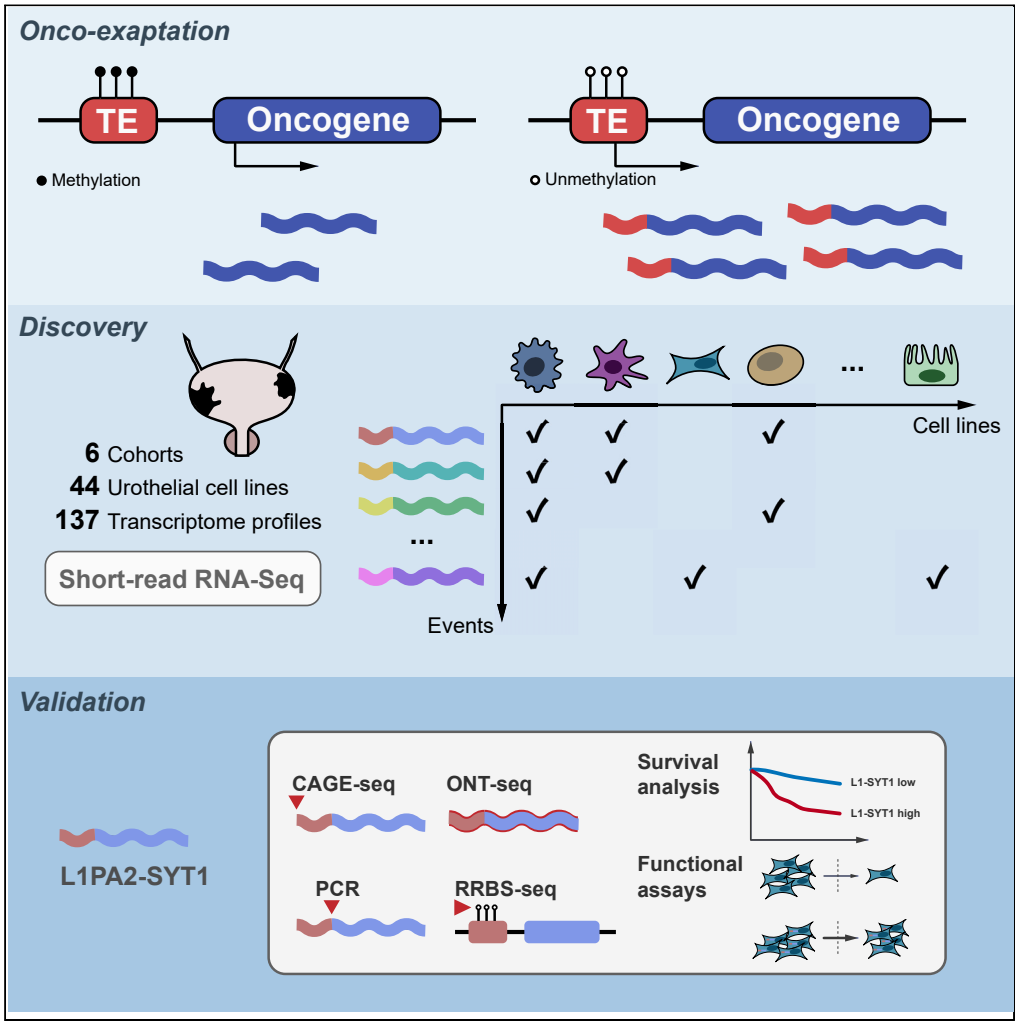
iScience

CellPress
OPEN ACCESS

Article

Comprehensive identification of onco-exaptation events in bladder cancer cell lines revealed L1PA2-SYT1 as a prognosis-relevant event



Ziwei Wang, Yidie Ying, Maoyu Wang, ..., Jing Li, Shuxiong Zeng, Chuanliang Xu

ljing@smmu.edu.cn (J.L.)
zengshuxiong@126.com (S.Z.)
chuanliang_xu@126.com (C.X.)

Highlights

A comprehensive identification of onco-exaptation events in bladder cancer cell lines

Transposable elements from LINE1 family contribute most events in bladder cancer

Multiomics analysis validated L1PA2-SYT1 as tumor specific and prognosis relevant

Wang et al., iScience 26, 108482
December 15, 2023
https://doi.org/10.1016/j.isci.2023.108482

iScience

CellPress
OPEN ACCESS

Article

Comprehensive identification of onco-exaptation events in bladder cancer cell lines revealed L1PA2-SYT1 as a prognosis-relevant event

Ziwei Wang,[1,4] Yidie Ying,[1,4] Maoyu Wang,[1,4] Qing Chen,[1] Yi Wang,[1] Xufeng Yu,[1] Wei He,[1] Jing Li,[2,3,*] Shuxiong Zeng,[1,*] and Chuanliang Xu[1,5,*]

SUMMARY

Transposable elements (TEs) can provide ectopic promoters to drive the expression of oncogenes in cancer, a mechanism known as onco-exaptation. Onco-exaptation events have been extensively identified in various cancers, with bladder cancer showing a high frequency of onco-exaptation events (77%). However, the effect of most of these events in bladder cancer remains unclear. This study identified 44 onco-exaptation events in 44 bladder cancer cell lines in 137 RNA-seq datasets from six publicly available cohorts, with L1PA2 contributing the most events. L1PA2-SYT1, L1PA2-MET, and L1PA2-XCL1 had the highest frequency not only in cell lines but also in TCGA-BLCA samples. L1PA2-SYT1 showed significant tumor specificity and was found to be activated by CpG island demethylation in its promoter. The upregulation of L1PA2-SYT1 enhances the *in vitro* invasion of bladder cancer and is an independent risk factor for patient's overall survival, suggesting L1PA2-SYT1 being an important event that promotes the development of bladder cancer.

INTRODUCTION

Transposable element (TE) is a type of repetitive sequence in human genes that accounts for about 45% of the whole genome length.[1] Although TEs, once considered as "junk DNA," have been discovered to exhibit regulatory functions, their precise roles remain largely unknown. TEs can move within the genome by a mechanism called transposition. Based on different transposition mechanisms, TE can be divided into class-I transposons and class-II transposons. Class-I transposons primarily transpose through RNA intermediate. Depending on whether their structure contains long terminal repeat (LTR) sequences, they can be divided into LTR elements and non-LTR elements. Class-I transposons can also be classified as autonomous retrotransposons or non-autonomous retrotransposons based on whether they have open reading frames (ORFs) that encode proteins required for retro-transposition. For example, in non-LTR elements, long interspersed element (LINE) is autonomous, while short interspersed element (SINE) is non-autonomous. Class-II transposons, known as DNA transposons (DNA), mainly transpose by a "cut-and-paste" mechanism. The proportion and composition of TE differ among different species. In humans, TE mainly includes four classes: DNA, LTR, SINE, and LINE, and each class can be further subdivided into different families and subfamilies.[2]

