Nucleic Acids Research, 2024, **52**, 5166–5178
https://doi.org/10.1093/nar/gkae273
Advance access publication date: 22 April 2024
Genomics

Comprehensive profiling of L1 retrotransposons in mouse

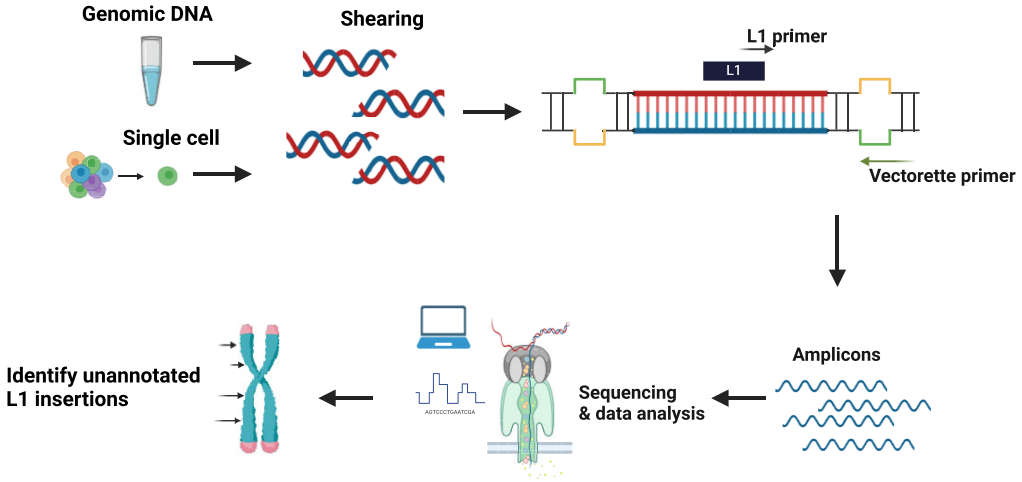
Xuanming Zhang[1,2], Ivana Celic[1,2], Hannah Mitchell[1,2], Sam Stuckert[1,2], Lalitha Vedula[1,2] and Jeffrey S. Han[1,2,*]

[1]Department of Biochemistry and Molecular Biology, Tulane University School of Medicine, New Orleans, LA 70112, USA
[2]Tulane Cancer Center, Tulane University School of Medicine, New Orleans, LA 70112, USA
[*]To whom correspondence should be addressed. Tel: +1 504 988 5920; Email: jhan5@tulane.edu

Abstract

L1 elements are retrotransposons currently active in mammals. Although L1s are typically silenced in most normal tissues, elevated L1 expression is associated with a variety of conditions, including cancer, aging, infertility and neurological disease. These associations have raised interest in the mapping of human endogenous *de novo* L1 insertions, and a variety of methods have been developed for this purpose. Adapting these methods to mouse genomes would allow us to monitor endogenous *in vivo* L1 activity in controlled, experimental conditions using mouse disease models. Here, we use a modified version of transposon insertion profiling, called nanoTIPseq, to selectively enrich young mouse L1s. By linking this amplification step with nanopore sequencing, we identified >95% annotated L1s from C57BL/6 genomic DNA using only 200 000 sequencing reads. In the process, we discovered 82 unannotated L1 insertions from a single C57BL/6 genome. Most of these unannotated L1s were near repetitive sequence and were not found with short-read TIPseq. We used nanoTIPseq on individual mouse breast cancer cells and were able to identify the annotated and unannotated L1s, as well as new insertions specific to individual cells, providing proof of principle for using nanoTIPseq to interrogate retrotransposition activity at the single-cell level *in vivo*.

Graphical abstract



Introduction

Transposable elements (TEs) are often referred to as 'genetic parasites' that have thrived and mobilized in mammalian genomes for millions of years (1,2). TEs have played a critical role in shaping human genomes, making up at least 45% of the human genome, with long interspersed elements (LINE-1, L1) comprising 17% (1). L1 has been extensively investigated due to not only its abundance in the genome but also its activity in humans, which can drive mutations and endanger the integrity of the genome (3–7). L1 mobilizes in the genome via a 'copy and paste' mechanism, utilizing two proteins encoded by the L1 ORF1 and ORF2 genes (8,9). L1 ORF1p possesses nucleic acid binding activity (10), presumably facilitating the binding of L1 messenger RNA (mRNA) in *cis* (11) to form the ribonucleoprotein (RNP) complex in the cytoplasm (12–15). The L1 RNP complex is then transported to the nucleus through an uncertain mechanism that may vary depending on cell type (16,17). ORF2p contains an endonuclease domain that nicks the genome at the 5′-TTTT/AA-3′ preferred insertion site (8,18), followed by converting L1 mRNA to complementary DNA using ORF2p reverse transcriptase (RT) (19). This process is called target-primed reverse transcription, and the mobilization of L1 is called retrotransposition (20,21). L1 RT is susceptible to early termination often resulting in

Received: November 13, 2023. **Revised:** March 25, 2024. **Editorial Decision:** April 1, 2024. **Accepted:** April 6, 2024

5′ truncated L1 insertions, which are unable to further retrotranspose in the genome. It has been reported that the majority of new L1 insertions are caused by a few 'hot' transposable full-length L1s (3). The human genome has ∼100 L1s capable of retrotransposition (3,5) and L1 dysregulation has been shown to be associated with many health-related issues (6,7,22–28). Mouse models exist for many of these conditions; for example, L1 overexpression is observed in piRNA mutant male mice germ cells, where massive DNA damage and meiotic arrest occur (29–31). The presumption is that L1 overexpression leads to these genotoxic effects due to endonuclease-mediated cutting of DNA and possibly new L1 insertions. However, to date there is no clear-cut, direct evidence for meiotic *de novo* retrotransposition of endogenous L1s in these mice. Tracking endogenous L1 insertion activity in various mouse disease models would be beneficial for understanding the role and impact of L1 on these conditions. In addition, L1 has been found to insert at CRISPR–Cas9 editing sites (32), demonstrating the potential impact of L1 on CRISPR-related gene therapy and the importance of monitoring new L1 retrotransposition events. A variety of methods have been used to profile human L1 insertions, including bioinformatic analysis of high-coverage whole-genome sequencing (WGS) and targeted L1 enrichment followed by sequencing (33–36). Early forays into profiling mouse L1s have begun but can be costly due to the amount of sequencing data required (37,38).

TIPseq (transposon insertion profiling by sequencing) is a promising method for comprehensive profiling of fixed, polymorphic and *de novo* transposon insertions (35,39). In TIPseq, a transposon-specific primer is used in a ligation-mediated polymerase chain reaction (PCR) to enrich for transposon sequences. The resulting amplicon mixture is subjected to paired-end short-read deep sequencing. This method has been used to successfully profile the currently active family of human L1 retrotransposons, L1Hs, in a variety of samples (35). The success of this protocol relies on the specificity of the L1H-specific primer. Adapting TIPseq to mouse genomes is desirable because it would allow us to monitor endogenous *in vivo* L1 activity in controlled, experimental conditions using mouse disease models. However, adapting L1 TIPseq to mouse genomes presents additional challenges. In mouse (as compared to humans), not only are there more active transposons (40), there are multiple active L1 subfamilies (Tf, Gf and A) (2,41). Within the active mouse L1 subfamilies, there are more young, potentially active family members. There are an estimated 30× more active L1s in mouse as compared to the active L1 population in humans (41). Thus, we would expect that comprehensive profiling of this larger and more complex population of young L1s in mice would require greater sequencing depth and specificity.

Here, we describe a modified TIPseq protocol to capture young, actively retrotransposing L1 elements in mouse. We focus on the L1MdTf subfamily, as this family is responsible for the vast majority (>80%) of *de novo* L1 insertions in mouse (38,42). This modified protocol utilizes long-read nanopore sequencing and allows the comprehensive, unambiguous identification of fixed, polymorphic and *de novo* L1MdTf elements from pooled or single cells. As little as 200 000 reads [115 million bases (MB)] of sequencing data can be used to identify 98% of L1MdTf elements in a genome, making this a cost-effective method of transposon profiling that can be routinely implemented even by smaller labs.

Materials and methods

Genomic DNA isolation

Mouse tail (1 cm) was used to extract the genomic DNA (gDNA) using NEB Monarch® Genomic DNA Purification Kit (NEB T3010L). DNA concentration was measured on Qubit dsDNA BR Assay Kit (cat Q32856) and visualized on 1% agarose gel.

Library preparation and sequencing

Purified gDNA was sheared to an average size of 1500 bp using ultrasonication (Covaris M220). Library input was examined on an Agilent Bioanalyzer 2100 using the high-sensitivity DNA kit (cat 5067-4626). The sheared DNA was end polished (NEBNext® Ultra™ II End Repair/dA-Tailing Module E7546S), followed by ligation to T-overhang vectorette adapters (39,43) (NEBNext® Quick Ligation Module E6056S; see Supplementary Table S1 for primer sequences). We recommend cleaning up the ligation product before PCR amplification to remove excessive adapters that minimize potential undesired concatemers and other nonspecific amplification. We used bead purification (AMPure XP Reagent A63880, 1× ratio) to remove small DNA fragments and excessive adapters. Touchdown PCR (Supplementary Table S2) with Platinum™ II Taq Hot-Start DNA Polymerase (cat 14966001) was performed to balance yield and specificity, using an L1-specific primer and vectorette primers. The PCR products were cleaned up with beads (0.7× ratio) and quality controlled using Qubit assays and the Agilent Bioanalyzer 2100. For short-read sequencing, PCR amplicons were sheared to a size of ∼400 bp and sequenced by BGI (150-bp paired-end reads, 30 million reads ∼ 10 Gb data).

