BMC Cancer

RESEARCH

Open Access

Long-read RNA sequencing enables full-length chimeric transcript annotation of transposable elements in lung adenocarcinoma

Yang Li[1], Yahui Liu[1], Yingxin Xie[2], Yaxuan Wang[3], Jing Wang[1], Huan Wang[1], Lin Xia[1*] and Dan Xie[1*]

Abstract

Background Transposable elements (TEs), which constitute nearly half of the human genome, have long been regarded as genomic "dark matter". However, their reactivation in tumor cells, resulting in the production of TE-chimeric transcripts (TCTs), has emerged as a potential driver of cancer progression. The complexity and full extent of these transcripts remain elusive, largely due to the limitations of short-read next-generation sequencing technologies. These methods have struggled to comprehensively capture the diversity and structure of TCTs, particularly those involving short interspersed nuclear elements (SINEs) or closely co-transcribed TEs.

Methods Leveraging full-length cDNA sequencing technology based on nanopore sequencing platform, we developed a customized pipeline for identifying and quantifying TCTs in 19 lung adenocarcinoma (LUAD) cell lines. The short-read RNA-seq dataset from a LUAD corhort (~ 200 tumor samples) was employed to validate the identified TCTs and explore their association with tumor progression. To assess the functional roles of a specific TCTs, cell migration and cell proliferation assays were performed.

Results We uncovered 208 unique TCT candidates in the LUAD cell lines. Our approach allowed for the identification of cryptic promoters and terminators within non-transposing TEs. Notably, we identified a chimeric transcript involving *MIR_HKDC1*, which appears to play a significant role in the progression of LUAD. Furthermore, the expression of these TCTs were associated with poor clinical outcomes in a cohort of LUAD patients, suggesting their potential as novel biomarkers for both LUAD progression and prognosis.

Conclusions Our study underscores the application of long-read sequencing to unravel the complex landscape of TCTs in LUAD. We provide a comprehensive characterization of TCTs in LUAD, exploring their potential regulatory roles in cancer progression. These findings contribute to a deeper understanding of the genomic intricacies underlying cancer, and offer new directions for the development of targeted therapies and personalized treatment strategies for LUAD. This research highlights the potential of TCTs as both biomarkers and therapeutic targets in the oncogenesis, offering new insights into the interplay between transposable elements and gene regulation in cancer.

Keywords Transposable elements, Long-read RNA sequencing, Lung adenocarcinoma

*Correspondence:
Lin Xia
xialin.phd@wchscu.edu.cn
Dan Xie
danxie@scu.edu.cn
Full list of author information is available at the end of the article

Background

Transposable elements (TEs) are mobile genetic sequences that constitute nearly half of the human genome [1]. Although TEs are generally epigenetically silenced in most tissues, a growing number of studies have shown that TE reactivation and insertion can cause human genetic disease [2]. Previous studies have demonstrated that TEs can lead to various universal alternative splicing events [3] or become alternative promoters especially in lung adenocarcinoma (LUAD) cells, where their aberrant expression is linked to tumor progression and metastasis. Two recent pancancer studies of TE expression demonstrated that chimeric TE-gene transcripts exist in the vast majority of tumors, and a list of shared tumor specific TE chimeric antigens (TS-TEAs) [3, 4] has promising therapeutic value for creating pancancer vaccines.

However, two limitations of short-read-based studies have prevented us from obtaining complete structures of TE-gene transcripts. First, the high copy number of TE sequences in the human genome leads to creates problems with multiple alignments of short reads, which poses a challenge for accurately identifying expressed TEs and novel junctions they constituted [5]. With methodological advancements, recent studies have led to a significant increase in the number of defined aberrantly expressed short interspersed nuclear elements (SINEs) [6]. However, the majority of short-read sequencing studies continue to focus primarily on the function of long TE chimeric transcripts in tumors, such as LINEs [7–9] and LTRs [10, 11], in tumors, often overlooking the function of short TEs, such as SINEs [12, 13], and the recombination of TEs [14]. Second, short reads are less than 1 kb long, preventing full-length transcripts from being sequenced, and complex algorithms are needed to accomplish transcript assembly for novel transcript identification. Therefore, short-read transcript assembly results contain more transcriptional noise than long-read transcript assembly [15, 16].

Long-read sequencing, a recent technological breakthrough capable of generating reads spanning millions of bases [17], offers a transformative approach to studying transposable elements (TEs) and their adjacent genomic regions. This advancement addresses longstanding challenges associated with accurately locating TEs within the genome, particularly considering the variability in TE copy numbers and lengths. Furthermore, long-read data enable the comprehensive capture of full-length transcripts [18], allowing a nuanced exploration of isoform diversity and the detection of novel transcripts that surpass the capabilities of short-read sequencing [19, 20]. Long-read RNA sequencing (lrRNA-seq) has been used in several studies to investigate TE expression across diverse species, including Arabidopsis [21], Drosophila [22] and locusts [23]. Notably, despite these advancements, the exploration of TE-chimeric transcripts (TCTs) has not been well studied in LUAD using lrRNA-seq data.

In this study, we developed a custom computational pipeline for detecting long-read assembled full-length TCTs. To study TCTs in LUAD, we first constructed a reference transcriptome for LUAD cells, which contained high confidence and full-length TCTs, using long-read RNA-seq data. Using this annotation, we assessed the role of these TCTs in promoting LUAD tumorigenesis. Furthermore, we applied this new annotation strategy to investigate the association between TCTs and clinical outcomes in LUAD patients.

Methods

RNA isolation and generation of long-read RNA-seq data

The RNA extraction protocol involved tissue grinding and lysis, where TRIzol was used as the lysis solution. After centrifugation, the upper aqueous phase containing RNA was separated. Further purification was achieved by adding chloroform, followed by additional centrifugation steps. The RNA was then precipitated with isopropanol, washed with ethanol, and centrifuged to obtain a purified RNA pellet. The final step involved dissolving the RNA precipitate in RNase-free water. This method ensures the isolation of high-quality RNA for downstream applications.

Oxford PromethION full-length transcript libraries were generated according to the Nanopore community protocol using the SQK-LSK109 library preparation kit and sequenced on R9 flow cells. For base calling of the raw data, we used Guppy (v.3.2.10) with the default parameters.

TCT identification

Alignment of long-read RNA-seq data

We employed minimap2 v2.2.17 [24] for read alignment against the GRCh38 reference human genome using the parameters -ax splice -secondary = no.

Assembly and annotation of transcripts

Transcript assembly was performed for each cell line using FLAIR v1.7.0 [25]. FLAIR-*correct* was applied to rectify misaligned splice sites using GENCODE v.42 annotations. FLAIR-*collapse* was executed per sample, generating high-confidence transcript sets supported by at least ten reads and MAPQ > 10.

A custom pipeline was used to annotate transcripts overlapping with the Repeatmasker TEs and GENCODE v.42 transcripts. Candidate transcripts that supported by a minimum of 5 reads spanning both TEs and adjacent

exons of target genes, were selected as TCTs for further analysis.

Generating a reference transcriptome including TCTs

The StringTie *merge function was* subsequently used to create a reference transcriptome that included the GENCODE v.42 transcripts as well as all TE-chimeric events that met our filtering criteria.

Transcript-level quantification and candidate selection

To determine the contribution of TCTs to overall gene expression, we used the StringTie *quantification* (-e -b) function with the merged transcriptome as the reference. The FPKM values generated by this command were extracted from the ballgown output files to obtain transcript-level expression.

Open-reading frame prediction and annotation

We used TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder/) to predict the coding reading frames of the TCTs and generate a FASTA file with potential transcript products. For in-frame proteins, we determined whether the start codon was predicted to make a protein in one of the following categories: normal, truncated, chimeric normal, frame shift or chimeric truncated.

Analysis of short-read RNA sequencing data

We downloaded 2 short-read RNA-seq datasets (GSE40419, GSE37765). Paired-end reads were aligned to the GRCh38 human genome reference by HISAT2 (v2.2.1). We used StringTie (v.2.1.7) to quantify the expression of each transcript with the parameter '-b -e', using the merged transcriptome, which contained both the Gencode V42 and TCT results, as a reference. The FPKM values generated by this command were extracted from the ballgown output files to obtain transcript-level expression. To assess statistical differences between tumor and paired normal samples, we employed a Linear Model Fit tool (lmFit from R package limma). Transcripts were considered significantly differentially expressed if their absolute log2-fold change (|log2(FC)|) was greater than 1 and the false discovery rate (FDR) was less than 0.05. Functional enrichment and gene set variation analyses were performed using the DAVID database.

Cellular culture

In this investigation, human cell lines procured from the American Type Culture Collection (Manassas, VA, USA) were used. The cells were incubated in Dulbecco's modified Eagle's medium (DMEM) or RPMI 1640 medium (Gibco; Thermo Fisher Scientific, Inc.). The culture medium was enriched with 10% fetal bovine serum (FBS) (Gibco; Thermo Fisher Scientific, Inc.), and the cells were maintained at a consistent temperature of 37 °C in an atmosphere containing 5% CO_2. Upon reaching 70–80% confluence, the cells were washed with phosphate-buffered saline (PBS) and subsequently detached from the plates using 0.25% trypsin/0.2% ethylenediaminetetraacetic acid (EDTA). The morphology of the cells was assessed utilizing a light microscope, and cells were resuspended to a density of 1×10^6 cells/mL.

