associated with response to innate immunity. For example, the top protein-coding genes that were positively associated with B2_Mm1t such as *H2-T23* ($P_{adj} < 6.2 \times 10^{-34}$) (Fig. 6B), *Malat1* ($P_{adj} < 2.99 \times 10^{-14}$), and *Tapbpl* ($P_{adj} < 1.09 \times 10^{-10}$) (Fig. 6C) are important in immune signaling pathways (Sarantopoulos et al. 2004; Ilca et al. 2018; Li et al. 2018), which are also elevated in response to SINE RNA (Kerur et al. 2013; Karijolich et al. 2015). In addition to innate immune genes, neuronal genes that respond to cellular stress such as *Inhba* ($P_{adj} < 1.0 \times 10^{-4}$) and *Lingo1* ($P_{adj} < 1.68 \times 10^{-7}$) were positively associated with B2 expression after multiple-testing correction. These results support further investigation into how B2 expression may be associated with an innate immune signaling following behaviorally relevant activity.

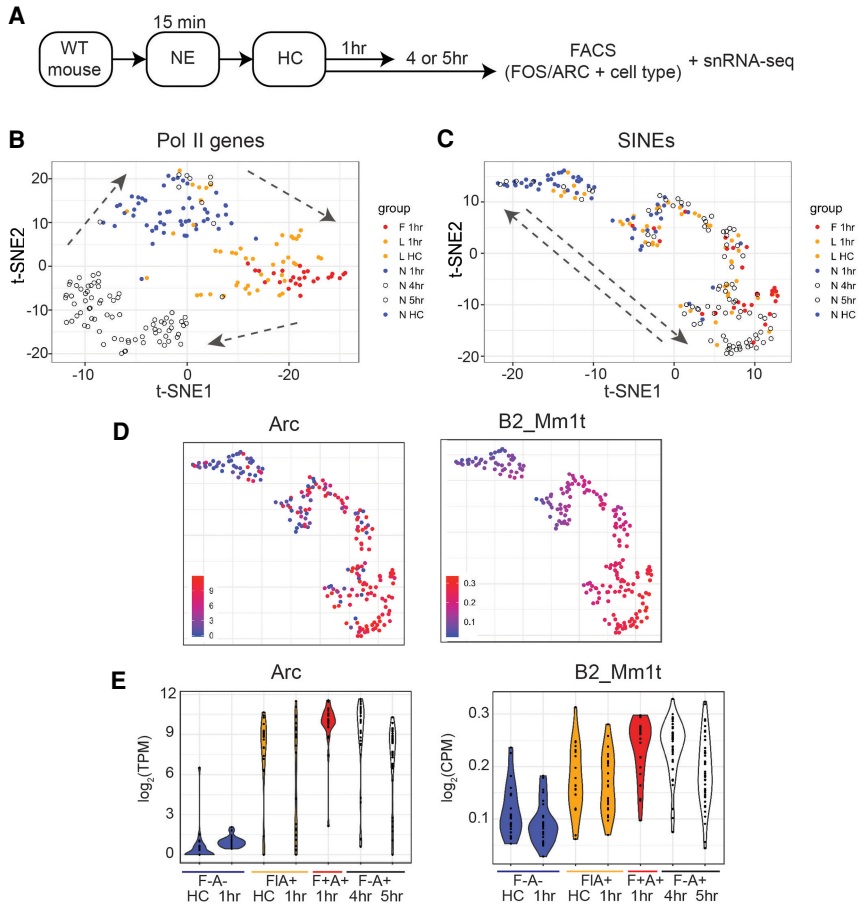

Figure 5. SINEs show IEG-like dynamics in DG neurons. (*A*) Study design from Jaeger et al. (2018), in which mice were raised in the HC, placed in a NE for 15 min, and then returned to the HC for 1, 4, or 5 h. The 4-h and 5-h nuclei were then sorted on FOS and ARC, as well as cell-type markers, and prepared for snRNA-seq. (*B*) T-SNE of DG nuclei from HC at the 1-, 4-, and 5-h time points using the standard RefSeq library of nonrepeat protein-coding and noncoding genes. Arrows denote the progression from an inactive state (blue; FOS^-) to an early gene response (orange and red; FOS^+) and then to a late gene response (open circles; FOS^-ARC^+). Orange, red = FOS^+ nuclei with low or high levels of FOS protein stain during sorting, respectively. (*C*) T-SNE of the same DG nuclei in *B*, but based on the expression of bona fide SINEs. Arrows indicate the progression from an inactive state (blue; FOS^-), through FOS^+ low nuclei (orange), to FOS^+ high (red), which coincides with late DG nuclei (FOS^-ARC^+). (*D*) The same t-SNE as in *C*, colored by *Arc* and B2_Mm1t expression. (*E*) Violin plots of Arc and B2_Mm1t expression by sorting group. (F) Fos, (A) Arc, (Fl) FOS low, (F+) FOS high.

