Published in final edited form as:
Nat Genet. 2019 April ; 51(4): 611–617. doi:10.1038/s41588-019-0373-3.

Transposable elements drive widespread expression of oncogenes in human cancers

Hyo Sik Jang[#1,2], **Nakul M. Shah**[#1,2], **Alan Y. Du**[1,2], **Zea Z. Dailey**[1,2], **Erica C. Pehrsson**[1,2], **Paula M. Godoy**[1,2], **David Zhang**[1,2], **Daofeng Li**[1,2], **Xiaoyun Xing**[1,2], **Sungsu Kim**[1,3], **David O'Donnell**[1,4], **Jeffrey I. Gordon**[1,4], and **Ting Wang**[1,2,*]

[1] Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, USA

[2] The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA

[3] Hope Center for Neurological Disease, Washington University School of Medicine, St. Louis, MO, USA

[4] Center for Gut Microbiome and Nutrition Research, Washington University School of Medicine, St. Louis, MO, USA

[#] These authors contributed equally to this work.

Abstract

Transposable elements (TEs) are an abundant and rich genetic resource of regulatory sequences[1–3]. Cryptic regulatory elements within TEs can be epigenetically reactivated in cancer to influence oncogenesis in a process termed onco-exaptation[4]. However, the prevalence and impact of TE onco-exaptation events across cancer types are poorly characterized. Here, we analyzed 7,769 tumors and 625 normal datasets from 15 cancer types, identifying 129 TE cryptic promoter activation events involving 106 oncogenes across 3,864 tumors. Furthermore, we interrogated the AluJb-LIN28B candidate: the genetic deletion of the TE eliminated oncogene expression, while dynamic DNA methylation modulated promoter activity, illustrating the necessity and sufficiency of a TE for oncogene activation. Collectively, our results characterize the global profile of TE onco-exaptation and highlight this prevalent phenomenon as an important mechanism for promiscuous oncogene activation and ultimately tumorigenesis.

[*]Corresponding author: twang@wustl.edu.

Contributions:
H.S.J., N.M.S. and T.W. conceived and implemented the study; N.M.S., H.S.J., E.C.P., D.L. and T.W. contributed to the computational analysis; H.S.J. generated transcriptomic and epigenomic profiles of cell lines; H.S.J., X.X. and D.Z. performed the CRISPR-mediated deletion experiments; H.S.J. and Z.Z.D. performed the promoter luciferase, motif mutagenesis, and let-7 qPCR experiments; H.S.J. and A.Y.D. performed the growth and migration assays; H.S.J., A.Y.D., D.O., and J.I.G. performed the xenograft experiments; H.S.J., P.G., and S.K. performed the targeted methylation experiments; H.S.J. performed the rescue experiments; and the manuscript was prepared and revised by H.S.J., N.M.S. and T.W. with input from all authors.

Competing Interests:
Authors declare no competing interests.

Data Availability Statement
Datasets generated and analyzed in this study are available on Gene Expression Omnibus (GEO) under accession code GSE113946.

The elucidation of mechanisms behind oncogene activation has been a long-standing goal in cancer biology. Genetic mutation, gene amplification, and chromosomal rearrangement are three classic genetic mechanisms that drive cancer progression and identity[5,6], but they provide an incomplete explanation for oncogene activation. Recently, a wave of discoveries has demonstrated how TEs change the gene expression landscape during evolution, development, and disease[1–3,7–9]. Although epigenetically silenced in somatic tissues, TEs can become active in cancer due to DNA hypomethylation, which can expose regulatory sequences and lead to functional consequences[10–12]. Indeed, some TEs are epigenetically reactivated as cryptic promoters to drive oncogene expression in cancer, a process known as onco-exaptation[4,13–18]. To our knowledge, no comprehensive study has investigated whether onco-exaptation is a widespread mechanism for oncogene activation across multiple cancer types.

To globally characterize onco-exaptation events, we canvassed RNA-seq data across 15 cancer types from the TCGA Research Network (http://cancergenome.nih.gov/) (Supplementary Fig. 1a). We constructed a computational pipeline that identifies TE-derived oncogene transcripts that are highly tumor-enriched (Supplementary Fig. 1b). A comprehensive list of 702 oncogenes was generated from previously annotated onco-exaptation examples[4,15] and ONGene[19] (Supplementary Table 1). Considering the technical limitations of RNA-seq data, we set stringent filters (Methods) to maximize the specificity for onco-exaptation events. In total, we analyzed 7,769 tumor samples and 625 tumor-matched normal samples (Supplementary Fig. 1b), which identified 625 TE-oncogene chimeric transcripts; this list includes five previously published onco-exaptation examples (Supplementary Table 2). After selecting further for high tumor-enrichment and expression contribution, we identified 129 high confidence onco-exaptation events across 106 oncogenes (Supplementary Table 3). In addition, we detected at least one onco-exaptation event in 49.7% of all tumors, with prevalence ranging from 12% to 87% across cancer types, indicating that onco-exaptation could be a promiscuous mechanism for oncogene activation (Fig. 1a). On average, each onco-exaptation event was discovered in 51 samples and often distributed across multiple cancer types. We report that the onco-exapted TEs strongly enrich for the long terminal repeat (LTR) class (Fig. 1b and Supplemental Fig. 2b). Examining the cancer-type distribution of onco-exaptation candidates (Fig. 1c) showed both cancer-type-specific events, such as THE1A-HMGA2 in SKCM[20], and highly prevalent events were present across multiple cancer types. Furthermore, for eight oncogenes, we observed various TEs activating an in-frame isoform of the same gene (Supplementary Table 4), a phenomenon that had only been described for one oncogene[16]. These additional examples support the cancer epigenetic evolution model as previously described[4]. In summary, we provide a global profile of onco-exaptation events across 15 cancer types and enumerate TEs' role in driving oncogene activation and upregulation.

Next, we examined transcript-level information for the top 10 most prevalent onco-exaptation candidates that on average accounted for greater than 50% of their target oncogenes expression (Fig. 1d). Eight of these candidates were predicted to form in-frame transcripts that conserve protein sequence, suggesting preservation of oncogene function. Onco-exaptation candidates include isoforms of genes such as *SALL4* and *LIN28B* that have recently emerged as potent cancer drivers[21–24]. Additionally, the L1PA2-derived isoform of *SYT1* occurs in more than 10% of all tumors, suggesting that it could be an important cancer marker. While investigating transcript-level abundance of candidates, we found that many of the onco-exaptation events were driving a significant fraction of oncogene expression; some greater than 90% (Fig. 1d & Supplementary Fig. 3). Furthermore, we report that half of the top candidates were associated with worse survival in at least 1 cancer type (Supplementary Fig. 4). For example, we show that the HERVH-SLCO1B3 transcript, a previously characterized onco-exaptation event, is abundant across various cancer types, highly expressed, and associated with worse prognosis[25]. These findings imply that TEs are not only associated with oncogene activation but also contribute significantly to overall oncogene expression and oncogenic potential.

For validation, we sought to confirm transcription initiation from a few exapted TEs. We queried the FANTOM5 promoter database[26] and discovered five out of the ten most prevalent onco-exaptation candidates show promoter signature. We validated a few FANTOM5 results by mapping transcription start sites (TSS) with Cap Analysis of Gene Expression (CAGE)-seq[26–28] in the H727 lung carcinoid cell line. Indeed, *SYT1* and *ARID3A* oncogenes are transcribed from alternative promoters located within TEs (Fig 2a and Supplemental Fig. 5). In addition, we analyzed 27 RNA-seq datasets from lung cancer cell lines[29] and detected 5 of the 10 most prevalent onco-exaptation candidates (Supplementary Table 5). One of the most highly expressed candidates was an AluJb-LIN28B fusion transcript that is present in the H1299, RERF-LC-OK, and H838 cell lines. Considering that *LIN28B* is a well-characterized and potent oncogene[22,24,30–32], we pursued this candidate for further functional validation.