Quantitative real-time polymerase chain reaction (qPCR)

Total cellular RNA was extracted from a panel of PC-9, A549, H1299, H1975 and Beas-2B cell lines. The extraction process was executed using a high-performance RNeasy Mini Kit (QIAGEN). The RNA integrity was validated through agarose gel electrophoresis, followed by the conversion of RNA into complementary DNA (cDNA) through the use of a sophisticated iScript cDNA Synthesis Kit (Bio-Rad). SYBR Green Supermix (Bio-Rad) was used for real-time quantitative PCR. The reaction conditions and PCR system used were in accordance with the instructions. All nucleotide sequences were designed and synthesized by TSINGKE (Chengdu, China). These sequences are comprehensively cataloged in the oligonucleotide table within the Key Resources section, including the GAPDH forward primer (AGATCCCTCCAA AATCAAGTGG). Target mRNA levels were measured using the 2-ΔΔCt method.

Cellular lipid transfection

Beas-2B cells, characterized by low HKDC1 expression, were seeded onto coverslips in 6-well plates at a density of 40×10^4 cells/well, and 80–90% monolayer confluence was achieved after a minimum 24-h culture period prior to transfection. Following incubation, the specimens were washed with sterile PBS, and 500 μL of serum-free DMEM containing 2 μg of MIRb-HKDC1-pcDNA3.1, HKDC1-pcDNA3.1 overexpression plasmid, empty vector pcDNA3.1, or lip3000 (Invitrogen) was added to each well. The cells were maintained in a serum-free environment for 4 h, followed by a PBS wash and subsequent addition of 1.5 mL of complete DMEM to each well.

Scratch wound healing assay

Cell migration was assessed utilizing a wound healing scratch assay. Beas-2B cells were seeded in 6-well plates (1×10^6/well) in culture medium and cultured until they reached confluence. Subsequently, the cells were treated with MIRb_HKDC1-pcDNA3.1, HKDC1-pcDNA3.1, pcDNA3.1, and lip3000. A controlled artificial scratch wound was introduced at the center of the well and photographed. After 24 h of incubation, the scratch wound was photographed again, and the migration distance was

quantified by the ratio of the healing width at 24 h relative to the initial wound width at 0 and 12 h. Each assay was conducted in triplicate and replicated three times.

Cell counting Kit 8 (CCK-8) assay

Cell viability was assessed using the Cell Counting Kit 8 (CCK-8) Assay. Beas-2B cells were seeded in 96-well plates at 5000 cells per well. Subsequently, the cells were cultured in a 37 °C, 5% CO_2 incubator for 24 h, 48 h, 72 h, or 96 h. After each time point, 10 μL of CCK-8 reagent was added to each well, and the cells were incubated for 1 h. The absorbance was measured at 450 nm using a microplate reader (BioTek Epoch2). Each experimental group was subjected to the assay three times.

RNA extraction and RT-PCR

RNA extraction (TRIzol, Invitrogen) and RT–PCR (Code No. RR047APrimeScript™/Code No. RR901APremix Taq™, Takara) were performed following the protocol provided by the kit. Primers were designed to bind regions between insertions and adjacent exons, and the primer sequences can be found in Table S6. The PCR conditions included an initial step of 2 min at 98 °C, followed by 30 cycles of 10 s at 98 °C, 15 s at 60 °C, and 60 s at 72 °C. The final extension step involved incubation at 72 °C for 2 min. To confirm that the bands detected in the PCR assay were the predicted insertions, we purified (Invitrogen) and sequenced the PCR products (Tsingke, Beijing, China). All experiments were performed in triplicate.

Statistical analysis

We used the Wilcoxon rank-sum test to determine significance of differences between two groups. Multiple testing correction was performed using the Benjamini and Hochberg method. A p-value < 0.05 was considered to indicate statistical significance.

Results

Identification of TE-chimeric transcripts in non-small cell lung cancer cell lines

The long-reads derived from the Oxford Nanopore sequencing platform have the capability to span the entire transposable element (TE) sequence, as well as its adjacent sequences, thereby facilitating enhanced precision in the localization of expressed TEs. This, in turn, enables a more refined elucidation of the structural attributes of TCTs. To globally characterize TCTs across non-small cell lung cancer (NSCLC) cell lines, we generated long-read RNA sequencing (lrRNA-seq) data for A549 cells and collected previously published lrRNA-seq data from 18 NSCLC cell lines [26] (Table S1) to obtain approximately 73.8 million long reads from these 19 samples (mean N50 of 2,197 bp), with an average mapping rate of 77.89% in the BAM files (Table S1).

We developed a computational pipeline to detect TCTs (Fig. 1A), incorporating full-length long reads to enhance the accuracy of TCT detection. First, we de novo-assembled the transcriptome using full-length long reads detected by pychopper (https://github.com/epi2me-labs/pychopper) using FLAIR [25]. Then, FLAIR *correct* was used to correct splice sites using the GENCODE V42 genome annotation. FLAIR *collapse* was used to define high-confidence transcripts from the corrected reads. In total, we identified 129,100 transcripts, with a prevalence ranging from 2,927 to 9,024 (Table S2, Fig. 1B), and 81.63% of the assembled transcripts were multi-exon. Compared to GENCODE version 42, 25.71% of the transcripts were non-GENCODE-annotated transcripts (Table S3, Figure S1A). We restricted our downstream analysis to novel transcripts harboring TE sequences supported by at least 5 full-length reads. In total, we identified 302 TCTs across all cell lines, with a prevalence ranging from 2 to 58. Only 5 TCTs involved single-exon (Figure S1B). These TCTs involved 188 genes, including 86.75% protein-coding genes, 10.93% lncRNAs, and 2.32% pseudogenes. A total of 19 genes had TCTs in at least two cell lines, involving 147 TCTs (Fig. 1C). One example is the oncogene *RING1*. We found that *RING1* had the TCT *AluJb_RING1* across 3 cell lines. *RING1* is a protein that can bind DNA and act as a transcriptional repressor. This suggested that a conserved mechanism of TE expression could contribute to the oncogenesis and progression of LUAD, though further validation is needed.

We merged TCTs identified across all cell lines using the StringTie *merge* to improve the robustness of the detected TCTs and improve the confidence in their characterization. Finally, we found 378 expressed TEs that created 208 TCTs, with 97.6% of them being multi-exon transcripts. Enrichment analysis identified RNA-binding and protein-binding genes as being enriched in these expressed TE-related genes (Figure S1C). The majority of expressed TEs (95.13%) were derived from intronic regions. This result may be due to the use of FLAIR *correct* (based on Gencode V42) for splice site correction, filtering out isoforms with large differences in intron chains. The expressed TEs contribute to the generation of alternative first exons (31.93%), cryptic alternative splice sites (2.37%), and alternative last exons (65.70%) of host genes (Fig. 1D). Most TE terminators appeared to be associated with intron retention caused by TE expression, potentially leading to gene truncation.

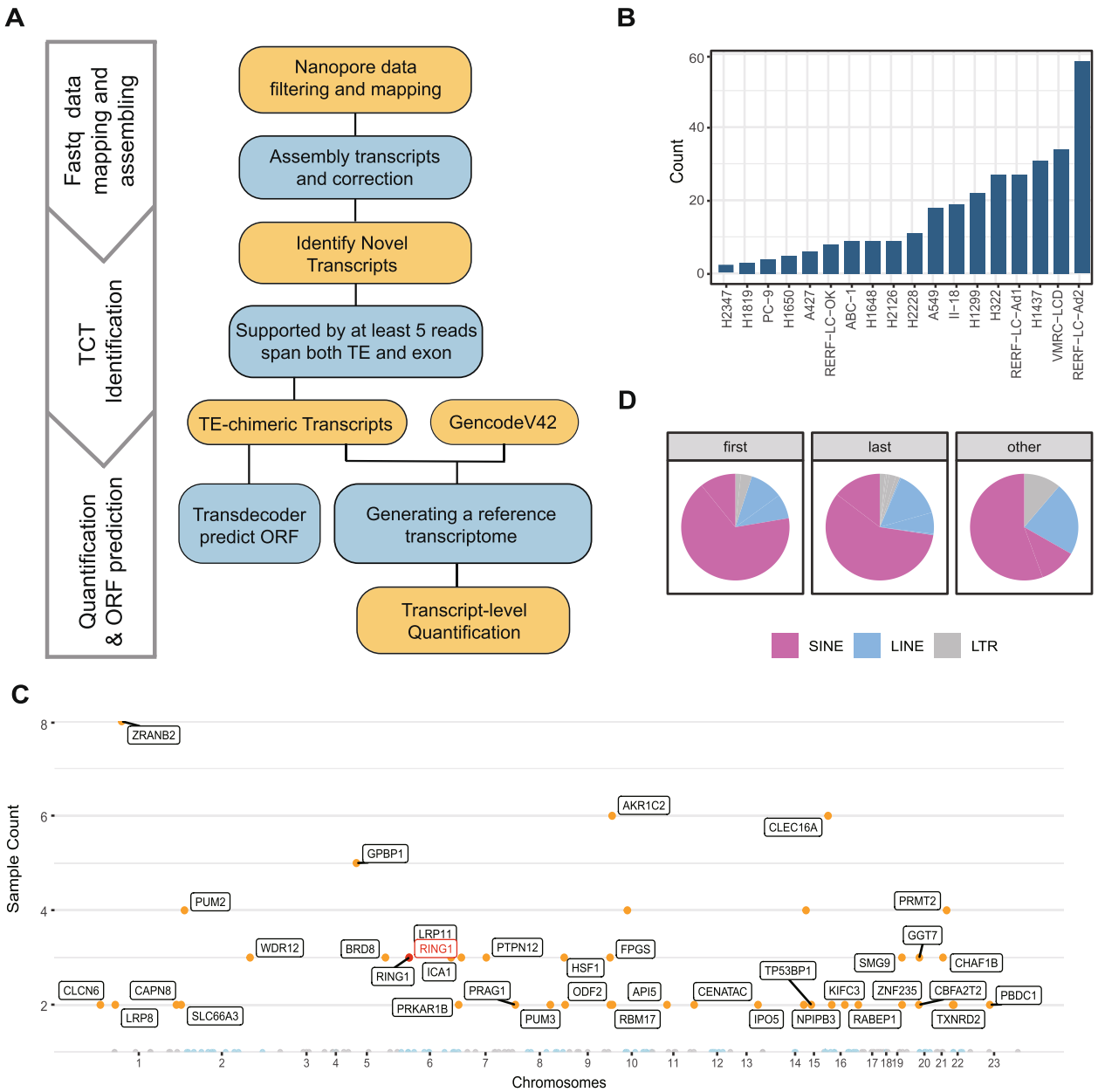

Fig. 1 Basic statistics of TE-chimeric transcripts defined with lrRNA-seq data. **A** The flowchart illustrates the computational pipeline for detecting TE-chimeric transcripts (TCTs), with each step outlined in detail in the Methods section. **B** Barplot shows the count of TCTs identified with lrRNA-seq data from each cell line. **C** The Manhattan plot illustrates the number of cell lines sharing TCTs on the y-axis, with chromosomes plotted on the x-axis. **D** The pie chart illustrates the distribution of different TE classes expressed. Three panels represents TE expression in the alternative first exon, alternative last exon and cryptic alternative splice sites, respectively