Discussion

Retrotransposons make up an extensive proportion of the mammalian genome; however, knowledge of their function lags behind that of protein-coding genes, due in part to methodological constraints in quantifying their expression. Here, we show that Smart-seq2 and the single-cell 5′ gene expression kit from 10x, which are commonly used for single-cell RNA-seq, can be used to examine bona fide retrotransposon expression. Importantly, our filtering approach does not require TE-specific enrichment strategies during library preparation and can therefore be applied to any publicly available data set. In our approach, we have mapped reads to the consensus sequence, which provides an efficient way to simultaneously detect the TE subfamily and the position of the read within the consensus, allowing the detection of TE promoters. In addition to the internal Pol III promoter of B2 SINEs, previous studies indicate that it is likely that additional Pol III promoters exist upstream of these expressed elements. However, given that this approach does not identify exact genomic coordinates, the upstream promoter sequences remain unknown. Although we have validated the findings of this approach with orthogonal methods, an equally suitable alternative approach is to align TSO-containing reads to the genome. Genome alignment is equally compatible with filtering on TSO-containing reads and provides the ability to detect the expression of elements when the promoter has deviated from the consensus sequence but is still capable of driving efficient transcription.

SINEs are transcribed by Pol III but do not contain a canonical Pol III terminator within the consensus sequence. Pol III will therefore continue to transcribe until a sufficient terminator is reached (Orioli et al. 2011). Although Smart-seq provides a method to capture the 5′-end of RNA, it does not directly capture the 3′-end. Through direct 3′ and 5′ ligation, we showed the B2 ending at a Pol III terminator; it is also possible that the element has been truncated by post-transcriptional processing. By capturing the 3′-end of RNA transcripts through direct RNA ligation, we can sequence the downstream, nonrepetitive sequence that is transcribed owing to a lack of a Pol III terminator within SINEs. This provides an advantage when mapping SINEs to their position in the genome. Future work using long-read high-throughput sequencing techniques can therefore use this method to not only identify the level of SINE expression in single cells but also identify their locus of expression in the genome.

As a proof-of-concept, we used the BonaFide-TEseq approach to examine SINE expression at single-nucleus resolution in neurons from the mouse hippocampus. A key finding to note was the difference in biological interpretations that were derived from passenger TE transcripts versus bona fide TE transcripts. For example, cell type–dependent and activity-dependent TE findings were only observed from bona fide elements, with passenger elements showing no separation along either of these variables. Given that many current studies rely on generalized methods to quantify TE expression, such as standard qRT-PCR and RNA-seq, these results should serve as a cautionary note to future work attempting to identify the role of TEs in a system of interest.

Our findings show that there is a cell type–specific and cell state–dependent expression profile of SINEs, indicating that these elements are under tight regulatory constraints in hippocampal neurons. This is intriguing given the relatively recent evolution of B2 SINEs; however, the function related to this expression is unknown. It is intriguing to speculate what the downstream consequences might be of increased B2 expression in the dentate gyrus. Previous studies examining the physiological role of SINE

