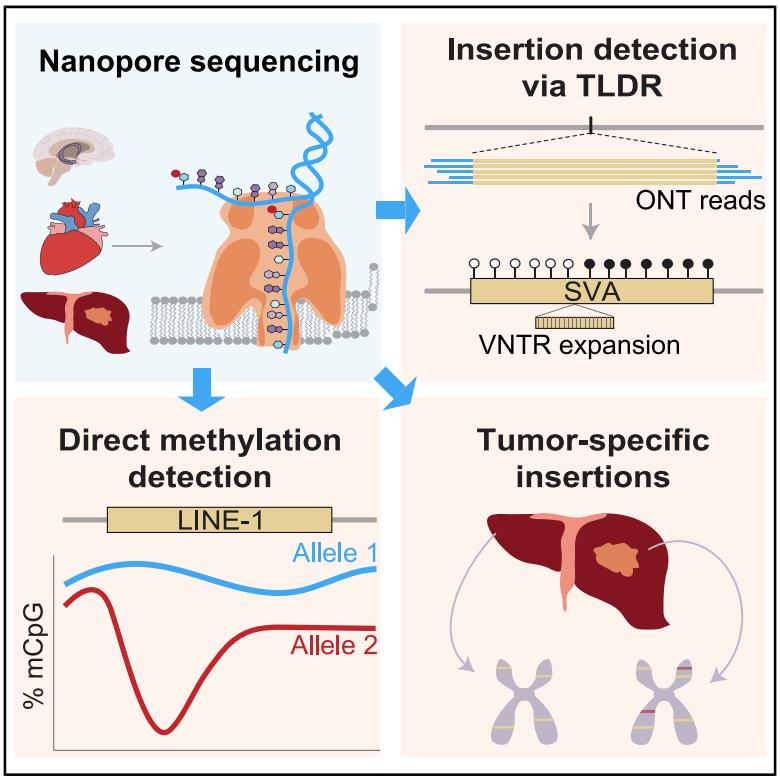
Molecular Cell

Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling

Graphical Abstract



Highlights

- Transposons from long DNA reads (TLDR) detects transposable element (TE) insertions
- TLDR resolves entire TE insertions, including SVA internal tandem repeat expansions
- Nanopore analysis finds aberrant and allele-specific TE methylation in normal tissues
- Young LINE-1s show highly dynamic locus- and element-specific methylation in cancer

Authors

Adam D. Ewing, Nathan Smits, Francisco J. Sanchez-Luque, ..., Sandra R. Richardson, Seth W. Cheetham, Geoffrey J. Faulkner

Correspondence

adam.ewing@mater.uq.edu.au (A.D.E.), seth.cheetham@mater.uq.edu.au (S.W.C.), faulknergj@gmail.com (G.J.F.)

In Brief

Ewing et al. report TLDR, a tool to fully resolve transposable element (TE) insertions with long-read sequencing. TLDR detects polymorphic and tumor-specific TE insertions in whole-genome nanopore sequencing data from normal and cancerous human tissues. Nanopore analysis reveals CpG methylation landscapes of young LINE-1, *Alu*, and SVA retrotransposon families.

Ewing et al., 2020, Molecular Cell *80*, 915–928
December 3, 2020
https://doi.org/10.1016/j.molcel.2020.10.024

CellPress

Molecular Cell

CellPress

Technology

Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling

Adam D. Ewing,[1,*] Nathan Smits,[1] Francisco J. Sanchez-Luque,[2,3] Jamila Faivre,[4] Paul M. Brennan,[5] Sandra R. Richardson,[1] Seth W. Cheetham,[1,*] and Geoffrey J. Faulkner[1,6,7,*]

[1]Mater Research Institute, University of Queensland, Woolloongabba, QLD 4102, Australia
[2]GENYO, Pfizer-University of Granada-Andalusian Government Centre for Genomics and Oncological Research, PTS Granada 18016, Spain
[3]MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine (IGMM), University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK
[4]INSERM, U1193, Paul-Brousse University Hospital, Hepatobiliary Centre, Villejuif 94800, France
[5]Translational Neurosurgery, Centre for Clinical Brain Sciences, Edinburgh EH16 4SB, UK
[6]Queensland Brain Institute, University of Queensland, St. Lucia, QLD 4067, Australia
[7]Lead Contact
*Correspondence: adam.ewing@mater.uq.edu.au (A.D.E.), seth.cheetham@mater.uq.edu.au (S.W.C.), faulknergj@gmail.com (G.J.F.)
https://doi.org/10.1016/j.molcel.2020.10.024

SUMMARY

Transposable elements (TEs) drive genome evolution and are a notable source of pathogenesis, including cancer. While CpG methylation regulates TE activity, the locus-specific methylation landscape of mobile human TEs has to date proven largely inaccessible. Here, we apply new computational tools and long-read nanopore sequencing to directly infer CpG methylation of novel and extant TE insertions in hippocampus, heart, and liver, as well as paired tumor and non-tumor liver. As opposed to an indiscriminate stochastic process, we find pronounced demethylation of young long interspersed element 1 (LINE-1) retrotransposons in cancer, often distinct to the adjacent genome and other TEs. SINE-VNTR-*Alu* (SVA) retrotransposons, including their internal tandem repeat-associated CpG island, are near-universally methylated. We encounter allele-specific TE methylation and demethylation of aberrantly expressed young LINE-1s in normal tissues. Finally, we recover the complete sequences of tumor-specific LINE-1 insertions and their retrotransposition hallmarks, demonstrating how long-read sequencing can simultaneously survey the epigenome and detect somatic TE mobilization.

INTRODUCTION

Transposable elements (TEs) pervade our genomic architecture and broadly influence human biology and disease (Chuong et al., 2017; Kazazian and Moran, 2017). Recently, Oxford Nanopore Technologies (ONT) long-read DNA sequencing has enabled telomere-to-telomere chromosome assembly at base-pair resolution, including high-copy-number TEs previously refractory to short-read mapping (Goerner-Potvin and Bourque, 2018; Jain et al., 2018; Miga et al., 2020). While most evolutionarily older TEs have accumulated sufficient nucleotide diversity to be uniquely identified, recent TE insertions are often indistinguishable from their source elements when assayed with short-read approaches (Lanciano and Cristofari, 2020; Philippe et al., 2016).

Each diploid human genome contains 80–100 potentially mobile long interspersed element 1 (LINE-1) copies, referred to here as L1Hs (L1 human specific) (Beck et al., 2010; Brouha et al., 2003). L1Hs elements are ~6 kbp long and encode proteins required to retrotranspose (Kazazian et al., 1988; Moran et al., 1996) *in cis* and *trans* mobilize shorter *Alu* (~300 bp) and composite SVA (SINE-VNTR-*Alu*, ~2 kbp) retrotransposons, as well as processed mRNAs (Dewannieux et al., 2003; Esnault et al., 2000; Hancks et al., 2011; Raiz et al., 2012; Wei et al., 2001). While the reference genome assembly contains thousands of human-specific TE copies, the vast majority of polymorphic TEs found in the global population are non-reference (Ewing and Kazazian, 2010; Sudmant et al., 2015). L1Hs-mediated germline insertional mutagenesis is a prominent source of disease, whereas somatic L1Hs retrotransposition can occur during early embryogenesis as well as in the committed neuronal lineage, and is a common feature of many epithelial cancers (Burns, 2017; Faulkner and Billon, 2018; Hancks and Kazazian, 2016).

A wide array of host factors have been implicated in mammalian TE regulation (Bruno et al., 2019; Goodier, 2016). Central among them is CpG methylation (Castro-Diaz et al., 2014; Greenberg and Bourc'his, 2019; Jönsson et al., 2019; Pehrsson et al., 2019; de la Rica et al., 2016; Walter et al., 2016). Most CpGs are located within TEs, and it has been posited that CpG methylation arose to limit the mobility of young TEs (Rollins et al., 2006; Yoder et al., 1997), whereas older TEs are controlled by repressive histone marks and other pathways (Imbeault et al., 2017; Jacobs et al., 2014; Rowe et al., 2010).

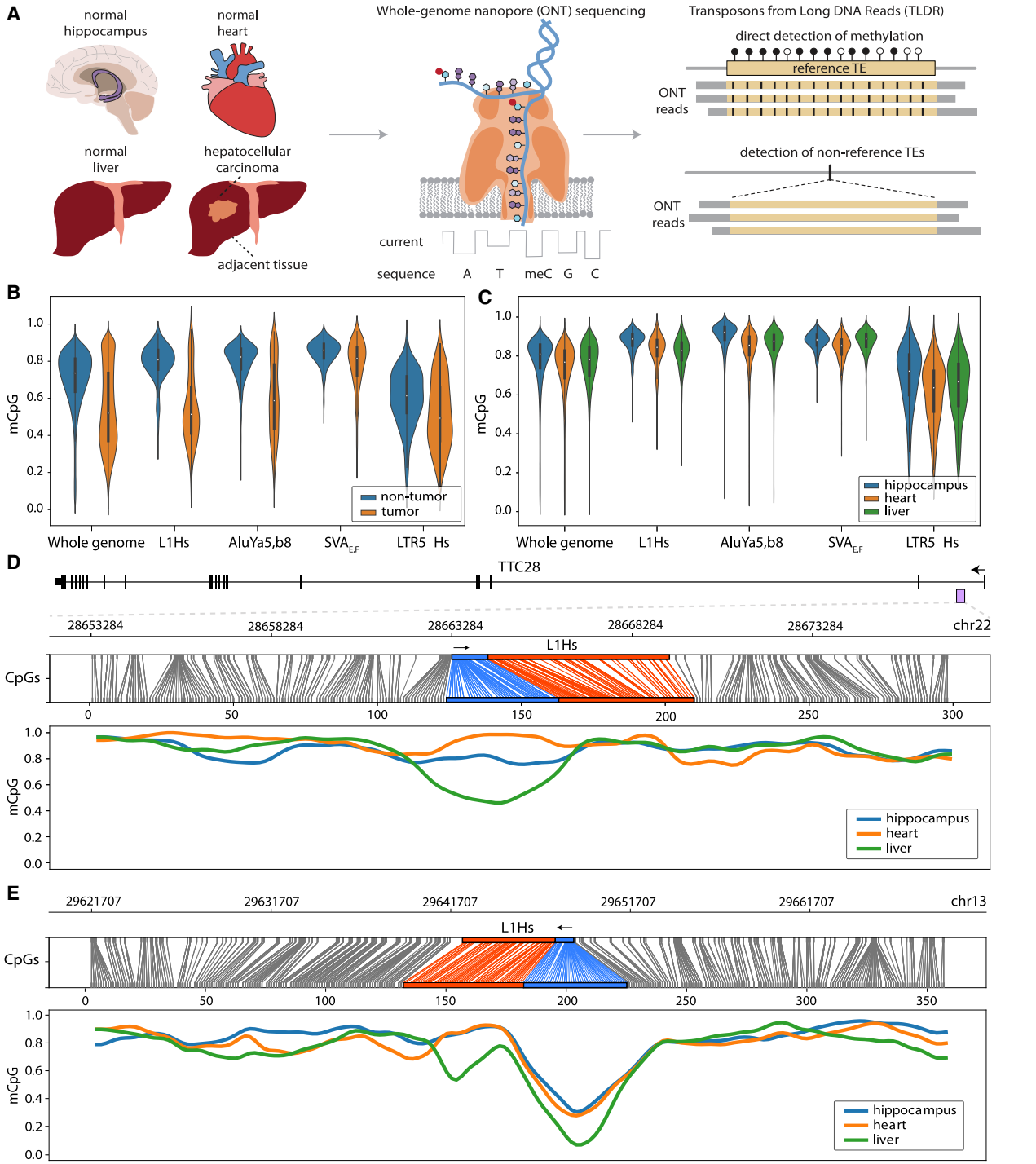


Figure 1. Measurement of CpG Methylation on TEs

(A) Hippocampus, heart, and liver tissue from a healthy individual (CTRL-5413), as well as tumor and adjacent liver tissue from a hepatocellular carcinoma patient (HCC33), was ONT sequenced. TLDR analysis identified TE insertions and quantified TE locus-specific CpG methylation.

(B) CpG methylation in HCC33 samples for the whole genome (6-kbp windows), L1Hs copies >5.9 kbp, young *Alu* copies >280 bp (AluYa5 and AluYb8), human-specific SVA copies >1 kbp (SVA_E, and SVA_F), and HERV-K flanking long terminal repeats >900 bp (LTR5_Hs).

(C) As for (B), except for CTRL-5413 normal tissues.