Prevalence of SINE-chimeric transcripts

We further characterized the TE sequences expressed in TCTs, identifying expression events from four major classes of transposable elements: short interspersed nuclear elements (SINEs), long interspersed nuclear elements (LINEs), long terminal repeat elements (LTRs), and DNA elements (DNA) (Table S4). Notably, in contrast to previous studies that reported LINEs as the dominant TE class [4], we found that the majority of TCT-associated TE sequence are SINEs (73.8%) (Fig. 2A and S2A). The proportion of SINEs in the expressed TE sequences identified by lrRNA-seq was more than double that observed in the srRNA-seq results (Fig. 2B). Long-read data provide evidence

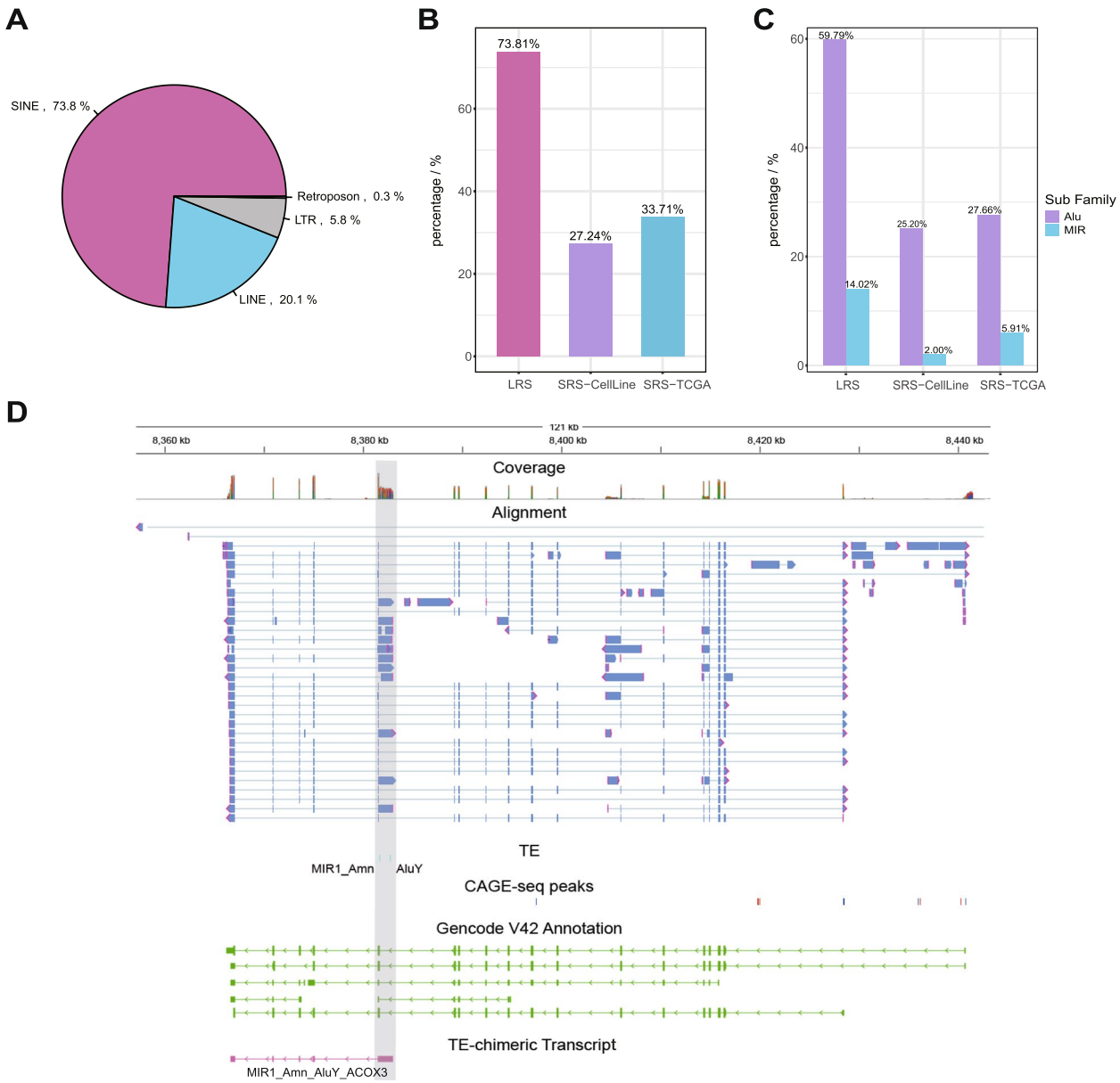

Fig. 2 TE classification of expressed TCT and comparison with SRS-data. **A** The pie chart illustrates the distribution of TE families contributing to TCTs as identified in our study. **B** The proportion of SINE-chimeric transcripts found across different studies. "LRS" indicates the proportion of SINE-chimeric transcripts detected in our research. "SRS-CellLine" indicates the percentage of SINE-chimeric transcripts discovered by Nakul M Shah et al. across 675 cancer cell lines using short-read sequencing data. "SRS-TCGA" indicates the presence of SINE-chimeric transcripts reported by Nakul M Shah et al. within 10,357 TCGA samples using short-read sequencing data. **C** The proportion of Alu-chimeric transcripts and MIR-chimeric transcripts found across different studies. **D** lrRNA-seq full-length reads span the entire TCT, enabling the detection of adjacent transcription between two neighboring TEs. The first track shows the coverage of genome. The second tracks shows lrRNA-seq full-length reads. The third track shows the location of TEs. The forth track shows the location of CAGE-seq peaks. The fifth track shows gene annotation contains transcripts of *ACOX3* from Gencode V42. The last trach shows the assembled TCT MIR1_Amn_AluY_ACOX3 (MSTRG.139.1)

supporting the widespread expression of SINEs in NSCLC.

When examining the subfamily proportions, we observed a substantially greater representation of Alu and MIR within the expressed TEs quantified by lrRNA-seq data than within those quantified by srRNA-seq data [4] (Fig. 2C). The proportions of AluJ, AluS and MIR in the lrRNA-seq-identified expressed TEs were approximately 1.46 times, 1.41 times and 1.6 times greater, respectively, than those observed in the srRNA-seq results. These

findings suggested that lrRNA-seq could potentially offer a more accurate quantification of SINE expression, which is often challenging to capture fully with short-read sequencing due to the high copy numbers and shorter sequences of SINEs compared to other TEs.

Furthermore, 56.3% of the TE-chimeric gene expression events detected via the lrRNA-seq data were consistent with previously reported findings from the srRNA-seq TCGA data analysis (Figure S2B), suggesting potential differences in sequencing sensitivity or alignment efficiency between the two platforms. Subsequently, we conducted a detailed examination of the characteristics of TCT sequences identified exclusively through lrRNA-seq data but overlooked in srRNA-seq analysis. Our analysis revealed that all 86 TCTs uniquely identified by lrRNA-seq involved adjacent pairs of contiguous TE sequences within the genome that were coexpressed within the same transcript. These coexpressed TE sequences, due to their extended length and high sequence repetition, may have posed challenges for accurate alignment with srRNA-seq, potentially leading to their underrepresentation in srRNA-seq analysis.

For instance, a previous study revealed four types of TCTs of the *ACOX3* gene (*AluY_ACOX3*, *MLT1D_ACOX3*, and two different types of *HERV3-int_ACOX3*; Table S4). Our lrRNA-seq data confirmed the presence of the TE-expressed event, *AluY_ACOX3,* and additionally provided a more complete transcript structure for this event. Upon examining this complete transcript structure, we observed that the AluY element was coexpressed with an adjacent MIR element (separated by 850 bp from AluY) within the same transcript of ACOX3 (Fig. 2D, Table S4), suggesting a potential interaction between these elements. These results suggested that lrRNA-seq may provide a more comprehensive view of the diversity of TCTs and provide insights into the characteristics of expressed TEs that might be underrepresented or overlooked by srRNA-seq data, highlighting the potential advantages of lrRNA-seq for capturing complex TCT structures.