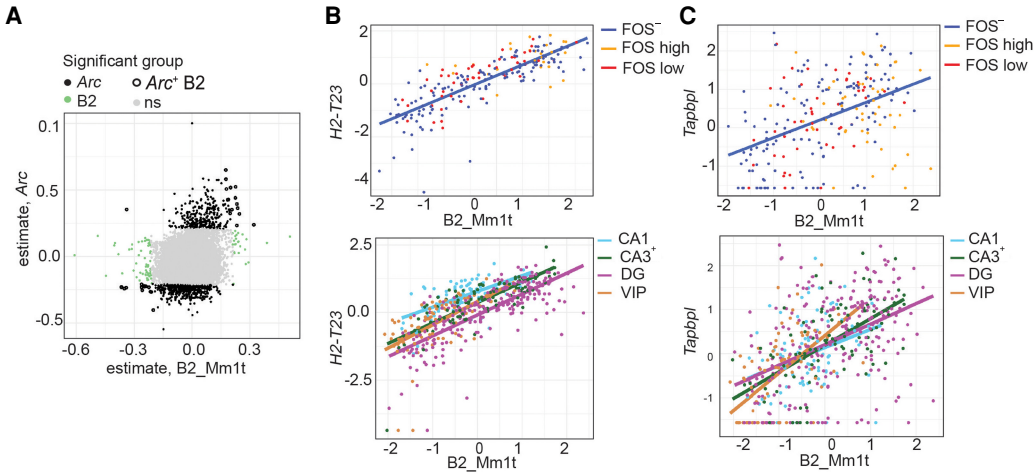

Figure 6. B2 element expression is associated with stress-response genes. (*A*) Correlation plot of the association of B2_Mm1t expression and a given gene versus the association of *Arc* expression and a given gene. Each dot represents a gene. (*B*,*C*) Association of B2_Mm1t expression with *H2-T23* (*B*) and *Tapbpl* (*C*) in inactive and activated DG neurons (*top*) or inactivated and activated neurons from all extracted hippocampal cell types (*bottom*).

RNA in other systems have shown that SINEs can impact gene regulation in *cis* and cellular function in *trans*. For example, B1 SINEs show *cis*-regulation of chromatin dynamics akin to enhancer RNAs, an action that is thought to facilitate activity-dependent Pol II transcription (Lunyak et al. 2007; Crepaldi et al. 2013; Policarpi et al. 2017). However, it is unclear if B2 elements serve a similar role. B2 elements are often studied in relationship to their regulatory role in *trans*. For example, B2, but not B1, SINEs can also have a global impact on transcription by nonspecifically blocking the binding of Pol II to DNA (Yakovchuk et al. 2009), and this effect is increased in response to the up-regulation of B2 after the cellular stress of heat shock (Allen et al. 2004). Furthermore, recent work indicates that B2 SINEs play a role in the transcription of stress-response genes by relieving transcriptional repression in response to cleavage by EZH2 (Zovoilis et al. 2016; Hernandez et al. 2020). Perhaps elevated SINE transcription is a way for the cell to temporally control activity-dependent changes.

In conclusion, we have used a novel method of analyzing Smart-seq2 RNA-seq data to examine bona fide TE expression at single-neuron resolution. Through this method, we have performed a proof-of-concept study in which we identified that SINE B2 elements are expressed in the healthy mouse hippocampus in a cell type–specific manner, with particularly elevated expression in DG neurons following neuronal activity induced by exploration of a NE. Together, these findings support the BonaFide-TEseq approach as a simple and robust method to assess TE expression at single-cell resolution.

Methods

Repeat element and gene expression estimation

For gene estimates, transcripts per million (TPM) values calculated by Jaeger et al. (2018), corresponding to the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) accession number GSE98679, were used. To obtain repeat element expression profiles, 50-bp reads from raw FASTQ files from GSE98679 or GSE67833 were trimmed using the dynamictrim algorithm (SolexaQA++ v3.1.3) (Cox et al. 2010). Reads were then aligned to the complete rodent RepeatMasker library using the Bowtie 2 algorithm (v2.3.5.1; k = 1) (Langmead and Salzberg 2012). Each read was matched to the 3′-end of the TSO sequence AACGCAG AGTAC. For each subfamily, a read was counted as bona fide if it contained the TSO sequence and was matched to the subfamily of interest. The distance to the start of the subfamily promoter was noted for each bona fide read for downstream promoter analysis. If a read was a match to a repetitive element but did not contain the TSO sequence, it was tagged as a passenger read. Reads were normalized by counts per million (CPM). Both gene TPM and repeat CPM values were $\log_2 + 1$ transformed for downstream analyses.

10x Analysis

5′ Gene expression data were downloaded from accession number GSE111360/SRR6798781 (Neal et al. 2018) using fastq-dump (v2.9.6). Reads were aligned to the human reference genome, GRCh38 (v1.2.0) using the count function within Cell Ranger (v3.0.2). The raw FASTQ file was then split into individual cells based on the barcode sequence "CB." Cells with less than 2500 reads were filtered from downstream analysis. Repeats abundance was estimated from each FASTQ file as described above.

TE abundance estimates with TEtranscripts and SQuIRE

TE abundance was estimated using TEtranscripts (v2.1.4) and SQuIRE (v0.9.9.92) with default settings. The general transfer format (GTF) file for TE annotations was downloaded from RepeatMasker, and the UCSC mm10 and RefSeq annotations were used for analysis.