Recent studies have revealed a close relationship between TE and the development of tumors. Epigenetic dysregulation has been found in many tumors, including changes in methylation levels and histone modifications, which are related to TE activation.[3–5] TE activation can further induce microsatellite instability and increased genomic alterations.[6,7] It can also affect DNA repair mechanisms mediated by genes such as BRCA1/2 and APC.[8,9] The expression products of TE can induce *in vivo* immune response, indicating potential therapeutic targets of tumor.[10,11]

In addition, TE can promote tumor development by participating in transcriptional regulation. TE insertions can introduce new splicing sites,[12] change ORFs,[13] or affect post-transcriptional regulation.[14] Furthermore, regulatory elements present in TE sequences can be reactivated in tumors to promote oncogene expression, a mechanism known as "onco-exaptation."[15] For example, LTR contains promoter and enhancer elements.[16] The THE1B subfamily from the MaLR-ERVL family of LTR class, can splice into oncogene CSF1R, driving full-length CSF1R transcription in Hodgkin's lymphoma.[17] This event is represented as "THE1B-CSF1R," indicating the combination of the TE subfamily and the gene. The 5-end untranslated region (5′-UTR) of LINE1 contains two oppositely oriented promoter sequences, with the forward promoter used for its own transposition and the reverse promoter driving the expression of adjacent genes.[18] For instance, the L1PA2 subfamily of LINE1 drives oncogene MET expression in bladder cancer and colorectal cancer.[5,19] SINE elements are known to recruit RNA polymerase

[1]Department of Urology, Changhai Hospital, Naval Medical University, Shanghai 200433, China
[2]Department of Bioinformatics, Center for Translational Medicine, Naval Medical University, Shanghai 200433, China
[3]Shanghai Key Laboratory of Cell Engineering, Shanghai, China
[4]These authors contributed equally
[5]Lead contact
*Correspondence: ljing@smmu.edu.cn (J.L.), zengshuxiong@126.com (S.Z.), chuanliang_xu@126.com (C.X.)
https://doi.org/10.1016/j.isci.2023.108482

(RNAP) III,[20] but they can accumulate mutations during evolution to produce new transcription factor-binding sites that recruit RNAP II, such as AluJb-LIN28B identified in lung cancer cell lines.[21]

Recently, a pan-cancer study across 15 cancer types identified onco-exaptation events comprehensively.[21] It is impressive that more than 1 onco-exaptation event exist in 77% of bladder tumors, second only to squamous cell lung cancer (89%), suggesting the significance of onco-exaptation in bladder cancers. However, the biological significance of only a few events has been experimentally elucidated. L1PA2-MET is one such event that involves a novel exon 1 located in the intron 2 of the MET gene, and spliced into the exon 3 of MET, encoding a LINE1-driven MET isoform. In a study that includes ten bladder cancer cell lines, L1PA2-MET expression was significantly negatively correlated with the methylation level of L1PA2,[5] and this event has also been identified in colon cancer.[3] However, most of the frequent events have not been studied, mainly due to the incomplete knowledge of onco-exaptation in bladder cancer cell lines, which limits the use of this fundamental biological tool. Although TE-derived novel transcripts, including onco-exaptation events, were identified in 675 cell lines recently, only six bladder cancer cell lines were analyzed.[22,23]

To better utilize cell lines to study onco-exaptation events, we collected 137 RNA-seq data for 44 bladder cancer cell lines in six datasets and identified 44 events. We found that the cell lines reflected frequent events in primary tumors, including L1PA2-SYT1, L1PA2-MET, and L1PA2-XCL1, whose alternative promoters were all provided by L1PA2. We verified the splicing site of L1PA2-SYT1 and found that its expression was associated with poor prognosis in patients in an external dataset, potentially regulated by DNA hypomethylation.

RESULTS

Cell line data collection and identity confirmation

We searched for untreated bladder cancer cell lines in the SRA and ENA databases. In 137 datasets, a total of 44 different named cell lines were identified (STAR Methods). 18 cell lines were sequenced in two or more datasets, with RT4 being sequenced in five datasets, and J82, T24, TCCSUP, and 5637 being sequenced in four datasets. HT1197, HT1376, SCABER, SW780, and UMUC3 were profiled in three datasets (Figure 1A). In addition, SVHUC1 is an immortalized non-malignant urothelial cell line. Considering that cell line contamination is a common issue, we confirmed the identity of the cell lines using two methods (STAR Methods). One was principal component analysis based on gene expression profiles to determine phenotypic similarities between cell lines. We found that the similarities between cell lines exceeded the effects of different sequencing strategies and experimental batches (Figures 1B and S1A). Another method was adapted from the CeL-ID approach to detect curated somatic mutations among cell lines (STAR Methods). The same cell lines profiled in different datasets were highly similar in terms of curated somatic mutation profile (Figure 1C). Besides, the background mutation correlation between different cell lines was very low, whereas derived cell lines and parental clones showed high similarity, such as RT4 and RT4v6, 253J and 253JBV, and 1A6 and 5637[24] (Figure 1C). Based on these results, we conclude that the possibility of cross-contamination in the studied cell lines is low.

Identification of onco-exaptation events in bladder cancer cell lines

Afterward, we used the TEPROF2 algorithm to identify onco-exaptation events in the *de novo* transcript assemblies. Considering the potential high noise in short-read sequencing data during transcript assembly, TEPROF2 applied multiple heuristic filters to reduce false positive events (STAR Methods). For the identified transcripts that met the criteria, we retained events that had TPM >1 in at least one sample and accounted for at least 25% of the total expression of the corresponding gene. 44 events were identified and summarized in Table S1. Overall, we identified at least 1 onco-exaptation event in each cell line. We then compared the number of events identified between low grade (grades 1 and 2) and high grade (grade 3 and 4), and between luminal and basal subtypes, but found no significant differences (Figures S1B and S1C). Interestingly, we found that the number of events identified in the low genomic instability (GIN) group was significantly higher than that in the high GIN group (Figure S1D). Of the 44 events, 9 TE sequences were from intergenic regions (20.45%) and 35 were from intronic regions (79.54%) (Figure S1E). 2 TE sequences were DNA class, 16 were LINE, 15 were LTR, and 11 were SINE, with events significantly enriched in LINE and LTR sequences (enrichment score: DNA 0.503, LINE 1.195, LTR 2.156, SINE 0.509) (Figure S1F). We presented 17 events defined in at least 5 samples (Figure 2A). We found that in cell lines where L1PA2-SYT1, AluJb-SALL4, L1PA2-XCL1, L1PA3-TYRP1, MLT2B1-VAV2, and THE1D-WNT5A exist, the overall expression of the corresponding genes was significantly elevated compared to cell lines without the events, suggesting that these TE sequences promoted strong transcriptional activity and played a role in driving the expression of oncogenes (Figure 2A). Although the expression of L1PA2-XCL1 correlated well with the involved L1PA2 subfamily (Pearson's $r = 0.49$), overall, we found that the onco-exaptation events did not show a stronger correlation with the involved TE subfamily than with uninvolved TE subfamilies ($p = 0.15$) (Figures S2A and S2B). For onco-exaptation events that exist in at least three cell lines in each gender, we found L1PA2-MET exhibited elevated expression in cell lines originating from female patients, suggesting a gender-related influence (Figure S2C). Besides, no evident association was found between germline mutations within the promoter region of the novel transcription start site (TSS) and the occurrence of an onco-exaptation event (Figure S3).