Abnormal activation of the cryptic promoter within the TE in LUAD

Previous research has highlighted the potential of transposable elements (TEs) as cryptic promoters [4]. By querying the FANTOM5 promoter database, we identified 19 expressed TEs that were associated with the first exon of the corresponding TCT and overlapped with promoter signals (Table S5). Notably, our list of TEs capable of acting as cryptic promoters included 3 TE promoters identified in TCGA tumors using srRNA-seq data (*AluSp_SP1*, *AluSc8_ETAA1*, and *MLT1K_BCAS1*, Fig. 3C and 3D, Figure S3C) (Table S5) [4]. Among these, only *AluSp_SP1* and *AluSc8_ETAA1* were expressed as major transcript variants of *SP1* and *ETAA1* (the transcript accounted for at least 25% of total gene expression), respectively, in more than 10 LUAD cell lines (Figs. 3A, 3B, S3A and S3B).

The transcription factor SP1, encoded by the *SP1 gene*, plays a crucial role in regulating the expression of genes involved in various cellular processes, including cell growth [27]. Using lrRNA-seq data, which provides full-length transcripts, we were able to delineate the complete structure of the *AluSp_SP1* TCT (Fig. 3C). The AluSp element is located in an intron between exon 1 and exon 2 of the *SP1* gene and overlapped with a CAGE-seq peak representing a promoter signal. In the *AluSp_SP1* TCT, the AluSp element and its proximal downstream region are co-transcribed as the first exon of the TCT. A previous study indicated that *AluSp_SP1* is ubiquitously expressed in tumor samples [4], with 4246/10,357 TCGA tumor samples containing *AluSp_SP1*, while only 357/729 TCGA normal samples exhibited this TCT. Among these, 205 TCGA LUAD tumors and 233 TCGA LUSC tumors harbored *AluSp_SP1*, with only 34 TCGA matched normal tissues expressed this TCT.

Similarly, the protein encoded by *ETAA1* is a replication stress response protein known to accumulate at DNA damage sites and promote replication fork progression and integrity [28–30]. AluSc8, co-transcribed with the *ETAA1* gene and overlapping with the CAGE-seq peak, is located in an intron between exon 1 and exon 2 of the *ETAA1* gene (Fig. 3D). Out of 10,357 TCGA tumors, 606 contained *AluSc8_ETAA1*, with 6/729 TCGA normal samples including this TCT [4]. In the LUAD and lung squamous cell carcinoma (LUSC) cohorts, there were 26 and 69 *AluSc8_ETAA1*-positive patients, respectively, with one matched normal sample showing expresssion. These findings suggested that while *AluSp_SP1* and *AluSc8_ETAA1* are commonly detected in tumors, they are not exlusive to tumors.

However, another 3 expressed TEs, identified by both lrRNA-seq and srRNA-seq, exhibited different alternative splicing patterns in the two datasets. While previous studies based on srRNA-seq suggested that these TEs could act as cryptic promoters, our analysis found that two of these TEs were transcribed within a TE-associated alternative last exon in two TCTs (*PABL_A_RNPS1* and *L2c_HSPBAP1*). Another TE, located in the intron of its co-transcribed gene, provided a cryptic splicing site for this gene (*MER4A_MLPH*). These discrepancies may reflect differences in alignment and assembly algorithms between lrRNA-seq and srRNA-seq, which may affect the accurate identification of transcript isoforms.

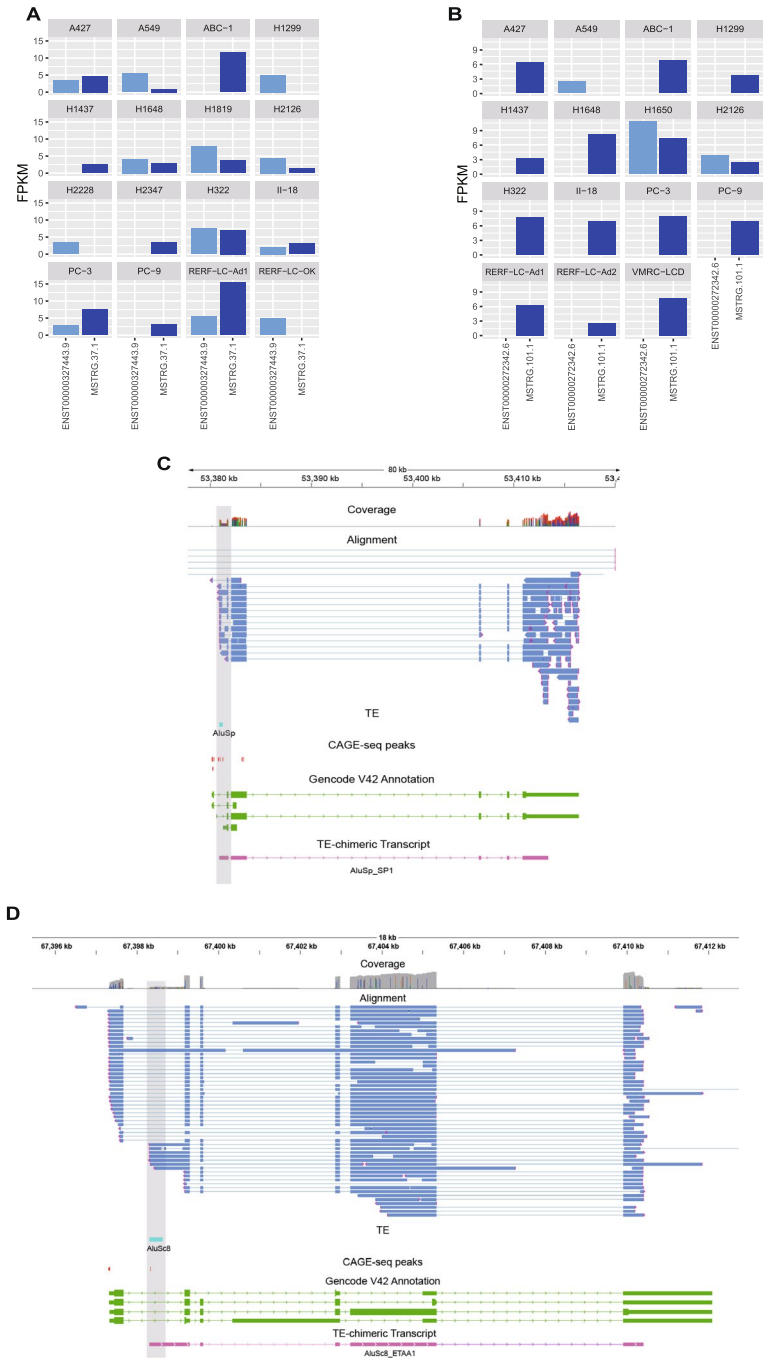

Fig. 3 Expression of transcripts of *SP1* and *ETAA1* in lrRNA-seq data and their IGV. **A** The histogram illustrates the expression of classical transcripts and TCT of *SP1* detected in the lrRNA-seq data in each cell line (FPKM). **B** The histogram illustrates the expression of classical transcripts and TCT of *ETA11* detected in the lrRNA-seq data in each cell line (FPKM). **C** lrRNA-seq full-length reads span the entire TCT. The first track shows the coverage of genome. The second tracks shows lrRNA-seq full-length reads. The third track shows the location of TEs. The forth track shows the location of CAGE-seq peaks. The last track shows gene annotation contains transcripts of *SP1* from Gencode V42 and assembled TCT (MSTRG.37.1). **D** lrRNA-seq full-length reads span the entire TCT. The first track shows the coverage of genome. The second tracks shows lrRNA-seq full-length reads. The third track shows the location of TEs. The forth track shows the location of CAGE-seq peaks. The last track shows gene annotation contains transcripts of *ETA11* from Gencode V42 and assembled TCT (MSTRG.101.1)

Chimeric truncations are abundantly detected in LUAD cells

Next, we predicted the protein products of these TCTs as previously described (https://github.com/TransDecoder/TransDecoder/, Methods). Interestingly, highly dynamic TCTs differ from the corresponding canonical gene isoforms in both structure and open reading frames (ORFs). Based on the location of the start codon, we categorized the candidates into one of five groups, as previously outlined: (1) normal, (3) truncated, (4) chimeric normal, (5) chimeric truncated or (6) frameshift (Fig. 4A). Among these, TCTs encoding alternative ORFs predominantly exhibited chimeric truncations (67.1%), truncations (19.6%), or frameshifts (10.1%) (Fig. 4B) with only a small fraction exhibiting chimeric normal (2.5%) or normal sequences (0.6%) (Fig. 4B).

Since 84.9% of the candidates were TE-related truncations (including chimeric truncations and truncations), we further explored whether any cancer-related genes exhibited TE-related truncation. Five genes included in the COSMIC oncogene database were found to have TE-related truncation. Specifically, *CBFB, EPAS1, FLNA* and *RABEP1* harbored chimeric truncations caused by expressed TEs, while *TRIP11* had a truncation resulting from expressed TEs. For example, *L1M4a1_CBFB* was generated by the L1M4a1 element, *which is* located in the intron between exon4 and exon5 of the canonical transcript of *CBFB* (Fig. 4C). The predicted ORF starts within exon1 and ends in the L1M4a1 element, resulting in a chimeric truncated protein with 135 amino acids. *CBFB* encodes the beta subunit of a core-binding transcription factor that belongs to the *PEBP2/CBF* family. It regulates genes specific to hematopoiesis (e.g., *RUNX1*) and osteogenesis (e.g., *RUNX2*). The beta subunit is a non-DNA binding regulatory subunit that promotes DNA binding allosterically when in complex with the alpha subunit, binding to various enhancers and promoters, including those of murine leukemia virus, polyomavirus, T-cell receptor, and GM-CSF. Previous studies have suggested that the mitochondrial translation dysregulation due to *CBFB* deficiency is assiciated with mutant *PIK3CA* and is vulnerable to breast cancer [31]. Another example is *AluSp_RABEP1*, which was created by an Alu element located in the intron between the exon 4 and exon 5 of the canonical transcript of *RABEP1* (Figure S4). The predicted ORF of *AluSp_RABEP1* starts from within exon 4 and ends in the intron between exon 4 and exon 5, with the ORF overlapping only with exon4 of *RABEP1*. The predicted peptide encoded by *AluSp_RABEP1* is 226 amino acids long. Multiple studies have revealed that dysregulation of the protein encoded by *RABEP1* promotes cancer invasion and metastasis [32, 33], though the biological role of *AluSp_RABEP1* remains unclear.