The AluJb TE is located 20 kb upstream of the canonical promoter of *LIN28B* and drives the majority of expression of LIN28B in a substantial number of tumors (Fig. 1d). To verify the existence of the AluJb-LIN28B isoform in lung cancer cell lines, we profiled TSSs in the H1299 and H838 cell lines using paired-end CAGE-seq. We confirmed a CAGE peak, composed of mate reads that align to *LIN28B*, that spans ~40 bp in the AluJb element in both cell lines (Fig 2b). Next, we profiled DNA methylation levels and chromatin accessibility using WGBS-seq and ATAC-seq, respectively (Fig. 2b). The AluJb TE is completely methylated in somatic tissues profiled by Roadmap (http://www.roadmapepigenomics.org/) (Supplementary Fig. 6a). In H1299, the region surrounding the AluJb promoter (AluJb-P) is unmethylated, whereas in H838, it is ~50% methylated. In both cell lines, the region displayed accessibility, indicating an open chromatin state. Together, these findings suggest that an AluJb TE is epigenetically reactivated as an alternative promoter to drive LIN28B expression in lung cancer cell lines.

Next, we dissected the genetic determinants behind the AluJb-LIN28B onco-exaptation event. In H1299 and H838, we discovered that active epigenetic marks encompassed two TEs, a truncated AluJb and MLT1B, upstream of AluJb-P (Fig. 2b). Since various TEs are known to harbor transcription factor binding sites that could have *cis*-regulatory function[2], we tested whether these upstream TEs impact AluJb-P promoter strength. Luciferase assays using various combinations of TEs before a luciferase reporter showed that vectors without AluJb-P displayed minimal activity (Fig. 2c). Furthermore, the luciferase activity did not diminish in the solo AluJb-P vector relative to other vectors. These results illustrate that

AluJb-P contains all the necessary sequences for strong promoter activity, and the upstream TEs have minimal *cis*-regulatory effect on AluJb-P transcription.

AluJb is a primate-specific subfamily within the short interspersed nuclear element (SINE) class of TEs. SINE elements are known to recruit RNA polymerase (RNAP) III to generate short transcripts that can potentially be retrotransposed[33]. However, majority of mRNAs are typically transcribed by RNAP II. We hypothesized that AluJb-P accumulated mutations through evolution that generated novel transcription factor binding sites that recruit RNAP II. To explore this hypothesis, we performed pair-wise sequence alignment using EMBOSS Needle[34] between the AluJb-P sequence and the AluJb consensus sequence from Dfam[35]. We then identified potential novel transcription factor motifs that were generated by mutations specific to AluJb-P with FIMO[36]. Previous work has demonstrated that NFYA binds to AluJb-P and knockdown of NFYA reduces promoter activity in Huh-7 cells[37]. However, the degree of NFYA's impact on AluJb promoter function is still unclear. Our analysis with FIMO detected four other transcription factor motifs that potentially arose from mutations: C/EBPD, SP1, SP4, and YY1 (Fig. 2d). To interrogate the functional importance of these motifs, we cloned AluJb-P sequences mutagenized for each motif into a luciferase reporter and assessed the change in promoter activity. In both H1299 and H838, mutating SP1, SP4, and YY1 sites significantly diminished relative luciferase expression, which is consistent with previous findings that SP transcription factors cooperate with YY1 to drive strong promoter expression (Fig. 2d)[38]. Furthermore, these results were recapitulated in the K562 leukemia cell line (Supplementary Fig. 8a,b), which does not express the AluJb-LIN28B transcript. This finding suggests that K562 cells have all the transcriptional machinery to transcribe from the AluJb-P, but DNA methylation might be suppressing the activity of the promoter (Supplementary Fig. 6a).

To evaluate the functional consequences of the AluJb-LIN28B onco-exaptation event, we first investigated whether the fusion transcript produces a protein product. Within the AluJb-P sequence, we detected a strong start codon 72 bp downstream of the TSS. This results in the addition of 22 amino acids at the N-terminus of exon 2 of LIN28B (Supplementary Fig. 6c), for a predicted protein size increase of 2.5 kDA compared to normal LIN28B. Western blots verified the expected size difference between the onco-exapted AluJB-LIN28B isoform present in H1299 and H838 cells compared to the canonical LIN28B protein present in K562 and HepG2 (Supplementary Fig. 6d). To confirm that the larger protein originated from AluJb-P, we performed CRISPR-Cas9-mediated deletion of AluJb-P in H1299 and H838 (Fig. 3a). In addition, we deleted a 1-kb sequence of the canonical *LIN28B* promoter (LIN28BP). The deletion of AluJb-P abolished the larger LIN28B protein, while the deletion of LIN28BP did not (Fig. 3b), verifying that AluJb-P produced the larger LIN28B isoform.

Since the AluJb-LIN28B protein is identical to canonical LIN28B, aside from the additional N-terminal amino acids, we examined whether AluJb-LIN28B retained normal LIN28B function. LIN28B represses let-7 miRNAs[30,31,39–41], ultimately contributing to oncogenesis through the upregulation of oncogenes such as MYC and RAS[22,24,32]. As anticipated, we observed an appreciable increase in the levels of let-7a, let-7b and let-7g in the AluJb-P knockout (KO) cells but not in LIN28BP KO cells of H1299 and H838 (Fig. 3c). We further assessed how the deletion of AluJb-P impacts cancer-specific attributes. In both H1299 and

H838, AluJb-P KO cells show much slower growth (Fig. 3d) and migration (Fig. 3e) relative to the parental cell lines and LIN28BP KO cells. Also, parental H1299 and LIN28BP KO clone established rapidly growing tumors *in vivo*, whereas AluJb-P KO cells exhibited a marked defect in tumor growth during the time of inspection (Fig. 3f), consistent with the necessity of LIN28B for tumor growth in murine xenograft models[23,37]. In contrast, the deletion of AluJb-P in K562 cells did not result in elevated let-7 levels (Supplementary Fig. 8e) or loss of proliferation (Supplementary Fig. 8f), implying that the loss of AluJb-LIN28B was causal for the decreased oncogenic attributes in H1299 and H838 cells and not due to an off-target effect. Additionally, re-expression of canonical LIN28B and AluJb-LIN28B in H1299 and H838 AluJb-P KO cells reduced let-7 miRNA levels and modestly rescued proliferation (Fig. 3g). Altogether, these results indicate that TE-induced oncogene expression can retain its canonical function, which contributes to cell proliferation, migration, and tumor formation.

Most tumors exhibit global DNA hypomethylation, which provides cancer cells with an opportunity to exploit the regulatory potential of TEs. However, whether the loss of DNA methylation is causal for spurring TE’s cryptic promoter activity has been underexplored due to a lack of efficient targeted methylation techniques. To directly assess how DNA methylation regulates AluJb-P activity, we utilized the CRISPR SUperNova tagging system (SunTag) to recruit either DNMT3A or TET1CD for targeted methylation or demethylation, respectively (Fig. 4a,b)[42–44]. This system allowed us to modestly increase DNA methylation of the AluJb TE by ~20–30% (Fig. 4c), which led to ~40% decrease in LIN28B expression in the H1299 (Fig. 4d), suggesting that DNA methylation of the TE is sufficient to decrease oncogene expression. Additionally, demethylation of the AluJb TE in K562 (Fig. 4e) led to the production of AluJb-LIN28B fusion protein (Fig. 4f). These results illustrate that dynamic DNA methylation is a driving epigenetic control that act as on-off switch for AluJb-P’s activity and moreover suggests that TE onco-exaption events arise in tumors due to the unique epigenetic landscape.

Discussion:

In conclusion, TEs provide an additional means by which cancer can activate oncogenes. Stochastic, global DNA hypomethylation of cancer cells indiscriminately resurrects TEs of varying regulatory ability, which, if they confer a fitness advantage, can be epigenetically inherited and selectively propagated during tumor progression. Here, we present a global profile of tumor-enriched, TE-derived oncogene transcripts across 15 cancer types and show that onco-exaptation is a highly prevalent and promiscuous mechanism that contributes to oncogene activation in close to half of all tumors. By dissecting the mechanisms behind AluJb-derived LIN28B expression, we describe how TEs may be epigenetically and transcriptionally activated to drive oncogene expression. Recently, this tumor-specific *LIN28B* alternative promoter usage in liver cancer has also been characterized by Guo et al. (2018)[37], but not in an onco-exaptation context. Our concomitant findings in lung cancer cell lines provide cross-cancer support of the robust oncogenic potential of AluJb-LIN28B. Recognizing onco-exaptation events can provide additional insights into potential genetic and epigenetic mechanisms that drive promoter activity in cancer. For example, we were able to identify additional putative transcription factors that might be controlling AluJb promoter activity by exploring the evolution of the SINE element. Furthermore, we provide evidence that these onco-exaptation events are potentially reversible through targeted epigenetic alterations, which could present a translational avenue for personalized epigenetic oncotherapy. In summary, TEs act as double-edged swords for cancer by offering additional mechanisms for oncogene activation but also providing a potential target for therapeutics.