Expressed sequence analysis

Reads from B2 or PB1 elements were extracted from the RepeatMasker aligned SAM files and aligned to one another through CLUSTAL Omega (dealign input = no, MBED-like clustering guide-tree = yes, MBED-like clustering iteration = yes, number of combined iterations = 0, max guide tree iterations = −1, MAX HMM iteration = −1) (Sievers et al. 2011). For comparison between the sequenced consensus and the reference sequences, PB1, B1, and B2 FASTA sequences were extracted from the rodent RepeatMasker reference set. The 7SL sequence was extracted from the mouse mm10 reference corresponding to gene name

Rn7s1. The consensus sequence and corresponding reference sequence were aligned using CLUSTAL Omega.

Cloning full-length B2 elements

Four mice (two HC, two pentylenetetrazol [PTZ] injected) were group housed in standard 75-sq-inch shoebox cages within a specific pathogen-free facility under a 12-h:12-h light–dark cycle with ad libitum access to food and water. Two female 8-wk-old wild-type C57BL/6 mice injected with 50 ng/g of PTZ were sacrificed by cervical dislocation 1 h following PTZ injection. The hippocampus was dissected following perfusion with PBS and placed in RNA-Bee for long-term storage. RNA was ethanol precipitated following phenol/chloroform separation. DNA adapters (/5Phos/AGTACTC TGCGTTGATACCACTGCTT/3ddC/) were adenylated on the 5′-end by incubating with 10× 5′-DNA adenylation reaction buffer (NEB), 1 mM ATP, and 100 pmol Mth RNA ligase (NEB) in nuclease-free water for 1 h at 65°C followed by heat inactivation for 5 min at 85°C. Adapters were recovered from solution by ethanol precipitation and stored for future use in nuclease-free water at −20°C. Adapters were ligated to the 3′-end of total RNA with T4 RNA ligase 2 (NEB) in the presence of 50% PEG 8000 and RNase inhibitor. Mixtures were incubated overnight at 4°C, and then the salts were removed by passing through a Qiagen RNeasy column and reconstituting in nuclease-free water. CDNA synthesis with template switching was performed by denaturing RNA in the presence of the ISPCR oligo (5′-AAGCAGTGGTATCAACGCAGAGT-3′), dNTPs, and water for 3 min at 72°C. The reverse transcription reaction (Protoscript RT, Protoscript Buffer, DTT, RNase inhibitor, $MgCl_2$, Betain, and TSO primer 5′-AAGCAGTGGTATCAACGC AGAGTACATrGrG + G-3′) was added to the denatured RNA and extended for 10 cycles with an extension time of 15 min at 70°C per cycle to ensure full-length cDNA generation. CDNA libraries were amplified for 32 cycles with KAPA HiFi HotStart ReadyMix with a primer complementary to the TSO adapter (ISPCR) and a primer complementary to the junction between the B2 promoter and TSO adapter (5′-AGAGTACGGGGGGCTGGTG-3′). The complete PCR reaction was then ligated into a pGEM-T easy vector using the standard protocol and transformed into top 10 competent cells with heat shock at 42°C and grown on ampicillin plates in the presence of IPTG and XGal. After overnight incubation, white colonies were placed in LB broth and incubated overnight followed by DNA extraction through miniprep. To check for insertions, plasmids were digested with EcoRI. Clones containing insertions were then Sanger sequenced (cf. Supplemental Data) and the full sequence analyzed for Pol III promoter and terminator sequences.

B2 northern blot

Two HC and two female 8-wk-old wild-type C57BL/6 mice injected with 50 ng/g of PTZ were sacrificed by cervical dislocation 1 h following PTZ injection. The hippocampus was dissected following perfusion with PBS and placed in RNA-Bee for long-term storage. QPCR validation of IEG expression was performed following cDNA synthesis with random hexamer primers using the SuperScript III RT. 2× SYBR Green was used, corresponding to the manufacturer's protocols with primers at a concentration of 5 µM.

To generate B2 probes, B2 sequence was extracted from mouse DNA using the forward primer 5′-GGGCTGGAGAGATGGCTC-3′ and the reverse primer 5′-TATTTATTATATGTGAGTACACTG-3′ originally published by Steck et al. (2010). The B2 probe was cloned into pGEM-T easy vector. The B2 probe was labeled with P32 CTP using the high prime DNA labeling kit with Klenow enzyme. A large 2% agarose gel was prepared with 10% MSE and 37% formaldehyde: running buffer = 1% MSE in DI water; sample buffer = formamide, formaldehyde, and 1% MSE. RNA was transferred onto the membrane in 10x SSC buffer overnight. Probes were hybridized overnight at 42°C with shaking.