To sequence using Oxford Nanopore Technologies (ONT), we used a MinION sequencer with R10.4.1 flow cells. Purified TIPseq PCR amplicons were used to construct the ONT library using Native Barcoding Kit 24 V14 (SQK-NBD114.24). Sequencing was carried out in-house to our desired sequence depth. For WGS, high molecular weight gDNA was used as the direct input for the Native Barcoding Kit 24 V14.

L1 annotation curation

TIPseqHunter utilizes a small set of 200 high-confidence 'fixed present' L1H insertions in humans, providing essential alignment characteristics for building an accurate machine learning model (35). We downloaded all L1MdTf_I to L1MdTf_III elements from the UCSC table browser using the filter option (mm39, filtered for L1MdTf*) for mouse L1 annotation. We manually curated the annotation file to better represent the active L1 population of interest to us. We removed L1s that did not contain the TuJH922 primer sequence, as they were presumably older elements not relevant for our purpose. We also separated individual elements that were bookended together erroneously as a single L1 element (e.g. Supplementary Table S3). This step effectively removed most L1MdTf_III elements (older elements) and left us with 3266 elements, which better represented younger and active L1s in the mouse genome.

TIPseqHunter data analysis

For short-read TIPseq data, we ran the TIPseqHunter pipeline as previously described (35,39). Briefly, paired-end reads were

quality controlled and trimmed to remove Illumina adapter sequences, vectorette adapter sequences and reads with poor quality scores. The reads were subsequently mapped to the mm39 reference genome where L1MdTf_I to L1MdTf_III elements were masked, using bowtie2/2.3.3 with the following settings: -X 1000 –local –phred33 –sensitive. l1hskey was changed to G(6833) to map the 5′ most position of TuJH922 primer in the mouse L1MdTf_I consensus sequence. We replaced the human L1 reference with our customized mouse L1 annotation file.

For downsampling, seqtk sample was used to scale down to 0.01×, 0.1×, 0.2× and 0.5× of the original fastq files in triplicates using different seeding keys. The scaledown fastq files were used to run TIPseqHunter as described above.

Long-read TIPseq data analysis

The raw sequencing data were processed during sequencing runs by the MinKNOW application (fast basecalling model, trim barcodes on). Reads were aligned to the mm39 reference genome using the integrated aligner, minimap2. Default settings were applied, producing alignments in the bam format. To extract the clipped sequences from each alignment, the SamExtractClip tool (Jvarkit) was utilized. These clipped sequences were then remapped to the L1MdTf_I consensus sequence (44) using bowtie2. A customized scoring system (–sensitive -N 1 –mp 1,1 –rdg 5,2 –rfg 5,2) was employed to achieve a balance between sensitivity and accuracy. Reads that mapped to the L1MdTf_I consensus sequences were filtered based on the presence of a poly(A) tail at the end of the alignment. By default, the filtering required at least five consecutive As at the end. The filtered reads were subsequently used to retrieve the original mapping location in mm39 using the read name, where the mapped intervals were added to the potential insertion list, which was formatted as a bed file. The bedtools cluster function (45) was employed to identify regions with individual reads. Regions with more than three supporting reads were merged using the bedtools merge function, resulting in the final set of insertion candidates. The precise insertion site was determined using an in-house script based on the clipped position obtained in the header region from the SamExtractClip. The full pipeline is available at https://github.com/JHanLab/NanoTipSeq.

Single-cell whole-genome amplification

Cultured 4226 cells (46) were washed with phosphate-buffered saline and stained with propidium iodide. Single live cells were sorted into individual wells of a 96-well LoBind plate by the Louisiana State University Cellular Immunology and Immune Metabolism Core, based on PI staining, forward scatter and side scatter. Freshly sorted single cells were used to perform single-cell whole-genome amplification by either MDA (multiple displacement amplification) or PTA (primary template-directed amplification) following manufacturer's recommendations. Qiagen REPLI-g Single Cell Kit (cat 150343) was used for MDA, and BioSkryb Genomics ResolveDNA® Kit was used for PTA. The amplified genomes were quality controlled using Qubit assays and the Agilent Bioanalyzer 2100. Amplified DNA was used in our TIPseq protocols, library construction and sequencing as described above.

Results and discussion

L1 primer design

The L1MdTf lineage contains three families, namely L1MdTf_I, L1MdTf_II and L1MdTf_III. These families are responsible for all of the reported spontaneous *de novo* germline insertions in the literature (42). Because we were unable to identify primers that would specifically amplify all three L1MdTf families, we chose to focus on targeting L1MdTf_I and L1MdTf_II, as these two families account for 80% of the reported spontaneous *de novo* germline insertions and are more closely related to each other than to L1MdTf_III. L1MdTf_I and L1MdTf_II are also younger than L1MdTf_III, with average age of 0.25 million and 0.27 million years, respectively (2). Thus, we reasoned that monitoring L1MdTf_I and L1MdTf_II insertions would be an acceptable proxy for L1 activity. Our strategy to target L1MdTf_I and L1MdTf_II elements involved designing L1 primers in the L1 3′ untranslated region (UTR), where the L1MdTf_I and L1MdTf_II subfamilies exhibited substantial differences from other subfamilies. One of the specific primers TuJH922 is shown in Figure 1A. We inspected the multi-sequence alignment of consensus sequences from each family. We found a conserved 17-bp region where an ATA→GTG change appears to be diagnostic for L1MdTf_I and L1MdTf_II. We then designed the TuJH922 primer spanning this region with 3′ end terminating at this diagnostic GTG trinucleotide. We performed an *in silico* search for TuJH922 sequence in all C57BL/6 L1 elements (total of 410 549 elements). A perfect match for TuJH922 was only found in 56.6% of L1MdTf_I (1479/2613) and 51.3% of L1MdTf_II (1950/3801) elements but rarely in other subfamilies (24 total instances from all other subfamilies combined), even though some subfamilies are closely related to L1MdTf_I and L1MdTf_II elements (Figure 1B). The phylogenetic tree shown was built using the L1 ORF2 sequence (2) (Figure 1B). Moreover, all previously identified 'hot' L1MdTf_I and L1MdTf_II elements or *de novo* germline insertions contain an exact match for TuJH922 in their 3′ UTR (when sequence spanning that region is available to examine) (42,47), indicating the suitability of TuJH922 to target young, potentially transposable L1 elements in the mouse genome.

Mouse L1 TIPseq using short-read sequencing

The original TIPseq protocol relies on ligation-mediated PCR to selectively amplify L1Hs from human DNA, involving five main steps:

1. gDNA extraction and digestion. Six 5- or 6-base cutter restriction enzymes are used individually to digest the genome to achieve even fragmentation of gDNA.
2. Vectorette adapter ligation. The vectorette adapters consist of two short oligonucleotides that are complementary on both ends but mismatched in the center. This structure allows amplifying a region of interest with knowing only one end of the target.
3. TIPseq PCR by using an L1-specific primer and vectorette primer. The vectorette primer has identity to the bottom mismatched strand of the vectorette adapter, which has no complementary sequence to anneal. The complement of the bottom strand of the vectorette adapter is only present after synthesis from the L1