Methods:

Data download

All patient sample RNA-seq data analysis was done on the GDC Data Release 9.0 of TGCA data (10/24/17). Normal and tumor RNA-seq BAM files for the following 15 cancers were downloaded using the gdc-client version 1.3.0: bladder urothelial carincoma (BLCA), breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), low grade glioma (LGG), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), ovarian serous cystadenocarcinoma (OV), prostate adenocarcinoma (PRAD), skin cutaneous melanoma (SKCM), stomach adenocarcinoma (STAD), thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC). In addition, normalized gene expression data (HTSeq-FPKM-Uq) and clinical metadata for all samples were downloaded using the gdc-client version 1.3.0. A total of 7,769 tumor samples and 625 matched-normal samples were used for analysis. 26 lung adenocarcinoma cancer cell line RNA-seq files were downloaded using sratools with the following accession: DRA001846. We included RNA-seq of the H838 lung cancer cell line, which has been previously generated in our laboratory and will be publicly available. GENCODE Version 25 was used as the transcript reference[45]. The GTF file of consensus transcripts was downloaded from https://www.gencodegenes.org/releases/25.html. Repeatmasker annotations were downloaded from the UCSC table browser for hg38[46,47]. FANTOM5 hg38-aligned peaks used for annotating the supplementary tables were downloaded from http://fantom.gsc.riken.jp/5/datafiles/reprocessed/hg38_latest/. 698 protein-coding oncogenes obtained from the ONGene database [19]. 4 genes from previous publications noting “onco-exaptation” were included in the list: IRF5, FABP7, SLCO1B3, IL33[4,15]. More details about the software used in our analysis can be found in the *Life Sciences Reporting Summary*.

Assembly and annotation of transcripts

BAM files were sorted and indexed and chr1–22, X, and Y were extracted. Stringtie version 1.3.3 was used to assemble the BAM files for all the RNA-seq samples (stringtie –m 100 –c 1)[48]. These transcripts were then annotated with features from GENCODE v25 with a custom script. Briefly, GENCODE v25 was first processed into a coordinate dictionary based on chromosome, start, and end location. Only the transcripts that were considered “appris_principal” were used so that alternative transcripts of the gene would not be excluded as potential TE-derived candidates. This set of principal transcripts as well as the Repeatmasker TE coordinates were used to annotate the transcripts generated from the stringtie assembly for each sample. The starting position of the transcript was annotated using the Repeatmasker table to find TE-derived transcription start sites. Then, the first exon of the transcript was annotated based on overlap with exonic or intronic features of

GENCODE v25. If the exon overlapped both an exon and intron, then the exon was selected as the annotation for that element. Then, all subsequent exons in the transcript were annotated until one overlapped with a protein-coding gene exon; this exon of the protein-coding gene was selected as the "splice target" of that transcript. After all transcripts were annotated, candidate transcripts were selected based on the following criteria: the start site of the transcript being within a TE, the TE being intergenic or intronic, the starting exon not overlapping with exon 1 of the canonical gene, and the transcript splicing into a protein-coding gene. We further limited our analysis to only include a list of 702 oncogenes to increase likelihood of finding candidates with tumorigenic impact.

Generating a reference transcriptome including onco-exaptation candidates

Aggregating annotation data across all tumor and normal RNA-seq data sets, we constructed a list of unique onco-exaptation candidates based on the subfamily of the TE, the chromosomal coordinates of the TE, and the exon of the gene that the transcript spliced into. To remove potential assembly artifacts and genomic contamination, we removed candidates that had an average exon 1 length greater than the 99th percentile of all GENCODE v25 transcript first exons (2,588 bp). Furthermore, transcripts with first exons that retained an intron were also removed. Finally, we only included candidates that were present in at least 2 samples.

To further increase confidence of promoter activity, we interrogated all reads that uniquely mapped to each candidate TE. We subsequently annotated the mate pair of those reads to see if any overlapped directly with oncogene exons. For single-end reads, we annotated the portion of the read mapping outside the TE to see if it overlapped with an oncogene exon. First, we removed candidates that had zero files where there were at least 10 uniquely mapped reads that started within the TE. In addition, these events were required to have at least 1 sample with uniquely mapped paired-end reads where one of the pairs mapped to the TE and the other to the splice target of the candidate. For intronic onco-exaptation events, we also removed candidates that had evidence of exonization (there were reads mapping to both an upstream exon and the TE) in more than 15% of samples. Finally, candidates that were exclusively in single-end RNA-seq files were removed. The remaining candidate TE-derived transcripts were then merged with the reference GENCODE v25 annotation file using Cuffmerge to create a reference transcriptome inclusive of potential onco-exaptation events that have not been previously annotated.

Transcript-level quantification and candidate selection

To determine the contribution of candidates to overall gene expression, we used stringtie (-e -b) with the merged transcriptome as the reference. For each sample, we labeled a candidate as being present if it met the following criteria: (1) the transcript accounted for at least 25% of total gene expression, (2) there was at least one read covering the splice junction between the TE and the splice target (candidates without unique splice junctions were removed), and (3) the target gene had at least 1 FPKM expression. Next, we filtered for candidates that were highly tumor enriched (> 10× enrichment in the tumor samples) and present in at least 4 tumor samples. For the two cancers where there were no normal samples (OV and LGG), we removed candidates that had > 75% of their samples in these tumor types to avoid simply enriching for tissue-specific alternative promoters. This gave us a master list of 129 tumor-enriched onco-exaptation candidates involving 106 oncogenes. We then explored the abundance of these 129 candidates across the various cancer types to determine the prevalence of this phenomenon.

Open-reading-frame (ORF) prediction and FANTOM5 annotation

After determining the predicted transcript sequences of our candidates, we used CPC2 which predicted which candidates were coding or non-coding[49]. For coding transcripts, we subsequently used the start codon identified by CPC2 for the longest open reading frame and evaluated if it was in-frame or out-of-frame in relation to the canonical isoform. For FANTOM5 promoter annotation, we first filtered the FANTOM5 peaks in hg38 for samples that were not part of exposure or time-course experiments. Subsequently, we evaluated if there were any peaks that overlapped with the onco-exapted TE that were on the same strand as our candidate transcript.

Code Availability

All custom scripts are available from the authors upon request.

Cell culture methods

All cell lines were grown in a humidified incubator with 95% CO_2 at 37°C. H1299, H838, H727 and K562 cell lines were cultured in RPMI 1640 media (Gibco, 11875–085) supplemented with 10% fetal bovine serum (Corning, 35–011-CV) and 100U/ml penicillin-streptomycin (Gibco, 15140–122). HEK293T cell line was cultured in DMEM media (Gibco, 11965–084) supplemented with 10% fetal bovine serum and 100U/ml penicillin-streptomycin. Adherent cells were passaged at 70–90% confluency with 0.05% Trypsin-EDTA (Gibco, 25300–54).

Epigenome and transcriptome profiling

H1299 and K562 whole-genome bisulfite (WGBS)-seq and Cap Analysis of Gene Expression (CAGE)-seq were obtained from previously published results[27,50]. To generate WGBS-seq of H838 cell lines, we extracted genomic DNA with *Quick*-DNA Miniprep Kit (Zymo, D3024) and bisulfite converted 200 ng of DNA using EZ DNA Methylation-Direct kit (Zymo, D5020). For WGBS-seq, we processed the bisulfite-converted DNA with TruSeq DNA Methylation Kit (Illumina, 15066014). To evaluate DNA methylation of targeted regions, we performed bisulfite-PCR using ZymoTaq PreMix (Zymo, E2003) following manufacturer's protocol. Illumina adapters were ligated onto the BS-PCR product and amplified for sequencing. WGBS-seq and targeted BS-PCR libraries were sequenced on Illumina NextSeq and MiSeq platforms, respectively. The sequencing reads were aligned to hg19 genome with Bismark and CpG methylation values were calculated using bismark_methylation_extractor function[51].