Differential expression

All differential expression tests were performed on raw counts using the edgeR algorithm (Robinson et al. 2010).

Dimensionality reduction

Nuclei were excluded as outliers if they had fewer than 100,000 total aligned reads to the transcriptome or fewer than 4000 genes expressed at a cutoff of logFC = 1. The Barnes–Hut implementation of t-distributed stochastic neighbor embedding was used to calculate T-SNE coordinates through the Rtsne library in R version 3.5.0 (van der Maaten and Hinton 2008; van der Maaten 2014). In all cases, two output dimensions were calculated with a theta = 0.5, max_iter = 1000. Parameters for cell-type t-SNE calculations based on passenger or bona fide CPM values: initial dimensions = 15, perplexity = 22. Parameters for activity-dependent cell type–specific calculations based on passenger or bona fide CPM values: initial dimensions = 12, perplexity = 18. Parameters for long-term dynamics based on total gene expression: initial dimensions = 13, perplexity = 17. Parameters for long-term dynamics based on bona fide SINE CPM: initial dimensions = 12, perplexity = 18.

Correlation of SINE and gene expression

For analysis of B2_Mm1t expression in the absence of recent activity, we selected FOS^- DG ($PROX1^+CTIP2^+$) neurons sorted from mice that were either retained in the HC or exposed to a NE for 15 min followed by 1 h in the HC. For the association with B2_Mm1t in the context of activity, we selected FOS^-, FOS low, and FOS high nuclei from CA1, $CA3^+$, VIP, and DG neurons. Repetitive element expression CPM values and gene expression TPM values were scaled and combined into a single data set. Nuclei were excluded as outliers if they had fewer than 100,000 total aligned reads to the transcriptome or fewer than 4000 genes expressed at a cutoff of logFC = 1. Repetitive elements or genes were excluded from analysis if they were not expressed in any nuclei. Correlations were calculated using the lm function in R version 3.5.0. For calculations in the absence of activity, B2_Mm1t was coded as the response variable and the query gene as the explanatory variable. For calculations in the presence of activity, B2_Mm1t was coded as the response and an additive mixture of the query gene, and *Arc* expression was used for the explanatory variables. *P*-values of the effect of the explanatory variable on B2_Mm1t expression were corrected for multiple-testing with the p.adjust method in R (R Core Team 2010).

Competing interest statement

The authors declare no competing interests.

Acknowledgments

Research reported in this publication was supported by the National Institutes of Health (National Institute of Mental Health) under award numbers MH114030 and MH106882; the AHA-Allen Initiative in Brain Health and Cognitive Impairment award was made jointly through the American Heart Association and The Paul G. Allen Frontiers Group 19PABH134610000 and by the JPB Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views

of the NIH. Salk core facilities including the next-generation sequencing core are supported by the Salk Cancer Center with a grant from the National Cancer Institute (P30 CA014195). We thank Mary Lynn Gage for her assistance in editing.

Author contributions: S.B.L. designed the project, prepared the manuscript, and performed computational analyses and cloning of B2 transcripts. S.B.L. and L.R.-M. performed molecular biology experiments. K.K. performed motif analyses in both mouse and human genomes. F.Q. calculated TE abundances using TEtranscript and SQuIRE. S.B.L., B.N.J., and J.B. performed mouse PTZ experiments. F.H.G. assisted with the conceptual design and manuscript preparation.

References

Ahl V, Keller H, Schmidt S, Weichenrieder O. 2015. Retrotransposition and crystal structure of an Alu RNP in the ribosome-stalling conformation. *Mol Cell* **60:** 715–727. doi:10.1016/j.molcel.2015.10.003

Allen TA, Von Kaenel S, Goodrich JA, Kugel JF. 2004. The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock. *Nat Struct Mol Biol* **11:** 816–821. doi:10.1038/nsmb813

Ardeljan D, Taylor MS, Ting DT, Burns KH. 2017. The human long interspersed element-1 retrotransposon: an emerging biomarker of neoplasia. *Clin Chem* **63:** 816–822. doi:10.1373/clinchem.2016.257444

Athanasiadis A, Rich A, Maas S. 2004. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLoS Biol* **2:** e391. doi:10.1371/journal.pbio.0020391

Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, Zagalak JA, Emmett W, Smith CWJ, Luscombe NM, et al. 2018. Heteromeric RNP assembly at LINEs controls lineage-specific RNA processing. *Cell* **174:** 1067–1081.e17. doi:10.1016/j.cell.2018.07.001