(legend continued on next page)

While the 5′ end of a CpG island present in the L1Hs 5′ untranslated region (UTR) is usually demethylated as a prerequisite for retrotransposition (Alves et al., 1996; Salvador-Palomeque et al., 2019; Sanchez-Luque et al., 2019; Scott et al., 2016; Thayer et al., 1993), the accompanying internal L1Hs methylation profile remains obscure. SVA methylation and transcriptional regulation are even less well understood, owing to the ambiguous SVA canonical promoter structure and a CpG island located in its internal variable number of tandem repeats (VNTRs) region. Young *Alu* subfamilies are, by contrast, CpG rich and more accessible to short-read approaches, yet they are difficult to analyze individually because of their high copy number (Lander et al., 2001). Although methylation of young TEs can be ascertained by locus-specific and genome-wide bisulfite sequencing assays, these approaches are currently limited in throughput and resolution, respectively. As a result, the methylation landscape of young TEs in human tissues, as well as how and why it changes in cancer, is unclear. Here, we demonstrate the capacity of ONT sequencing to concurrently assess TE methylation and resolve polymorphic and somatic TE insertions.

Design

Robust and well-established computational tools are available for the detection of TE insertions from whole-genome short-read sequencing data (Ewing, 2015; O'Neill et al., 2020). In most cases, these approaches report TE insertion genomic coordinates and adjacent sequence information relevant to retrotransposition, the molecular process by which new TE insertions are generated (Boeke et al., 1985; Jurka, 1997; Luan et al., 1993). To our knowledge, there are no general-purpose tools available for TE insertion detection that are suitable for ONT data, although software for detecting LINE-1 insertions from PacBio reads has been recently developed (Zhou et al., 2020). The ability to detect modified DNA bases with ONT long-read sequencing presented an opportunity to study both the complete internal structure of reference and non-reference TE insertions, alongside their CpG methylation profiles. We therefore developed TLDR (transposons from long DNA reads), a software tool to detect, assemble, and annotate non-reference TE insertions via long-read alignments, including ONT and PacBio sequencing data. A major feature of TLDR is that it can resolve entire TE insertions, along with transductions, 5′ inversions, target site duplications (TSDs), 3′ poly(A) tracts, and other sequence hallmarks of LINE-1-mediated retrotransposition (Jurka, 1997; Kazazian and Moran, 2017). Another advantage of using long reads to identify TE insertions with TLDR is that in general fewer reads have to be considered per insertion, leading to much shorter computational processing times, compared to typical short-read methods. We also developed software to interrogate the methylation patterns of both non-reference and reference TE insertions (see STAR Methods), which can in principle be applied to any set of genomic coordinates.

RESULTS

Genome-wide Methylation Profiles of Young TE Subfamilies in Human Tissues

We employed an ONT PromethION platform to sequence five human samples at ~15× genome-wide depth. Samples consisted of hippocampus, heart, and liver tissue, representing each of the three germ layers, from an individual (CTRL-5413; female, 51 years) without post-mortem pathology and paired tumor/non-tumor (T/NT) liver tissue from a hepatocellular carcinoma patient (HCC33; female, 57 years) (Figure 1A; Table S1). ONT analysis allowed us to compare CpG methylation among genomes (Simpson et al., 2017) and between haplotypes within samples (Gigante et al., 2019). Examining TE subfamilies *en masse*, we observed significant tumor-specific L1Hs demethylation in HCC33 (29.4% median difference, $p < 1.35e\text{-}39$, Mann-Whitney test) that was more pronounced than demethylation of other young TEs and the genome overall (Figures 1B and S1). Comparing CTRL-5413 normal hippocampus, heart, and liver, we found L1Hs methylation decreased significantly in that order (Figure 1C; hippocampus versus heart $p < 4.11e\text{-}14$, heart versus liver $p < 0.042$, Mann-Whitney test with Bonferroni correction), and this effect appeared more marked among older LINE-1 subfamilies (Figure S2). Older *Alu* subfamilies were also generally less methylated on average than younger *Alu*s (Figure S2). SVA methylation, by contrast, did not significantly vary among the three CTRL-5413 samples or with subfamily age (Figure S2). Long terminal repeat (LTR5_Hs) regions flanking the likely immobile human endogenous retrovirus K (HERV-K) family were less methylated than other TEs in normal tissues and non-tumor liver (Figures 1B and 1C). Genome-wide and TE subfamily methylation was slightly lower in HCC33 non-tumor liver than CTRL-5413 normal liver (Figures 1B and 1C). Composite methylation profiles spanning the previously inaccessible interiors of full-length TEs revealed a clear trough adjacent to the L1Hs 5′ UTR CpG island in all samples (Figure 2), whereas the CpG-rich VNTR core of the youngest SVA_F subfamily was more consistently methylated than its flanking SINE-R and *Alu*-like sequences (Figure 2).

Locus-, Element- and Haplotype-Specific Young Reference TE Methylation States

While most TEs were constitutively methylated (Figures 1C and S3A), we identified striking patterns of differential methylation for individual reference genome TEs among the CTRL-5413 normal tissues (Figures S3B–S3D; Table S2). For example, an L1Hs located intronic to the TTC28 gene and known to be mobile in liver and other cancers (Pradhan et al., 2017; Rodriguez-Martin et al., 2020; Schauer et al., 2018) was hypomethylated in CTRL-5413 liver (Figure 1D). A slightly 5′ truncated L1Hs situated on chromosome 13 and found, thus far, to cause somatic

(D) Methylation profile of a reference L1Hs intronic to TTC28. A purple rectangle indicates the L1Hs position within the TTC28 locus. Upper panel: relationship between CpG positions in genome space and CpG space. The L1Hs 5′ UTR and body are highlighted in blue and orange, respectively. Lower panel: fraction of methylated CpGs for CTRL-5413 tissues across CpG space.

(E) Similar to (D), except for an intergenic L1Hs located on chromosome 13 and known to be demethylated and mobile during neurodevelopment (Evrony et al., 2015; Sanchez-Luque et al., 2019).

Please see Figures S1–S3 and Tables S1 and S2 for further reference TE methylation data.

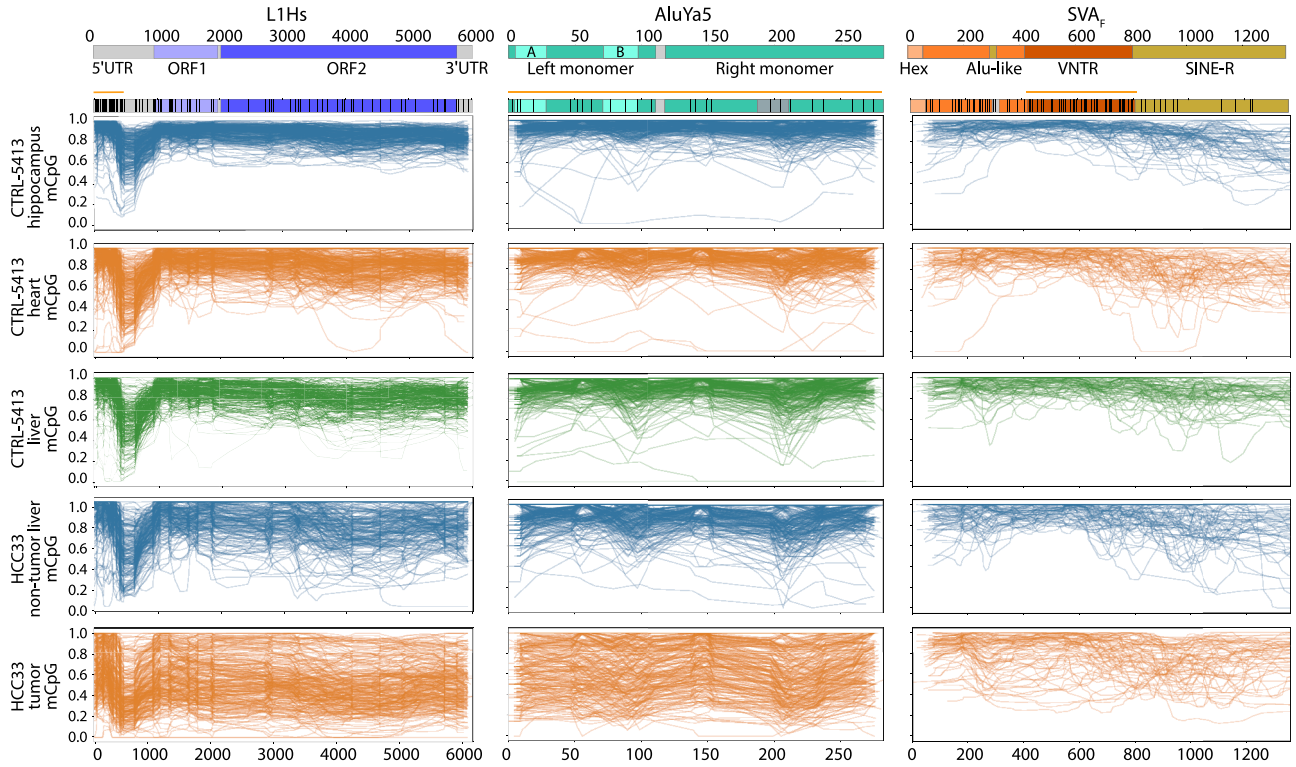


Figure 2. Composite Methylation Profiles for Representative Mobile Human TE Subfamilies

Data are shown for L1Hs, AluYa5, and SVA_F in CTRL-5413 and HCC33 samples. Each graph displays up to 300 profiles for the specified TE subfamily. Annotated TE consensus sequences are provided at top, with CpG positions (black bars) and CpG islands (orange lines) indicated.

retrotransposition during neurodevelopment in two unrelated individuals (Evrony et al., 2015; Sanchez-Luque et al., 2019) was strongly demethylated in each CTRL-5413 sample (Figure 1E). An L1Hs situated antisense and intronic to ZNF638 was similarly demethylated, particularly in CTRL-5413 heart tissue, and from its 5′ UTR promoted transcription of a previously described alternative ZNF638 transcript (Wheelan et al., 2005) (Figure S3C). A chromosome 1 L1Hs that is mobile in the germline and cancer (Gardner et al., 2017), and highlighted as expressed in senescent fibroblasts (De Cecco et al., 2019), was hypomethylated in CTRL-5413 heart and liver, but not hippocampus (Figure S3D).

We also noted exceptions to the subfamily-wide methylation trends observed when comparing HCC33 tumor and non-tumor liver (Figure 3A; Table S2). While SVA methylation was broadly unchanged in HCC33 tumor (Figures 1B and S2B), individual SVAs, such as an intergenic SVA_F located on chromosome X, were demethylated (Figure 3B). Reciprocally, despite L1Hs subfamily-wide tumor-specific demethylation in HCC33 (Figures 1B and S2B), some individual TEs were hypermethylated in tumor relative to non-tumor liver, such as an L1Hs copy intronic to PGAP1 (Figure 3C). Overall, we found 4.7%, 2.6%, and 1.8% of L1Hs, AluY, and SVA copies, respectively, were discordantly hypermethylated compared to the adjacent genome in HCC33 tumor (Z score > 1) versus 1.9%, 3.0%, and 4.1% that were hypomethylated (Figure 3A). As well as such cases of exceptional element-specific methylation, we found TEs apparently demethylated by virtue of their genomic location (Figure 3D). Dynamic locus- and element-specific methylation, combined with patterns and exceptions seen at the subfamily level, is therefore characteristic of young TEs in the cancer epigenome.

Mammalian TEs can exhibit distinct methylation states depending on their parent of origin (Brind'Amour et al., 2018; Ferguson-Smith and Bourc'his, 2018; Kazachenka et al., 2018). By generating long-read-backed phased methylation profiles (Gigante et al., 2019; Patterson et al., 2015; Wu et al., 2015), we found haplotype-specific differentially methylated regions within known imprinted genes, such as GNAS (Davies and Hughes, 1993) (Figure 4A) and PEG3 (Kaneko-Ishino et al., 1995) (Figure 4B). A survey of full-length reference TE insertions revealed 1 L1Hs, 7 L1PA2, 4 L1PA3, 3 SVA_B, 1 SVA_D, and two AluYa5 copies exhibiting both element- and haplotype-specific methylation patterns (Figure S4; Table S2). The single L1Hs example (Figure 4C) was located on chromosome 7 and is moderately mobile in cancer and the germline (Brouha et al., 2003; Rodriguez-Martin et al., 2020). L1PA2, the second youngest LINE-1 subfamily, provided more examples than L1Hs of haplotype-specific methylation, as expected based on the comparatively higher L1PA2 genomic copy number and proportion of fixed elements (Khan et al., 2006; Lander et al., 2001; Mills et al., 2007). These results altogether highlight how haplotype-specific TE regulation can be studied, and placed amid a wider genomic context, via ONT analysis.