To generate chromatin accessibility profiles for H1299 and H838, we followed the published Omni-ATAC-seq protocol[52]. Omni-ATAC-seq libraries were sequenced on Illumina NextSeq platform and reads were mapped to hg19 genome using bwa-mem[53].

Total RNA was extracted using TRIzol Reagent (ThermoFisher Scientific, 15596026) following manufacturer's protocol with few modifications. We performed an extra chloroform wash after transferring the aqueous phase. Furthermore, we added 5 μg of glycogen and 750 μl of isopropanol to the aqueous phase and incubated the solution overnight at −20°C to precipitate the RNA. Total RNA was treated with TURBO Dnase (ThermoFisher Scientific, AM2238). H838 RNA-seq library was generated using TruSeq RNA Library Prep Kit v2 (Illumina, RS-122–2001).

To annotate transcription start site locations, we generated CAGE-seq libraries using CAGE Preparation Kit (DNAFORM). In brief, 10 μg of total RNA was reverse transcribed using SuperScript III (ThermoFisher Scientific, 18080093) and 5' cap of mRNA was biotinylated. Biotinylated RNA/cDNA hybrid was purified using Dynabeads M-280 Streptavidine beads (ThermoFisher Scientific, 11205D) and processed to be sequenced on the Illumina sequencing platforms. For H727, we generated nanoCAGE-seq libraries[54]. In summary, polyA mRNA was extracted using Dynabeads™ mRNA DIRECT™ Purification Kit (ThermoFisher Scientific, 61011). The mRNA was enriched for 5' capped mRNA via Terminator exonuclease (Lucigen, TER51020) digestion. Then we followed standard nanoCAGE protocol to generate the cDNA via template-switching technology. H1299 and H838 CAGE-seq reads were aligned to the hg19 genome while H727 nanoCAGE-seq was aligned to hg38 genome with HISAT and processed using CAGEr package in R statistics[55]. All browser tracks are visualized with the WashU Epigenome Browser[56].

Quantitative PCR (qPCR) of let-7 miRNA and LIN28B

Let-7 miRNA levels were profiled using a published real-time PCR-based platform[57]. To summarize, 500 ng of total RNA was reverse transcribed using SuperScript IV First-Strand Synthesis System (ThermoFisher Scientific, 18091050) with primers specific to let-7a, let-7b, let-7g and U6-snRNA transcripts. For *LIN28B*, *GAPDH* and *β-actin* qPCR, we processed 500 ng of total RNA with iScript Reverse Transcription Supermix (Bio-Rad, 1708840). Afterwards, we performed quantitative PCR on 1 μl of cDNA using PerfeCTa SYBR Green SuperMix (Quantabio, 95053–100). qPCR primers are listed in Supplemental Table 6. Results from qPCR were normalized to house-keeping gene to obtain ΔC_T and ΔC_T of samples were normalized to WT values to obtain $\Delta\Delta C_T$ values. Relative fold-change is calculated as $2^{-\Delta\Delta CT}$.

Western blot and antibodies

Whole-cell lysates for Western blots were extracted with Blue Loading Buffer Pack (Cell Signaling Technology, 7722S). Protein lysates were loaded into Novex 16% Tris-Glycine Mini Gels (Thermo Fisher Scientific, XP99165BOX) and separated by gel electrophoresis at 125V for 4 hours. LIN28B and β-actin were detected using an anti-LIN28B antibody (Cell Signaling Technology, #4196) and an anti-ACTB mouse monoclonal antibody (GenScript, A00702), respectively. More details about the antibodies can be found in the *Life Sciences Reporting Summary*. The Western blot was imaged with Thermo Scientific myECL Imager (Thermo Scientific, 62236).

Promoter and mutagenesis luciferase assay

Various promoter sequences derived from TEs were amplified and extended using primers listed in Supplementary Table 6 from H1299 genomic DNA. The minimal promoter sequence of pGL4.23 luciferase plasmid (Addgene, E8411) was removed with HindIII & NcoI restriction enzyme digest and the TE-derived promoters were cloned into pGL4.23 plasmid via Gibson Assembly following manufacture's protocol (New England Biolabs, E2661S). For mutagenesis assay, we mutated specific motifs within the AluJb promoter-luciferase vector with QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent, 210518). We used the Neon transfection system (MPK5000) to deliver 400 ng of promoter-luciferase vector and 200 ng of pRL-TK *Renilla* vector (Addgene, E2241) into 3×10^4 H1299 cells, 3×10^4 H838 cells or 5×10^4 K562 cells. Luciferase levels were measured after 24 hours of incubation with Dual-Glo Luciferase Assay System (Promega, E2940).

CRISPR-Cas9-mediated Deletion of AluJb and LIN28B Promoter

We selected CRISPR-Cas9 sgRNAs by using both CRISPOR[58] and CRISPRscan[59] to identify sequences that have minimal off-targets and are highly efficient. We purchased pU6-(BbsI)_CBh-Cas9-T2A-BFP plasmid (Addgene, 64323) & pU6-(BbsI)_CBh-Cas9-T2A-mCherry plasmid (Addgene, 64324) as the CRISPR delivery vectors. For each sgRNA, we designed and annealed pairs of oligonucleotides that can be cloned into a BbsI-digested CRISPR vector through standard ligation techniques. We constructed BFP-CRISPR vectors that express sgRNAs targeting upstream and mCherry-CRISPR vectors that express sgRNAs targeting downstream of the region we want to delete. BFP-CRISPR vector and mCherry-CRISPR vector are co-transfected into H1299, H838 and K562 cells via Neon transfection system. After 24 hours of incubation, the transfected cells are analyzed by flow-cytometry (Beckman Coulter MoFlo) for BFP-positive and mCherry-positive fluorescence. We sorted double-positive fluorescent cells into 96-well plates for single-cell clone expansion. Genomic DNA from CRISPR clones was extracted using *Quick*-DNA Miniprep Kit for genotyping and validated with Sanger sequencing.

Cell proliferation assay

We seeded 2,500 wild-type cells or CRISPR-deletion clones in 100 µl of culture media into each well of 96-well plates. Ten µl of Cell Counting Kit-8 (Dojindo Molecular Technologies, CK04–01) were added to each well at appropriate time points. After 1 hour of incubation in humidified incubator with 95% CO_2 at 37°C, we recorded O.D. at 450 nm using BioTek Synergy H1 Hybrid Reader.

In vitro scratch migration assay

Wild-type cells and CRISPR-deletion clones were seeded into 6-well plates and grown to 100% confluency. We made straight scratches in middle of the well using 200-µl pipette tips and gently washed the well with culture media twice to remove free floating cells. Then, we imaged the scratch with Leica DMIL microscope and measured the width of the scratch using Leica Application Suite X software at the time of the scratch and 8 hours after the scratch.

Mouse xenograft experiment

All experiments were approved by the Institutional Animal Care and Use Committee of Washington University in St. Louis (Protocol #20170204) and conducted in accordance with the National Institutes of Health Guidelines for the Care and Use of Laboratory Animals. All experiments complied with the ethical regulations and considerations outlined in protocol. For H1299 xenografts, 3×10^6 wild-type cells or CRISPR-KO clones were collected and resuspended in 75 μl of chilled RPMI1640. Then, 75 μl of Matrigel Basement Membrane Matrix (Corning, 354234) was mixed into the cell solution and the sample was kept on ice until injection. The samples were injected subcutaneously into the right flank of four nude mice for WT and six nude mice for CRISPR KO clones (Jackson Lab, 002019, 4 weeks old homozygous NU/J females). Length (longer diameter) and width (shorter diameter) of the tumors were recorded and tumor volume was calculated by (Length × Width × Width)/2 equation.