Belancio VP, Roy-Engel AM, Pochampally RR, Deininger P. 2010. Somatic expression of LINE-1 elements in human tissues. *Nucleic Acids Res* **38:** 3909–3922. doi:10.1093/nar/gkq132

Brouha B, Schustak J, Badge RM, Lutz-Prigge S, Farley AH, Moran JV, Kazazian HH Jr. 2003. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci* **100:** 5280–5285. doi:10.1073/pnas.0831042100

Buckley PT, Lee MT, Sul JY, Miyashiro KY, Bell TJ, Fisher SA, Kim J, Eberwine J. 2011. Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons. *Neuron* **69:** 877–884. doi:10.1016/j.neuron.2011.02.028

Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. *BMC Bioinformatics* **11:** 485. doi:10.1186/1471-2105-11-485

Craig NL, Chandler M, Gellert M, Lambowitz A, Rice PA, Sandmeyer S. 2015. *Mobile DNA III*. ASM Press, Washington, DC.

Crepaldi L, Policarpi C, Coatti A, Sherlock WT, Jongbloets BC, Down TA, Riccio A. 2013. Binding of TFIIIC to sine elements controls the relocation of activity-dependent neuronal genes to transcription factories. *PLoS Genet* **9:** e1003699. doi:10.1371/journal.pgen.1003699

Criscione SW, Zhang Y, Thompson W, Sedivy JM, Neretti N. 2014. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics* **15:** 583. doi:10.1186/1471-2164-15-583

De Cecco M, Ito T, Petrashen AP, Elias AE, Skvir NJ, Criscione SW, Caligiana A, Brocculi G, Adney EM, Boeke JD, et al. 2019. L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature* **566:** 73–78. doi:10.1038/s41586-018-0784-9

Deininger P, Morales ME, White TB, Baddoo M, Hedges DJ, Servant G, Srivastav S, Smither ME, Concha M, DeHaro DL, et al. 2017. A comprehensive approach to expression of L1 loci. *Nucleic Acids Res* **45:** e31. doi:10.1093/nar/gkw1067

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet* **41:** 563–571. doi:10.1038/ng.368

Gao Z, Herrera-Carrillo E, Berkhout B. 2018. Delineation of the exact transcription termination signal for type 3 polymerase III. *Mol Ther Nucleic Acids* **10:** 36–44. doi:10.1016/j.omtn.2017.11.006

Geiduschek EP, Tocchini-Valentini GP. 1988. Transcription by RNA polymerase III. *Annu Rev Biochem* **57:** 873–914. doi:10.1146/annurev.bi.57.070188.004301

Gogolevskaya IK, Kramerov DA. 2010. 4.5S_I RNA genes and the role of their 5′-flanking sequences in the gene transcription. *Gene* **451:** 32–37. doi:10.1016/j.gene.2009.11.007

Hasler J, Strub K. 2006. *Alu* RNP and *Alu* RNA regulate translation initiation *in vitro*. *Nucleic Acids Res* **34:** 2374–2385. doi:10.1093/nar/gkl246

Hernandez AJ, Zovoilis A, Cifuentes-Rojas C, Han L, Bujisic B, Lee JT. 2020. B2 and ALU retrotransposons are self-cleaving ribozymes whose activity is enhanced by EZH2. *Proc Natl Acad Sci* **117:** 415–425. doi:10.1073/pnas.1917190117

Higgins DG, Sharp PM. 1988. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* **73:** 237–244. doi:10.1016/0378-1119(88)90330-7

Ilca FT, Neerincx A, Wills MR, de la Roche M, Boyle LH. 2018. Utilizing TAPBPR to promote exogenous peptide loading onto cell surface MHC I molecules. *Proc Natl Acad Sci* **115:** E9353–E9361. doi:10.1073/pnas.1809465115

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921. doi:10.1038/35057062

Jaeger BN, Linker SB, Parylak SL, Barron JJ, Gallina IS, Saavedra CD, Fitzpatrick C, Lim CK, Schafer ST, Lacar B, et al. 2018. A novel environment-evoked transcriptional signature predicts reactivity in single dentate granule neurons. *Nat Commun* **9:** 3084. doi:10.1038/s41467-018-05418-8

Jang KL, Latchman DS. 1989. HSV infection induces increased transcription of Alu repeated sequences by RNA polymerase III. *FEBS Lett* **258:** 255–258. doi:10.1016/0014-5793(89)81667-9