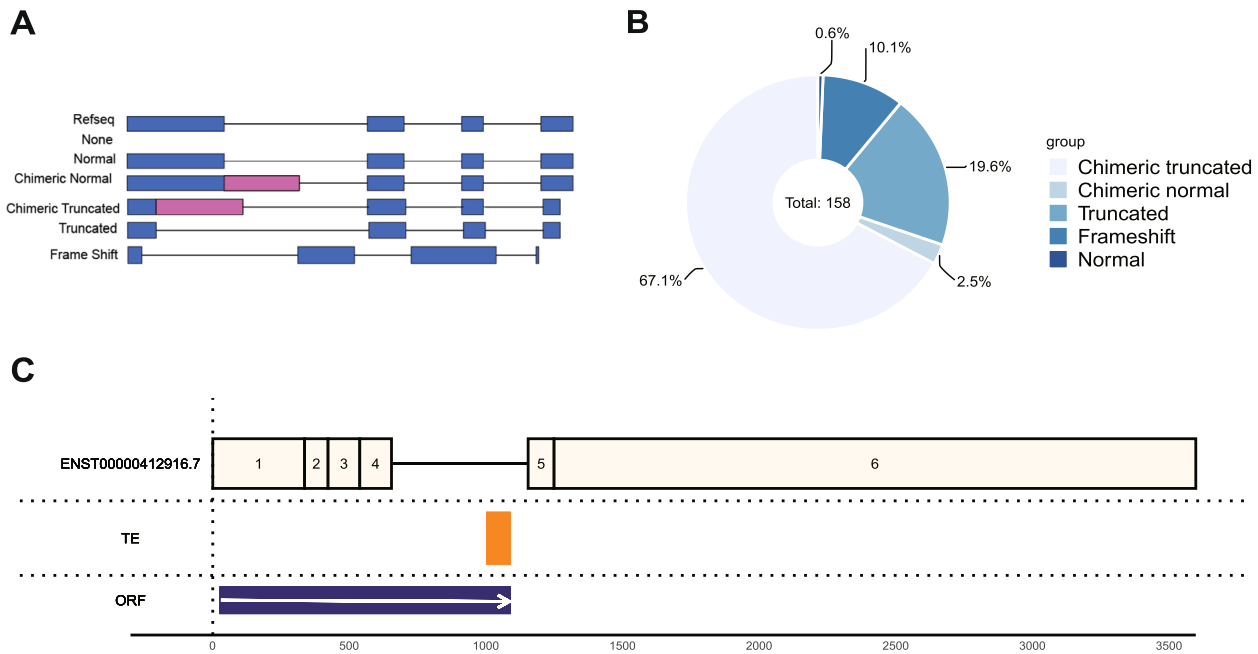

Fig. 4 Different kinds of predicted protein products of TCTs. **A** Schematic representation of the different kinds of potential protein products of TE-chimeric transcripts structures: None, Normal, Chimeric Normal, Chimeric Truncated, Truncated, Frame Shift. **B** Pie plot represents the distribution of potential protein products of TE-chimeric transcripts. **C** Diagram of *L1M4a1_CBFB* transcript exons and highlighting the predicted open reading frame ending in a end codon within the L1P2 TE

Exploring the Role of TE-chimeric transcripts in cancer progression

Next, we aimed to assess the validity of the TE-chimeric transcripts identified and explore their potential biological functions. Initially, we validated 20 TE-chimeric transcripts in four types of LUAD cell lines using RT-PCR (Fig. 5A, Methods). Among these, the *MIRb_HKDC1* was notably highle expressed in PC-9 cells (Fig. 5A, 5B). Further expression analysis, utilizing both long-read and short-read sequencing methods, confirmed that *MIRb_HKDC1* is the predominant transcript of the *HKDC1* gene in PC-9 cells, with expression levels significantly higher than those observed in other cell lines (Figure S5A, S5B). The *HKDC1* gene, a member of the hexokinase protein family, plays a role in glucose metabolism. Decreased expression of *HKDC1* has been linked to gestational diabetes mellitus, while elevated expression has been associated with poor prognosis in liver cancer [34] and gastric cancer [35].

To elucidate the molecular mechanisms underlying MIR-mediated gene regulation and its impact on cellular proliferation within the context of cell line models, we performed targeted overexpression of *HKDC1* and *MIRb-HKDC1* variants in the Beas-2B cell line utilizing a recombinant eukaryotic expression plasmid. After validation of the overexpression of *HKDC1* and *MIRb-HKDC1* in Beas-2B cells compare to the control cells transfected with an empty vector (pcDNA3.1), we observed that the presence of plasmids encoding the MIRb sequence led to increased transcription levels of the *HKDC1* gene (Fig. 5C).

To further delineate the functional role of MIRb in cancer etiology, we used Cell Counting Kit-8 (CCK-8) assays to quantitatively assess the proliferative capacity of Beas-2B-OE-MIRb-HKDC1 cells and compare it to that of Beas-2B-OE-HKDC1 cells and the respective control cell lines. The results indicated a pronounced increase in the proliferation rate of Beas-2B-OE-MIRb-HKDC1 cells (Fig. 5D). Concurrently, the migratory potential of Beas-2B-OE-HKDC1 and Beas-2B-OE-MIRb-HKDC1 cells was evaluated through wound healing assays. The rate of wound closure in Beas-2B-OE-MIRb-HKDC1 cells was significantly greater than that in Beas-2B-OE-HKDC1 cells and control cells (Fig. 5E). Collectively, these results suggested that MIRb expressed via *HKDC1* can notably promote gene expression, augment cell proliferation, indicating a potential role in facilitating tumor progression.

TCTs as potential biomarkers in LUAD

TEs have garnered attention for their potential regulatory roles in cancer progression [7, 36]. To study whether TE events similar to those observed in cancer cell lines can be detected in human patient tumors, we downloaded 2 publicly available short-read transcriptome datasets [37, 38] on a total of 178 samples, including 95 LUAD tissue samples and 83 matched normal tissue samples. We used a gene annotation file that included the GENCODE v.42 transcript annotations as well as 208 TCTs to evaluate the expression level of each transcript. Before identifing differentially expressed transcripts (DETs), we evaluated whether the expression level obtained from lrRNA-seq data were consistent with those from srRNA-seq data in the same cell line. We found that the srRNA-seq data were strongly correlated with those based on the lrRNA-seq data (Pearson correlation, $R = 0.7969$, p-value = 2.2e-16, Figure S6A).

By employing a Linear Model Fit (FDR < 0.05, fold change > 2 or fold change < 0.5, Table S7), we identified a total of 144 DETs, consisting of 103 downregulated transcripts and 41 upregulated transcripts in the LUAD samples compared to normal samples (Table S7). Among these DETs, 23 were TCTs (20 downregulated and 3 upregulated transcripts) (Fig. 6A). Notably, *MER4A-MLPH* (MSTRG.109.1), *MIR-KRT7* (MSTRG.36.1), and *AluJb-CAPN8* (MSTRG.12.1) emerged as the most significantly down-regulated genes, while *AluSc-N*EDD9 (MSTRG.148.1) stood out as the most significantly up-regulated gene (|fold change| < 9 and FDR < 0.001).

Within these subgroups, we observed notable variations in the expression patterns of TCTs among different patient subgroups. We found that 3 TCTs showed higher expression in never-smokers (Fig. 6B), while 5 TCTs showed higher expression in stage I patients (Fig. 6C). Two TCTs of the *CAPN8* gene, *AluJb_CAPN8* (MSTRG.12.1) and *AluSz_CAPN8* (MSTRG.12.2), were major contributors to *CAPN8* expression in tumor samples (Figure S6B). The expression of these 2 TCTs was significantly greater in never-smokers than in smokers (Student's t test, p-value < 0.05). Furthermore, *AluJb_CAPN8* and *AluSz_CAPN8* had higher expression in stage I than in other stages. Moreover, normal samples exhibited significantly higher expression of *AluJb_CAPN8* and *AluSz_CAPN8* compared to tumor samples (Figure S6C). *AluJb_CAPN8* and *AluSz_CAPN8* may generate truncated protein products that are dysfunctional. The protein product of *CAPN8* is a calcium-regulated non-lysosomal thiol-protease, and the E2F-targets pathway has been verified as the downstream signaling pathway of *CAPN8*, which is a well-acknowledged pathways that promotes cancer metastasis and proliferation [39]. Taken together, these results indicated the potential regulatory involvement of *AluSz_CAPN8* and *AluJb_CAPN8* in LUAD suppression and growth inhibition.