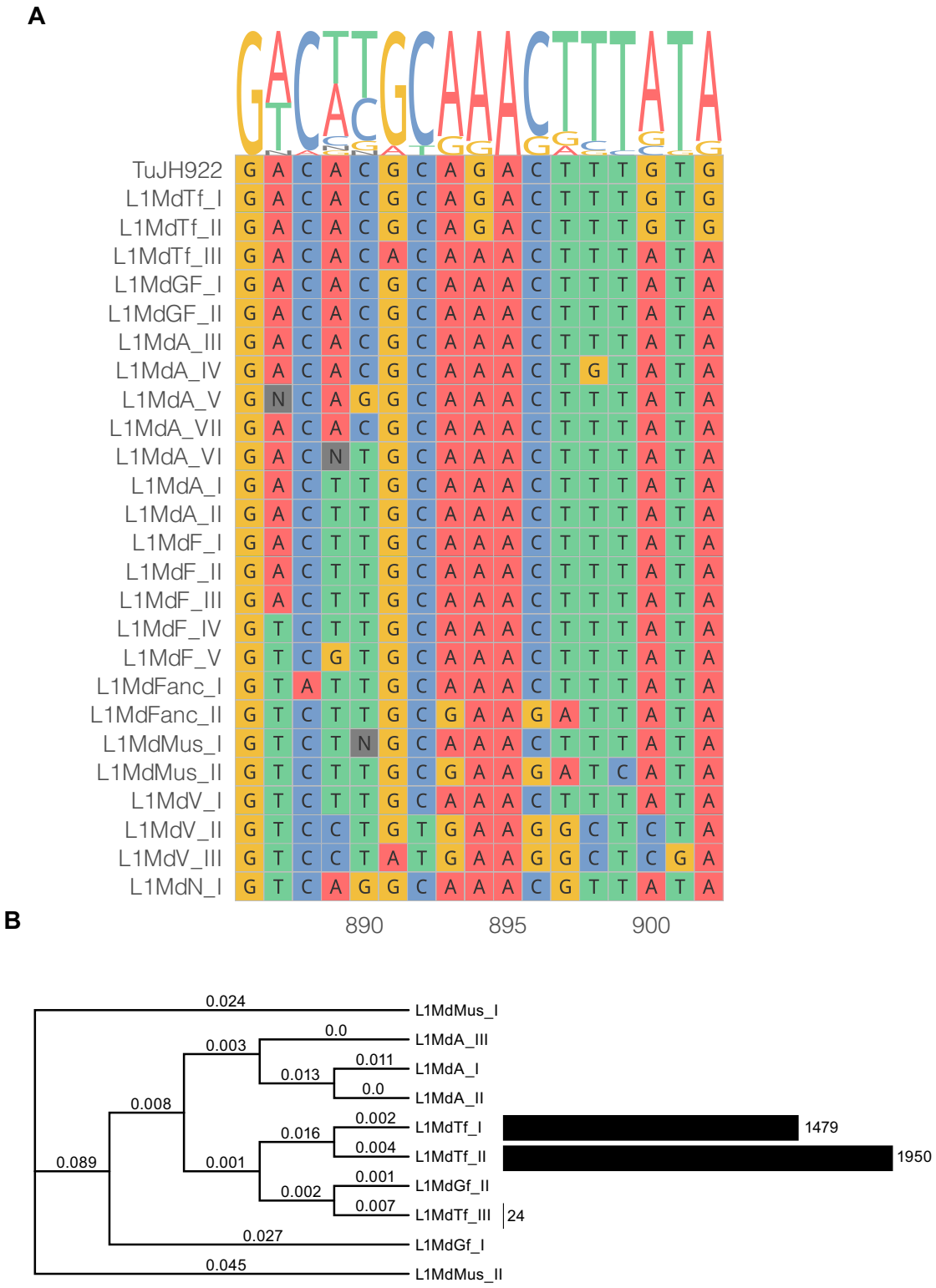


Figure 1. Design of an L1MdTf-specific primer. (**A**) Mouse L1-specific primer design based on alignment of consensus sequences from major L1 subfamilies. We selected TuJH922 for its diagnostic trinucleotides 'GTG' in the 3′ end, which distinguish young active L1s from inactive old L1s. (**B**) Number of elements containing an exact match of TuJH922 primer sequence for each L1 subfamily.

primer, allowing specific amplification of L1 adjacent sequence.
4. Amplicon shearing and 150-bp paired-end sequencing.
5. Machine learning algorithm to identify annotated L1s and new L1 insertions.

We used C57BL/6 gDNA to evaluate the sensitivity of the TIPseq pipeline for identifying L1 elements in the mouse genome through short-read sequencing (Figure 2A). TIPseq PCR amplicons were sequenced to a depth of 30 million reads (150 bp × 2, 10 Gb of data) for each sample. We employed alignment evidence such as genomic–genomic read pairs and L1–genomic junction reads (Figure 2B and C) to detect L1 elements in the short-read data. TIPseqHunter (35,39) was used to analyze the short-read data, and we successfully identified over 90% (∼3000) of pre-existing L1MdTf_I and L1MdTf_II (annotated L1s) using different designs of L1-specific primers (Figure 2D). The primers performed similarly with respect to annotated L1 identification (Figure 2D) and the primers had similar coverage of L1 3′ UTR (Figure 2E). To assess primer specificity, we aligned reads to the L1MdTf_I consensus sequence and found that 99% of reads contained the diagnostic 'GTG' nucleotides (Supplementary Figure S1). In contrast, TIPseq amplicons generated with a less specific primer (TuJH801) resulted in only 25% of reads containing the GTG nucleotides and were largely comprised of older mouse L1s. To test how much data we needed to achieve good recovery of annotated L1s, we employed multiple rounds of randomly selecting 50%, 20%, 10% and 1% of the original data and analyzing with the TipseqHunter pipeline. With as little as 10% of the original data, we were able to detect over 79% of annotated L1MdTf_I and L1MdTf_II L1s from sample using the TuJH922 primer (Figure 2D and Supplementary Figure S2). For samples made from TuJH922, the average annotated L1 recovery rate was 90%, 88%, 84%, 79% and 44% for 100%, 50%, 20%, 10% and 1% of the original data, respectively. These results demonstrate that mouse TIPseq coupled with short-read sequencing can detect up to 80% of pre-existing L1 elements using as few as 2 million reads (1 Gb).

Our goal in using TIPseq is to be able to detect new, unannotated mouse L1 insertions. In our original TuJH922 dataset, TIPseqHunter predicted a total of 869 candidate L1 insertions absent from the reference genome. Many of these candidate insertions were predicted to be within an existing repetitive sequence (e.g. Figure 2F). We filtered out these candidates within repetitive sequence because we had low confidence in the accuracy and uniqueness of the alignments. This left 113 candidates for further validation (Supplementary Table S4). We visually inspected alignments of these remaining candidates in IGV and checked for partial primer sequence that may have caused mispriming incidents. After examining all candidates, we identified only 17 high-confidence potential insertions (e.g. Figure 2G) where no existing TE is near the prediction site. In contrast, we found 37 mispriming cases where only the 3′ region of TuJH922 matched the mouse reference genome (e.g. Figure 2H), 26 candidates located near an annotated L1 with very low read coverage (e.g. Figure 2I) and 36 candidates with low mapping quality (Supplementary Table S4 and Supplementary Figure S3). Thus, although TIPseqHunter had the sensitivity to identify most existing, annotated young mouse L1s, the high false-positive rate for unannotated insertions led us to search for improvements to identify unannotated mouse L1s.

Mouse TIPseq with long-read sequencing (nanoTIPseq)

Short-read sequencing data have several limitations to identify L1 insertions: (i) low mappability to highly repetitive regions; (ii) sophisticated algorithms needed to reconstruct and identify potential nonreference (polymorphic or *de novo*) insertions; and (iii) greater chance of artifacts and need to infer insertion site structure due to short reads requiring additional validation. To overcome these limitations, we tested TIPseq with long-read sequencing. ONT sequencing is a third-generation sequencing method that produces long continuous readout of DNA strands ranging from a couple hundred to millions of bases (48,49). The advantages of ONT include portability, affordability and instantaneity, which are ideal for real-time applications (49–51). Although ONT is prone to high error rates as compared to short-read sequencing methods, the long reads still usually allow unambiguous alignment, even within repetitive regions. High-coverage ONT WGS has been used to detect L1 insertions (34,37,38); however, WGS may not be cost-effective when examining a large number of samples, such as studies involving single-cell analysis. Sequencing an L1-enriched TIPseq reaction is expected to decrease costs considerably. We used the same TIPseq amplicon mixtures previously sequenced with short reads (Figure 2D) to prepare a library for ONT. We needed only 200 000 reads (129 Mb) of the long-read data to identify, with at least 5 support reads, 89.3% of our annotated L1MdTf_I and L1MdTf_II L1s in the C57BL/6 genome (Supplementary Table S5). Thus, a similar number of elements can be detected with 81× less data using long-read sequencing, when comparing short-read TIPseq. In most cases, individual reads encompass an entire amplicon, starting from the TuJH922 primer followed by the L1 3′ UTR, poly(A) tail, flanking genomic sequence and the vectorette linker (Figure 3A). The original TIPseq protocol entails digestion of gDNA with various restriction enzymes before vectorette adapter ligation, and long reads can be grouped into multiple populations with discrete endpoints based on restriction site positions (Figure 3B). To directly compare TIPseq short-read and long-read data, we normalized the number of reads mapping to annotated L1 3′ UTR (starting from TuJH922 sequence to end of the element) by counts per million mapped reads (CPM) (Figure 3C). Long reads cover annotated L1 3′ UTR 2.44× better than short reads, despite having less sequencing depth compared to short reads. In addition, long-read data allowed us to detect 59% (192/326) of those annotated L1s that were not found by TIPseqHunter. To distinguish TIPseq with short reads and long reads, we call TIPseq with long reads 'nanoTIPseq'. NanoTIPseq allowed us to detect L1s in highly repetitive regions, a major shortcoming of short-read sequencing. We found 3.26× better coverage in the 300-bp downstream region of the annotated L1s using long-read sequencing, demonstrating the capability of nanoTIPseq to detect L1s in a complex region (Figure 3D and Supplementary Figure S4).

To discover L1 insertions not present in the reference genome, we developed a customized bioinformatic pipeline for nanoTIPseq (Figure 3E). Briefly, at a new L1 insertion site we anticipate that the inserted L1 sequence will be missing from the reference genome, thus resulting in L1 sequence being clipped from the reads during alignment. The reads will still map to the insertion site owing to the genomic sequence in the same read. We extracted the clipped sequences and aligned them to the L1MdTf_I consensus using Bowtie 2.