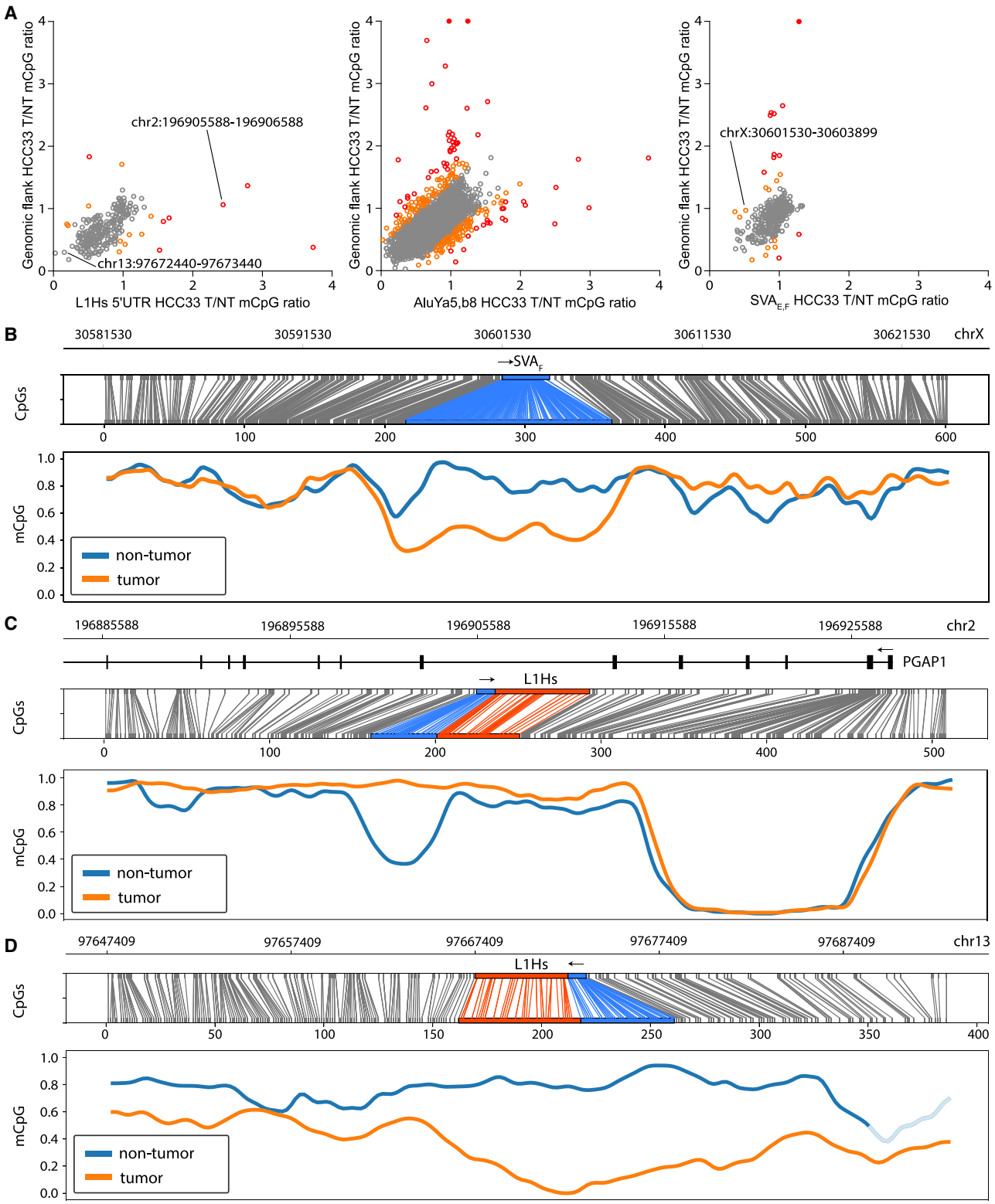


Figure 3. Locus- and Element-Specific Reference TE Methylation in Cancer

(A) Differential methylation of individual TEs in patient HCC33 tumor versus non-tumor liver, relative to the adjacent genome. Only TEs with ≥10 CpGs and ≥20 methylation calls were included, with the genomic flank defined as an equal number of CpGs including at least 1 kbp upstream. Filled circles represent points with

(legend continued on next page)

Long-Read Detection of Polymorphic and Somatic TE Insertions with TLDR

We developed TLDR to study non-reference TE insertions (Figure 5; Table S3) while achieving sensitivity similar to our short-read TE insertion detection method, TEBreak (Carreira et al., 2016). To assess the sensitivity of TLDR, we compiled a high-confidence set of known non-reference (KNR) TE insertions reported by TLDR in CTRL-5413 or HCC33 and by at least 2 of 17 previous datasets aggregated here (see STAR Methods; Table S4). We then applied TEBreak to ~45× Illumina whole-genome sequencing generated from CTRL-5413 heart and HCC33 non-tumor liver and annotated KNR insertions as for TLDR. Of 2,842 KNR TE insertions called by either TLDR or TEBreak, 2,643 were detected by TLDR (Table S3) and 2,644 were found by TEBreak (Table S5). A total of 2,445 KNR insertions were reported by both TLDR and TEBreak (Jaccard metric ~0.86), indicating high concordance. Consistent with the recent use of PacBio long-read sequencing to resolve LINE-1 insertions in difficult to map genomic regions (Zhou et al., 2020), we found non-reference insertions called only by TLDR covered a much broader spectrum of mappability scores than those found only by TEBreak or by both methods (Figure S5). An additional source of discrepancy between short- and long-read analyses is that, unlike TEBreak, TLDR requires insertions to be spanned by at least one ONT read, whereas TEBreak requires only 5′ and 3′ junction coverage.

TLDR reports informative sequence features of TE insertions (Kazazian and Moran, 2017). In total, 2,798 known and unknown non-reference insertions were called by TLDR and passed filtering in CTRL-5413 or HCC33, including 2,359 *Alu*, 322 L1Hs, 108 SVA, and 9 LTR5_Hs (HERV-K) insertions (Figure 6A; Table S3). The median TSD size was 14 bp (Figure 6B), consistent with prior results (Stewart et al., 2011). 1,073 (38.3%) insertions were intragenic. *Alu* supplied all 15 (12 UTR and 3 coding sequence) identified exonic events (Figure 6A). L1Hs presented the lowest fraction (25.2%, 81/322) of intronic insertions (Figure 6A). TE insertion length distributions, while consistent with family consensus sequences, were long-tailed toward insertions of greater than consensus length (Figures 6C–6E). This was due to TLDR being able to resolve entire insertions, including transductions and TEs embedded in other non-reference insertions. 5′ and 3′ transductions carried by L1Hs (Goodier et al., 2000; Moran et al., 1999; Pickeral et al., 2000) and SVA insertions (Hancks et al., 2009), and attributable to their putative source elements (Figure 6F), ranged in length from 31 to 2,072 bp (Table S6). Three (0.9%) and 15 (4.7%) L1Hs insertions were accompanied by 5′ and 3′ transductions, respectively, with SVAs carrying 5 (4.6%) 5′ and 12 (11.1%) 3′ transductions (Figures S6A–S6H; Tables S3 and S6). As 5′ transductions are relatively rare (Lander et al., 2001), we confirmed all eight of these examples by PCR and capillary sequencing (Table S6). Overall, 23.6% (76/322) of L1Hs insertions were 5′ inverted (Ostertag and Kazazian, 2001) and 16.8% (54/322) were full-length (>6 kbp). The latter result, and the number of L1Hs 3′ transductions, were potential under-ascertainments relative to previous studies (Beck et al., 2010; Ewing and Kazazian, 2011; Watkins et al., 2020) owing to the TLDR requirement of at least one spanning read per insertion. Internal polymorphisms, such as VNTR, and CCCTCT hexamer length increases within SVAs (Figure 6G) were also resolved by TLDR. For example, we identified 145 polymorphic SVA VNTR expansions, ranging in size from 37 to 4,039 bp (1 to ~100 repeat units, 8 units on average) (Figures S6I–S6K; Table S3), as well as three polymorphic AluY insertions within VNTR sequences (Table S3). Altogether, TLDR was able to consistently recover the hallmarks of LINE-1-mediated retrotransposition, and resolve all reported non-reference TE insertions *in toto*.

Highlighting its capacity to detect somatic retrotransposition events, TLDR successfully re-identified both PCR-validated tumor-specific L1Hs insertions previously found by us with short-read sequencing of patient HCC33 samples (Shukla et al., 2013). TLDR accurately recapitulated their LINE-1 insertion features, including TSDs, and now revealed the internal breakpoint of the ~2-kb 5′ inversion of the EFHD1 insertion (Figure 6H; Table S3). No additional HCC33 tumor-specific TE insertions were found by TLDR.

Genome-wide Resolution of Non-reference TE Methylation

As for individual reference TEs, TLDR can be used to generate element-specific methylation profiles for non-reference TE insertions (Figures 5B, 7A, and S7; Table S7), including retrotransposition-competent source L1Hs copies. For example, the non-reference element L1.2, responsible for the first report of LINE-1 mobility and pathogenesis in humans (Dombroski et al., 1991; Kazazian et al., 1988), was ~15% less methylated in CTRL-5413 liver and heart than in hippocampus (Figure S7A; Table S7). As non-reference L1Hs copies tend to retrotranspose more efficiently when tested in cultured cells than reference L1Hs elements and the vast majority of mobile L1Hs copies in the global population are absent from the reference genome (Badge et al., 2003; Beck et al., 2010; Seleme et al., 2006), the capacity of TLDR to find non-reference L1Hs alleles and survey their methylation state in parallel is notable.

While somatic methylation appears less ubiquitous as some TE subfamilies age (Figure S2), the uniformity and initial duration

T/NT ratio > 4. Outliers strongly (Z score > 2) and moderately ($2 \geq Z$ score > 1) distant from y = x are colored red and orange, respectively. Genomic coordinates indicate examples from (B)–(D).

(B) Methylation profile for an intergenic SVA_F located on chromosome X. A region of reduced methylation specific to the SVA sequence was observed in HCC33 tumor (orange) when compared to non-tumor liver (blue).

(C) An L1Hs insertion located in an intron of the PGAP1 gene and methylated distinct to the surrounding locus in HCC33 non-tumor liver and not the matched tumor. The demethylated region to the right of the L1Hs corresponds to the PGAP1 promoter region.

(D) An intergenic L1Hs demethylated along with its surrounding genomic region in HCC33 tumor. The non-tumor sample smoothed plot line (blue) is colored to appear faded for a short lower-confidence region (<20 methylation calls within a 30-CpG window).

Please see Table S2 for supporting data.