Jin Y, Tam OH, Paniagua E, Hammell M. 2015. TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics* **31:** 3593–3599. doi:10.1093/bioinformatics/btv422

Kaer K, Branovets J, Hallikma A, Nigumann P, Speek M. 2011. Intronic L1 retrotransposons and nested genes cause transcriptional interference by inducing intron retention, exonization and cryptic polyadenylation. *PLoS One* **6:** e26099. doi:10.1371/journal.pone.0026099

Kalkkila J-P, Sharp FR, Kärkkäinen I, Reilly M, Lu A, Solway K, Murrel M, Honkaniemi J. 2004. Cloning and expression of short interspersed elements B1 and B2 in ischemic brain. *Eur J Neurosci* **19:** 1199–1206. doi:10.1111/j.1460-9568.2004.03233.x

Karijolich J, Abernathy E, Glaunsinger BA. 2015. Infection-induced retrotransposon-derived noncoding RNAs enhance herpesviral gene expression via the NF-κB pathway. *PLoS Pathog* **11:** e1005260. doi:10.1371/journal.ppat.1005260

Karijolich J, Zhao Y, Alla R, Glaunsinger B. 2017. Genome-wide mapping of infection-induced SINE RNAs reveals a role in selective mRNA export. *Nucleic Acids Res* **45:** 6194–6208. doi:10.1093/nar/gkx180

Kerur N, Hirano Y, Tarallo V, Fowler BJ, Bastos-Carvalho A, Yasuma T, Yasuma R, Kim Y, Hinton DR, Kirschning CJ, et al. 2013. TLR-independent and P2X7-dependent signaling mediate *Alu* RNA-induced NLRP3 inflammasome activation in geographic atrophy. *Invest Ophthalmol Vis Sci* **54:** 7395–7401. doi:10.1167/iovs.13-12500

Koval AP, Gogolevskaya IK, Tatosyan KA, Kramerov DA. 2012. Complementarity of end regions increases the lifetime of small RNAs in mammalian cells. *PLoS One* **7:** e44157. doi:10.1371/journal.pone.0044157

Lacar B, Linker SB, Jaeger BN, Krishnaswami S, Barron J, Kelder M, Parylak S, Paquola A, Venepally P, Novotny M, et al. 2016. Nuclear RNA-seq of single neurons reveals molecular signatures of activation. *Nat Commun* **7:** 11022. doi:10.1038/ncomms11022

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9:** 357–359. doi:10.1038/nmeth.1923

Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. 2017. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res* **45:** e17. doi:10.1093/nar/gkx334

Li Z, Zhang Q, Wu Y, Hu F, Gu L, Chen T, Wang W. 2018. lncRNA Malat1 modulates the maturation process, cytokine secretion and apoptosis in airway epithelial cell-conditioned dendritic cells. *Exp Ther Med* **16:** 3951–3958. doi:10.3892/etm.2018.6687

Liu WM, Chu WM, Choudary PV, Schmid CW. 1995. Cell stress and translational inhibitors transiently increase the abundance of mammalian SINE transcripts. *Nucleic Acids Res* **23:** 1758–1765. doi:10.1093/nar/23.10.1758

Llorens-Bobadilla E, Zhao S, Baser A, Saiz-Castro G, Zwadlo K, Martin-Villalba A. 2015. Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell* **17:** 329–340. doi:10.1016/j.stem.2015.07.002

Lunyak VV, Prefontaine GG, Nunez E, Cramer T, Ju BG, Ohgi KA, Hutt K, Roy R, Garcia-Diaz A, Zhu X, et al. 2007. Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* **317:** 248–251. doi:10.1126/science.1140871

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562. doi:10.1038/nature01262

Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH. 2005. Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature* **435:** 903–910. doi:10.1038/nature03663

Neal JT, Li X, Zhu J, Giangarra V, Grzeskowiak CL, Ju J, Liu IH, Chiou SH, Salahudeen AA, Smith AR, et al. 2018. Organoid modeling of the tumor immune microenvironment. *Cell* **175:** 1972–1988.e16. doi:10.1016/j.cell.2018.11.021

Oler AJ, Traina-Dorge S, Derbes RS, Canella D, Cairns BR, Roy-Engel AM. 2012. Alu expression in human cell lines and their retrotranspositional potential. *Mob DNA* **3:** 11. doi:10.1186/1759-8753-3-11