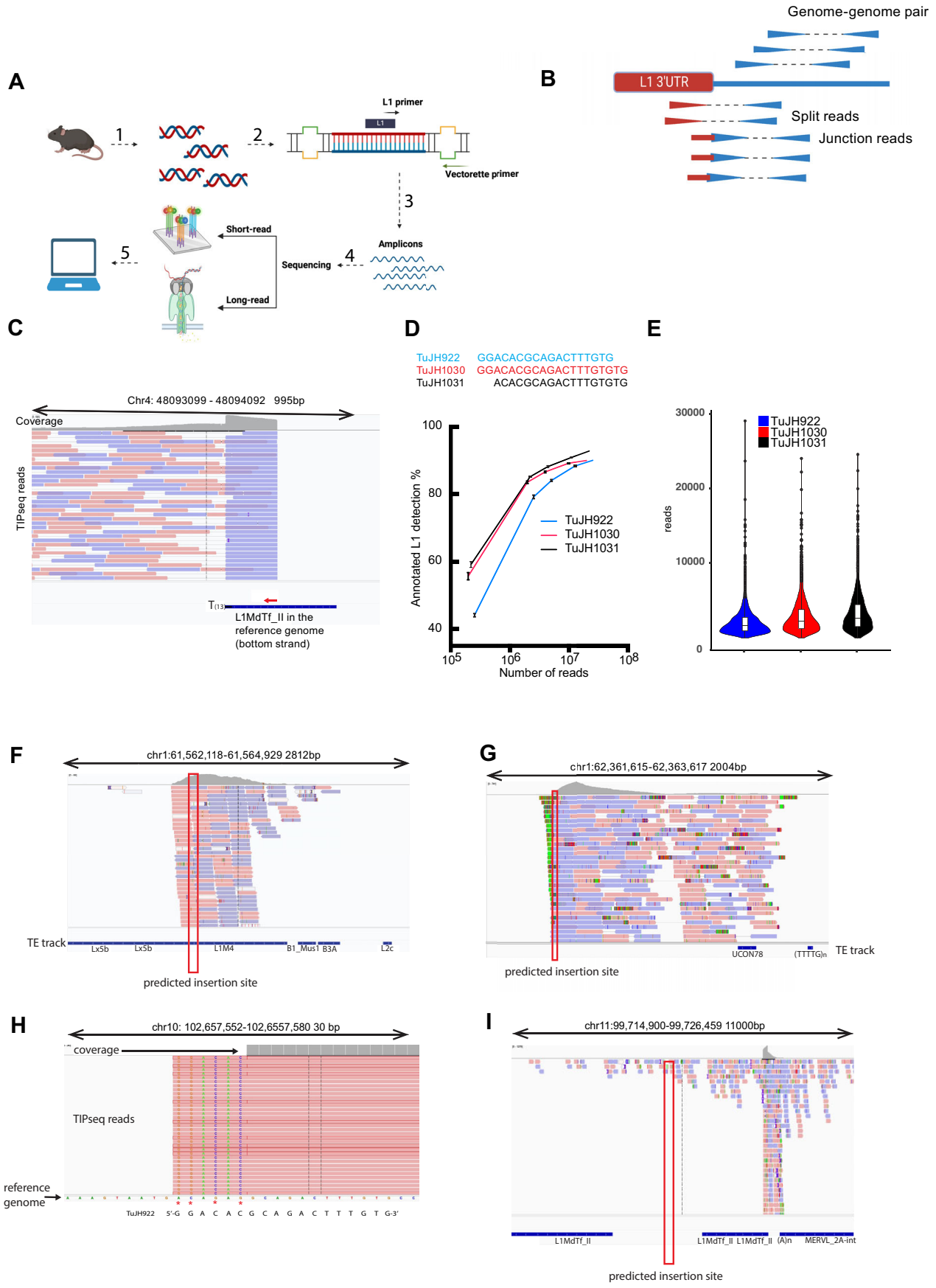


Figure 2. Short-read TIPseq identifies annotated L1s in the mouse genome. (**A**) TIPseq workflow consists of five main steps: (1) gDNA extraction and fragmentation; (2) vectorette adapter ligation; (3) TIPseq PCR using mouse L1-specific primers; (4) sequencing of the amplicons; and (5) customized

Subsequently, we selected the clipped reads that mapped to the L1 3′ UTR and possessed a poly(A) tail. By utilizing alignment of flanking genomic sequence from the corresponding original reads, we identified putative new L1 insertion sites (Figure 3E and F). Employing a minimum of five support reads as a cutoff for potential insertions, our pipeline predicted a total of 54 insertions that were present in our C57BL/6 mouse but absent from the C57BL/6 reference genome (Supplementary Table S6). We manually curated the potential insertion sites in the IGV and called 43 out of 54 insertions as likely true insertions based on the poly(A) tail, number and quality of supporting reads, alignments and putative endonuclease cleavage site. We observed a bias toward the endonuclease cleavage preference (5′-TTTT/AA-3′ on the bottom strand) at the putative insertion sites (Figure 3G) (52). The majority of the false-positive cases were caused by poor sequencing quality. Of the 43 nonreference L1s identified by nanoTIPseq, only 6 were found when TIPseq was coupled with short reads (Figure 4).

In our initial experiments, we digested gDNA with restriction enzymes to adhere closely to the original TIPseq protocols published for human L1 (35,39). To eliminate the need for restriction enzymes, we replaced enzymatic digestion with shearing, A-tailing and ligation to a universal vectorette with a 3′ T-overhang. These steps drastically reduced sample processing time, and with 200 000 reads (115 Mb), the sheared DNA method identified over 98.2% annotated L1 ('Sheared_B6_230217' from Supplementary Table S5) and had 4× more coverage than enzyme-digested TIPseq (Figure 3C) and identified 88% (288/326) of annotated L1s not found by short reads. The advantage observed from shearing is presumably due to less bias due to the variable distance of L1 loci to the closest relevant restriction enzyme recognition site. Our pipeline also predicted 93 nonreference insertions (Supplementary Table S7). We used manual examination of alignments at the locations of these predicted insertions on IGV to exclude 11 as false positives, leaving a total of 82 nonreference insertions identified. We picked 18 of these nonreference insertions and all were confirmed by genomic PCR or low-coverage WGS with ONT (Supplementary Table S8). Notably, shearing also resulted in a continuous distribution of read lengths, as opposed to the discrete read lengths obtained when digesting the gDNA with restriction enzymes (Supplementary Figure S5). This lifted the concern that many of the reads were PCR duplicates. These results suggest that shearing has similar, if not higher, effectiveness compared to enzymatic digestion for our purpose. The large number of nonreference insertions present in the analyzed C57BL/6 genome is likely due to multiple reasons. Some nonreference insertions could be present in the original DNA used for the reference genome, but absent in the reference genome sequence due to errors in genome assembly, perhaps due to repeated regions. Other nonreference insertions may be polymorphic L1s that have appeared over time in mouse colonies at vendors, other labs or in our lab. We would not necessarily expect to find complete overlap of these nonreference insertions in a C57BL/6 mouse from another lab, or even another C57BL/6 mouse from our lab. Although it is possible that some of the nonreference insertions are *de novo* events unique to the specific mouse analyzed, the relatively short length (8–36 bases) of the poly(A) tails (Supplementary Table S7) suggests that most of the insertions have been circulating in the population for some time.

NanoTIPseq is capable of detecting insertion events at less than one copy per cell

Somatic mutations can lead to catastrophic consequences such as cancer and neurodevelopmental diseases (53,54). While L1 insertions have traditionally been considered to occur primarily in the germline, it has been increasingly recognized that somatic L1 insertions can be detected in human and mouse genomes (6,24–26,55). To test the performance of nanoTIPseq when insertions are present at less than one copy per cell, we mixed FVB mouse gDNA with C57BL/6 gDNA at a 1:4 ratio. We used this mixture to prepare a library for nanoTIPseq. Additionally, we prepared a separate nanoTIPseq library using 100% FVB gDNA to serve as a reference for existing FVB L1 insertions. In the 100% FVB nanoTIPseq library, we discovered 1937 possible FVB strain-related insertions that are absent from the C57BL/6 genome. In the 1:4 FVB:C57BL/6 nanoTIPseq library, we identified 1465 of the FVB-specific L1s (76%) (Supplementary Figure S6A). Upon manual examination of our sequencing data for the 472 nondetected FVB-specific L1s, we found that 410/472 had supporting reads in the 1:4 mixture data (e.g. Supplementary Figure S6B). Thus, in total, 97% (1875/1937) of the FVB-specific L1s were present in the 1:4 library, but only 76% were 'called' by the nanoTIPseq pipeline due to lower number of support reads. When we downsampled the pure FVB sample to 20% reads, the pipeline called 1395 of the FVB-specific insertions, a comparable result to the 1:4 mixture. This suggests that diluting the FVB DNA did not introduce biases against FVB insertions at the amplification/library preparation steps, and

bioinformatic pipeline analysis. (**B**) Alignment evidence supporting the detection of annotated L1 elements, including genome–genome pairs, junction reads and split reads. (**C**) Example of typical mapping evidence supporting the identification of an existing L1MdTf element. Integrative Genomics Viewer (IGV) is used to visualize alignments. Shown is an IGV screenshot. At the top, read coverage is shown in gray. TIPseq reads are colored by read strand with respect to the reference genome. The position of an existing L1 in the reference sequence is shown at the bottom of the panel. The poly(A) tract is shown as poly(T) because the L1 is on the bottom strand. The position of the TuJH922 primer in the L1MdTf_II is indicated by the arrow. (**D**) Scaling-down experiments. Original sequencing files were scaled down to 50%, 20%, 10% and 1% in triplicates and analyzed using TIPseqHunter. Percentage of annotated L1s identified is shown on the *Y*-axis and number of reads scaled is shown on the *X*-axis. (**E**) Coverage of L1 3′ 300-bp flank region. The box plot displays 25% to 75% percentiles with median shown as the middle line in the box. (**F**) Example of a TIPseqHunter predicted nonreference insertion that falls inside another TE. Because we have low confidence in short reads mapped to a repetitive region, we discarded this prediction. (**G**) Potential nonreference L1 insertion site predicted by TIPseqHunter is highlighted with the box. To the left of the box, sequence that does not align to the reference genome is present and corresponds to L1 3′ UTR and poly(A) sequence. The lack of L1 in the reference sequence, along with L1–genome junction reads, poly(A) sequence and pile-up of adjacent genomic reads suggest the presence of a nonreference L1. (**H**) False-positive predictions resulting from mispriming of the TuJH922 primer. Shown are alignments of TIPseq reads to the mouse reference genome. The TuJH922 primer amplified this region even though there were four mismatches at the 5′ end of the primer (shown by the asterisks). Where sequence of the TIPseq reads is not explicitly shown, the sequence is identical to the mouse reference. (**I**) False-positive prediction with low coverage and prediction between two annotated L1s, possibly caused by mismapping or nonspecific mapping.