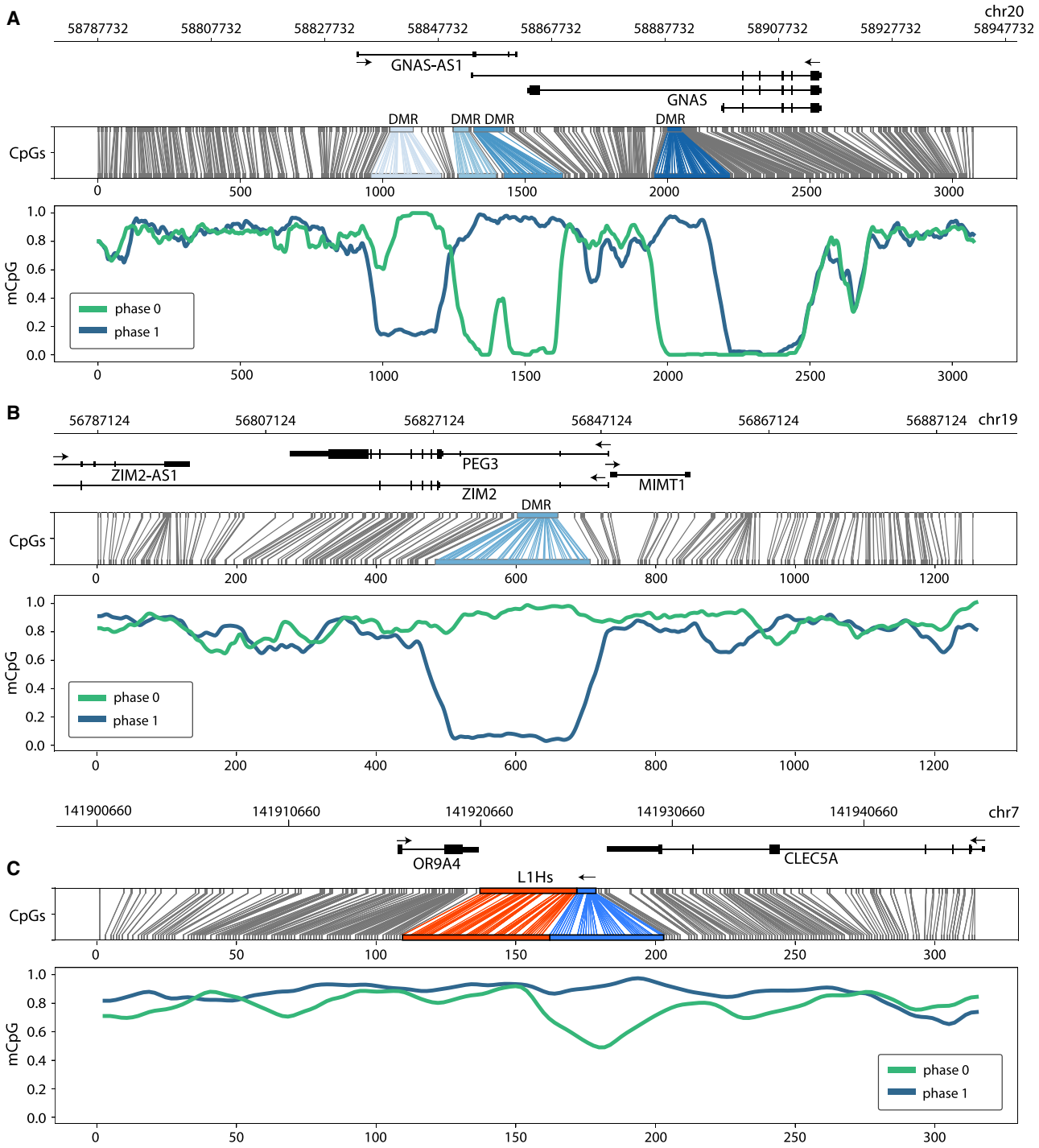


Figure 4. Haplotype-Specific TE Methylation Detected via ONT Analysis

(A) Read-backed phasing was used to identify haplotype-specific differences in CTRL-5413 hippocampus methylation for the GNAS gene, where differentially methylated regions (DMRs) specific to the paternal and maternal alleles were identified. This panel includes, from top to bottom, (1) the genomic position of DMRs, (2) a diagram showing the relationship between genome space and CpG space, and (3) the fraction of methylated CpGs across the region where one haplotype (phase 0, teal) is compared to the other (phase 1, blue). Data are shown via a sliding window plot.

(B) A DMR found in the PEG3 gene, as per (A). PEG3 is known to be expressed from only the paternal allele (Kaneko-Ishino et al., 1995).

(C) As for (A), except for a full-length intergenic L1Hs located on chromosome 7. The L1Hs 5′ UTR and body are highlighted in blue and orange, respectively. Please see Figure S4 and Table S2 for additional examples of haplotype-specific TE methylation.

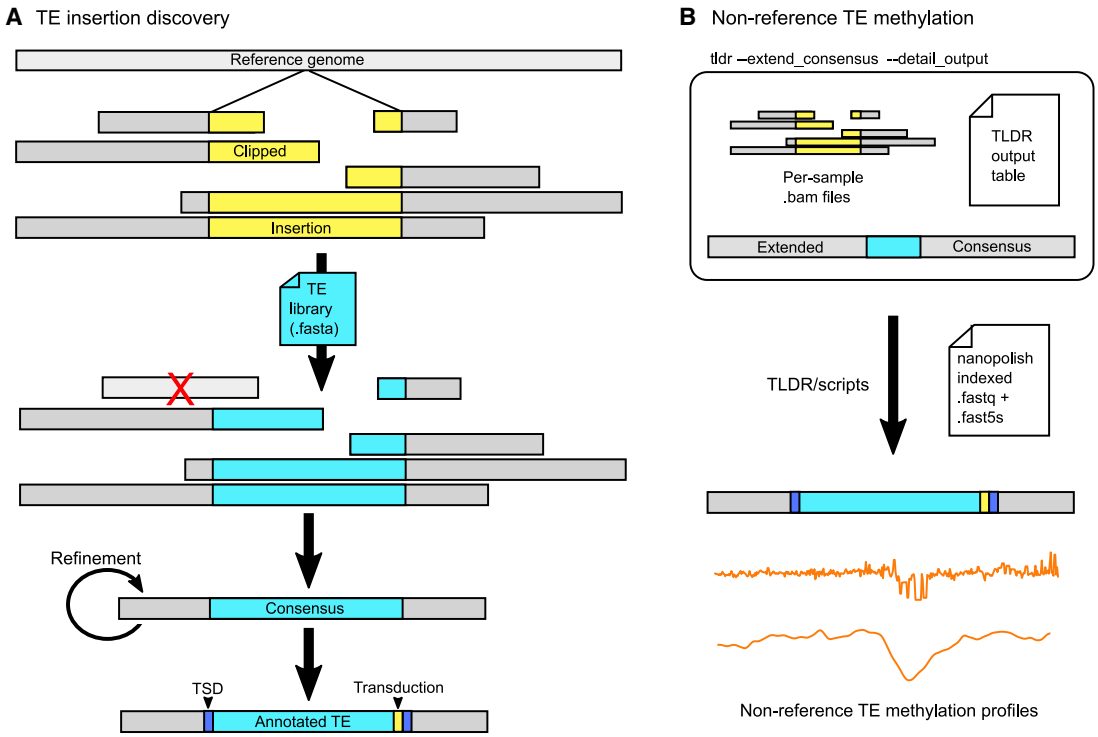


Figure 5. Schematic Depiction of TLDR Operation

(A) Insertion discovery with TLDR begins by identifying clusters of reads with a profile of mapped (gray) and unmapped (yellow) regions consistent with an insertion. The unmapped portion of reads is aligned to a TE reference library (TE mappings in blue), and a consensus is built from multiple sequence alignment of the reads supporting a TE insertion, which is refined through realignment and pileup-based assessment of the supporting reads. Subsequent refinements define TSDs if present and assist with final annotation of the insertion.

(B) Assessment of non-reference methylation profiles is enabled by generation of an optional per-insertion .bam file, which, along with the consensus sequence and flanking regions, can be used to determine the methylation status of the insertion.

Please see Figure S5 and STAR Methods for further details.

required for new TE insertions to be strictly methylated is unclear. Using TLDR, we established that non-reference TE insertions (Figures 7B and 7C) appeared to be, on average, less methylated than reference elements (Figures 1B and 1C) in each of the CTRL-5413 tissues and HCC33 non-tumor liver. Among the analyzed TE subfamilies, L1Hs exhibited the largest and most significant ($p < 5.76e-30$, Mann-Whitney test) difference in mean methylation between reference (82.5%) and non-reference (70.6%) elements. To speculate, assuming non-reference genetic variants are on average younger than reference variants, the observed pattern for young TEs could suggest their methylation level collectively increases for some time subsequent to their integration into the germline, before gradually weakening with age.

DISCUSSION

When interrogated with TLDR, long-read ONT sequencing can robustly detect and characterize somatic and polymorphic TE insertions, including in genomic regions refractory to reliable short-read mapping. ONT analysis provides end-to-end resolution of TE insertions, without generating molecular artifacts associated with PCR amplification. Hallmark features of LINE-1-mediated retrotransposition are therefore readily recovered by TLDR, including relatively long transductions and internal rearrangements. These attributes mean that TLDR has the potential to, for instance, resolve and characterize somatic TE insertions arising during development (Erwin et al., 2016; Evrony et al., 2015; van den Hurk et al., 2007; Kano et al., 2009; Richardson et al., 2017; Sanchez-Luque et al., 2019) and at the same time infer methylation of the inserted TE, as well as its integration site and source locus.

TLDR requires at least one read to completely span any non-reference TE insertion to report that event. As such, the entire annotated nucleotide sequence of a TE insertion is fully resolved in TLDR output, as shown in Table S3. TLDR is also computationally efficient, requiring only ~1 h to process a 15× ONT genome with default parameters and 32 CPUs. These two properties mean that TLDR can be applied to ONT sequencing data generated for population-scale surveys of human genetic variation (Beyter et al., 2019) and, as opposed to catalogs based on short-read sequencing (Sudmant et al., 2015), resolve the internal sequence regions of all reported TE insertions. This is a major advantage for studies of TE-driven structural variation, as sequence polymorphisms internal to TEs are common and, for example, involve SVA VNTR expansions (Sulovari et al., 2019),

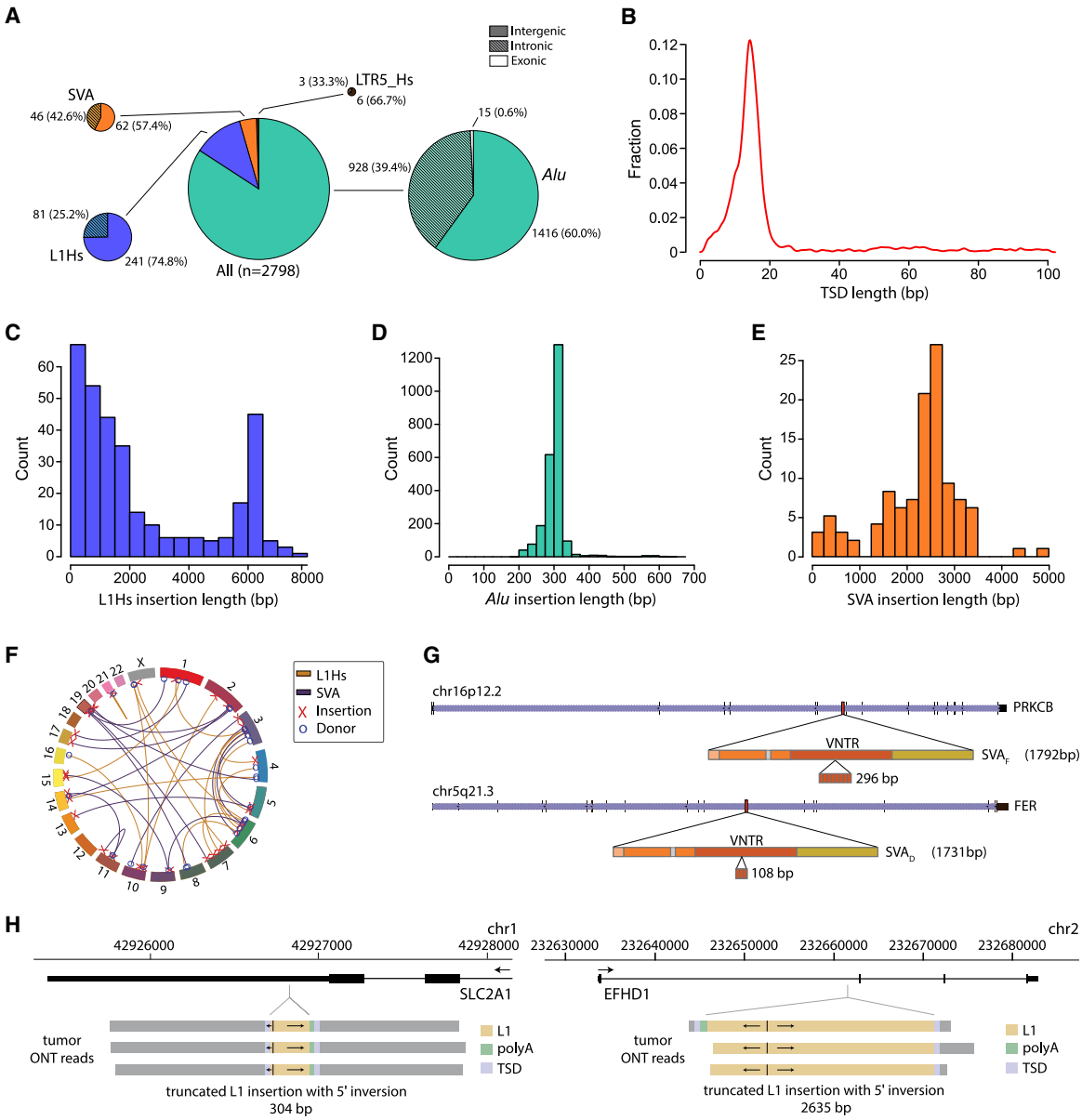


Figure 6. Detection and Characterization of Non-reference TE Insertions

(A) Composition and genomic distribution of TE families. TLDR identified 2,798 non-reference TE insertions in CTRL-5413 and HCC33, and these were annotated by family (central pie chart) and grouped as intergenic, intronic or exonic (satellite pie charts) with respect to protein-coding genes (Ensembl genes version 97). Values next to pie charts represent the counts for each group and their fraction of the total for that TE family.