To verify the significance of those onco-exaptation events, we quantified their expression in 22 cases of paired bladder tumor and normal samples. Notably, the most prevalent events, including L1PA2-SYT1, AluJb-SALL4, L1PA2-MET, and L1PA2-XCL1, demonstrated upregulated expression within the tumor samples (Figure S4). L1PA2-MET is also expressed in adjacent normal samples, which is consistent with the previous result that L1PA2-MET is activated in pre-malignant normal tissue, a phenomenon named "field defect."[5] Altogether, we found L1PA2-SYT1 expression nearly absent in adjacent normal samples, exhibiting high tumor specificity.

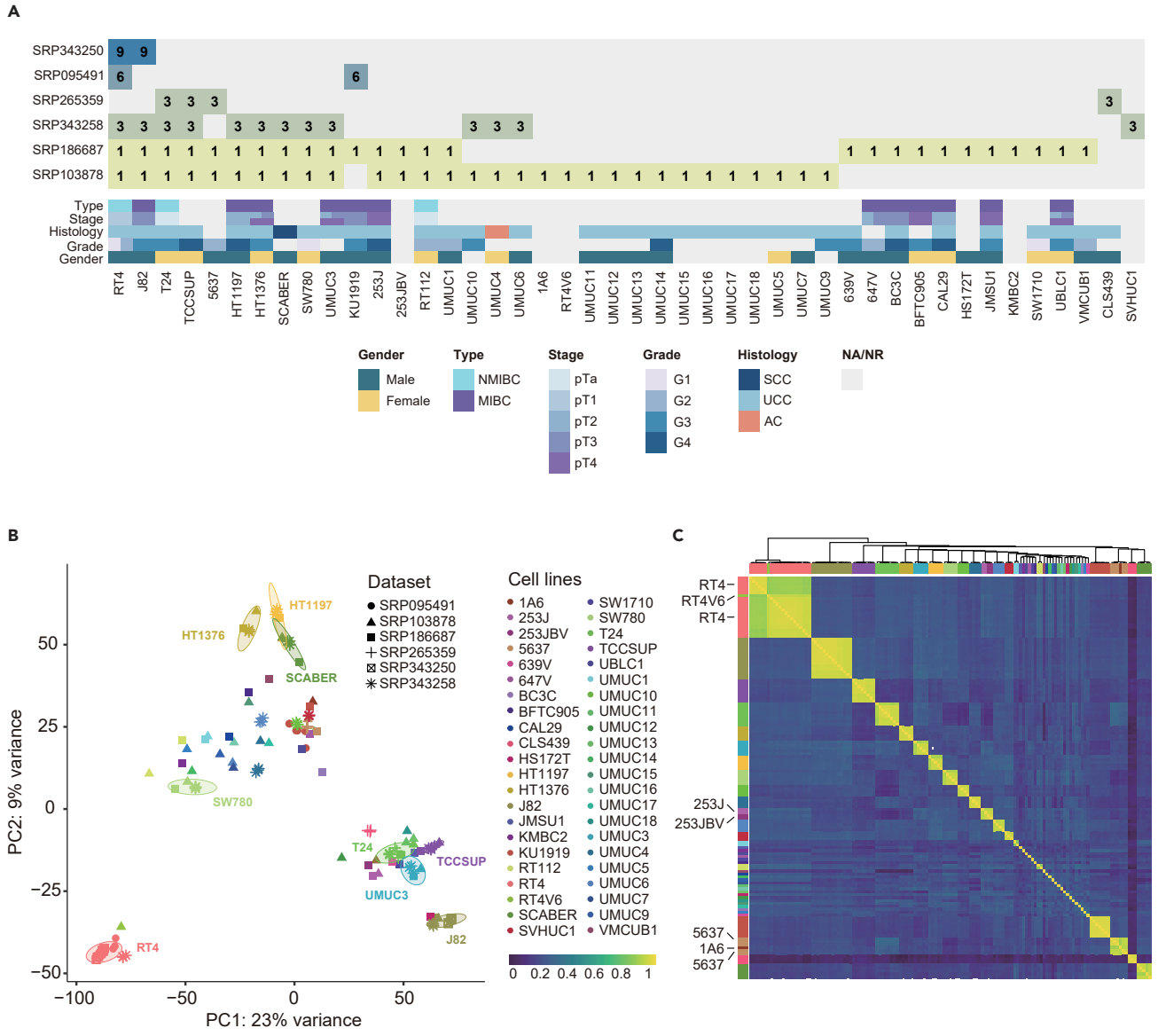
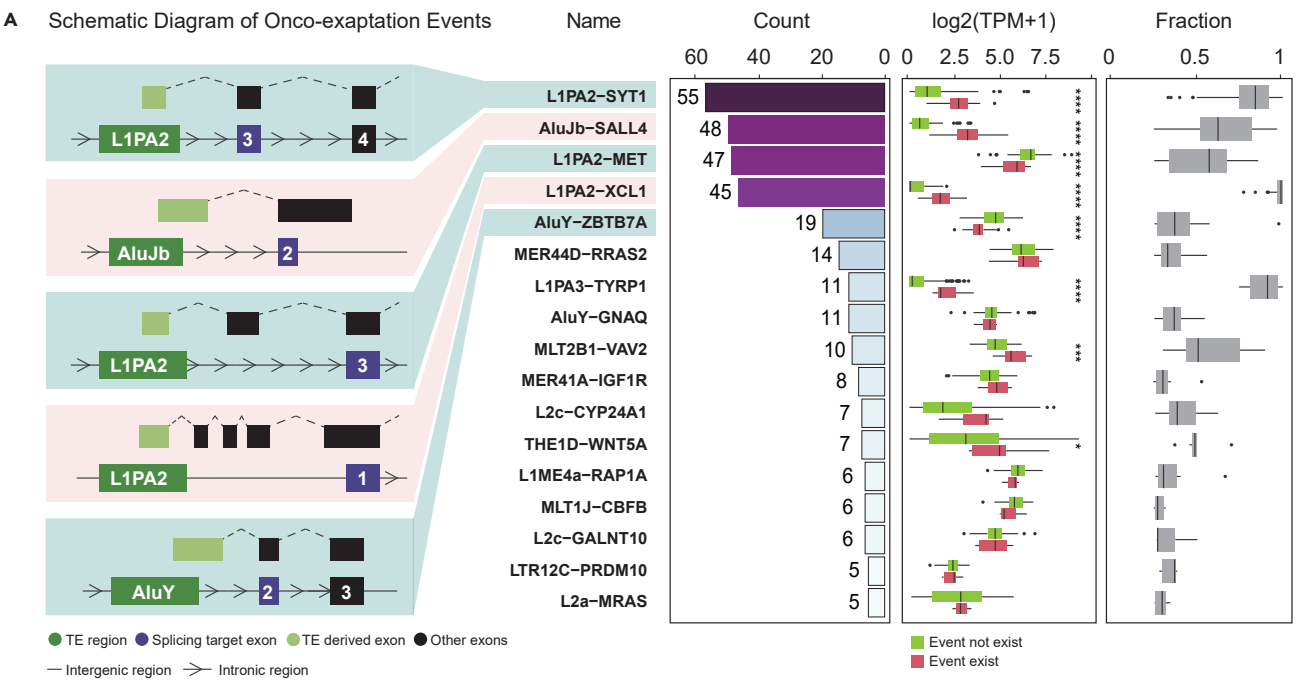