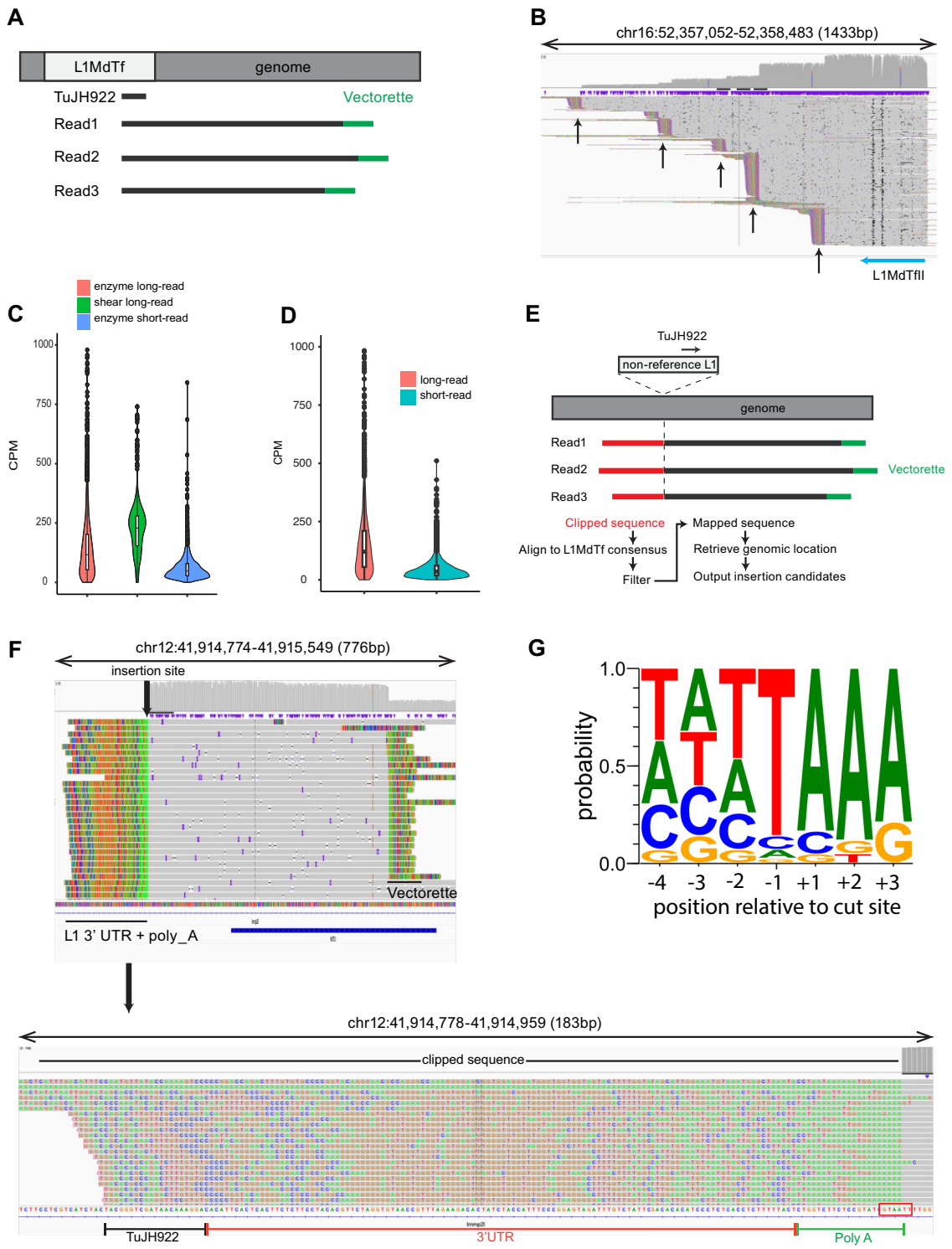


Figure 3. NanoTIPseq detection of nonannotated L1s in the mouse genome. (**A**) Typical alignment evidence for annotated L1 elements. One single alignment should contain L1 primer (TuJH922), L1 3′ UTR sequence, adjacent genomic sequence and vectorette adapter sequence. (**B**) Representative image showing an L1MdTf_II element on chr16. Vertical arrows indicate the location of restriction enzyme sites used to digest gDNA. (**C**) Annotated L1 3′ UTR coverage among C57BL/6 samples processed with enzyme-digested long reads, enzyme-digested short reads and sonicated long reads. CPM, counts per million mapped reads. (**D**) Annotated L1 3′ adjacent genomic region coverage among C57BL/6 sample processed with enzymatic digestion and sequenced with either short reads or long reads. The long-read method demonstrated 3.26× better coverage than the short-read method. (**E**) Scheme showing the basic idea behind nanoTIPseq. The nonreference L1 that is absent from the reference is shown in an overhang box and TuJH922 mapping location is shown. The clipped sequence should contain TuJH922, L1 3′ UTR and a poly(A) tract. (**F**) A typical nonreference insertion site should consist of a clipped sequence region (nonaligned bases shown in color) at the 3′ integration site (zoomed-in image is shown below) that contains TuJH922 sequence, L1 3′ UTR sequence and a poly(A) tract. Presumptive integration site is highlighted by the box (5′-AATT/AC-3′ on the bottom strand; sequence shown in the figure is the top strand). There is no annotated L1 at or nearby this insertion site. This alignment evidence suggests that this is a nonreference L1 in the C57BL/6 mouse we tested. (**G**) Sequence logo of nonreference L1 insertion site recapitulates L1 ORF2p preference site. Shown is the sequence of the top strand, although cleavage occurs on the bottom strand.

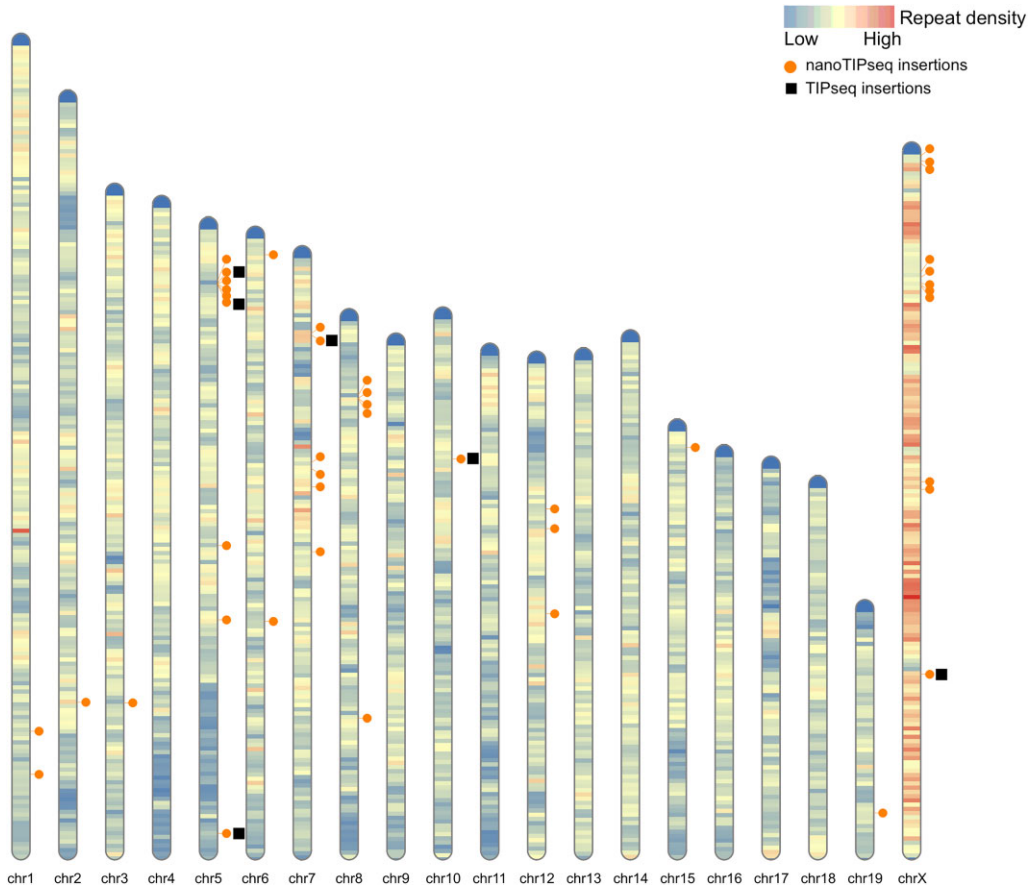


Figure 4. Genomic distribution of nonreference L1 insertions identified by long and short reads. Chromosome idiogram showing the identified nonreference L1s in our C57BL/6 mouse (circles). Squares indicate insertions also found by short-read TIPseqHunter. Each chromosome is color coded to show the density of repetitive elements.

reliably detecting low frequency event insertions with nanoTIPseq should be a matter of increasing sequencing depth.