(B) Combined TSD size distribution for all detected TEs.

(C–E) L1Hs (C), *Alu* (D), and SVA (E) insertion length distributions.

(F) L1Hs and SVA source element (donor) to insertion relationships, defined based on transductions of >30bp.

(G) Examples of SVA internal sequence variation. A VNTR sequence expansion of 296 bp was detected in an SVA_F located in an intron of the PRKCB gene (top; UUID de2c116b in Table S3d) and a 108bp VNTR expansion was detected in an SVA_D intronic to the FER gene (bottom; UUID 9b7ad27c in Table S3d).

(H) Detection of HCC33 tumor-specific 5′ truncated L1Hs insertions. Arrows within L1Hs sequences indicate 5′ inversions.

Please see Figure S6 for additional information regarding TLDR analysis of non-reference TE insertions, Tables S3 and S4 for supporting data, and Table S6 for annotated transductions.

non-allelic homologous recombination of proviral LTRs (Mager and Goodchild, 1989), L1Hs 5′ inversions/deletions (Kazazian et al., 1988), and L1Hs-mediated transductions (Goodier et al., 2000; Moran et al., 1999; Pickeral et al., 2000). Long-read sequencing analyzed with TLDR greatly lessens the need for PCR validation of polymorphic and somatic TE insertions, as variant calling is to a standard akin to capillary sequencing of "filled site" PCR reaction products (Erwin et al., 2016; Evrony

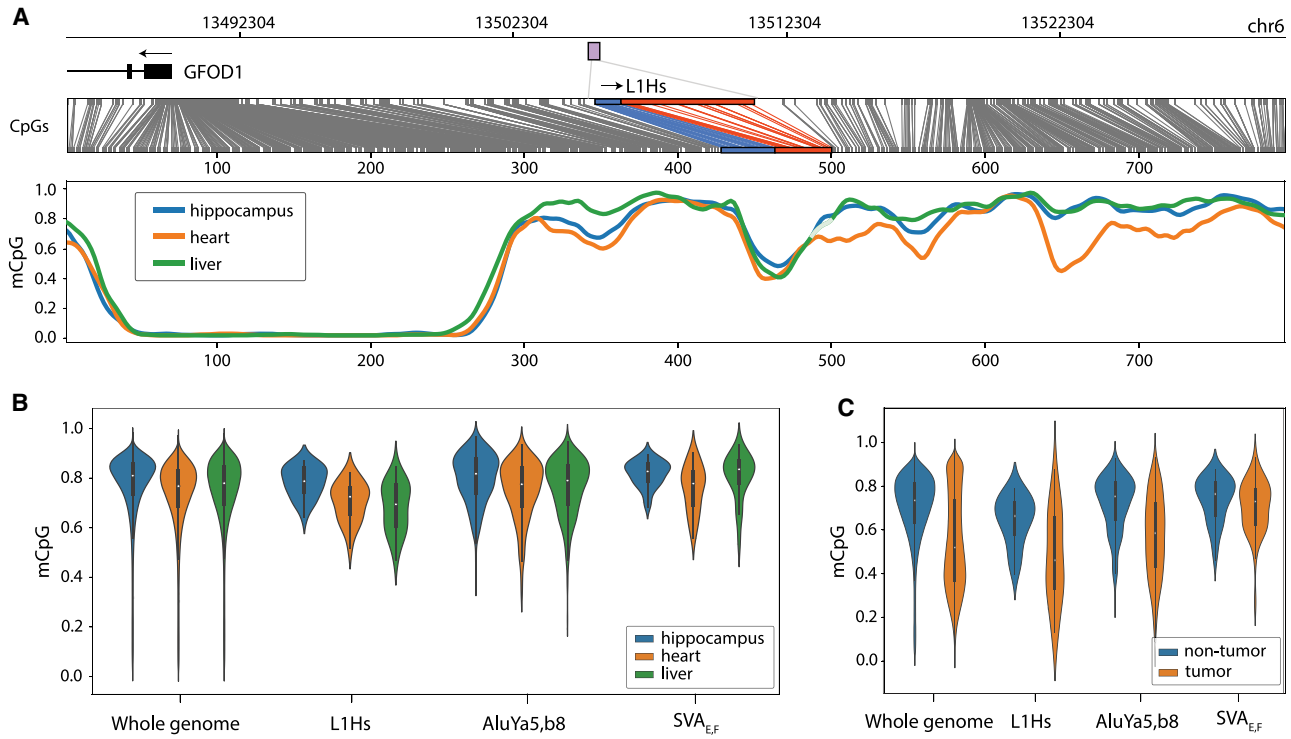


Figure 7. Non-reference TE Methylation Profiles

(A) Exemplar methylation profile for a non-reference L1Hs insertion (UUID d11b3baf; Table S3). A purple box indicates the genomic position of the L1Hs upstream of GFOD1 on chromosome 6. The liver smoothed plot line (green) is colored to appear faded for a short lower-confidence region (<20 methylation calls within a 30-CpG window) at the 3′ end of the L1Hs. Panels are otherwise as described for Figures 1D and 1E. The demethylated region to the left of the L1Hs sequence corresponds to the GFOD1 promoter.

(B) Fraction of methylated CpGs in CTRL-5413 tissues, as in Figure 1C, except for non-reference TE insertions.

(C) As for (B), except for non-reference TE methylation in HCC33 samples.

Please see Figure S7 and Table S7 for additional results and examples of non-reference TE methylation.

et al., 2015; Sanchez-Luque et al., 2019). The requirement for at least one insertion-spanning read is a potential caveat, as full-length (~6 kbp) L1Hs insertion detection sensitivity will particularly decline for datasets with shorter average read lengths. The error rate of long-read sequencing is higher than that of Illumina sequencing, meaning that TSDs, 3′ poly(A) tracts, and other insertion features are only likely to be resolved exactly via the consensus of multiple ONT sequences or, as for PacBio sequencing, through strategies that permit error correction of input reads (Zhou et al., 2020).

Long-read nanopore sequencing greatly illuminates the methylation landscape of young TEs in cancer. Consistent with prior observations for older TEs based on short reads (Pehrsson et al., 2019), demethylation of young *Alu* and SVA elements was here less than or equivalent to that of the remaining HCC33 liver tumor genome. L1Hs methylation was, by contrast, highly dynamic and exceeded genome-wide changes. We found numerous examples of TEs where methylation was disjointed with the surrounding genome, in cancer and in normal tissues, and including haplotype-specific hypomethylation. For instance, differential methylation discordant with the adjacent HCC33 tumor epigenome was observed for 6.6% of reference L1Hs copies. We also identified individual cases of TEs where methylation increased in HCC33 tumor compared to non-tumor liver, despite reduced methylation of nearby genomic regions. Hence, as opposed to potentially stochastic tumor-associated demethylation (Jang et al., 2019), we found that some TE subfamilies appeared more likely than others to lose methylation in tumor cells, especially L1Hs. Consistent with this finding, the vast majority of somatic TE insertions detected in cancer genomes to date have been L1Hs, with comparatively very few *Alu* or SVA insertions (Rodriguez-Martin et al., 2020).

ONT analysis yields a direct methylation readout throughout TE sequences, including previously inaccessible epigenomic regions, such as the SVA VNTR. SVA insertions and their internal structural variants are a significant source of genetic disease (Bragg et al., 2017; Hancks and Kazazian, 2016; Kim et al., 2019; Taniguchi-Ikeda et al., 2011), and it remains to be seen whether spreading of strong VNTR methylation from SVAs can disrupt nearby genes. As shown here for the SVA VNTR and 3′ end of the L1Hs 5′ UTR, CpG methylation can vary greatly within mobile TEs. These methylation "sloping shores" around recently inserted TE CpG islands (Grandi et al., 2015) have the potential to mislead assays that mainly access TE termini. While locus-specific bisulfite sequencing can accurately assess terminal TE methylation for a limited number of loci (Sanchez-Luque

et al., 2019), ONT analysis is far higher throughput and encompasses all human TE subfamilies and their internal regions. Coupled with phased genotypes, long reads can often be assigned to haplotypes, allowing exploration of allele-specific TE methylation. Long-read sequencing and TLDR therefore make a broad range of questions relating to TE biology more accessible.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and Code Availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Sequencing data generation
  - Reference insertions
  - Finding non-reference TE insertions from short reads
  - Finding non-reference TE insertions from long reads
  - 5′ transduction PCR validation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.molcel.2020.10.024.

ACKNOWLEDGMENTS

The authors thank the human subjects of this study who donated tissues to the MRC Edinburgh Brain and Tissue Bank and the Centre Hépatobiliaire, Paul-Brousse Hospital. The authors thank P. Gerdes for helpful discussions and C. James for technical assistance, as well as the University of Queensland Genome Innovation Hub for continuing support. The authors acknowledge the Translational Research Institute (TRI) for research space, equipment, and core facilities that enabled this research. This study was funded by the Australian Department of Health Medical Frontiers Future Fund (MRFF) (MRF1175457 to A.D.E.), the Australian National Health and Medical Research Council (NHMRC) (GNT1125645, GNT1138795, and GNT1173711 to G.J.F.; GNT1173476 to S.R.R.; and GNT1161832 to S.W.C.), a CSL Centenary Fellowship to G.J.F., and the Mater Foundation (Equity Trustees/AE Hingeley Trust).

AUTHOR CONTRIBUTIONS

A.D.E., S.W.C., and G.J.F. designed the research project. A.D.E., N.S., S.W.C., and G.J.F. wrote the manuscript. S.W.C. conducted sample preparation and quality control. J.F. and P.M.B. provided resources. A.D.E. and N.S. developed software tools and analyzed the data with F.J.S.-L., S.R.R., S.W.C., and G.J.F.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 29, 2020
Revised: October 14, 2020
Accepted: October 15, 2020
Published: November 12, 2020

REFERENCES

Alves, G., Tatro, A., and Fanning, T. (1996). Differential methylation of human LINE-1 retrotransposons in malignant cells. Gene *176*, 39–44.

Badge, R.M., Alisch, R.S., and Moran, J.V. (2003). ATLAS: a system to selectively identify human-specific L1 insertions. Am. J. Hum. Genet. *72*, 823–838.

Beck, C.R., Collier, P., Macfarlane, C., Malig, M., Kidd, J.M., Eichler, E.E., Badge, R.M., and Moran, J.V. (2010). LINE-1 retrotransposition activity in human genomes. Cell *141*, 1159–1170.

Beyter, D., Ingimundardottir, H., Eggertsson, H.P., Bjornsson, E., Kristmundsdottir, S., Mehringer, S., Jonsson, H., Hardarson, M.T., Magnusdottir, D.N., Kristjansson, R.P., et al. (2019). Long read sequencing of 1,817 Icelanders provides insight into the role of structural variants in human disease. bioRxiv. https://doi.org/10.1101/848366.

Boeke, J.D., Garfinkel, D.J., Styles, C.A., and Fink, G.R. (1985). Ty elements transpose through an RNA intermediate. Cell *40*, 491–500.

Bragg, D.C., Mangkalaphiban, K., Vaine, C.A., Kulkarni, N.J., Shin, D., Yadav, R., Dhakal, J., Ton, M.-L., Cheng, A., Russo, C.T., et al. (2017). Disease onset in X-linked dystonia-parkinsonism correlates with expansion of a hexameric repeat within an SVA retrotransposon in *TAF1*. Proc. Natl. Acad. Sci. USA *114*, E11020–E11028.

Brind'Amour, J., Kobayashi, H., Richard Albert, J., Shirane, K., Sakashita, A., Kamio, A., Bogutz, A., Koike, T., Karimi, M.M., Lefebvre, L., et al. (2018). LTR retrotransposons transcribed in oocytes drive species-specific and heritable changes in DNA methylation. Nat. Commun. *9*, 3331.

Brouha, B., Schustak, J., Badge, R.M., Lutz-Prigge, S., Farley, A.H., Moran, J.V., and Kazazian, H.H., Jr. (2003). Hot L1s account for the bulk of retrotransposition in the human population. Proc. Natl. Acad. Sci. USA *100*, 5280–5285.

Bruno, M., Mahgoub, M., and Macfarlan, T.S. (2019). The arms race between KRAB-zinc finger proteins and endogenous retroelements and its impact on mammals. Annu. Rev. Genet. *53*, 393–416.

Burns, K.H. (2017). Transposable elements in cancer. Nat. Rev. Cancer *17*, 415–424.