Orioli A, Pascali C, Quartararo J, Diebel KW, Praz V, Romascano D, Percudani R, van Dyk LF, Hernandez N, Teichmann M, et al. 2011. Widespread occurrence of non-canonical transcription termination by human RNA polymerase III. *Nucleic Acids Res* **39:** 5499–5512. doi:10.1093/nar/gkr074

Ostertag EM, DeBerardinis RJ, Goodier JL, Zhang Y, Yang N, Gerton GL, Kazazian HH Jr. 2002. A mouse model of human L1 retrotransposition. *Nat Genet* **32:** 655–660. doi:10.1038/ng1022

Percharde M, Lin CJ, Yin Y, Guan J, Peixoto GA, Bulut-Karslioglu A, Biechele S, Huang B, Shen X, Ramalho-Santos M. 2018. A LINE1-nucleolin partnership regulates early development and ESC identity. *Cell* **174:** 391–405.e19. doi:10.1016/j.cell.2018.05.043

Picelli S, Faridani OR, Björklund AK, Winberg G, Sagasser S, Sandberg R. 2014. Full-length RNA-seq from single cells using smart-seq2. *Nat Protoc* **9:** 171–181. doi:10.1038/nprot.2014.006

Policarpi C, Crepaldi L, Brookes E, Nitarska J, French SM, Coatti A, Riccio A. 2017. Enhancer SINEs link Pol III to Pol II transcription in neurons. *Cell Rep* **21:** 2879–2894. doi:10.1016/j.celrep.2017.11.019

R Core Team. 2010. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Rudin CM, Thompson CB. 2001. Transcriptional activation of short interspersed elements by DNA-damaging agents. *Genes Chromosomes Cancer* **30:** 64–71. doi:10.1002/1098-2264(2000)9999:9999<::AID-GCC1066>3.0.CO;2-F

Sarantopoulos S, Lu L, Cantor H. 2004. Qa-1 restriction of $CD8^+$ suppressor T cells. *J Clin Invest* **114:** 1218–1221. doi:10.1172/JCI23152

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol Syst Biol* **7:** 539. doi:10.1038/msb.2011.75

Steck E, Burkhardt M, Ehrlich H, Richter W. 2010. Discrimination between cells of murine and human origin in xenotransplants by species specific genomic in situ hybridization. *Xenotransplantation* **17:** 153–159. doi:10.1111/j.1399-3089.2010.00577.x

Teissandier A, Servant N, Barillot E, Bourc'his D. 2019. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mob DNA* **10:** 52. doi:10.1186/s13100-019-0192-1

Thomas CA, Tejwani L, Trujillo CA, Negraes PD, Herai RH, Mesci P, Macia A, Crow YJ, Muotri AR. 2017. Modeling of TREX1-dependent autoimmune disease using human stem cells highlights L1 accumulation as a source of neuroinflammation. *Cell Stem Cell* **21:** 319–331.e8. doi:10.1016/j.stem.2017.07.009

Treangen TJ, Salzberg SL. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13:** 36–46. doi:10.1038/nrg3117

van der Maaten LJP. 2014. Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res* **2014:** 3221–3245.

van der Maaten LJP, Hinton GE. 2008. Visualizing high-dimensional data using t-SNE. *J Mach Learn Res* **9:** 2579–2605.

Yakovchuk P, Goodrich JA, Kugel JF. 2009. B2 RNA and Alu RNA repress transcription by disrupting contacts between RNA polymerase II and promoter DNA within assembled complexes. *Proc Natl Acad Sci* **106:** 5569–5574. doi:10.1073/pnas.0810738106

Yang WR, Ardeljan D, Pacyna CN, Payer LM, Burns KH. 2019. SQuIRE reveals locus-specific regulation of interspersed repeat expression. *Nucleic Acids Res* **47:** e27. doi:10.1093/nar/gky1301

Zarnack K, Konig J, Tajnik M, Martincorena I, Eustermann S, Stevant I, Reyes A, Anders S, Luscombe NM, Ule J. 2013. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell* **152:** 453–466. doi:10.1016/j.cell.2012.12.023

Zhang Y, Romanish MT, Mager DL. 2011. Distributions of transposable elements reveal hazardous zones in mammalian introns. *PLoS Comput Biol* **7:** e1002046. doi:10.1371/journal.pcbi.1002046

Zovoilis A, Cifuentes-Rojas C, Chu HP, Hernandez AJ, Lee JT. 2016. Destabilization of B2 RNA by EZH2 activates the stress response. *Cell* **167:** 1788–1802.e13. doi:10.1016/j.cell.2016.11.041

Received February 10, 2020; accepted in revised form September 29, 2020.