Similarly, *L1ME3G_PRMT2* exhibited greater expression in never smokers and stage I patients. Conversely,

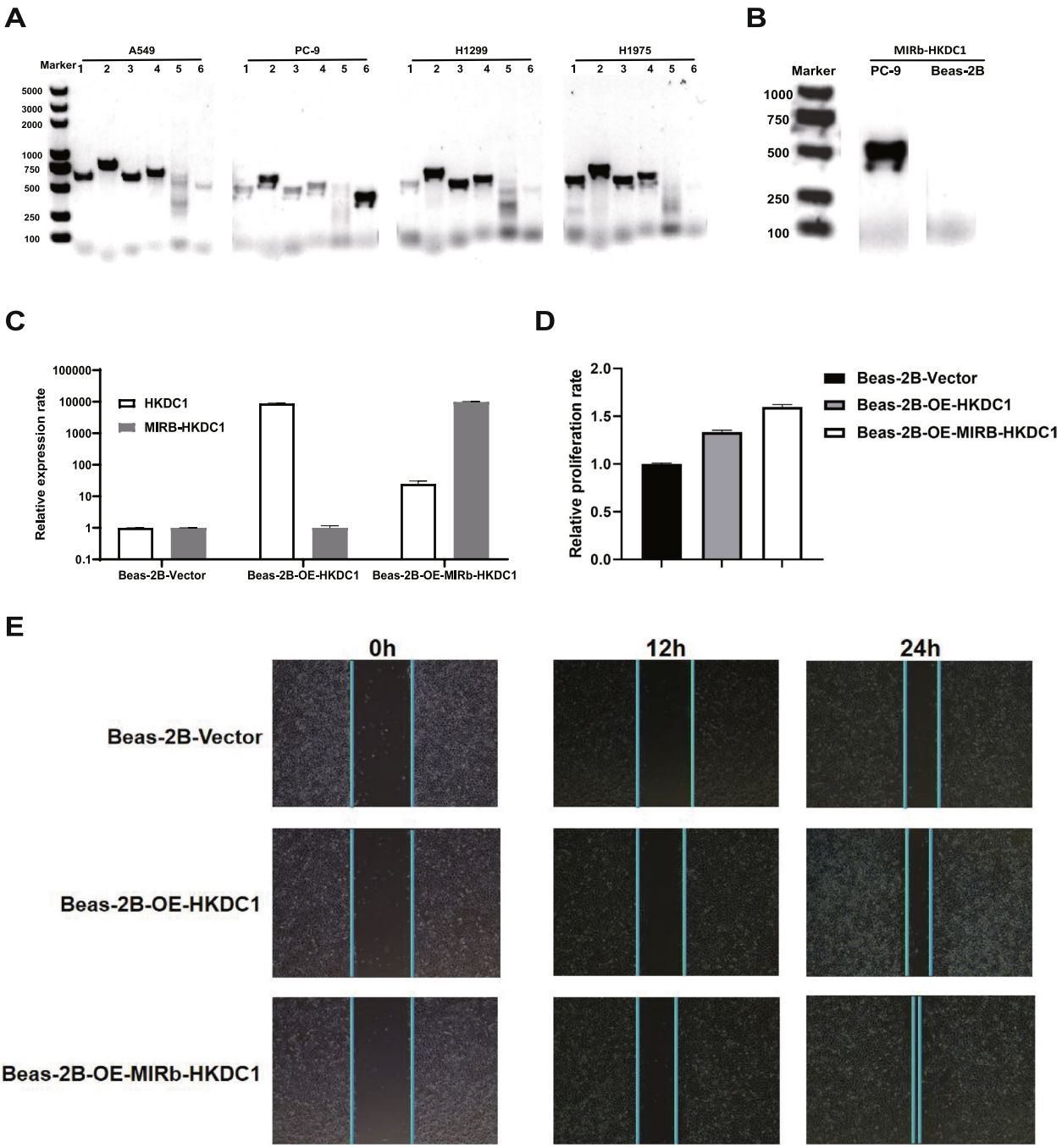

Fig. 5 Experimental verification of TCTs expression. **A**. Agarose gel analysis of MIRs expression in A549, PC-9, H1299 and H1975 cells. MIRs: Lane 1. *AluSg_PUM2*(1); Lane 2. *AluSg_PUM2*(2); Lane 3. *AluSc8_ETAA1*(1); Lane 4. *AluSc8_ETAA1*(2); Lane 5. *AluJb_NFS1*; 6. *MIRb_HKDC1*. **B**. Agarose gel analysis of *MIRb_HKDC1* expression in PC-9 cell line.The Beas-2B cell line was used as control. **C**. *HKDC1* and *MIRb_HKDC1* were overexpressed in the Beas-2B cell line, and the transcript levels of *HKDC1* and *MIRb_HKDC1* were examined by qPCR. MIRb increased the transcription levels of the *HKDC1* gene. **D**. A CCK8 kit was used to detect cell proliferation after overexpression. **E** Representative images of wound-healing assays after overexpression

normal tissues displayed greater *PRMT2* expression (FDR < 0.05 and log2(Fold Change) > 2) than paired tumor tissues. *L1ME3G_PRMT2* might be the predominant contributor of *PRMT2* expression (Figure S6D). Recent studies have correlated *PRMT2* expression with the progression of breast cancer, glioblastoma, and renal cell carcinoma [40, 41].

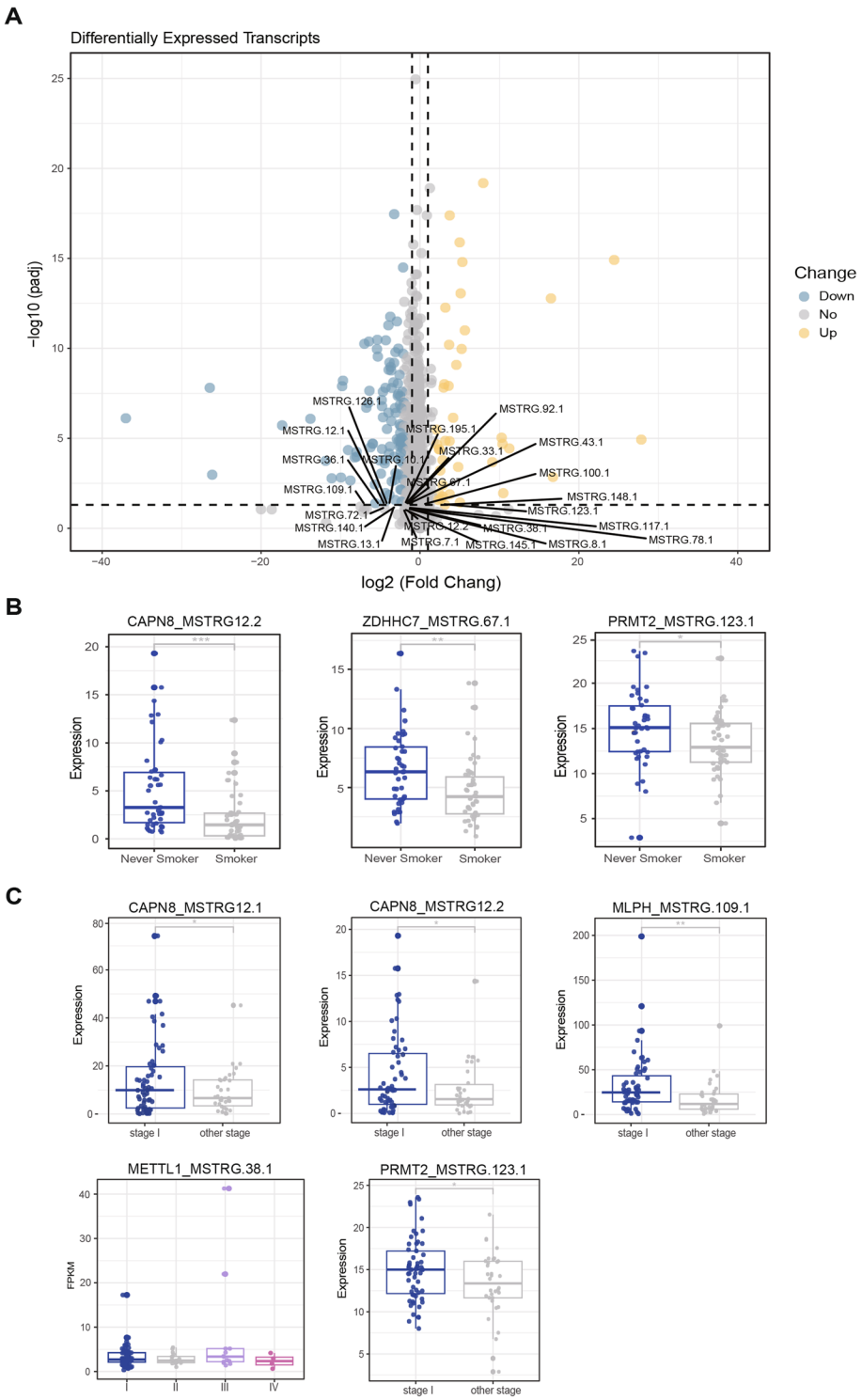

Fig. 6 Differentially expressed TCTs between tumor and matched normal samples. A. Volcano map shows the expression of 144 DETs. The yellow dots (Up) represent significantly upregulated genes, the blue dots (Down) represent significantly downregulated genes, and the grey dots (No) represent insignificantly differentially expressed transcripts (DETs). **B** The BoxPlot shows the differential expression of *AluSz_CAPN8* (MSTRG.12.2), *AluSg7_ZDHHC7* (MSTRG.67.1) and *L1ME3G_PRMT2* (MSTRG.123.1), respectively, between smoking and non-smoking patients. **C** The BoxPlot shows the differential expression of TCTs of *AluJb_CAPN8* (MSTRG.12.1), *AluSz_CAPN8*(MSTRG.12.2), *MER4A_MLPH* (MSTRG.109.1), *MIRc_AluSz_METTL1* (MSTRG.38.1) and *L1ME3G_PRMT2* (MSTRG.123.1) in patients at different stages

Discussion

In cancer, TEs are known to undergo widespread activation, promoting the expression of non-canonical transcripts, which may show high expression across diverse tumor samples [9]. However, conventional short-read sequencing technologies have limitations in providing a comprehensive and accurate depiction of the structure of TCTs. To address this issue, we present a computational pipeline for the thorough identification and analysis of TCTs in 19 LUAD cell lines, leveraging long-read RNA-seq data. Our analysis identified 208 high-confidence, full-length TCTs in these LUAD cell lines, revealing a significant presence of expressed SINEs, with 73.8% of identified TCTs belonging to SINE-chimeric transcript. While the role of LINEs and LTRs in tumor development has been extensively discussed in large-scale srRNA-seq data, a study of 678 tumor samples identified SINE-related alternative splicing events as the primary type of TE-related splicing [3]. However, due to their short length (approximately 300 bp) and high copy number relative to LINEs and LTRs, SINEs present challenges in precise genomic alignment of short reads, thus limiting a thorough investigation into the functional significance and overall impact of SINE chimeric transcripts in tumors. Our work significantly enhances the detailed characterization of the structural features of SINE chimeric transcripts and fosters further investigation into their roles in cancer.