Carreira, P.E., Ewing, A.D., Li, G., Schauer, S.N., Upton, K.R., Fagg, A.C., Morell, S., Kindlova, M., Gerdes, P., Richardson, S.R., et al. (2016). Evidence for L1-associated DNA rearrangements and negligible L1 retrotransposition in glioblastoma multiforme. Mob. DNA *7*, 21.

Castro-Diaz, N., Ecco, G., Coluccio, A., Kapopoulou, A., Yazdanpanah, B., Friedli, M., Duc, J., Jang, S.M., Turelli, P., and Trono, D. (2014). Evolutionally dynamic L1 regulation in embryonic stem cells. Genes Dev. *28*, 1397–1409.

Chuong, E.B., Elde, N.C., and Feschotte, C. (2017). Regulatory activities of transposable elements: from conflicts to benefits. Nat. Rev. Genet. *18*, 71–86.

Cingolani, P., Platts, A., Wang, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.

Davies, S.J., and Hughes, H.E. (1993). Imprinting in Albright's hereditary osteodystrophy. J. Med. Genet. *30*, 101–103.

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.

de la Rica, L., Deniz, Ö., Cheng, K.C.L., Todd, C.D., Cruz, C., Houseley, J., and Branco, M.R. (2016). TET-dependent regulation of retrotransposable elements in mouse embryonic stem cells. Genome Biol. *17*, 234.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). LINE-mediated retrotransposition of marked Alu sequences. Nat. Genet. *35*, 41–48.

Dombroski, B.A., Mathias, S.L., Nanthakumar, E., Scott, A.F., and Kazazian, H.H., Jr. (1991). Isolation of an active human transposable element. Science *254*, 1805–1808.

Erwin, J.A., Paquola, A.C.M., Singer, T., Gallina, I., Novotny, M., Quayle, C., Bedrosian, T.A., Alves, F.I.A., Butcher, C.R., Herdy, J.R., et al. (2016). L1-associated genomic regions are deleted in somatic cells of the healthy human brain. Nat. Neurosci. *19*, 1583–1591.

Esnault, C., Maestre, J., and Heidmann, T. (2000). Human LINE retrotransposons generate processed pseudogenes. Nat. Genet. *24*, 363–367.

Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S., Park, P.J., and Walsh, C.A. (2015). Cell lineage analysis in human brain using endogenous retroelements. Neuron *85*, 49–59.

Ewing, A.D. (2015). Transposable element detection from whole genome sequence data. Mob. DNA *6*, 24.

Ewing, A.D., and Kazazian, H.H., Jr. (2010). High-throughput sequencing reveals extensive variation in human-specific L1 content in individual human genomes. Genome Res. *20*, 1262–1270.

Ewing, A.D., and Kazazian, H.H., Jr. (2011). Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. *21*, 985–990.

Faulkner, G.J., and Billon, V. (2018). L1 retrotransposition in the soma: a field jumping ahead. Mob. DNA *9*, 22.

Ferguson-Smith, A.C., and Bourc'his, D. (2018). The discovery and importance of genomic imprinting. eLife *7*, e42368.

Gardner, E.J., Lam, V.K., Harris, D.N., Chuang, N.T., Scott, E.C., Pittard, W.S., Mills, R.E., and Devine, S.E.; 1000 Genomes Project Consortium (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Res. *27*, 1916–1929.

Gigante, S., Gouil, Q., Lucattini, A., Keniry, A., Beck, T., Tinning, M., Gordon, L., Woodruff, C., Speed, T.P., Blewitt, M.E., and Ritchie, M.E. (2019). Using long-read sequencing to detect imprinted DNA methylation. Nucleic Acids Res. *47*, e46.

Goerner-Potvin, P., and Bourque, G. (2018). Computational tools to unmask transposable elements. Nat. Rev. Genet. *19*, 688–704.

Goodier, J.L. (2016). Restricting retrotransposons: a review. Mob. DNA *7*, 16.

Goodier, J.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2000). Transduction of 3′-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. *9*, 653–657.

Grandi, F.C., Rosser, J.M., Newkirk, S.J., Yin, J., Jiang, X., Xing, Z., Whitmore, L., Bashir, S., Ivics, Z., Izsvák, Z., et al. (2015). Retrotransposition creates sloping shores: a graded influence of hypomethylated CpG islands on flanking CpG sites. Genome Res. *25*, 1135–1146.

Greenberg, M.V.C., and Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. Nat. Rev. Mol. Cell Biol. *20*, 590–607.

Hancks, D.C., and Kazazian, H.H., Jr. (2016). Roles for retrotransposon insertions in human disease. Mob. DNA *7*, 9.

Hancks, D.C., Ewing, A.D., Chen, J.E., Tokunaga, K., and Kazazian, H.H., Jr. (2009). Exon-trapping mediated by the human retrotransposon SVA. Genome Res. *19*, 1983–1991.

Hancks, D.C., Goodier, J.L., Mandal, P.K., Cheung, L.E., and Kazazian, H.H., Jr. (2011). Retrotransposition of marked SVA elements by human L1s in cultured cells. Hum. Mol. Genet. *20*, 3386–3400.

Imbeault, M., Helleboid, P.-Y., and Trono, D. (2017). KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. Nature *543*, 550–554.

Jacobs, F.M.J., Greenberg, D., Nguyen, N., Haeussler, M., Ewing, A.D., Katzman, S., Paten, B., Salama, S.R., and Haussler, D. (2014). An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature *516*, 242–245.

Jain, M., Koren, S., Miga, K.H., Quick, J., Rand, A.C., Sasani, T.A., Tyson, J.R., Beggs, A.D., Dilthey, A.T., Fiddes, I.T., et al. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. *36*, 338–345.

Jang, H.S., Shah, N.M., Du, A.Y., Dailey, Z.Z., Pehrsson, E.C., Godoy, P.M., Zhang, D., Li, D., Xing, X., Kim, S., et al. (2019). Transposable elements drive widespread expression of oncogenes in human cancers. Nat. Genet. *51*, 611–617.

Jönsson, M.E., Ludvik Brattås, P., Gustafsson, C., Petri, R., Yudovich, D., Pircs, K., Verschuere, S., Madsen, S., Hansson, J., Larsson, J., et al. (2019). Activation of neuronal genes via LINE-1 elements upon global DNA demethylation in human neural progenitors. Nat. Commun. *10*, 3182.

Jurka, J. (1997). Sequence patterns indicate an enzymatic involvement in integration of mammalian retroposons. Proc. Natl. Acad. Sci. USA *94*, 1872–1877.

Kaneko-Ishino, T., Kuroiwa, Y., Miyoshi, N., Kohda, T., Suzuki, R., Yokoyama, M., Viville, S., Barton, S.C., Ishino, F., and Surani, M.A. (1995). Peg1/Mest imprinted gene on chromosome 6 identified by cDNA subtraction hybridization. Nat. Genet. *11*, 52–59.

Kano, H., Godoy, I., Courtney, C., Vetter, M.R., Gerton, G.L., Ostertag, E.M., and Kazazian, H.H., Jr. (2009). L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev. *23*, 1303–1312.

Kazachenka, A., Bertozzi, T.M., Sjoberg-Herrera, M.K., Walker, N., Gardner, J., Gunning, R., Pahita, E., Adams, S., Adams, D., and Ferguson-Smith, A.C. (2018). Identification, characterization, and heritability of murine metastable epialleles: implications for non-genetic inheritance. Cell *175*, 1259–1271.e13.

Kazazian, H.H., Jr., and Moran, J.V. (2017). Mobile DNA in health and disease. N. Engl. J. Med. *377*, 361–370.

Kazazian, H.H., Jr., 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.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res. *12*, 996–1006.

Khan, H., Smit, A., and Boissinot, S. (2006). Molecular evolution and tempo of amplification of human LINE-1 retrotransposons since the origin of primates. Genome Res. *16*, 78–87.

Kim, J., Hu, C., Moufawad El Achkar, C., Black, L.E., Douville, J., Larson, A., Pendergast, M.K., Goldkind, S.F., Lee, E.A., Kuniholm, A., et al. (2019). Patient-customized oligonucleotide therapy for a rare genetic disease. N. Engl. J. Med. *381*, 1644–1652.

Lanciano, S., and Cristofari, G. (2020). Measuring and interpreting transposable element expression. Nat. Rev. Genet. Published online June 23, 2020. https://doi.org/10.1038/s41576-020-0251-y.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al.; International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718–719.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, 1303.3997 https://arxiv.org/abs/1303.3997.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

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.

Mager, D.L., and Goodchild, N.L. (1989). Homologous recombination between the LTRs of a human retrovirus-like element causes a 5-kb deletion in two siblings. Am. J. Hum. Genet. *45*, 848–854.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The GEM mapper: fast, accurate and versatile alignment by filtration. Nat. Methods *9*, 1185–1188.

Miga, K.H., Koren, S., Rhie, A., Vollger, M.R., Gershman, A., Bzikadze, A., Brooks, S., Howe, E., Porubsky, D., Logsdon, G.A., et al. (2020). Telomere-to-telomere assembly of a complete human X chromosome. Nature *585*, 79–84.

Mills, R.E., Bennett, E.A., Iskow, R.C., and Devine, S.E. (2007). Which transposable elements are active in the human genome? Trends Genet. *23*, 183–191.

Moran, J.V., Holmes, S.E., Naas, T.P., DeBerardinis, R.J., Boeke, J.D., and Kazazian, H.H., Jr. (1996). High frequency retrotransposition in cultured mammalian cells. Cell *87*, 917–927.

Moran, J.V., DeBerardinis, R.J., and Kazazian, H.H., Jr. (1999). Exon shuffling by L1 retrotransposition. Science *283*, 1530–1534.

Nakamura, T., Yamada, K.D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics *34*, 2490–2492.

O'Neill, K., Brocks, D., and Hammell, M.G. (2020). Mobile genomics: tools and techniques for tackling transposons. Philos. Trans. R. Soc. Lond. B Biol. Sci. *375*, 20190345.

Ostertag, E.M., and Kazazian, H.H., Jr. (2001). Twin priming: a proposed mechanism for the creation of inversions in L1 retrotransposition. Genome Res. *11*, 2059–2065.

Patterson, M., Marschall, T., Pisanti, N., van Iersel, L., Stougie, L., Klau, G.W., and Schönhuth, A. (2015). WhatsHap: weighted haplotype assembly for future-generation sequencing reads. J. Comput. Biol. *22*, 498–509.

Pehrsson, E.C., Choudhary, M.N.K., Sundaram, V., and Wang, T. (2019). The epigenomic landscape of transposable elements across normal human development and anatomy. Nat. Commun. *10*, 5640.

Philippe, C., Vargas-Landin, D.B., Doucet, A.J., van Essen, D., Vera-Otarola, J., Kuciak, M., Corbin, A., Nigumann, P., and Cristofari, G. (2016). Activation of individual L1 retrotransposon instances is restricted to cell-type dependent permissive loci. eLife *5*, e13926.

Pickeral, O.K., Makałowski, W., Boguski, M.S., and Boeke, J.D. (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. *10*, 411–415.

Pradhan, B., Cajuso, T., Katainen, R., Sulo, P., Tanskanen, T., Kilpivaara, O., Pitkänen, E., Aaltonen, L.A., Kauppi, L., and Palin, K. (2017). Detection of subclonal L1 transductions in colorectal cancer by long-distance inverse-PCR and Nanopore sequencing. Sci. Rep. *7*, 14521.

Raiz, J., Damert, A., Chira, S., Held, U., Klawitter, S., Hamdorf, M., Löwer, J., Strätling, W.H., Löwer, R., and Schumann, G.G. (2012). The non-autonomous retrotransposon SVA is trans-mobilized by the human LINE-1 protein machinery. Nucleic Acids Res. *40*, 1666–1683.

Richardson, S.R., Gerdes, P., Gerhardt, D.J., Sanchez-Luque, F.J., Bodea, G.-O., Muñoz-Lopez, M., Jesuadian, J.S., Kempen, M.H.C., Carreira, P.E., Jeddeloh, J.A., et al. (2017). Heritable L1 retrotransposition in the mouse primordial germline and early embryo. Genome Res. *27*, 1395–1405.

Rodriguez-Martin, B., Alvarez, E.G., Baez-Ortega, A., Zamora, J., Supek, F., Demeulemeester, J., Santamarina, M., Ju, Y.S., Temes, J., Garcia-Souto, D., et al.; PCAWG Structural Variation Working Group; PCAWG Consortium (2020). Pan-cancer analysis of whole genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. Nat. Genet. *52*, 306–319.