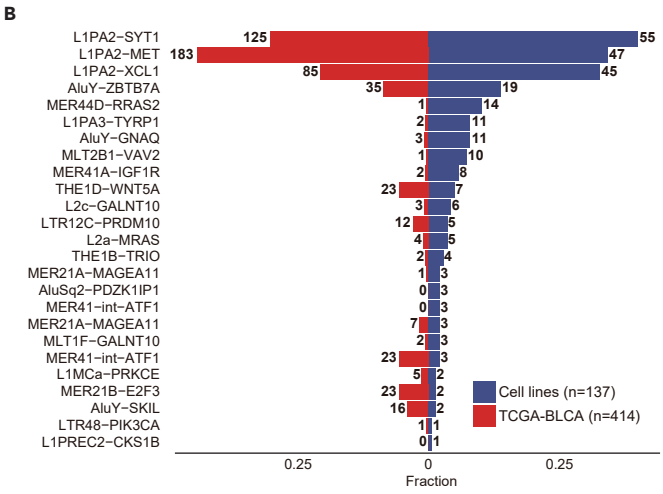
Figure 1. Overview of 44 bladder cancer cell lines from six publicly available dataset

(A) Description of the dataset containing cell line data, where rows represent dataset identifiers, columns represent Cellosaurus cell line names (STAR Methods), and values represent the number of biological replicates; (B) Principal component analysis based on the top 5000 highly variable genes, where the x and y axis represent PC1 and PC2, respectively. The same cell lines are annotated by the same color, and the same datasets are annotated by the same shape. Ellipse was estimated for cell lines profiled in at least three datasets; (C) Heatmap showing the pairwise Pearson correlation of somatic mutations curated in COSMIC v97 database between cell lines, where the distance metric is represented by 1-Pearson correlation coefficient. The heatmap color represents the correlation coefficient, and the cell line color coding is the same as in Figure 1B. See also Figure S1A.

Cell lines can reflect the onco-exaptation patterns of tumor tissues

Based on the TE's location and splicing target exon, 25 candidates were identified as consistent between TCGA-BLCA cohort[21] and the cell lines, including 13 frequent events shown in Figure 2A and 8 events strictly defined as "tumor specific" in TCGA-BLCA cohort (not exist in adjacent normal samples and exist in at least 4 samples or 10-fold enriched in tumor samples),[21] although it has been proven that onco-exaptation events can occur in adjacent normal tissues under the influence of "field defect."[5] By comparing the frequency of the 25 shared events, it was found that L1PA2-SYT1, L1PA2-MET, and L1PA2-XCL1 had the highest frequency in both datasets (Figure 2B). Interestingly, the three most frequent candidates all came from L1PA2 of the LINE1 family, which has been shown to provide abundant transcription factor-binding sites in other cancer cell lines.[25]





C

	TCGA-BLCA (n=414)	TCGA-Pan (n=7769)	Enrichment Score
L1PA2–SYT1	**125**	**901**	**2.60**
L1PA2–MET	**183**	**1018**	**3.73**
L1PA2–XCL1	**85**	**559**	**2.85**

Figure 2. Identification of onco-exaptation events in bladder cancer cell lines

(A) Description of 17 onco-exaptation events identified in five or more samples. From left to right: 1. schematic structure of the top 5 high-frequency events, 2. event names, 3. frequency of the event in 137 cell line samples, 4. oncogene expression levels between cell line groups with and without the corresponding event, and 5. the proportion of gene expression accounted for by the event in cell lines where it is present. The error bar indicated the interquartile range. interqu*p < 0.05, **p < 0.01, ***p < 0.001, ***p < 0.0001, Wilcoxon rank sum test; *p < 0.05, **p < 0.01, ***p < 0.001. (B) Frequency distribution bar plot of 25 events identified in both TCGA and cell line data, arranged from top to bottom based on the frequency of detection in cell lines, with the frequency indicated at the top of each bar. The x axis representing the fraction of the event; (C) Enrichment scores of three high-frequency events originating from L1PA2 in the TCGA BLCA bladder cancer cohort against the TCGA pan-cancer background. See also Figures S2–S4.

In addition, we found that L1PA2-SYT1, L1PA2-MET, and L1PA2-XCL1 are not only highly frequent events in bladder cancer but are also enriched in bladder cancer against the background of pan-cancer[21] (Figure 2C). L1PA2-SYT1 is predicted to be an in-frame event, widely exist in various cancers (11.6%), and is found in 44% of bladder cancers.[21] The SYT1 gene is located on chromosome 12q21.2 and is a type of synaptic vesicle membrane fusion protein containing two protein kinase C homology sequences. *In vitro* experiments have shown that SYT1 overexpression can promote proliferation, migration, and invasion of colon cancer cell lines.[26] However, the function of the SYT1 gene in bladder cancer is still unclear. L1PA2-MET is predicted to be an in-frame event,[21] and LINE1-driven MET transcript isoforms have been demonstrated