Single-cell nanoTIPseq in mouse

Next, we asked how well mouse TIPseq performs with single cells. Short-read TIPseq has been reported to work with human single-cell samples, with limited detection sensitivity and a high false-positive rate (56). In this case, MDA was used to amplify single-cell genomes. Although MDA has advantages such as having high yield and low error rate, MDA also has disadvantages such as large dropout rate, chimeric reads and large variation between read lengths (57,58). PTA is a modified version of MDA, with incorporation of a reaction terminator (57). PTA decreases the formation of the 'branches' seen in MDA, and ultimately leads to more uniform amplification of the target genome (57). We have tried both MDA and PTA whole-genome amplification on single cells from the 4226 cell line, a breast cancer line derived from the MMTV-Wnt mouse model (46). As expected, MDA samples showed larger molecular weight amplification products, while PTA samples produced precise products around 1.5 kb (Supplementary Figure S7). As a rough quality control measure, we designed primers to amplify L1 adjacent regions on each of mouse chromosomes 1–19 and X (primers are listed in Supplementary Table S1; example of an amplified region is shown in Supplementary Figure S8). One hundred percent of the PTA and MDA samples were positive for each chromosome (Supplementary Table S9). We performed nanoTIPseq on five PTA-amplified 4226 cells and one MDA-amplified 4226 cell. From PTA-amplified DNA, we identified on average 95.4% (~3115/3266) of annotated C57BL/6 L1 insertions (Supplementary Table S5) and 64/1937 (~3%) L1 insertions from our reference of FVB L1s. Because the MMTV-Wnt mice were originally made in the FVB background, but subsequently backcrossed and maintained in the C57BL/6 background, these results are consistent with the backgrounds of the mice from which the 4226 cell line was derived. It is worth noting that the MDA-amplified samples have a slightly higher mean CPM (215) compared to PTA-amplified samples (192, 196, 199, 187 and 220 for samples 1–5, respectively); however, the overall number of annotated L1s recovered from MDA-amplified genomes is less than the PTA group (88% versus 95.4%) at 200 000 reads. Furthermore, MDA-amplified samples have higher variance than PTA samples, where the L1 3′ UTR coverage is consistent across PTA-amplified single-cell samples and resembles nanoTIPseq libraries made from unamplified bulk gDNA (Figure 5A), with an average coverage of ~192 CPM. In addition, in all five single-cell nanoTIPseq reactions, we found a total of 130 nonreference L1s and 58 of them were not previously observed in either FVB or C57BL/6 background, indicating those insertion events occurred in a common progenitor cell or were polymorphic insertions in the original mice where the cancer originated (Figure 5B–D and Supplementary Table S10) The consistency among the data from five individual single cells suggests that PTA-coupled

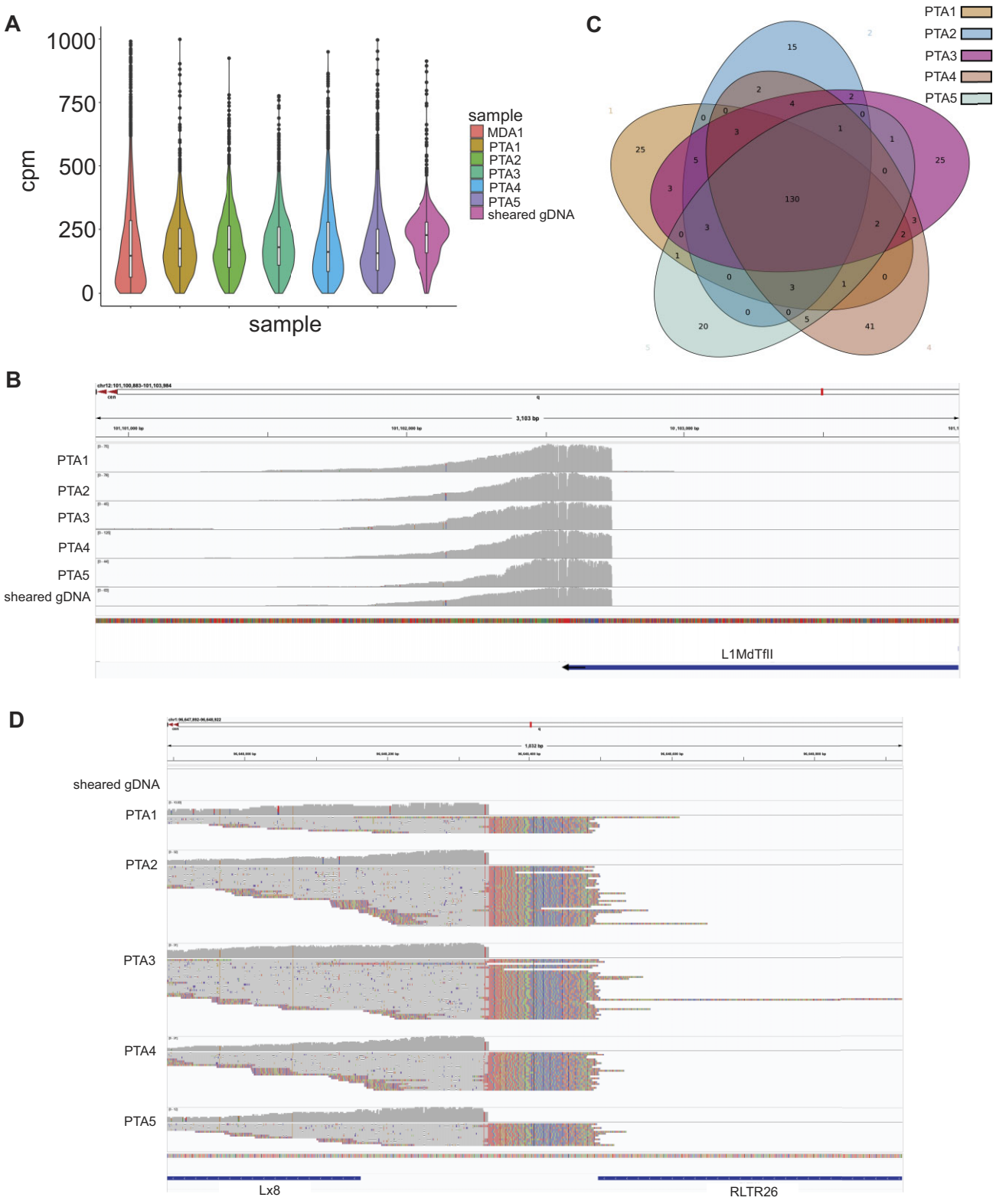


Figure 5. NanoTIPseq in 4226 single-cell L1 detection. (**A**) PTA single-cell amplification showed consistent annotated L1 coverage among five different single cells. (**B**) Annotated L1s were readily detected in all five tested single-cell nanoTIPseq samples. (**C**) Nonreference insertions from each single cell were intersected, and 130 common nonreference L1 insertions were identified in all five single-cell samples, demonstrating good consistency and specificity in nanoTIPseq. (**D**) A representative example of a common nonreference L1 insertion found only in 4226 cells, but not C57BL/6 gDNA ('sheared gDNA').

nanoTIPseq can be used for reliable identification of new L1 insertions from single cells.

Because nanoTIPseq only captures the 3′ junction of L1 elements, if desired other methods would be required to obtain the rest of the insertion, including the 5′ flank. Thus, nanoTIPseq is most useful for instances where we want to quickly and cheaply identify the location of new and existing L1 insertions, but do not necessarily need the complete insertion structure. Aberrant expression of L1 has been correlated with many abnormal states in mammals (6,7,22–28). However, in most of these cases it is unclear whether L1 expression contributes to the underlying pathophysiology of disease. Mouse models are critical tools for studying human health and disease and will be valuable for evaluating the role of L1 in disease. One key aspect to characterizing the role of L1 is monitoring new L1 insertions. WGS and targeted L1 enrichment followed by sequencing have previously been used to identify human L1 insertions (33–36), but the sequencing depth required (ranging from 5 to 120 GB) can create a cost limitation for smaller labs when a large number of samples or single cells need to be processed. WGS has also been used to identify mouse L1 insertions, but again the sequencing depth required poses an obstacle for many labs. NanoTIPseq allows us to identify >95% of mouse L1s from single cells or bulk tissue while requiring <150 MB of sequencing data per sample, a 30-fold reduction in sequencing data required per sample as compared to other methods. This opens the possibility of designing large studies with mouse disease models and then isolating tissues and/or single cells at various times during disease progression to pinpoint the cell type, timing and frequency of L1 integration. For example, the disruption of piRNA biogenesis in the mouse germline leads to massive L1 upregulation, with arrest during meiosis and eventual germline failure and infertility (29–31). NanoTIPseq can allow precise identification of the cell types where retrotransposition is occurring in piRNA-deficient mice and the locations of such insertions. Mouse models of aging (aged wild-type or SIRT6-deficient) suggest that L1 may contribute to the sterile inflammation associated with aging (22,59). Although L1 expression is elevated in certain tissues in these aged mice, and current models suggest that L1 intermediates contribute to inflammation, no studies have convincingly characterized retrotransposition activity in these mice. L1 expression is also elevated in many human cancers, and although mouse cancer models have not been extensively characterized for L1 activation, MMTV-PyVT (mouse mammary tumor virus-polyomavirus middle T antigen) mice, which express PyVT in the breast, are reported to upregulate L1 early in breast cancer progression (60). L1 upregulation has also been reported in many neurodegenerative diseases, including mouse models of amyotrophic lateral sclerosis/frontotemporal dementia (61), Alzheimer's disease (62), Parkinson's disease (63), Huntington's disease (64) and ataxia telangiectasia (65). In some of these mouse models, an increase of 'L1 DNA content' is measured by quantitative PCR. Quantitative PCR to measure L1 DNA content is difficult to independently reproduce and often gives results inconsistent with more conclusive methods such as sequencing. Thus, nanoTIPseq would be useful to monitor the landscape of L1 retrotransposition events as disease progresses in these various mouse models. Single-cell nanoTIPseq should also enable the unambiguous determination of the timing of L1 retrotransposition during normal development. Finally, the human genome has far less potentially active L1s than the mouse genome (~100 human potentially active L1s versus ~3000 potentially active mouse L1s) (3,47), suggesting that the population of young L1s in mice is much larger than that in humans. We expect that performing nanoTIPseq on the human genome would require amplifying a less complex population of elements and would require less sequencing data. For example, a Bowtie search for perfect matches to an L1H-specific primer [the L1 primer used in (39)] found 475 matches in the reference human genome, while a Bowtie search for perfect matches to TuJH922 in the reference mouse genome found 3450 matches. Thus, we would expect that using nanoTIPseq to profile young human L1s could require as little as 20 MB of sequencing data per sample, allowing low-cost screening for retrotransposition events in the clinic (e.g. during prenatal testing).