Rollins, R.A., Haghighi, F., Edwards, J.R., Das, R., Zhang, M.Q., Ju, J., and Bestor, T.H. (2006). Large-scale structure of genomic methylation patterns. Genome Res. *16*, 157–163.

Rowe, H.M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P.V., Layard-Liesching, H., Verp, S., Marquis, J., et al. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature *463*, 237–240.

Salvador-Palomeque, C., Sanchez-Luque, F.J., Fortuna, P.R.J., Ewing, A.D., Wolvetang, E.J., Richardson, S.R., and Faulkner, G.J. (2019). Dynamic methylation of an L1 transduction family during reprogramming and neurodifferentiation. Mol. Cell. Biol. *39*, e00499, e18.

Sanchez-Luque, F.J., Kempen, M.H.C., Gerdes, P., Vargas-Landin, D.B., Richardson, S.R., Troskie, R.-L., Jesuadian, J.S., Cheetham, S.W., Carreira, P.E., Salvador-Palomeque, C., et al. (2019). LINE-1 evasion of epigenetic repression in humans. Mol. Cell *75*, 590–604.e12.

Schauer, S.N., Carreira, P.E., Shukla, R., Gerhardt, D.J., Gerdes, P., Sanchez-Luque, F.J., Nicoli, P., Kindlova, M., Ghisletti, S., Santos, A.D., et al. (2018). L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Res. *28*, 639–653.

Scott, E.C., Gardner, E.J., Masood, A., Chuang, N.T., Vertino, P.M., and Devine, S.E. (2016). A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. *26*, 745–755.

Seleme, M.C., Vetter, M.R., Cordaux, R., Bastone, L., Batzer, M.A., and Kazazian, H.H., Jr. (2006). Extensive individual variation in L1 retrotransposition capability contributes to human genetic diversity. Proc. Natl. Acad. Sci. USA *103*, 6611–6616.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Shukla, R., Upton, K.R., Muñoz-Lopez, M., Gerhardt, D.J., Fisher, M.E., Nguyen, T., Brennan, P.M., Baillie, J.K., Collino, A., Ghisletti, S., et al. (2013). Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. Cell *153*, 101–111.

Simpson, J.T., Workman, R.E., Zuzarte, P.C., David, M., Dursi, L.J., and Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. Nat. Methods *14*, 407–410.

Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics *6*, 31.

Stewart, C., Kural, D., Strömberg, M.P., Walker, J.A., Konkel, M.K., Stütz, A.M., Urban, A.E., Grubert, F., Lam, H.Y.K., Lee, W.-P., et al.; 1000 Genomes Project (2011). A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. *7*, e1002236.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Fritz, M.H., et al.; 1000 Genomes Project Consortium (2015). An integrated map of structural variation in 2,504 human genomes. Nature *526*, 75–81.

Sulovari, A., Li, R., Audano, P.A., Porubsky, D., Vollger, M.R., Logsdon, G.A., Warren, W.C., Pollen, A.A., Chaisson, M.J.P., and Eichler, E.E.; Human Genome Structural Variation Consortium (2019). Human-specific tandem repeat expansion and differential gene expression during primate evolution. Proc. Natl. Acad. Sci. USA *116*, 23243–23253.

Taniguchi-Ikeda, M., Kobayashi, K., Kanagawa, M., Yu, C.-C., Mori, K., Oda, T., Kuga, A., Kurahashi, H., Akman, H.O., DiMauro, S., et al. (2011). Pathogenic exon-trapping by SVA retrotransposon and rescue in Fukuyama muscular dystrophy. Nature *478*, 127–131.

Thayer, R.E., Singer, M.F., and Fanning, T.G. (1993). Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. Gene *133*, 273–277.

van den Hurk, J.A.J.M., Meij, I.C., Seleme, M.C., Kano, H., Nikopoulos, K., Hoefsloot, L.H., Sistermans, E.A., de Wijs, I.J., Mukhopadhyay, A., Plomp, A.S., et al. (2007). L1 retrotransposition can occur early in human embryonic development. Hum. Mol. Genet. *16*, 1587–1592.

Walter, M., Teissandier, A., Pérez-Palacios, R., and Bourc'his, D. (2016). An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. eLife *5*, e11418.

Watkins, W.S., Feusier, J.E., Thomas, J., Goubert, C., Mallick, S., and Jorde, L.B. (2020). The Simons Genome Diversity Project: a global analysis of mobile element diversity. Genome Biol. Evol. *12*, 779–794.

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.

Wheelan, S.J., Aizawa, Y., Han, J.S., and Boeke, J.D. (2005). Gene-breaking: a new paradigm for human retrotransposon-mediated gene evolution. Genome Res. *15*, 1073–1078.

Wu, H., Xu, T., Feng, H., Chen, L., Li, B., Yao, B., Qin, Z., Jin, P., and Conneely, K.N. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Res. *43*, e141.

Yoder, J.A., Walsh, C.P., and Bestor, T.H. (1997). Cytosine methylation and the ecology of intragenomic parasites. Trends Genet. *13*, 335–340.

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.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
Bacterial and Virus Strains		
One Shot TOP10 Chemically Competent cells	Invitrogen	C404006
Biological Samples		
Snap frozen hippocampus, liver and heart tissue from a post-mortem individual.	Edinburgh Sudden Death Brain and Tissue Bank	CTRL-5413
Snap frozen hepatocellular carcinoma and matched non-tumor liver tissue.	Centre Hépatobiliaire, Paul-Brousse Hospital	HCC33
Chemicals, Peptides, and Recombinant Proteins		
Phenol saturated with 10mM Tris, pH 8.0, 1mM EDTA	Sigma-Aldrich	P4557
Phenol:Chloroform:Isoamyl Alcohol 25:24:1 saturated with 10mM Tris, pH 8.0, 1mM EDTA	Sigma-Aldrich	P2069
Chloroform:Isoamyl alcohol 24:1	Sigma-Aldrich	C0549
Sodium acetate	Sigma-Aldrich	S2889
Ethanol	Sigma-Aldrich	E7023
Isopropanol	Sigma-Aldrich	I9516
Agarose	Bioline	BIO-41026
SYBR Safe DNA Gel Stain	Invitrogen	S33102
Proteinase-K	NEB	P8107S
RNase A	Thermo Scientific	EN0531
SDS	Sigma-Aldrich	L3771
Tris-EDTA buffer solution	Sigma-Aldrich	93283
Critical Commercial Assays		
Qubit dsDNA HS Assay Kit	Invitrogen	Q32851
Ligation Sequencing Kit	Oxford Nanopore Technologies	SQK-LSK109
TOPO XL-2 Complete PCR Cloning Kit	Invitrogen	K8040-20
MyTaq DNA Polymerase	Bioline	BIO-21105
Deposited Data		
Nanopore WGS for hippocampus, liver and heart tissue from CTRL-5413, as well as hepatocellular carcinoma and non-tumor liver from individual HCC33. Illumina WGS for CTRL-5413 heart and HCC33 non-tumor liver.	This paper	PRJNA629858
Oligonucleotides		
Oligonucleotide sequences are shown in Table S6.	Integrated DNA Technologies	N/A
Recombinant DNA		
pGEMT®-Easy	Promega	A1360
TOPO XL-2	Invitrogen	K8040-20
Software and Algorithms		
TLDR	https://github.com/adamewing/tldr	This paper
TEBreak	https://github.com/adamewing/tebreak	Carreira et al., 2016
Nanopolish	https://github.com/jts/nanopolish	Simpson et al., 2017
Whatshap	https://github.com/whatshap/whatshap	Patterson et al., 2015
Minimap2	https://github.com/lh3/minimap2	Li, 2018
SAMtools	https://github.com/samtools/	Li et al., 2009
GATK	https://github.com/broadinstitute/gatk	DePristo et al., 2011

(*Continued on next page*)

Continued

REAGENT OR RESOURCE	SOURCE	IDENTIFIER
SnpSift	https://github.com/pcingola/SnpSift	Cingolani et al., 2012
BWA	https://github.com/lh3/bwa	Li., 2013
DSS	https://www.bioconductor.org/packages/release/bioc/html/DSS.html	Wu et al., 2015

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Geoffrey J. Faulkner (faulknergj@gmail.com).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

Illumina and Oxford Nanopore Technologies sequencing data generated by this study were deposited in the Sequence Read Archive (SRA), under BioProject SRA: PRJNA629858. TLDR and TEBreak, and instructions for their use and application, are available at https://github.com/adamewing/TLDR and https://github.com/adamewing/tebreak, respectively.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Snap frozen hippocampus, heart and liver tissue from one post-mortem individual (CTRL-5413, female, 51yrs) without neurological disease was provided by the Edinburgh Sudden Death Brain and Tissue Bank with ethical approval to be used as described in the study (East of Scotland Research Ethics Service, Reference: LR/11/ES/0022). Further ethics approvals were provided by the Mater Health Services Human Research Ethics Committee (Reference: HREC-15-MHS-52) and the University of Queensland Medical Research Review Committee (Reference: 2014000221). Liver tumor and non-tumor samples were previously obtained from a patient (HCC33, female, 57yrs) who underwent surgical resection at the Centre Hépatobiliaire, Paul-Brousse Hospital and were analyzed with approval from the French Institute of Medical Research and Health (Reference: 11-047).

METHOD DETAILS

Sequencing data generation

Tissues were subjected to phenol-chloroform DNA extraction. 50mg of tissue was shaved on dry ice into very fine pieces. The shaved tissue was dissolved at 65°C in TE with 2% SDS and 100μg/ml Proteinase K at a ratio of 1:10 weight/volume. Once dissolved, the sample was cooled to room temperature and RNase A was added to a to a final concentration of 20μg/ml and incubated for 30min at 37°C. An equal volume of phenol (phenol equilibrated with 10nM TRIS pH 8) was added and the sample was mixed by gentle inversion until homogeneous. The sample was centrifuged for 10min at 14,000 g. The aqueous phase was transferred to a new tube and an equal volume of phenol:chloroform:isoamyl alcohol (Tris saturated) was added and the sample was homogenized and centrifuged as above. An equal volume of chloroform:isoamyl was added to the aqueous phase and the sample was homogenized and centrifuged as above. 0.1 volumes of 3M NaOAc and 2 volumes of isopropanol were added, and the tube gently inverted until a white DNA precipitate formed. The DNA was spooled with a pipette tip and transferred to a new tube. The DNA was washed with 70% ethanol and briefly air-dried, then resuspended in 100μl of TE. The DNA was incubated for three days at 4°C before gentle resuspension. DNA concentration was quantified using a Qubit. DNA libraries were prepared at the Australian Genome Research Facility (AGRF) using the genomic DNA by ligation kit (SQK-LSK109) and were sequenced on an Oxford Nanopore Technologies (ONT) PromethION platform (r9.4.1 chemistry). Yield and read N50s varied among samples (Table S1). Bases were called using guppy version 1.8.5 (Oxford Nanopore Technologies) and aligned to the reference genome build hg38 using minimap2 version 2.16 (Li, 2018) and samtools version 1.9 (Li et al., 2009). Reads were indexed and per-CpG methylation calls generated using nanopolish version 0.11.0 (Simpson et al., 2017). Methylation likelihood data were sorted by position and indexed using tabix version 1.9 (Li, 2011).

To generate short reads for comparison of TE insertion detection methods and for the generation of high-quality variant calls necessary for haplotyping, DNA for one sample from each individual (HCC33 non-tumor liver and CTRL-5413 heart), was sequenced to 45x depth by AGRF on an Illumina NovaSeq 6000. Reads were aligned to hg38 using bwa mem version 0.7.12 (Li, 2013) and samtools version 1.9 and duplicate reads were marked using picard tools version 2.18.0 (http://broadinstitute.github.io/picard). Variants were called using GATK Haplotype Caller version 3.7 (DePristo et al., 2011) and known variants annotated via SnpSift version 4.3t

(Cingolani et al., 2012) using dbSNP build 146 (Sherry et al., 2001). Read-backed phasing of ONT reads was done using whatshap version 0.18 (Patterson et al., 2015).

Reference insertions

Per-element methylation statistics for reference TEs were generated using a python script, segmeth.py, available in the te-nanopore-tools GitHub repository at https://github.com/adamewing/te-nanopore-tools. Reads mapping completely within TEs were excluded from the reference TE methylation analysis to minimize the possibility of mismapping. Reference TE locations were derived from the RepeatMasker (http://www.repeatmasker.org/) .out files available for hg38 from the UCSC Genome Browser (Kent et al., 2002). As SVA elements are often broken into multiple adjacent SVA annotations (Jacobs et al., 2014), we merged adjacent similarly oriented SVAs prior to analysis and considered elements annotated as longer than 1000bp. LINE-1 elements were considered if annotated as greater than 5900bp in length. Only *Alu* elements greater than 280bp were considered. We required at least 5 methylation calls i.e., abs(log-likelihood ratio) > 2.5 across all samples to include an element in the survey (Table S2).