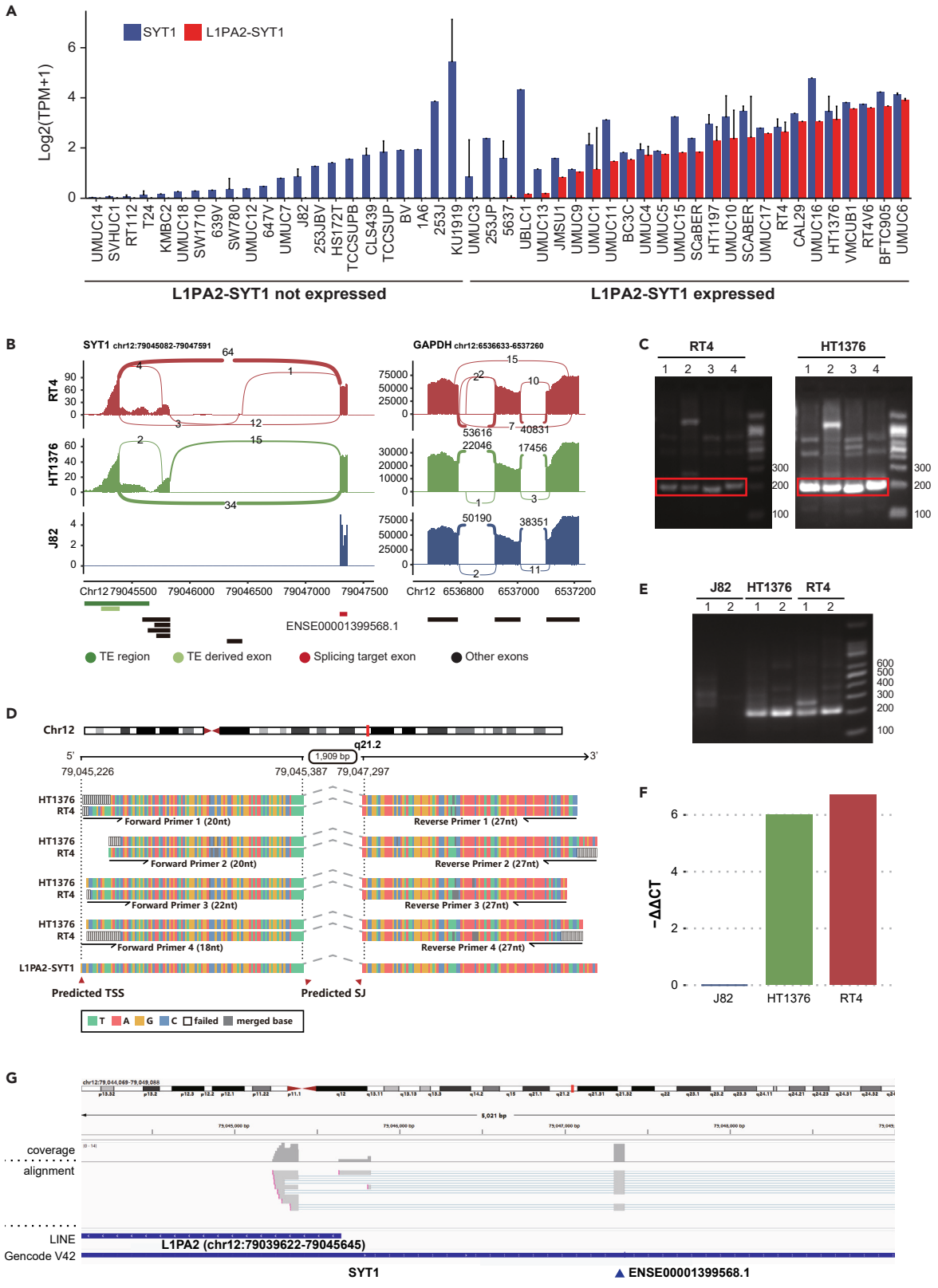

A
SYT1
L1PA2-SYT1
Log2(TPM+1)
UMUC14
SVHUC1
RT112
T24
KMBC2
UMUC18
SW1710
639V
SW780
UMUC12
647V
UMUC7
J82
253JBV
HS172T
TCCSUPB
CLS439
TCCSUP
BV
1A6
253J
KU1919
UMUC3
253JP
5637
UBLC1
UMUC13
JMSU1
UMUC9
UMUC1
UMUC11
BC3C
UMUC4
UMUC5
UMUC15
SCaBER
HT1197
UMUC10
SCABER
UMUC17
RT4
CAL29
UMUC16
HT1376
VMCUB1
RT4V6
BFTC905
UMUC6
L1PA2-SYT1 not expressed
L1PA2-SYT1 expressed
B
SYT1 chr12:79045082-79047591
GAPDH chr12:6536633-6537260
RT4
HT1376
J82
ENSE00001399568.1
TE region
TE derived exon
Splicing target exon
Other exons
C
RT4
HT1376
E
J82
HT1376
RT4
D
Chr12
q21.2
1,909 bp
79,045,226
79,045,387
79,047,297
Forward Primer 1 (20nt)
Reverse Primer 1 (27nt)
Forward Primer 2 (20nt)
Reverse Primer 2 (27nt)
Forward Primer 3 (22nt)
Reverse Primer 3 (27nt)
Forward Primer 4 (18nt)
Reverse Primer 4 (27nt)
L1PA2-SYT1
Predicted TSS
Predicted SJ
failed
merged base
F
−ΔΔCT
J82
HT1376
RT4
G
coverage
alignment
LINE
Gencode V42
L1PA2 (chr12:79039622-79045645)
SYT1
ENSE00001399568.1

Figure 3. Heterogeneous L1PA2-SYT1 expression across different cell lines

(A) The L1PA2-SYT1 expression group is arranged by L1PA2-SYT1 expression level, while the non-expression group is arranged by the overall expression level of the SYT1 gene. Data are represented as mean ± SEM.

(B) Sashimi plot showing the number of reads supporting the presence of the L1PA2-SYT1 splicing site. RT4 and HT1376 were selected as cell lines expressing L1PA2-SYT1, while J82 was selected as a control cell line that does not express L1PA2-SYT1. GAPDH was selected as a positive control. The x axis represents genomic coordinates. The bar plot shows the reads coverage. The curve connects the acceptor and donor sites of the splicing event, with the numbers on the curve indicating the number of reads supporting the splicing event; (C) Gel electrophoresis results of RT-PCR products using primers specific to the L1PA2-SYT1 splicing site. DNA lengths are labeled on the right. The products in the red box were cut and used for Sanger sequencing; (D) Sanger sequencing results of the PCR product shown in Figure 3C. The position of the primers, the predicted TSS, and the predicted SJ was shown; (E) Gel electrophoresis results of RT-PCR products using primers specific to the L1PA2-SYT1 splicing site. J82 as a negative control that does not express L1PA2-SYT1. DNA lengths are labeled on the right; (F) The relative expression of L1PA2-SYT1 in J82, HT1376, and RT4 using qPCR. GAPDH was used as reference; (G) The alignment of 5′RACE fragments captured by a gene-specific primer that target the exon3 of L1PA2-SYT1. See also Figure S5.