Data availability

The data underlying this article are available in BioProject, at https://www.ncbi.nlm.nih.gov/bioproject, and can be accessed with BioProject ID PRJNA1038748.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We thank Kathy Burns, Wilson McKerrow and David Fenyo for initial discussions on TIPseq and the use of TIPseqHunter.

Funding

National Institutes of Health [GM141381 to J.S.H.]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement

None declared.

References

1. International Human Genome Sequencing Consortium, Whitehead Institute for Biomedical Research, Center for Genome Research, Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
2. Sookdeo,A., Hepp,C.M., McClure,M.A. and Boissinot,S. (2013) Revisiting the evolution of mouse LINE-1 in the genomic era. *Mobile DNA*, **4**, 3.
3. Brouha,B., Schustak,J., Badge,R.M., Lutz-Prigge,S., Farley,A.H., Moran,J.V. and Kazazian,H.H. (2003) Hot L1s account for the bulk of retrotransposition in the human population. *Proc. Natl Acad. Sci. U.S.A.*, **100**, 5280–5285.
4. Huang,C.R.L., Schneider,A.M., Lu,Y., Niranjan,T., Shen,P., Robinson,M.A., Steranka,J.P., Valle,D., Civin,C.I., Wang,T., *et al.* (2010) Mobile interspersed repeats are major structural variants in the human genome. *Cell*, **141**, 1171–1182.
5. Sassaman,D.M., Dombroski,B.A., Moran,J.V., Kimberland,M.L., Naas,T.P., DeBerardinis,R.J., Gabriel,A., Swergold,G.D. and Kazazian,H.H. (1997) Many human L1 elements are capable of retrotransposition. *Nat. Genet.*, **16**, 37–43.
6. Lee,E., Iskow,R., Yang,L., Gokcumen,O., Haseley,P., Luquette,L.J., Lohr,J.G., Harris,C.C., Ding,L., Wilson,R.K., *et al.* (2012)

Landscape of somatic retrotransposition in human cancers. *Science*, **337**, 967–971.
7. Iskow,R.C., McCabe,M.T., Mills,R.E., Torene,S., Pittard,W.S., Neuwald,A.F., Van Meir,E.G., Vertino,P.M. and Devine,S.E. (2010) Natural mutagenesis of human genomes by endogenous retrotransposons. *Cell*, **141**, 1253–1261.
8. Feng,Q., Moran,J.V., Kazazian,H.H. and Boeke,J.D. (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell*, **87**, 905–916.
9. Moran,J.V., Holmes,S.E., Naas,T.P., DeBerardinis,R.J., Boeke,J.D. and Kazazian,H.H. (1996) High frequency retrotransposition in cultured mammalian cells. *Cell*, **87**, 917–927.
10. Kolosha,V.O. and Martin,S.L. (1997) *In vitro* properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc. Natl Acad. Sci. U.S.A.*, **94**, 10155–10160.
11. Wei,W., Gilbert,N., Ooi,S.L., Lawler,J.F., Ostertag,E.M., Kazazian,H.H., Boeke,J.D. and Moran,J.V. (2001) Human L1 retrotransposition: *cis* preference versus *trans* complementation. *Mol. Cell. Biol.*, **21**, 1429–1439.
12. Hohjoh,H. and Singer,M.F. (1996) Cytoplasmic ribonucleoprotein complexes containing human LINE-1 protein and RNA. *EMBO J.*, **15**, 630–639.
13. Kolosha,V.O. and Martin,S.L. (2003) High-affinity, non-sequence-specific RNA binding by the open reading frame 1 (ORF1) protein from long interspersed nuclear element 1 (LINE-1). *J. Biol. Chem.*, **278**, 8112–8117.
14. Kulpa,D.A. and Moran,J.V. (2005) Ribonucleoprotein particle formation is necessary but not sufficient for LINE-1 retrotransposition. *Hum. Mol. Genet.*, **14**, 3237–3248.
15. Martin,S.L. and Bushman,F.D. (2001) Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol. Cell. Biol.*, **21**, 467–475.
16. Horn,A.V., Celic,I., Dong,C., Martirosyan,I. and Han,J.S. (2017) A conserved role for the ESCRT membrane budding complex in LINE retrotransposition. *PLoS Genet.*, **13**, e1006837.
17. Mita,P., Wudzinska,A., Sun,X., Andrade,J., Nayak,S., Kahler,D.J., Badri,S., LaCava,J., Ueberheide,B., Yun,C.Y., *et al.* (2018) LINE-1 protein localization and functional dynamics during the cell cycle. *eLife*, **7**, e30058.
18. Cost,G.J. and Boeke,J.D. (1998) Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*, **37**, 18081–18093.
19. Mathias,S.L., Scott,A.F., Kazazian,H.H., Boeke,J.D. and Gabriel,A. (1991) Reverse transcriptase encoded by a human transposable element. *Science*, **254**, 1808–1810.
20. Cost,G.J., Feng,Q., Jacquier,A. and Boeke,J.D. (2002) Human L1 element target-primed reverse transcription *in vitro*. *EMBO J.*, **21**, 5899–5910.
21. Luan,D.D., Korman,M.H., Jakubczak,J.L. and Eickbush,T.H. (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*, **72**, 595–605.
22. De Cecco,M., Ito,T., Petrashen,A.P., Elias,A.E., Skvir,N.J., Criscione,S.W., Caligiana,A., Brocculi,G., Adney,E.M., Boeke,J.D., *et al.* (2019) L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, **566**, 73–78.
23. Gorbunova,V., Seluanov,A., Mita,P., McKerrow,W., Fenyö,D., Boeke,J.D., Linker,S.B., Gage,F.H., Kreiling,J.A., Petrashen,A.P., *et al.* (2021) The role of retrotransposable elements in aging and age-associated diseases. *Nature*, **596**, 43–53.
24. Hancks,D.C. and Kazazian,H.H. (2016) Roles for retrotransposon insertions in human disease. *Mobile DNA*, **7**, 9.
25. Kazazian,H.H., Wong,C., Youssoufian,H., Scott,A.F., Phillips,D.G. and Antonarakis,S.E. (1988) Haemophilia A resulting from *de novo* insertion of L1 sequences represents a novel mechanism for mutation in man. *Nature*, **332**, 164–166.
26. Morse,B., Rotherg,P.G., South,V.J., Spandorfer,J.M. and Astrin,S.M. (1988) Insertional mutagenesis of the myc locus by a LINE-1 sequence in a human breast carcinoma. *Nature*, **333**, 87–90.
27. Rodić,N., Sharma,R., Sharma,R., Zampella,J., Dai,L., Taylor,M.S., Hruban,R.H., Iacobuzio-Donahue,C.A., Maitra,A., Torbenson,M.S., *et al.* (2014) Long interspersed element-1 protein expression is a hallmark of many human cancers. *Am. J. Pathol.*, **184**, 1280–1286.
28. Suarez,N.A., Macia,A. and Muotri,A.R. (2018) LINE-1 retrotransposons in healthy and diseased human brain. *Dev. Neurobiol.*, **78**, 434–455.
29. Carmell,M.A., Girard,A., van de Kant,H.J.G., Bourc'his,D., Bestor,T.H., de Rooij,D.G. and Hannon,G.J. (2007) MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell*, **12**, 503–514.
30. Newkirk,S.J., Lee,S., Grandi,F.C., Gaysinskaya,V., Rosser,J.M., Vanden Berg,N., Hogarth,C.A., Marchetto,M.C.N., Muotri,A.R., Griswold,M.D., *et al.* (2017) Intact piRNA pathway prevents L1 mobilization in male meiosis. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E5635–E5644.
31. Soper,S.F.C., van der Heijden,G.W., Hardiman,T.C., Goodheart,M., Martin,S.L., de Boer,P. and Bortvin,A. (2008) Mouse maelstrom, a component of nuage, is essential for spermatogenesis and transposon repression in meiosis. *Dev. Cell*, **15**, 285–297.
32. Tao,J., Wang,Q., Mendez-Dorantes,C., Burns,K.H. and Chiarle,R. (2022) Frequency and mechanisms of LINE-1 retrotransposon insertions at CRISPR/Cas9 sites. *Nat. Commun.*, **13**, 3685.
33. Baillie,J.K., Barnett,M.W., Upton,K.R., Gerhardt,D.J., Richmond,T.A., De Sapio,F., Brennan,P.M., Rizzu,P., Smith,S., Fell,M., *et al.* (2011) Somatic retrotransposition alters the genetic landscape of the human brain. *Nature*, **479**, 534–537.
34. Ivančić,D., Mir-Pedrol,J., Jaraba-Wallace,J., Rafel,N., Sanchez-Mejias,A. and Güell,M. (2022) INSERT-seq enables high-resolution mapping of genomically integrated DNA using Nanopore sequencing. *Genome Biol.*, **23**, 227.
35. Tang,Z., Steranka,J.P., Ma,S., Grivainis,M., Rodić,N., Huang,C.R.L., Shih,I.-M., Wang,T.-L., Boeke,J.D., Fenyö,D., *et al.* (2017) Human transposon insertion profiling: analysis, visualization and identification of somatic LINE-1 insertions in ovarian cancer. *Proc. Natl Acad. Sci. U.S.A.*, **114**, E733–E740.
36. Zhou,W., Emery,S.B., Flasch,D.A., Wang,Y., Kwan,K.Y., Kidd,J.M., Moran,J.V. and Mills,R.E. (2020) Identification and characterization of occult human-specific LINE-1 insertions using long-read sequencing technology. *Nucleic Acids Res.*, **48**, 1146–1163.
37. Ewing,A.D., Smits,N., Sanchez-Luque,F.J., Faivre,J., Brennan,P.M., Richardson,S.R., Cheetham,S.W. and Faulkner,G.J. (2020) Nanopore sequencing enables comprehensive transposable element epigenomic profiling. *Mol. Cell*, **80**, 915–928.e5.
38. Gerdes,P., Lim,S.M., Ewing,A.D., Larcombe,M.R., Chan,D., Sanchez-Luque,F.J., Walker,L., Carleton,A.L., James,C., Knaupp,A.S., *et al.* (2022) Retrotransposon instability dominates the acquired mutation landscape of mouse induced pluripotent stem cells. *Nat. Commun.*, **13**, 7470.
39. Steranka,J.P., Tang,Z., Grivainis,M., Huang,C.R.L., Payer,L.M., Rego,F.O.R., Miller,T.L.A., Galante,P.A.F., Ramaswami,S., Heguy,A., *et al.* (2019) Transposon insertion profiling by sequencing (TIPseq) for mapping LINE-1 insertions in the human genome. *Mobile DNA*, **10**, 8.
40. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
41. Goodier,J.L., Ostertag,E.M., Du,K. and Kazazian,H.H. (2001) A novel active L1 retrotransposon subfamily in the mouse. *Genome Res.*, **11**, 1677–1685.
42. Gagnier,L., Belancio,V.P. and Mager,D.L. (2019) Mouse germ line mutations due to retrotransposon insertions. *Mobile DNA*, **10**, 15.