We present a concise analysis of the experimental data obtained from a series of gene validation experiments on *MIRb_HKDC1*. Using PCR and qPCR, we confirmed the transcription and upregulated expression of *MIRb_HKDC1*. The CCK-8 assay revealed that alterations in *MIRb_HKDC1* expression significantly influence cell proliferation rates, suggesting a potential role for *MIRb_HKDC1* in cell cycle regulation. Additionally, the scratch wound healing assay indicated that *MIRb_HKDC1* may modulate cell migration, a critical factor in metastasis and wound repair. These findings collectively imply that *MIRb_HKDC1* could serve as a potential therapeutic target for diseases characterized by abnormal cellular activities. However, the mechanisms by which *MIRb_HKDC1* exerts its effects require further exploration, and in vivo studies are necessary to validate these in vitro results. The implications of our research lay the groundwork for future investigations into the functional significance of *MIRb_HKDC1* and its potential applications in clinical settings.

Integrating these findings with published transcriptome data from LUAD tumor and matched normal samples, we identified 23 TE-chimeric DETs. Notably, certain TE-chimeric DETs exhibited elevated expression in never smokers and stage I samples. The significance of these DETs in the progression of lung adenocarcinoma or their impact on patient prognosis remains to be confirmed through additional experimental validation. Nonetheless, our findings suggest that TE chimeric DETs may have roles in tumor progression.

There are some important caveats to consider with respect to the TCTs we identified in the landscape of alternative splicing events caused by TEs in cancer. Firstly, it is crucial to acknowledge the limitations inherent in our computational pipeline. Guided by the splice junctions from the Gencode V42 genome annotation, we employed Flair *correct* to rectify primary alignments and eliminate reads containing noncanonical splice junctions. However, this correction resulted in the exclusion of TCTs with noncanonical splice junctions. Specifically, while the expressed TEs we identified were predominantly located near host genes or within host gene introns, rather than in intergenic regions, we did not observe any expressed TEs that extended the protein products of host genes. A more comprehensive understanding of TCTs involving noncanonical splice junctions, particularly in LUAD, is still lacking and warrants further targeted research. Despite these limitations, we identified several TCTs with significant clinical-pathological associations with LUAD. For instance, *AluJb_CAPN8* and *AluSz_CAPN8* could lead to truncation of the *CAPN8* protein in non-smokers and stage I patients, resulting in the inactivation of the E2F target pathway, thereby suppressing cancer metastasis and proliferation, and ultimately inhibiting cancer progression. Similarly, *L1ME3G_PRMT2*, which is upregulated in normal tissues, non-smokers, and stage I patients, primarily driving the expression of *PRMT2* in tumor samples and likely leads to the production of a truncated isoform, potentially suppressing *PRMT2* expression. These TEs, by suppressing the expression of cancer-related genes and producing truncated, non-functional proteins, could significantly influence cancer biology. However, additional functional experiments, including in vivo studies, are necessary to confirm the role of these TCTs in tumor progression and their potential as therapeutic targets.

Secondly, our study was limited to a small number of cell lines and clinical samples from LUAD patients, which may restrict the generalizability of our findings. LUAD, influenced by genetics, environmental exposure, and lifestyle, exhibits significant genetic diversity. The small size of our sample set may not fully capture this variability, meaning the identified TCTs could represent only a subset of those present in the broader patient population. Although useful for controlled studies, cell lines do not always perfectly replicate the in vivo tumor enviroment. Over time in the lab, cell lines can undergo changes that may cause them to lose certain original

tumor characteristics, potentially affecting the relevance of our findings. To validate our findings and enhance their clinical relevance, future research should include a larger and more diverse set of samples, covering various stages of the disease. Increasing the sample size could not only confirm our initial findings but also reveal additional TCTs and mechanisms that were not observed in our smaller corhort. This would strengthen the potential clinical application of our research in LUAD.

In summary, we present a comprehensive dataset of LUAD TCTs, with a focus on their structural characteristics. Notably, many of these transcripts are novel and previously unannotated, even within the scope of cancer research, likely due to differences in sequencing technologies. By employing long-read sequencing technology coupled with customizable data analysis pipeline, we achived extensive coverage of most transcripts, which is critical for the accurate identification of TCTs, especially those associated with tandem TEs. Our research significantly expand the existing knowledge base, revealing previously overlooked TCTs that could markedly improve our understanding of tumor progression and introduce new dimensions to cancer biology. Furthermore, we offer a curated list of actionable TCTs, which can serve as a foundation for future benchmark studies. This work has the potential to advance precision oncology, contributing to improved cancer monitoring, diagnosis, and therapeutic strategies.

Conclusions

Overall, our study provides a comprehensive characterization of TCTs in NSCLC, revealing they may have diverse roles in cancer progression and highlighting their potential as biomarkers and therapeutic targets in LUAD.

Abbreviations

TEs	Transposable elements
TCTs	TE-chimeric transcripts
LUAD	Lung Adenocarcinoma
TS-TEAs	Tumor specific TE chimeric antigens
SINEs	Short interspersed nuclear elements
LINEs	Long interspersed nuclear elements
LTRs	Long terminal repeat elements
lrRNA-seq	Long read RNA sequencing
srRNA-seq	Short read RNA sequencing
NSCLC	Non-small cell lung cancer
LUSC	Lung squamous cell carcinoma
DETs	Differentially expressed transcripts

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-13888-5.

Additional file1: FigureS1 Classification, structure and corresponding gene enrichment of TCTs. A The pie chart illustrates the distribution of assembled transcripts compared to the reference Gencode V42. B Barplots depicts the number of exons within each TCT. The numerical values atop each bar denote the frequency of transcripts containing the respective number of exons. C Lollipop chart demonstrates the GO enrichment of these expressed TEs.

Additional file 2: FigureS2 Characteristics of expressed TEs. A. Barplot shows the fraction of each subfamily within the total expressed SINEs, as represented on the y-axis. B The Venn diagram illustrates the intersection and distribution of lrRNA-seq-identified TE-chimeric gene expression events and srRNA-seq-identified TE-chimeric gene expression events. "SRS-TCGA" indicates events reported by Nakul M. Shah et al. within 10,357 TCGA samples using short-read sequencing data.

Additional file 3: FigureS3 Expression of transcripts of *SP1* and *ETAA1* in SRS-data and IGV of *MLT1K_BCAS1*. A The histogram illustrates the expression of classical transcripts and TCT of the SP1 gene detected in the srRNA-seq data in each cell line (FPKM). B The histogram illustrates the expression of classical transcripts and TCT of the ETAA1 gene detected in the srRNA-seq data in each cell line (FPKM). C lrRNA-seq full-length reads span the entire TCT. The first track shows the coverage of genome. The second tracks shows lrRNA-seq full-length reads. The third track shows the location of TEs. The forth track shows the location of CAGE-seq peaks. The last track shows gene annotation contains transcripts of BCAS1 from Gencode V42 and assembled TCT (MSTRG.116.1).

Additional file 4: FigureS4 Diagram of *AluSp_RABEP1* transcript exons and highlighting the predicted open reading frame ending in a end codon within the AluSp TE.

Additional file 5: FigureS5 Expression of all TCTs in different data and experimental validation. A The histogram illustrates the expression of classical and TCT of *HKDC1* detected in the lrRNA-seq data in each cell line (FPKM). B The histogram illustrates the expression of classical and TCT *of HKDC1 detected* in the srRNA-seq data in each cell line (FPKM). C Agarose gel analysis of MIRs expression in Lane 1. A549; Lane 2. PC-9; Lane 3. H1299; Lane 4. H1975 cells. D Agarose gel analysis of MIRs expression in Lane 1. A549; Lane 2. PC-9; Lane 3. H1299; Lane 4. H1975 cells. E Agarose gel analysis of MIRs expression in Lane 1. A549; Lane 2. PC-9; Lane 3. H1299; Lane 4. H1975 cells. F The transcript levels of MIRs were examined by qPCR.

Additional file 6: FigureS6 Expression of TCTs. A The scatter plot shows the pearson correlations between the gene-expression profiles in srRNA-seq and lrRNA-seq data. B The boxplot illustrates two TCTs and annotated transcript of the *CAPN8* gene, *AluJb-CAPN8* (MSTRG.12.1) and *AluSz-CAPN8* (MSTRG.12.2), were major contributors to *CAPN8* expression in tumor sample. C Heatmap shows the expression of the TCT chimeric transcripts we found in the patients' srRNA-seq data and its relation to the pathologic information. D The boxplot illustrates one TCT and annotated transcript of the PRMT2 gene, *L1ME3G_PRMT2* (MSTRG123.1) was the predominant driver of *PRMT2* expression.

Additional file 7.

Acknowledgements

We would like to thank Ranlei Wei for his help in server operation and maintenance, and Yan Huang for her help with the experiments.