CRISPR-SunTag vector construction

We obtained scFv-sfGFP-DNMT3A1 vector (Addgene, 102278) for targeted methylation vector. We purchased pHRdSV40-dCas9–10xGCN4_v4-P2A-BFP plasmid (Addgene, 60903) and pLKO5.sgRNA.EFS.tRFP657 plasmid (Addgene, 57824). For targeted demethylation, we replaced DNMT3A sequence with TET1 catalytic domain (CD) sequence, which was amplified from pPlatTET-gRNA2 plasmid (Addgene, 82559). Recent work revealed that dCas9-SunTag with 22aa linkers between GCN4 had higher demethylation efficiency[43]. In pHRdSV40-dCas9–10xGCN4_v4-P2A-BFP plasmid, we excised the 10xGCN4 sequence and cloned in GCN4–22aa sequence from pPlatTET-gRNA2 via Gibson Assembly. sgRNAs were cloned into pLKO5.sgRNA.EFS.tRFP657 plasmid.

Lentivirus production and transduction of CRISPR-SunTag vectors

HEK293T cells were seeded in 2 ml of DMEM complete media and grown to 50% confluency in 6-well plates. We co-transfected CRISPR-SunTag plasmids with pMD2.G and psPAX2 following polyethylenimine (PEI) transfection protocol. In brief, 6 μg of PEI and 2 μg of combined plasmids was added to 200 μl of Opti-MEM (ThermoFisher Scientific, 31985062) and incubated at room temperature for 30 minutes. The incubated PEI-vector mixture was added directly to HEK293T cells. After 48 hours, the viral supernatant was collected and filtered through 0.45-μm PES filter (Sigma-Aldrich, SLHV033RS). Then, polybrene (Sigma-Aldrich, TR-1003-G) was supplemented to the viral supernatant to a concentration of 5 μg/ml. The polybrene-viral supernatants of dCas9-SunTag-BFP, scFv-sfGFP-DNMT3A1/TET1CD and sgRNA.tRFP657 were added directly on top of H1299 and K562 cells in 6-well plates. After 2 days of incubation, the transduced cells were rinsed with PBS and analyzed by flow-cytometry (Beckman Coulter MoFlo) for BFP, GFP and farRFP657 fluorescence. Individual triple-positive fluorescent cells were sorted into 96-well plates and expanded. Once sufficiently expanded, the CRISPR-SunTag clones are resorted on the MoFlo for strong fluorescence and collected for downstream analysis of DNA methylation, gene expression and peptide expression.

Human LIN28B and AluJb-LIN28B rescue

We purchased pBABE-hLin28B plasmid (Addgene, 26358) that expresses FLAG-tagged human LIN28B protein[24]. We generated AluJb-LIN28B CDS from H1299 mRNA and cloned AluJb-LIN28B CDS in lieu of FLAG-hLIN28B sequence into the pBABE vector. We co-transfected pBABE plasmids with pMD2.G and pUMVC following polyethylenimine (PEI) transfection protocol into HEK293T cells. AluJb KO clones were transduced with viral supernatant supplemented with polybrene (5 μg/ml) for two days. Successfully infected cells were selected by 2 μg/ml puromycin (A.G. Scientific, P-1033-sol) treatment for 5 days before subsequent analysis.

Statistical analysis

Kaplan-Meier distributions between samples with or without candidate expression were compared using the logrank test. All statistics for in vitro experiments were performed using two-tailed Welch's *t* test. Enrichment for TE class was calculated with this formula: ((# of TE family onco-exapted / # of total TEs onco-exapted) / (# of total TE family / # of all TEs)).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank J. Hoisington-López and M.L. Jaeger from The Edison Family Center for Genome Sciences & Systems Biology (CGSSB) for assistance with sequencing; B. Koebbe and E. Martin from CGSSB for data processing; M. Savio, M. Patana & D. Schweppe from the Siteman Flow Cytometry core for FACS-related expertise; Dr. M. Goodell and Y. Huang for valuable expertise with the SunTag-DNMT3A system; Dr. L. Maggi (Washington University School of Medicine) for generously gifting the H838 cell line. We would like to thank anonymous reviewers for constructive and insightful suggestions. This work was funded by NIH grants 5R01HG007175, U24ES026699, U01HG009391 and the American Cancer Society Research Scholar grant RSG-14-049-01-DMC. HSJ was supported by a grant from NIGMS (T32 GM007067). NMS is a Howard Hughes Medical Institute (H.H.M.I.) Medical Research Fellow, http://www.hhmi.org/. The xenograft work cited in this publication was performed in a facility supported by NCRR grant C06 RR015502.

References:

1. Xie M et al. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. Nat. Genet 45, 836–841 (2013). [PubMed: 23708189]
2. Sundaram V et al. Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. 24, 1963–1976 (2014). [PubMed: 25319995]
3. Rebollo R, Romanish MT & Mager DL Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. Annu. Rev. Genet 46, 21–42 (2012). [PubMed: 22905872]
4. Babaian A & Mager DL Endogenous retroviral promoter exaptation in human cancer. Mob. DNA 7, 1–21 (2016). [PubMed: 26779288]
5. Botezatu A et al. Mechanisms of Oncogene Activation. in New Aspects in Molecular and Cellular Mechanisms of Human Carcinogenesis (ed. Bulgin D) 1–52 (InTech, 2016). doi:10.5772/61249
6. Pierotti MA, Sozzi G & Croce CM Mechanisms of oncogene activation in Holland-Frei Cancer Medicine (ed. Kufe DW, Pollock RE, Weichselbaum RR, et al.) (BC Decker, 2003).
7. Batut P, Dobin A, Plessy C, Carninci P & Gingeras TR High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. Genome Res. 23, 169–180 (2013). [PubMed: 22936248]

8. Chuong EB, Elde NC & Feschotte C Regulatory activities of transposable elements: From conflicts to benefits. Nat. Rev. Genet 18, 71–86 (2017). [PubMed: 27867194]

9. Chuong EB, Elde NC & Feschotte C Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science (80-.). (2016). doi:10.1126/science.aad5497

10. Hon GC et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. Genome Res. 22, 246–258 (2012). [PubMed: 22156296]

11. Baylin SB & Jones PA A decade of exploring the cancer epigenome — biological and translational implications. Nat. Rev. Cancer 11, 726–734 (2011). [PubMed: 21941284]

12. Esteller M Cancer epigenomics: DNA methylomes and histone-modification maps. Nat. Rev. Genet 8, 286–298 (2007). [PubMed: 17339880]

13. Babaian A et al. Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. Oncogene 35, 2542–2546 (2016). [PubMed: 26279299]

14. Lamprecht B et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat. Med 16, 571–579 (2010). [PubMed: 20436485]

15. Lock FE et al. A novel isoform of IL-33 revealed by screening for transposable element promoted genes in human colorectal cancer. PLoS One 12, 1–30 (2017).

16. Scarf I et al. Identification of a new subclass of ALK-negative ALCL expressing aberrant levels of ERBB4 transcripts. Blood 127, 221–233 (2016). [PubMed: 26463425]

17. Wiesner T et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature 526, 453–457 (2015). [PubMed: 26444240]

18. Wolff EM et al. Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet. 6, (2010).

19. Liu Y, Sun J & Zhao M ONGene: A literature-based database for human oncogenes. J. Genet. Genomics 44, 119–121 (2017). [PubMed: 28162959]

20. Raskin L et al. Transcriptome profiling identifies HMGA2 as a biomarker of melanoma progression and prognosis. J. Invest. Dermatol 133, 2585–2592 (2013). [PubMed: 23633021]

21. Zhang X, Yuan X, Zhu W, Qian H & Xu W SALL4: An emerging cancer biomarker and target. Cancer Letters (2015). doi:10.1016/j.canlet.2014.11.037

22. Wang T et al. Aberrant regulation of the LIN28A/LIN28B and let-7 loop in human malignant tumors and its effects on the hallmarks of cancer. Mol. Cancer 14, 125 (2015). [PubMed: 26123544]

23. Nguyen LH et al. Lin28b is sufficient to drive liver cancer and necessary for its maintenance in murine models. Cancer Cell 26, 248–261 (2014). [PubMed: 25117712]

24. Viswanathan SR et al. Lin28 promotes transformation and is associated with advanced human malignancies. Nat. Genet 41, 843–848 (2009). [PubMed: 19483683]

25. Babaian A et al. Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. Oncogene 35, 2542–2546 (2016). [PubMed: 26279299]