43. Arnold,C. and Hodgson,I.J. (1991) Vectorette PCR: a novel approach to genomic walking. *PCR Methods Appl.*, **1**, 39–42.
44. Storer,J., Hubley,R., Rosen,J., Wheeler,T.J. and Smit,A.F. (2021) The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA*, **12**, 2.
45. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
46. Tonnessen-Murray,C.A., Frey,W.D., Rao,S.G., Shahbandi,A., Ungerleider,N.A., Olayiwola,J.O., Murray,L.B., Vinson,B.T., Chrisey,D.B., Lord,C.J., *et al.* (2019) Chemotherapy-induced senescent cancer cells engulf other cells to enhance their survival. *J. Cell Biol.*, **218**, 3827–3844.
47. DeBerardinis,R.J., Goodier,J.L., Ostertag,E.M. and Kazazian,H.H. (1998) Rapid amplification of a retrotransposon subfamily is evolving the mouse genome. *Nat. Genet.*, **20**, 288–290.
48. Clarke,J., Wu,H.-C., Jayasinghe,L., Patel,A., Reid,S. and Bayley,H. (2009) Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, **4**, 265–270.
49. Wang,Y., Zhao,Y., Bollas,A., Wang,Y. and Au,K.F. (2021) Nanopore sequencing technology, bioinformatics and applications. *Nat. Biotechnol.*, **39**, 1348–1365.
50. Ashton,P.M., Nair,S., Dallman,T., Rubino,S., Rabsch,W., Mwaigwisya,S., Wain,J. and O'Grady,J. (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. *Nat. Biotechnol.*, **33**, 296–300.
51. Delahaye,C. and Nicolas,J. (2021) Sequencing DNA with nanopores: troubles and biases. *PLoS One*, **16**, e0257521.
52. Wheeler,T.J., Clements,J. and Finn,R.D. (2014) Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. *BMC Bioinformatics*, **15**, 7.
53. Greenman,C., Stephens,P., Smith,R., Dalgliesh,G.L., Hunter,C., Bignell,G., Davies,H., Teague,J., Butler,A., Stevens,C., *et al.* (2007) Patterns of somatic mutation in human cancer genomes. *Nature*, **446**, 153–158.
54. Poduri,A., Evrony,G.D., Cai,X. and Walsh,C.A. (2013) Somatic mutation, genomic variation, and neurological disease. *Science*, **341**, 1237758.
55. Scott,E.C. and Devine,S.E. (2017) The role of somatic L1 retrotransposition in human cancers. *Viruses*, **9**, 131.
56. McKerrow,W., Tang,Z., Steranka,J.P., Payer,L.M., Boeke,J.D., Keefe,D., Fenyö,D., Burns,K.H. and Liu,C. (2020) Human transposon insertion profiling by sequencing (TIPseq) to map LINE-1 insertions in single cells. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **375**, 20190335.
57. Gonzalez-Pena,V., Natarajan,S., Xia,Y., Klein,D., Carter,R., Pang,Y., Shaner,B., Annu,K., Putnam,D., Chen,W., *et al.* (2021) Accurate genomic variant detection in single cells with primary template-directed amplification. *Proc. Natl Acad. Sci. U.S.A.*, **118**, e2024176118.
58. Huang,L., Ma,F., Chapman,A., Lu,S. and Xie,X.S. (2015) Single-cell whole-genome amplification and sequencing: methodology and applications. *Annu. Rev. Genom. Hum. Genet.*, **16**, 79–102.
59. Simon,M., Van Meter,M., Ablaeva,J., Ke,Z., Gonzalez,R.S., Taguchi,T., De Cecco,M., Leonova,K.I., Kogan,V., Helfand,S.L., *et al.* (2019) LINE1 derepression in aged wild-type and SIRT6-deficient mice drives inflammation. *Cell Metab.*, **29**, 871–885.e5.
60. Gualtieri,A., Andreola,F., Sciamanna,I., Vallebona,P.S., Serafino,A. and Spadafora,C. (2013) Increased expression and copy number amplification of LINE-1 and SINE B1 retrotransposable elements in murine mammary carcinoma progression. *Oncotarget*, **4**, 1882–1893.
61. Zhang,Y.-J., Guo,L., Gonzales,P.K., Gendron,T.F., Wu,Y., Jansen-West,K., O'Raw,A.D., Pickles,S.R., Prudencio,M., Carlomagno,Y., *et al.* (2019) Heterochromatin anomalies and double-stranded RNA accumulation underlie C9orf72 poly(PR) toxicity. *Science*, **363**, eaav2606.
62. El Hajjar,J., Chatoo,W., Hanna,R., Nkanza,P., Tétreault,N., Tse,Y.C., Wong,T.P., Abdouh,M. and Bernier,G. (2019) Heterochromatic genome instability and neurodegeneration sharing similarities with Alzheimer's disease in old Bmi1+/− mice. *Sci. Rep.*, **9**, 594.
63. Blaudin de Thé,F.-X., Rekaik,H., Peze-Heidsieck,E., Massiani-Beaudoin,O., Joshi,R.L., Fuchs,J. and Prochiantz,A. (2018) Engrailed homeoprotein blocks degeneration in adult dopaminergic neurons through LINE-1 repression. *EMBO J.*, **37**, e97374.
64. Tan,H., Wu,C. and Jin,L. (2018) A possible role for long interspersed nuclear elements-1 (LINE-1) in Huntington's disease progression. *Med. Sci. Monit.*, **24**, 3644–3652.
65. Coufal,N.G., Garcia-Perez,J.L., Peng,G.E., Marchetto,M.C.N., Muotri,A.R., Mu,Y., Carson,C.T., Macia,A., Moran,J.V. and Gage,F.H. (2011) Ataxia telangiectasia mutated (ATM) modulates long interspersed element-1 (L1) retrotransposition in human neural stem cells. *Proc. Natl Acad. Sci. U.S.A.*, **108**, 20382–20387.

Received: November 13, 2023. **Revised:** March 25, 2024. **Editorial Decision:** April 1, 2024. **Accepted:** April 6, 2024