Authors' contributions

Y.L. and L.X. designed the study and wrote the manuscript with input from all coauthors. Y.L., J.W., Y.H. performed laboratory work and generated the sequencing data. L.X., Y.H.L., Y.X.X and Y.X.W. contributed to bioinformatics analysis. D.X. and L.X. edited the manuscript, supervised the research and secured funding for the project. All authors approve this version of the manuscript and agree to be accountable for their contributions.

Funding

This work was supported by grants from the National Natural Science Foundation of China (32200508 to L. Xia, 82173383 to D. Xie), the Sichuan Province Science and Technology Program (2022NSFSC1553 to L. Xia, 2024NSFSC1187 to Y. Li, 2022NSFSC1762 to H. Wang), the 1·3·5 project for disciplines of excellence, West China Hospital, Sichuan University (ZYYC23024 to D. Xie), and the Sichuan University postdoctoral interdisciplinary Innovation Foundation (2022SCU12045 to L. Xia).

Data availability
The RNA sequencing data used in this publication are accessible through GEO Series accession number GSE40419 and GSE37765. The lrRNA-seq data of A549 have been deposited in the Genome Sequence Archive in BIG Data Center, under accession number HRA007608, which can be accessed at https://bigd.big.ac.cn/gsa-human. Our custom computational pipeline can be accessed at https://github.com/LinXialab/TCT_TGS.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Laboratory of Omics Technology and Bioinformatics, Frontiers Science Center for Disease-Related Molecular Network, State Key Laboratory of Biotherapy, West China Hospital, Sichuan University, Chengdu 610041, Sichuan, China. [2]West China School of Basic Medical Sciences and Forensic Medicine, Sichuan University, Chengdu 610041, Sichuan, China. [3]West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu 610041, China.

Received: 29 May 2024 Accepted: 7 March 2025
Published online: 15 March 2025

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
2. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. Mob DNA. 2016;7:9.
3. Clayton EA, Rishishwar L, Huang T-C, Gulati S, Ban D, McDonald JF, et al. An atlas of transposable element-derived alternative splicing in cancer. Philos Trans R Soc Lond B Biol Sci. 2020;375:20190342.
4. Shah NM, Jang HJ, Liang Y, Maeng JH, Tzeng S-C, Wu A, et al. Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. Nat Genet. 2023;55:631–9.
5. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nat Rev Genet. 2020;21:721–36.
6. Grillo G, Keshavarzian T, Linder S, Arlidge C, Mout L, Nand A, et al. Transposable Elements Are Co-opted as Oncogenic Regulatory Elements by Lineage-Specific Transcription Factors in Prostate Cancer. Cancer Discov. 2023;13:2470–87.
7. Liang Y, Qu X, Shah NM, Wang T. Towards targeting transposable elements for cancer therapy. Nat Rev Cancer. 2024;24:123–40.
8. Scott EC, Gardner EJ, Masood A, Chuang NT, Vertino PM, Devine SE. A hot L1 retrotransposon evades somatic repression and initiates human colorectal cancer. Genome Res. 2016;26:745–55.
9. Jang HS, Shah NM, Du AY, Dailey ZZ, Pehrsson EC, Godoy PM, et al. Transposable elements drive widespread expression of oncogenes in human cancers. Nat Genet. 2019;51:611–7.
10. Attig J, Young GR, Hosie L, Perkins D, Encheva-Yokoya V, Stoye JP, et al. LTR retroelement expansion of the human cancer transcriptome and immunopeptidome revealed by de novo transcript assembly. Genome Res. 2019;29:1578–90.
11. Wu Y, Zhao Y, Huan L, Zhao J, Zhou Y, Xu L, et al. An LTR Retrotransposon-Derived Long Noncoding RNA lncMER52A Promotes Hepatocellular Carcinoma Progression by Binding p120-Catenin. Can Res. 2020;80:976–87.
12. Sorek R, Ast G, Graur D. *Alu* -Containing Exons are Alternatively Spliced. Genome Res. 2002;12:1060–7.
13. Lev-Maor G, Sorek R, Shomron N, Ast G. The Birth of an Alternatively Spliced Exon: 3′ Splice-Site Selection in *Alu* Exons. Science. 2003;300:1288–91.
14. Kakkos SK, Kirkilesis GI, Tsolakis IA. Re: "Re Efficacy and Safety of the New Oral Anticoagulants Dabigatran, Rivaroxaban, Apixaban, and Edoxaban in the Treatment and Secondary Prevention of VTE: A Systematic Review and Meta-analysis of Phase III Trials." Eur J Vasc Endovasc Surg. 2015;50:127.
15. Babarinde IA, Ma G, Li Y, Deng B, Luo Z, Liu H, et al. Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. Nucleic Acids Res. 2021;49:9132–53.
16. Babarinde IA, Hutchins AP. The effects of sequencing depth on the assembly of coding and noncoding transcripts in the human genome. BMC Genomics. 2022;23:487.
17. Method of the Year. long-read sequencing. Nat Methods. 2022;2023(20):1.
18. Sakamoto Y, Sereewattanawoot S, Suzuki A. A new era of long-read sequencing for cancer genomics. J Hum Genet. 2020;65:3–10.
19. Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proc Natl Acad Sci U S A. 2014;111:9869–74.
20. Singh M, Al-Eryani G, Carswell S, Ferguson JM, Blackburn J, Barton K, et al. High-throughput targeted long-read single cell sequencing reveals the clonal and transcriptional landscape of lymphocytes. Nat Commun. 2019;10:3120.
21. Panda K, Slotkin RK. Long-Read cDNA Sequencing Enables a "Gene-Like" Transcript Annotation of Transposable Elements. Plant Cell. 2020;32:2687–98.
22. Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, et al. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. Nat Commun. 2022;13:1948.
23. Jiang F, Zhang J, Liu Q, Liu X, Wang H, He J, et al. Long-read direct RNA sequencing by 5′-Cap capturing reveals the impact of Piwi on the widespread exonization of transposable elements in locusts. RNA Biol. 2019;16:950–9.
24. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
25. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun. 2020;11:1438.
26. Oka M, Xu L, Suzuki T, Yoshikawa T, Sakamoto H, Uemura H, et al. Aberrant splicing isoforms detected by full-length transcriptome sequencing as transcripts of potential neoantigens in non-small cell lung cancer. Genome Biol. 2021;22:9.
27. O'Connor L, Gilmour J, Bonifer C. The Role of the Ubiquitously Expressed Transcription Factor Sp1 in Tissue-specific Transcriptional Regulation and in Disease. Yale J Biol Med. 2016;89:513–25.
28. Haahr P, Hoffmann S, Tollenaere MAX, Ho T, Toledo LI, Mann M, et al. Activation of the ATR kinase by the RPA-binding protein ETAA1. Nat Cell Biol. 2016;18:1196–207.
29. Feng S, Zhao Y, Xu Y, Ning S, Huo W, Hou M, et al. Ewing Tumor-associated Antigen 1 Interacts with Replication Protein A to Promote Restart of Stalled Replication Forks. J Biol Chem. 2016;291:21956–62.
30. Bass TE, Luzwick JW, Kavanaugh G, Carroll C, Dungrawala H, Glick GG, et al. ETAA1 acts at stalled replication forks to maintain genome integrity. Nat Cell Biol. 2016;18:1185–95.
31. Malik N, Kim Y-I, Yan H, Tseng Y-C, du Bois W, Ayaz G, et al. Dysregulation of Mitochondrial Translation Caused by CBFB Deficiency Cooperates with Mutant PIK3CA and Is a Vulnerability in Breast Cancer. Cancer Res. 2023;83:1280–98.
32. Yan G-R, Xu S-H, Tan Z-L, Liu L, He Q-Y. Global identification of miR-373-regulated genes in breast cancer by quantitative proteomics. Proteomics. 2011;11:912–20.
33. Park MH, Choi K-Y, Min DS. The pleckstrin homology domain of phospholipase D1 accelerates EGFR endocytosis by increasing the expression of the Rab5 effector, rabaptin-5. Exp Mol Med. 2015;47:e200.
34. Dong L, Lu D, Chen R, Lin Y, Zhu H, Zhang Z, et al. Proteogenomic characterization identifies clinically relevant subgroups of intrahepatic cholangiocarcinoma. Cancer Cell. 2022;40:70-87.e15.

35. Zhao P, Yuan F, Xu L, Jin Z, Liu Y, Su J, et al. HKDC1 reprograms lipid metabolism to enhance gastric cancer metastasis and cisplatin resistance via forming a ribonucleoprotein complex. Cancer Lett. 2023;569:216305.
36. Lee M, Ahmad SF, Xu J. Regulation and function of transposable elements in cancer genomes. Cell Mol Life Sci. 2024;81:157.
37. Kim SC, Jung Y, Park J, Cho S, Seo C, Kim J, et al. A high-dimensional, deep-sequencing study of lung adenocarcinoma in female never-smokers. PLoS ONE. 2013;8:e55596.
38. Seo J-S, Ju YS, Lee W-C, Shin J-Y, Lee JK, Bleazard T, et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. 2012;22:2109–19.
39. Zhong X, Xu S, Wang Q, Peng L, Wang F, He T, et al. CAPN8 involves with exhausted, inflamed, and desert immune microenvironment to influence the metastasis of thyroid cancer. Front Immunol. 2022;13:1013049.
40. Dong F, Li Q, Yang C, Huo D, Wang X, Ai C, et al. PRMT2 links histone H3R8 asymmetric dimethylation to oncogenic activation and tumorigenesis of glioblastoma. Nat Commun. 2018;9:4552.
41. Li Z, Chen C, Yong H, Jiang L, Wang P, Meng S, et al. PRMT2 promotes RCC tumorigenesis and metastasis via enhancing WNT5A transcriptional expression. Cell Death Dis. 2023;14:322.

Publisher's Note