26. Forrest ARR et al. A promoter-level mammalian expression atlas. Nature 507, 462–470 (2014). [PubMed: 24670764]

27. Brocks D et al. DNMT and HDAC inhibitors induce cryptic transcription start sites encoded in long terminal repeats. Nat. Genet 49, 1052–1060 (2017). [PubMed: 28604729]

28. Shiraki T et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc. Natl. Acad. Sci 100, 15776–15781 (2003). [PubMed: 14663149]

29. Suzuki A et al. Aberrant transcriptional regulations in cancers: Genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. Nucleic Acids Res. 42, 13557–13572 (2014). [PubMed: 25378332]

30. Johnson CD et al. The let-7 microRNA represses cell proliferation pathways in human cells. Cancer Res. 67, 7713–7722 (2007). [PubMed: 17699775]

31. Newman MA, Thomson JM & Hammond SM Lin-28 interaction with the Let-7 precursor loop mediates regulated microRNA processing. RNA 14, 1539–1549 (2008). [PubMed: 18566191]

32. Zhou J, Ng SB & Chng WJ LIN28/LIN28B: An emerging oncogenic driver in cancer stem cells. Int. J. Biochem. Cell Biol 45, 973–978 (2013). [PubMed: 23420006]

33. Moqtaderi Z et al. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nat. Struct. Mol. Biol 17, 635–640 (2010). [PubMed: 20418883]
34. Rice P, Longden L & Bleasby A EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 16, 276–277 (2000). [PubMed: 10827456]
35. Hubley R et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 44, D81–D89 (2016). [PubMed: 26612867]
36. Grant CE, Bailey TL & Noble WS FIMO: Scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018 (2011). [PubMed: 21330290]
37. Guo W et al. A LIN28B Tumor-Specific Transcript in Cancer. Cell Rep. 22, 2094–2106 (2018). [PubMed: 29466736]
38. Beketaev I et al. cis-regulatory control of Mesp1 expression by YY1 and SP1 during mouse embryogenesis. Dev. Dyn 245, 379–387 (2016). [PubMed: 26384464]
39. Heo I et al. Lin28 Mediates the Terminal Uridylation of let-7 Precursor MicroRNA. Mol. Cell 32, 276–284 (2008). [PubMed: 18951094]
40. Viswanathan SR, Daley GQ & Gregory RI Selective blockade of microRNA processing by Lin28. Science (80-.). 320, 97–100 (2008).
41. Rybak A et al. A feedback loop comprising lin-28 and let-7 controls pre-let-7 maturation during neural stem-cell commitment. Nat. Cell Biol (2008). doi:10.1038/ncb1759
42. Tanenbaum ME, Gilbert LA, Qi LS, Weissman JS & Vale RD A protein-tagging system for signal amplification in gene expression and fluorescence imaging. Cell 159, 635–646 (2014). [PubMed: 25307933]
43. Morita S et al. Targeted DNA demethylation in vivo using dCas9-peptide repeat and scFv-TET1 catalytic domain fusions. Nat. Biotechnol 34, 1060–1065 (2016). [PubMed: 27571369]
44. Huang YH et al. DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. Genome Biol. 18, 1–11 (2017). [PubMed: 28077169]

Methods References:

45. Harrow J et al. GENCODE: The reference human genome annotation for the ENCODE project. Genome Res. 22, 1760–1774 (2012). [PubMed: 22955987]
46. Tarailo-Graovac M & Chen N Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma 25, 1–14 (2009).
47. Karolchik D The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 32, D493–496 (2004). [PubMed: 14681465]
48. Pertea M et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol 33, 290–295 (2015). [PubMed: 25690850]
49. Kang YJ et al. CPC2: A fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 45, W12–W16 (2017). [PubMed: 28521017]
50. Dunham I et al. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]
51. Krueger F & Andrews SR Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572 (2011). [PubMed: 21493656]
52. Corces MR et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat. Methods 14, 959–962 (2017). [PubMed: 28846090]
53. Bayat A, Gaëta B, Ignjatovic A & Parameswaran S Improved VCF normalization for accurate VCF comparison. Bioinformatics 33, 964–970 (2017). [PubMed: 27993787]
54. Salimullah M, Mizuho S, Plessy C & Carninci P NanoCAGE: A high-resolution technique to discover and interrogate cell transcriptomes. Cold Spring Harb. Protoc 6, 96–111 (2011).
55. Haberle V, Forrest ARR, Hayashizaki Y, Carninci P & Lenhard B CAGEr: Precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. 43, (2015).

56. Zhou X et al. The human epigenome browser at Washington University. Nature Methods (2011). doi:10.1038/nmeth.1772

57. Wang X Primer sequences for 96 cancer-related miRNA assays. RNA 15, 716–723 (2009). [PubMed: 19218553]

58. Haeussler M et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. 17, 1–12 (2016). [PubMed: 26753840]

59. Moreno-Mateos MA et al. CRISPRscan: Designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat. Methods 12, 982–988 (2015). [PubMed: 26322839]

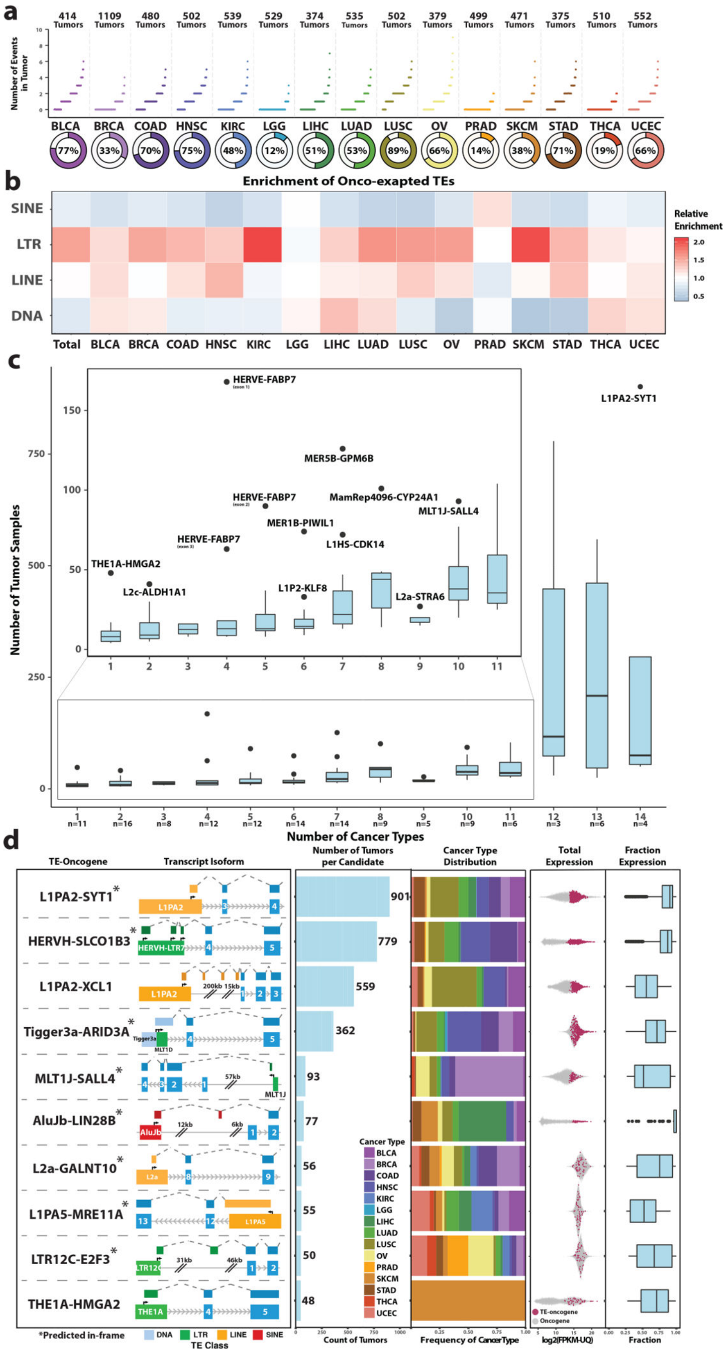

Fig. 1: The TE onco-exaptation landscape across cancer types.

a, Frequency of onco-exaptation events per tumor across cancer types. Donut plot reports the percent of tumor samples with at least one event. **b,** Enrichment of TE class in onco-exapted TEs across cancer types. **c,** A series of boxplots that highlight the distribution of the total number of tumor samples per candidate that is present in a certain number of cancer types. We have zoomed in on 1–11 so that the distribution can be more clearly seen. Each box represents the median and interquartile range, and the whiskers are 1.5× the IQR. Below each boxplot, we have labeled the number of candidates. We have also labeled all the outlier candidates. **d,** The top 10 most prevalent onco-exaptation candidates are presented. The left-most panel gives the TE-oncogene candidate label as well as a diagram of the transcript structure of the candidate. The next two panels display the number of tumor samples each candidate is present in as well as the distribution of the candidate across cancer types. The “Total Expression” panel displays the expression of the oncogene across all the tumor samples as grey dots, and the samples with the onco-exaptation candidate are highlighted in red. The “Fraction Expression” panel displays a boxplot of the percent of total expression of the oncogene contributed by the onco-exaptation candidate across the samples in which the candidate is present. Each box represents the median and interquartile range, and the whiskers are 1.5× the IQR.