to exist in bladder cancer cell lines (GenBank: BF208095.1).[5] The promoter activity of this LINE1 is activated by hypomethylation and chromatin remodeling, and L1PA2-MET has been observed to be highly expressed in both cancer and adjacent tissues, but lowly expressed in normal bladder epithelium,[5] which has been demonstrated as a diagnostic biomarker for bladder cancer.[27] Recently, the role of MET in bladder cancer has been systematically reviewed.[28] In addition, in colorectal cancer, the expression of MET is driven by LINE1, promoting tumor metastasis.[3] Although XCL1 is a protein-coding oncogene, L1PA2-XCL1 is predicted to be a noncoding transcript, thus may have a different function compared with its coding counterpart.[21]

Heterogeneous L1PA2-SYT1 expression across different cell lines

Considering the prevalence of L1PA2-SYT1 in bladder cancer cell lines and the unclear expression and function in bladder cancer, we focused on L1PA2-SYT1 in our next analysis. We arranged the bladder cancer cell lines first by the expression level of L1PA2-SYT1 and then by the total TPM of all SYT1 isoforms (Figure 3A). Typical expression patterns of SYT1 gene and L1PA2-SYT1 transcript in bladder cancer cell lines could be categorized into three groups: 1) almost all the expression of SYT1 gene was driven by L1PA2, such as UMUC6, RT4, and HT1376; 2) SYT1 was highly expressed and not driven by L1PA2, such as 5637, 253JP, and KU1919; and 3) SYT1 was lowly expressed, such as J82, T24, and immortalized normal cell line SVHUC1. The raw alignment confirmed the reads across the splicing junction between L1PA2 and the known exon (ENSE00001399568.1) of SYT1 (Figure 3B). Considering the repetitive nature of L1PA2, we designed four pairs of primers to validate the novel splice junction (SJ) of L1PA2-SYT1 in RT4 and HT1376 (Table S4A). Gel electrophoresis proved the target products are located at 200 bp (Figure 3C). The sequence was further confirmed identical as the prediction by Sanger sequencing of the PCR product (Figure 3D). Then, we selected J82 as a negative control and demonstrated that the SJ did not exist in J82 using both RT-PCR and qPCR (Figures 3E and 3F).

TEPROF2 predicts the TSS and the full length of onco-exaptation using short-read sequencing data. Considering the low coverage at the ends of transcripts and the uncertainty in transcript assembly associated with short-read sequencing, we validated the TSS and full length of L1PA2-SYT1 using 3rd generation, long-read transcriptome sequencing (ONT-seq), which revealed that the novel SJ of L1PA2-SYT1 exists in HT1376 and RT4 but is absent in J82 (Figure S5A). Compared with the prediction of TEPROF2, the full-length L1PA2-SYT1 from both HT1376 and RT4 exhibited longer 5′ ends (Figure S5B). However, TEPROF2 accurately predicts all exons of L1PA2-SYT1 (Figures S5C and S5D). Besides, using a gene-specific primer, we confirmed with 5′RACE sequencing that the TSS of L1PA2-SYT1 indeed originates from the L1PA2 element (chr12:79039622-79045645) (Figures 3G and S5C).

The expression of L1PA2-SYT1 is associated with DNA hypomethylation

The activity of LINE1 sequences is regulated by various epigenetic mechanisms in the host, including methylation and histone modifications. Dysregulation of methylation in tumors may activate LINE1 sequences that are normally suppressed by methylation, thereby exhibiting promoter activity. We searched for CpG islands located in the promoter region, i.e., 2000 bp upstream to 500 bp downstream of the TSS of L1PA2-SYT1 and found that CpG:_22 (chr12:79045269-79045493) met the standard. Considering that low methylation of CpG islands in promoters is related to transcriptional activity, we hypothesized that the low methylation of this CpG island accounts for the heterogeneous expression of L1PA2-SYT1 across bladder cancer cell lines. The paired RRBS data in the bladder cancer cell line RNA-seq dataset in the CCLE dataset (SRA: SRP186887) were used to analyze the relationship between the CpG:_22 methylation level and L1PA2-SYT1 expression in 25 cell lines. The CpG:_22 contains four measurable CpG sites, i.e., chr12:79045341, chr12:79045345, chr12:79045353, and chr12:79045362. After filtering out 6 cell lines with a minimum read coverage <5, we performed hierarchical clustering and defined two methylation groups. We found the group with overall low methylation level showed high level of L1PA2-SYT1 expression (Figure 4A). We then calculated the average methylation value of four CpG sites and found that the average CpG:_22 methylation level was significantly negatively correlated with the expression of this transcript ($R = -0.73$, $p < 0.001$) (Figure 4B). In contrast, in cell lines that did not express L1PA2-SYT1 ($n = 11$), the expression level of the SYT1 gene was not related to the methylation level of this CpG island (Figure 4C). These results suggested that demethylation of LINE1 is one of the mechanisms that specifically regulates the upregulation of L1PA2-SYT1 expression.

High expression of L1PA2-SYT1 is associated with poor prognosis of MIBC

We further explored the function of L1PA2-SYT1 in bladder cancer by ectopically expressing the fusion protein in J82 cells. It is interesting to find that the anti-synaptotagmin-1 monoclonal antibody (mAb) detects two bands of about 50 kDa, which could be two L1PA2-SYT1 protein