Methylation plots for individual elements can be generated using the plotmeth_ref_multi.py or plotmeth_ref_hap.py scripts; the former generates plots for one or more samples and the latter generates plots for .bams that have been haplotype tagged using whatshap. These plotting scripts generate plots with four panels: a positional panel which includes a rudimentary depiction of gene models, a panel which shows the conversion from genome space (i.e., A,C,T,G) into CpG space while optionally highlighting one or more segments, a plot of log-likelihood ratios via seaborn (https://seaborn.pydata.org/), and a plot showing the fraction of methylated CpGs which is windowed and stepped according to user parameters and smoothed using a Hann function. Methylation data displayed here were plotted using a 30bp sliding window with a 2bp step, and smoothed with a window size of 8 for the Hann function.

To survey instances of haplotype-specific divergence in methylation limited to individual TEs and not driven by the surrounding genomic regions, we applied DSS (Wu et al., 2015) to CTRL-5413 hippocampus ONT sequencing data. This generated a set of differentially methylated loci (DMLs) between the two alleles for each CpG. For each TE copy, we aggregated the Wald statistics across DMLs output by DSS within the genomic coordinates of the TE and normalized by the number of DMLs (NormAreaStat). We did the same for 6kbp regions upstream and downstream of each element containing at least as many CpGs as the associated TE. We required at least 40 CpGs for LINE-1 s and SVAs, and at least 20 CpGs for *Alu*s, with at least 20 DMLs having a genome-wide FDR < 0.05. TEs whose up- and downstream regions were within a 1-fold difference in NormAreaStat were then selected, with the cutoff for the minimum TE NormAreaStat ratio set empirically through visualization of haplotype-specific methylation plots. TEs with a predicted haplotype-specific methylation profile were visualized, and those passing visual inspection were included in Table S2.

Finding non-reference TE insertions from short reads

Non-reference TE insertions were detected from Illumina data using TEBreak (https://github.com/adamewing/tebreak) with recommended parameters, apart from the following: -d tebreak/lib/hg38.chr.disctgt.txt -m tebreak/lib/hg38.chr.centromere_telomere.bed–min_sr_per_break 2–skip_chroms chroms.txt. The file 'chroms.txt' was used to limit insertion calls to only canonical chromosomes (chr1-22, X, Y). Known non-reference insertions were annotated in the TEBreak output if also found by at least two of the 17 previous studies listed in Table S4) and otherwise filtered using the script provided (tebreak/scripts/general_filter.py). Insertion calls from TEBreak were compared against insertion calls from TLDR using the script "compare_tebreak_TLDR.py" included in the te-nanopore-tools repository. In this script, mappability of insertion sites (Figure S5) was determined via an index derived from mappability output from the GEM read mapper (Marco-Sola et al., 2012).

Finding non-reference TE insertions from long reads

We developed TLDR to analyze non-reference TE content from long reads. TLDR is available at https://github.com/adamewing/TLDR. While this study focuses on ONT long reads, TLDR can in principle use any accurately-aligned reads long enough to span TE insertions, including PacBio data. TLDR is a python application utilizing multiprocessing that can complete analysis of a 30x long-read genome in under 1 hour (walltime) when utilizing 32 cores. The operation of TLDR proceeds through two phases: clustering and insertion resolution. The required inputs consist of one or more sorted, indexed .bam files (recommended aligner is minimap2), a reference genome .fasta indexed with samtools faidx, and a set of reference TE sequences in fasta format (reference for human is included in distribution). Recommended options include–color_consensus for clear annotation of insertion features using ANSI colors and -p to specify a number of processes for multiprocessing. The results in this study were generated using parameters -e TLDR/ref/teref.human.fa -n TLDR/ref/nonref.collection.hg38.chr.bed.gz–color_consensus -c chroms.txt -o all_extend–extend_consensus 20000–detail_out. The file 'chroms.txt' was used to limit insertion calls to only canonical chromosomes (chr1-22, X, Y).

In the clustering phase, clusters are seeded by identifying insertions (i.e., long indels) completely embedded in long reads. One read containing a completely embedded insertion bound by a minimum and a maximum length (default 200-10000bp) is required to seed generation of a cluster. Additional reads are added to the cluster if they have an apparent breakpoint at either end of the seeding insertion, allowing for some ambiguity (default 200bp) around the junction. Nanopore reads can be arbitrarily long, bounded by the length of the chromosome, so one read can have membership in multiple clusters. In principle, this step could be further

informed by input from more accurate short read mappings. As reads are allowed to be up to one chromosome in length, the clustering step can be parallelized on a per-chromosome basis (i.e., all chromosomes run at once).

The processing phase is parallelized on a per-cluster basis and begins with trimming clusters around the seeding insertion based on a user-specified flank size (*F*, default 500bp). For a given cluster each read is aligned against a set of reference TEs, required as input to TLDR, using an external program (exonerate with an affine:local model) (Slater and Birney, 2005) Up to three non-overlapping alignments at a minimum of 80% identity are reported for each read to allow for internal rearrangement of TEs versus the input reference (e.g., LINE-1 5′ inversions). Per-read alignments are assigned to groups which contain non-overlapping sets of alignments based on the best aligning reference TE. The group with the highest alignment score indicates the identity of the reference TE. Within each cluster *C*, only reads with alignments corresponding to this reference TE are used going forward. We refer to this subset as $C_{useable}$. At this point clusters meeting the minimum read count requirements are retained (default is 3 reads total with at least 1 read fully containing the insertion). Supporting reads are filtered by transforming the TE-aligning fraction of each supporting read into z scores and rejecting reads where abs(z) > 2. A cluster is rejected If a reference TE cannot be assigned or less than 50% of the reads in a cluster align to the majority reference TE.

A consensus sequence is generated for each cluster ($C_{useable}$) using multiple sequence alignment (MSA) via MAFFT (Nakamura et al., 2018). For each column in the MSA, the majority base (A,T,C,G, or gap) is returned where a gap majority can be overridden by two votes from one of (A,T,C,G) (i.e., two "A" bases would override a gap). Gaps are removed from the final consensus. This consensus sequence is then further refined through the following procedure. For each consensus sequence, reads from $C_{useable}$ are aligned back to the consensus sequence using minimap2 and the pileup (i.e., output from samtools mpileup -B -Q 1) is examined. For each pileup column, the consensus sequence base is changed if greater than 50% of reads in the column vote for a different base with at least 3 votes for the same non-consensus base. For each column, if the pileup depth is less than 10% of the number of usable supporting reads in the cluster, the consensus base is not changed.

TE insertion breakpoints ($b_1 \leq b_2$) are initially defined by fitting Gaussian mixture models (GMM) with 1 or 2 means (i.e., an insertion can have 1 or 2 breakpoints depending on the presence of TSDs or not) and choosing the best-fit model based on Akaike information criterion. These initial estimates are used to extract the insertion region from the reference chromosome from position b_1-*F* to position b_2+*F*. The consensus is aligned against this reference sequence using minimap2 to refine breakpoint locations, and to annotate bases as aligned to the reference or part of the insertion. All alignments (primary, and supplementary) are considered where the gap-compressed per-base divergence (i.e., "de" tag) is less than 0.12. Alignments are converted to a per-read profile where matched bases, inserted bases, and soft-clipped bases are encoded. These profiles are merged vertically to yield an overall mask where reference bases covered by a matched base are encoded as 1, and 0 otherwise. This mask is applied to the consensus, marking reference bases in upper case (A,T,C,G) and inserted bases in lower case (a,t,c,g). In each masked consensus, the longest inserted segment is aligned against the reference TEs using exonerate (Slater and Birney, 2005). The initial subfamily designation is allowed to change based on this new alignment. In cases where there is > 1 unmapped segment, if the TE alignment spans both segments, the segments are merged and intervening reference bases are presumed inserted. The refined breakpoints are used to define an initial TSD, which is then expanded to maximize the number of bases in the TSD. If the TSD is extendable, the TE alignment is repeated after TSD extension.

If the user has requested detailed output (–detail_output), part of which is useful for assessing the methylation status of non-reference TE insertions, additional per-insertion information is compiled in a directory named based on the input .bam file(s). This includes a per-insertion file containing information on supporting read mappings and a per-insertion file containing the consensus sequence. If the consensus sequence is extended (–extend_consensus), which is recommended for calling methylation on non-reference insertions, the consensus will be integrated into a larger segment of the reference genome sequence. The detailed output also includes a per-insertion, per-sample .bam file where the reads in the region defined by the extended consensus are aligned against the extended consensus, which includes the insertion sequence. Given the detailed output and a nanopolish indexed fastq (and associated fast5 files), non-reference methylation likelihoods can be obtained using nanopolish call-methylation (Simpson et al., 2017) via the script included in scripts/TLDR_callmeth.sh. Additional auxiliary scripts are included for plotting and tabulating methylation data for non-reference insertions.

The output of TLDR is a tab-delimited file with columns as described in Table S4. Insertions are annotated as KNR if they are present in at least two of the 17 datasets listed in Table S4. The consensus sequence output includes upper case characters representing reference bases and lower case characters representing the inserted sequence. The consensus can optionally be colored (ANSI colors, viewable in compatible terminals via commands such as "cat" and "less -R") via the–color_consensus command. If this is enabled, the inserted TE sequence will appear blue, inserted non-TE sequence (e.g., untemplated bases and transductions) will appear yellow and TSDs will appear red.

5′ transduction PCR validation

To confirm the eight non-reference L1Hs and SVA insertions determined by TLDR to carry a 5′ transduction, we designed primers to PCR amplify a contiguous sequence spanning the 5′ genomic flank of the insertion, 5′ transduction and donor TE 5′ end (Table S6). Each reaction was performed using a T100 Thermal Cycler (Bio-Rad) and MyTaq HS DNA polymerase, with 1X MyTaq Reaction Buffer, 20pmol of each primer, 50ng of template DNA and 1U of enzyme, in a 20μL final volume. PCR cycling conditions were as follows: (95°C, 1min) × 1; (95°C, 15sec; 55°C, 15sec; 72°C, 1min) × 35; (72°C, 5min; 4°C, hold) × 1. Amplicons were visualized

on a 1% agarose gel stained with SYBR SAFE (Invitrogen) using a GelDoc (Bio-Rad) and gel extracted using a QIAquick® Gel Extraction Kit. Amplicons were then inserted into pGEMT®-Easy (Promega) according to the manufacturer's instructions. Regions that did not amplify with MyTaq were instead amplified with Platinum SuperFi Green PCR Master Mix (Invitrogen) using the following cycling conditions: (98°C, 30sec) × 1; (95°C, 10sec; 55°C, 10sec; 72°C, 1min) × 35; (72°C, 5min; 4°C, hold) × 1. Amplicons were visualized and extracted as described above, inserted into TOPO XL-2 (Invitrogen) and transformed into One Shot TOP10 Chemically Competent *E. coli* (Invitrogen) according to the manufacturer's instructions. pGEMT®-Easy and TOPO XL-2 clones were grown in 2mL of LB with 100μg/mL Ampicillin. Plasmid DNA was extracted with a QIAprep Spin Miniprep Kit and capillary sequenced by the Australian Genomics Research Facility (Brisbane).

QUANTIFICATION AND STATISTICAL ANALYSIS

Methylation log-likelihood ratios (LLRs) were calculated using nanopolish call-methylation version 11.0 (Simpson et al., 2017). CpGs are considered methylated if the LLR was above 2.5, unmethylated if the LLR was below −2.5, and ambiguous otherwise. Statistical assessment of differential methylation between groups (Figures 1B, 1C, 7B, and 7C) was carried out via Mann-Whitney tests as implemented in SciPy 1.4.1. Statistical assessment of differential methylation of the same TE between samples (Tables S2 and S7) was carried out via Fisher's Exact Test using methylation and non-methylation counts as implemented in SciPy 1.4.1. Correction for multiple testing was done via Bonferroni's method. Significance was defined as a corrected p value of less than 0.05. Other details regarding parameters pertaining to results shown in figures and tables can be found in the associated legends and in the section labeled Method Details. The survey of haplotype-specific methylation in TEs was accomplished by comparing methylated and non-methylated CpG counts between haplotypes via DSS 2.30 (Wu et al., 2015) as described under "Reference insertions" in Method Details. Python and R scripts used for the analyses presented in this paper are available at the GitHub repositories https://github.com/adamewing/te-nanopore-tools (reference TEs) and https://github.com/adamewing/tldr (non-reference TEs).