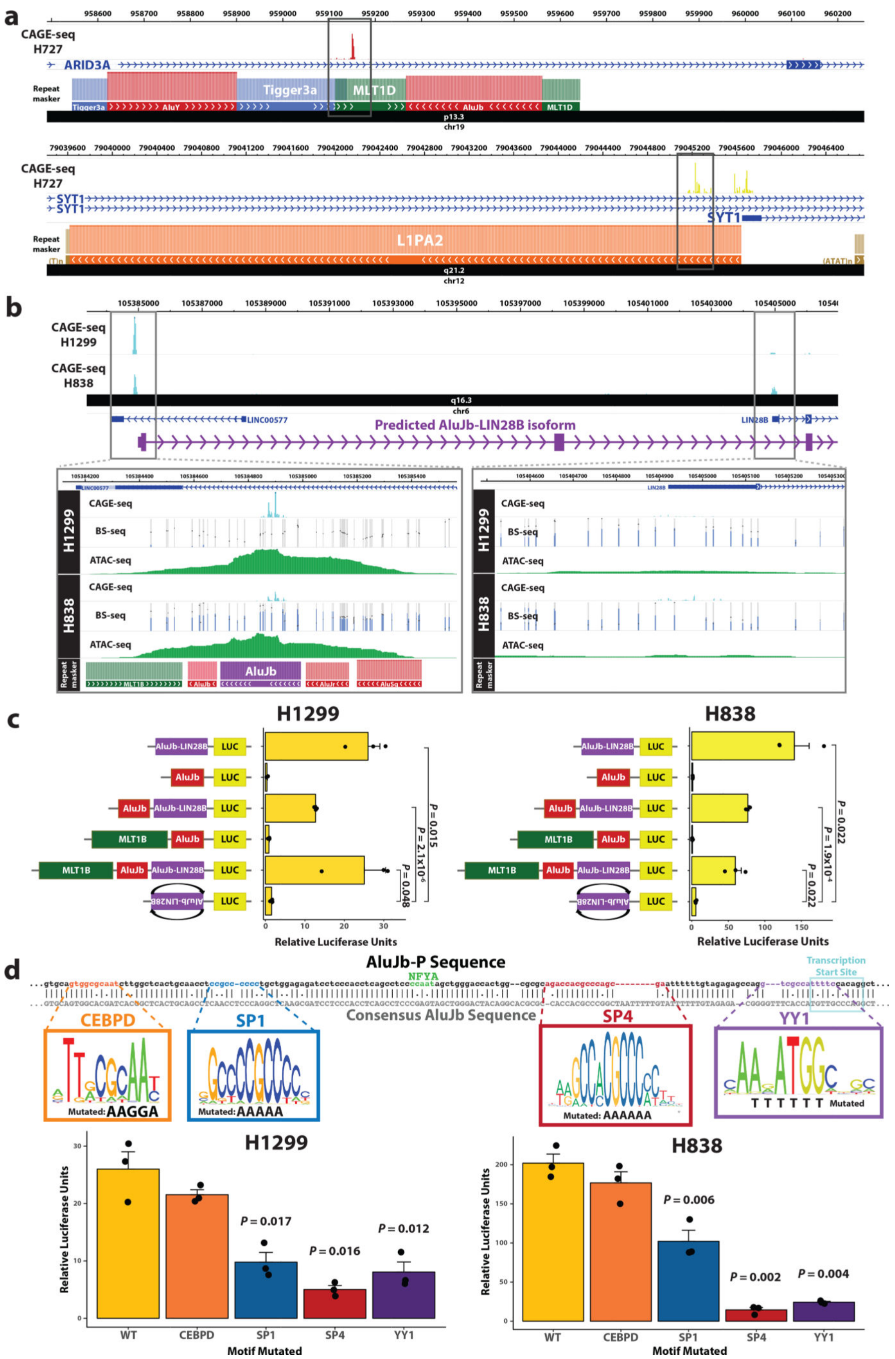

Fig. 2: TEs provide *bona fide* promoters for oncogenes in lung cancer cell lines.

a, CAGE-seq profile of H727 across onco-exaptation candidates (*ARID3A* & *SYT1*) visualized on WashU Epigenome Browser. Signals in CAGE-seq represent TSS locations. **b,** CAGE-seq and epigenetic profiles of the AluJb TE in the H1299 and H838. Signal in ATAC-seq represent open chromatin regions. Grey bars in the BS-seq track represent CpG locations while the height of blue bars indicate methylation %. **c,** Luciferase assays for transcriptional activity of various TE arrangements in H1299 (left) and H838 (right) (n = 3 independent experiments). **d,** Luciferase assays for promoter activity in H1299 (left) and H838 (right) with mutagenized transcription factor motifs in AluJb-P (n = 3 independent experiments). **c, d,** *P* values were derived from two-tailed Welch *t* test. All data are represented as means ± standard error (SE).