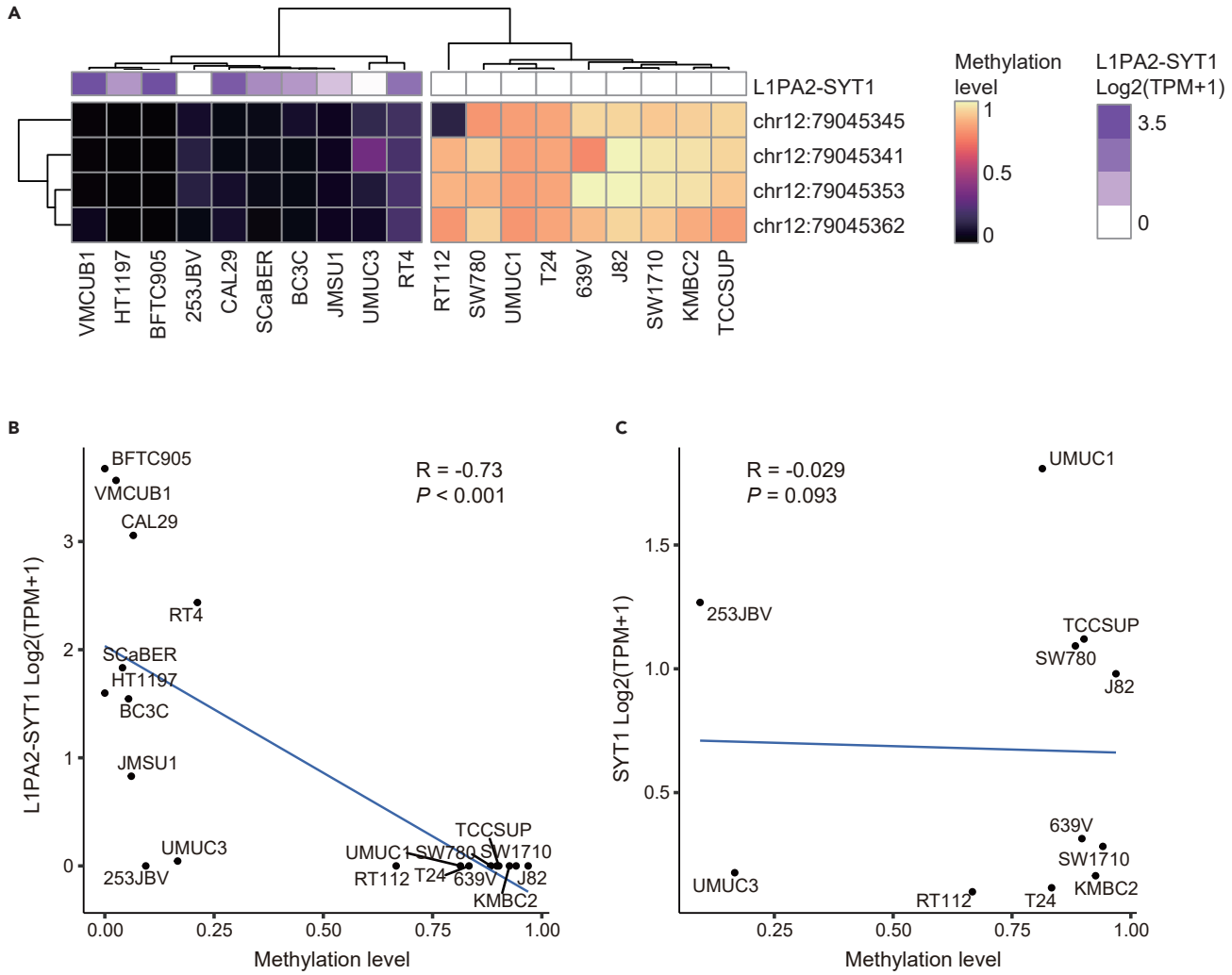


Figure 4. The expression of L1PA2-SYT1 is associated with DNA hypomethylation

(A) Heatmap showing the methylation levels of four measurable CpG sites within the L1PA2-SYT1 promoter region CpG_22 in 19 bladder cancer cell lines from the CCLE project, with L1PA2-SYT1 expression levels labeled at the top of the heatmap; (B) Pearson correlation between the average methylation level of CpG_22 and L1PA2-SYT1 expression levels across the 19 cell lines in the CCLbE project; (C) Pearson correlation between the average methylation level of CpG_22 in 11 cell lines that do not express L1PA2-SYT1 and the total expression level of SYT1.

isoforms (Figures 5A, 5B, S6A, and S6B). Besides, L1PA2-SYT1-overexpressed J82 exhibits higher invasion capability compared with the negative control (Figure 5C).

To investigate whether the high expression of L1PA2-SYT1 is associated with poor prognosis in cancer, we retrieved RNA-seq data (SRA: PRJNA891747) from 52 muscle-invasive bladder cancer patients with follow-up information as the validation group.[29] Again, we that found patients having L1PA2-SYT1 showed higher level of SYT1 expression, which validated the driving effect of L1PA2 on SYT1 (Figures 2A and 5D). Also, the existence of L1PA2-SYT1 is associated with poor overall survival of patients (Figure 5E). After adjusting for TP53 mutation, tumor grade, and lymph node metastasis stage, we found that the expression level of L1PA2-SYT1 remained an independent risk factor for overall survival of patients (Figure 5F).

Finally, we analyzed the differentially expressed genes between exist group and not exist group of L1PA2-SYT1 (Figure 5G). The results of Gene Ontology enrichment analysis showed that the exist group of L1PA2-SYT1 enriched genes related to cell cycle, methylation, retinoid metabolism, sodium ion transport, and the ERBB signaling pathway (Figure 5H).

DISCUSSION

Bladder cancer is a common tumor in urinary system, and many studies have suggested TEs are activated in bladder cancer and may be involved in the development and progression of bladder cancer. Interspersed repeat elements, including LINEs and SINEs, have been shown