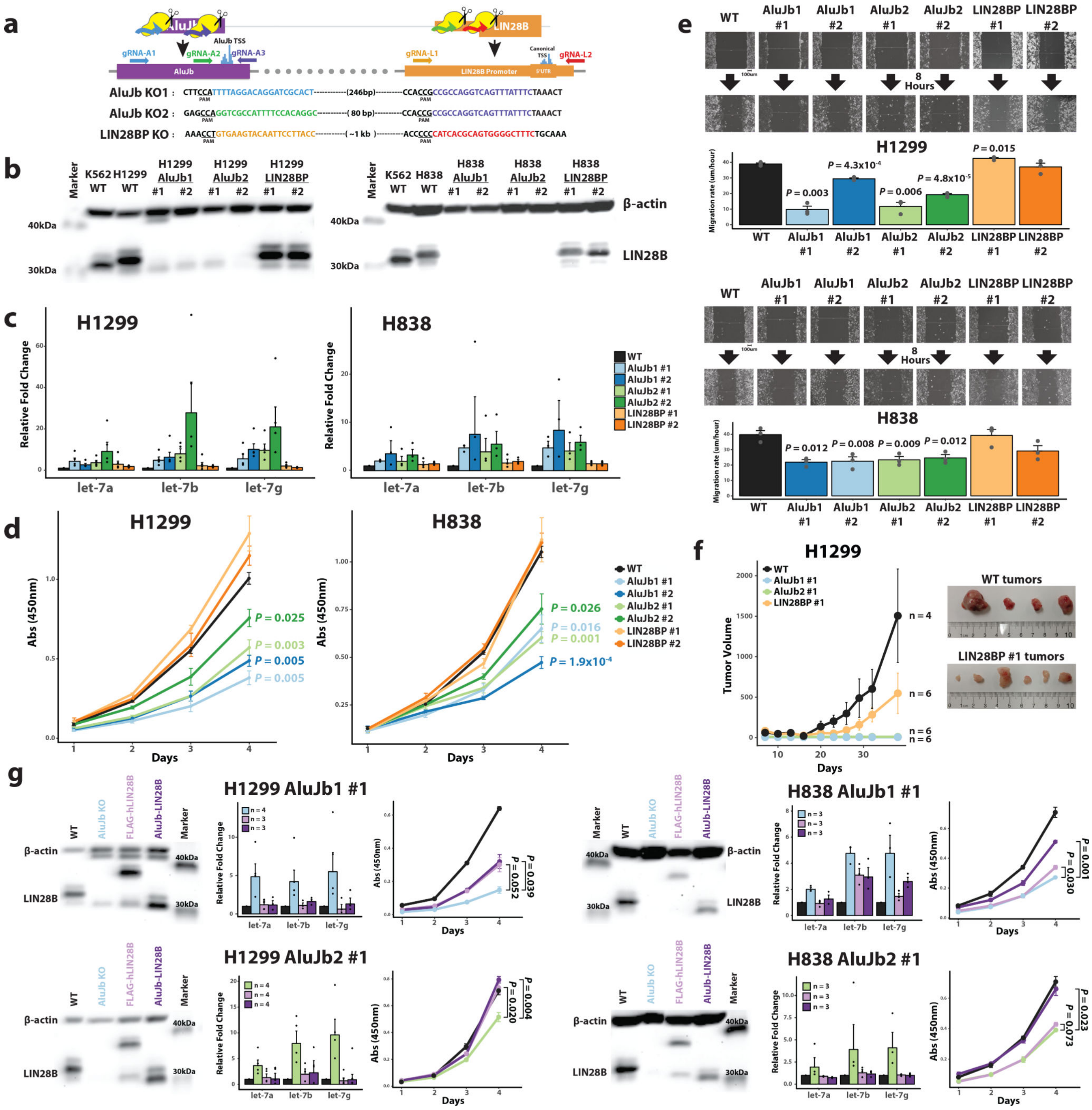

Fig. 3. AluJb drives LIN28B expression and contributes to oncogenesis in lung cancer cell lines.

a, Schematic describing sgRNA locations and sequence targets within AluJb-P and LIN28BP. **b,** Cropped Western blot for LIN28B protein in H1299 (top) and H838 (bottom) CRISPR clones. This experiment was repeated twice with similar results. **c,** Relative let-7a, let-7b, and let-7g miRNA levels compared to WT in CRISPR-knockout clones of H1299 (n = 4 independent experiments) and H838 (n = 3 independent experiments) as measured by qPCR. **d,** The effect of AluJb-P or LIN28BP deletion on cell growth rate as determined by CCK-8 assay in H1299 and H838 cells (n = 3 independent experiments). **e,** The effect of

AluJb-P or LIN28BP deletion on cell migration in H1299 (top) and H838 (bottom) as measured by scratch migration assay (n = 3 independent experiments). **f,** Tumor growth of H1299 WT and H1299 CRISPR-knockout clones injected in nude mouse. Resected tumors of WT and LIN28BP #1 xenografts. **g,** Cropped Western blot (repeated twice with similar results) of re-expression of human FLAG-LIN28B or AluJb-LIN28B in AluJb KO clones and its effect on relative let-7 miRNA levels (number of independent experiments indicated in figure as n) and growth rate (n = 3 independent experiments). **d,e,g,** *P* values from CCK-8 growth assays and scratch migration assays were derived from comparing to WT with two-tailed Welch *t* test. All data are represented as means ± SE.

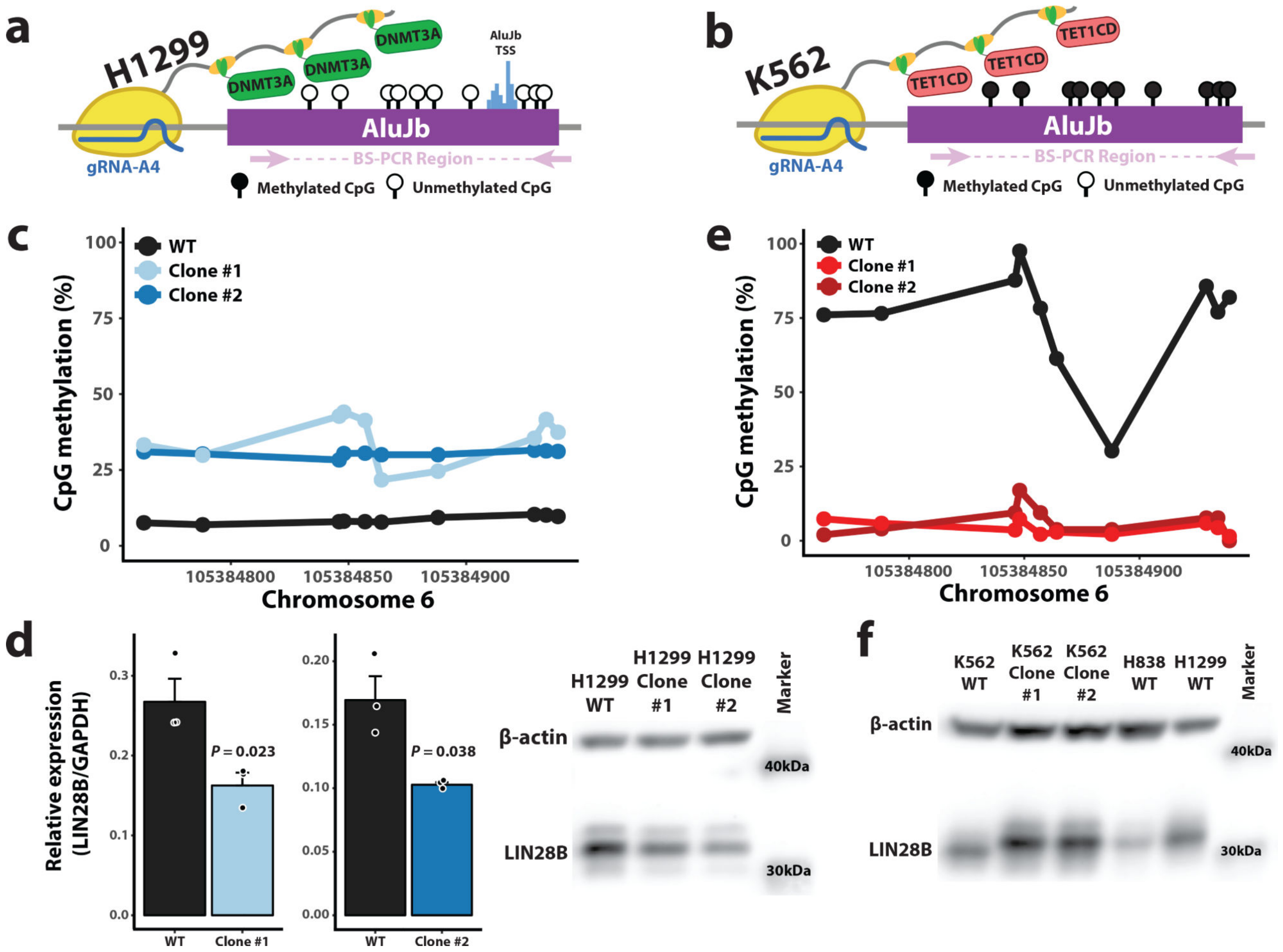

Fig. 4. Targeted DNA methylation dynamics uncover epigenetic control of AluJb promoter activity.

a Schematics illustrating CRISPR-SunTag models for targeted de/methylation of AluJb. DNMT3A was recruited to AluJb loci in H1299 to increase methylation. **b,** TET1CD was recruited to AluJb in K562 to remove DNA methylation from the TE. **c,** Methylation levels of AluJb in WT and CRISPR-SunTag-DNMT3A clones of H1299 measured by BSPCR-seq. **d,** Relative abundance of LIN28B in H1299 CRISPR-SunTag-DNMT3A Clone #1 (left) and Clone #2 (right) compared to WT as measured by qPCR (n = 3 independent experiments) and cropped Western blot (repeated twice with similar results). *P* values were derived from two-tailed Welch *t* test. All data are represented as means ± SE. **e,** Methylation levels of AluJb in WT and CRISPR-SunTag-TET1CD clones of K562. **f,** Cropped Western blot (repeated twice with similar results) illustrating the presence of larger LIN28B protein, similar size as AluJb-LIN28B in H1299 and H838, in K562 CRISPR-SunTag-TET1CD clones.