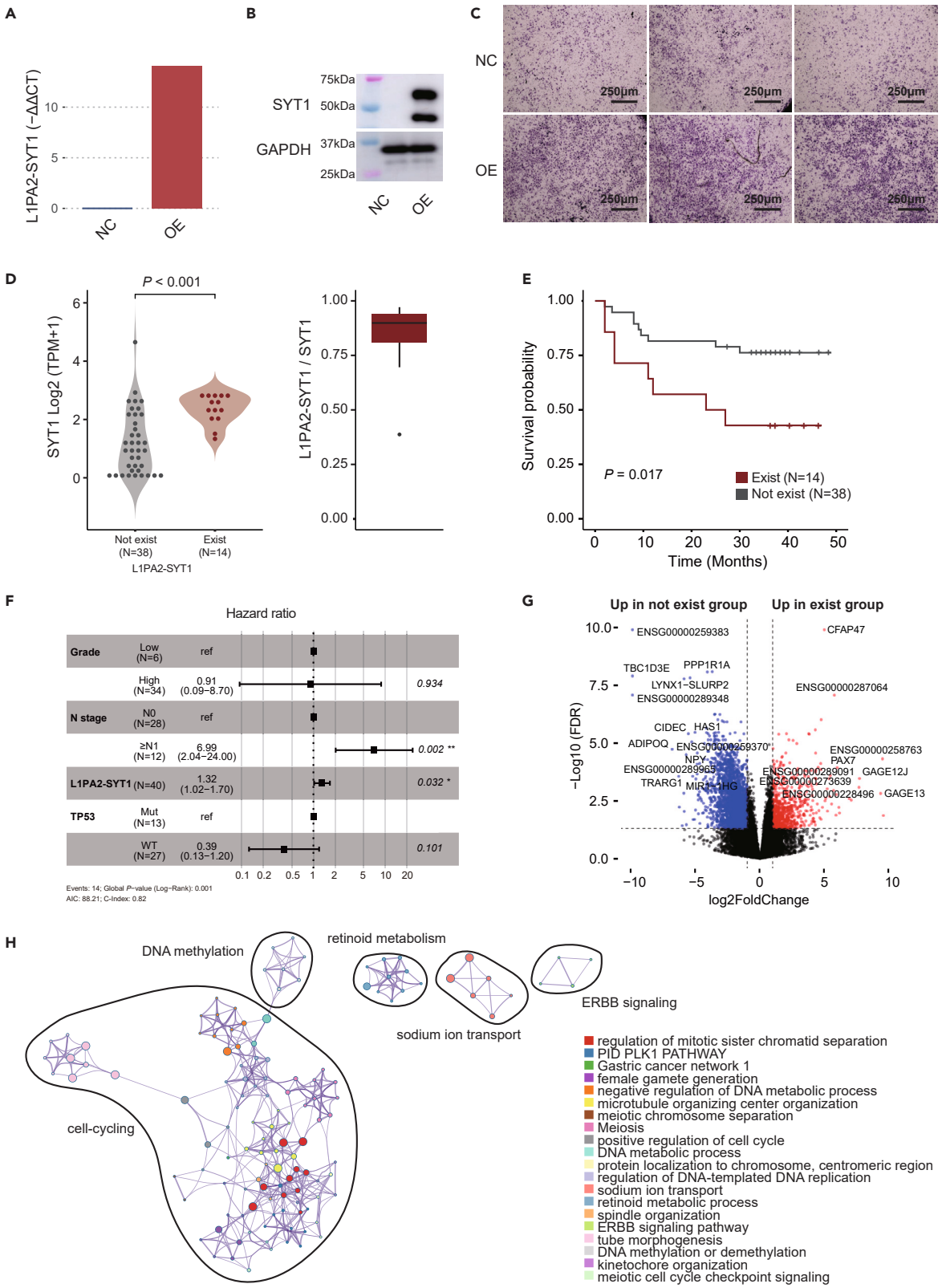

A
L1PA2-SYT1 (−ΔΔCT)
10
5
0
NC
OE
B
SYT1
GAPDH
75kDa
50kDa
37kDa
25kDa
NC
OE
C
NC
OE
250μm
D
P < 0.001
SYT1 Log2 (TPM+1)
Not exist (N=38)
Exist (N=14)
L1PA2-SYT1
L1PA2-SYT1 / SYT1
E
Survival probability
P = 0.017
Exist (N=14)
Not exist (N=38)
Time (Months)
F
Hazard ratio
Grade
Low (N=6)
ref
High (N=34)
0.91 (0.09−8.70)
0.934
N stage
N0 (N=28)
ref
≥N1 (N=12)
6.99 (2.04−24.00)
0.002 **
L1PA2-SYT1 (N=40)
1.32 (1.02−1.70)
0.032 *
TP53
Mut (N=13)
ref
WT (N=27)
0.39 (0.13−1.20)
0.101
Events: 14; Global P−value (Log−Rank): 0.001
AIC: 88.21; C-Index: 0.82
G
Up in not exist group
Up in exist group
−Log10 (FDR)
log2FoldChange
ENSG00000259383
CFAP47
TBC1D3E
PPP1R1A
LYNX1−SLURP2
ENSG00000289348
ENSG00000287064
CIDEC
HAS1
ADIPOQ
NPY
TRARG1
MIR1−1HG
PAX7
GAGE12J
GAGE13
H
DNA methylation
retinoid metabolism
ERBB signaling
sodium ion transport
cell-cycling
regulation of mitotic sister chromatid separation
PID PLK1 PATHWAY
Gastric cancer network 1
female gamete generation
negative regulation of DNA metabolic process
microtubule organizing center organization
meiotic chromosome separation
Meiosis
positive regulation of cell cycle
DNA metabolic process
protein localization to chromosome, centromeric region
regulation of DNA-templated DNA replication
sodium ion transport
retinoid metabolic process
spindle organization
ERBB signaling pathway
tube morphogenesis
DNA methylation or demethylation
kinetochore organization
meiotic cell cycle checkpoint signaling

Figure 5. High expression of L1PA2-SYT1 is associated with poor prognosis of MIBC

(A) qPCR validates the overexpression of L1PA2-SYT1 in transfected J82 cell line; (B) Western blot found L1PA2-SYT1 could be translated into two protein isoforms near 50 kDa; (C) Invasion assay suggests enhanced invasion capability with increased level of L1PA2-SYT1. The scale bar indicated 250μm; (D) Comparison of the overall expression levels of the SYT1 gene between the L1PA2-SYT1-present and L1PA2-SYT1-absent groups in the validation cohort (n = 52). The dot denotes the expression level of SYT1 for individual samples (Wilcoxon rank-sum test). Boxplot shows the fraction of SYT1 expression contributed by L1PA2-SYT1 expression in the L1PA2-SYT1-present group. The error bar indicated the interquartile range; (E) Comparison of OS between the L1PA2-SYT1 exist and L1PA2-SYT1 not exist groups in the validation cohort (n = 52), Log-rank test; (F) Forest plot showing the multivariable Cox analysis including tumor grade, N stage, L1PA2-SYT1 expression, and TP53 mutation in the validation cohort (n = 52). The error bar indicated the 95% confidence interval; (G) Volcano plot showing the differential gene expression analysis between the L1PA2-SYT1-present and L1PA2-SYT1-absent groups in the validation cohort (n = 52), with differentially expressed genes (|log2FoldChange| >5 and FDR <1E-3) labeled; (H) Network plot of GO enrichment analysis for the upregulated genes in the L1PA2-SYT1-present group in Figure 5D, with similar semantic terms labeled with the same color. Black circles indicating further manual summarization. See also Figure S6.

to undergo abnormal hypomethylation in bladder cancer.[30] LINE1 is the most studied TE in bladder cancer, and studies have found that the abnormal hypomethylation of LINE1 in tumors may be activated by reactive oxygen metabolism pathways, which further promotes the loss of tumor suppressor genes (e.g., CDKN2A[31]) and the activation of oncogenes (e.g., MET[5]), promoting the progression of bladder cancer. MET was first discovered to be activated by an ectopic promoter derived from LINE1 in leukemia, promoting the development of tumors, and was subsequently confirmed to be driven by an antisense promoter derived from LINE1 in bladder cancer.[5] In our research, we found that L1PA2-MET, other than L1PA2-XCL1 and L1PA2-SYT1, exhibited a gender-related expression pattern. In a recent study, it was observed that L1PA2 elements contained abundant transcription factor-binding motifs, including estrogen receptor 1, in the breast cancer cell line MCF7.[32] Therefore, we postulate the existence of transcriptional regulatory heterogeneity among L1PA2-derived transcripts. These results suggest the potential value of transposons in the diagnosis and prognosis of bladder cancer. In fact, LINE1-MET has been included in a diagnostic panel for bladder cancer.[27]

Some studies have identified onco-exaptation events in the TCGA-BLCA cohort,[21] but the function of most events has not been fully studied. One important reason is that whether the corresponding events exist or how frequently they occur is poorly understood in bladder cancer cell lines. Therefore, a comprehensive identification of events in bladder cancer cell lines can help us better utilize them to understand the biological importance of onco-exaptation. In this study, we collected 137 RNA-seq data from 44 bladder cancer cell lines from six different datasets publicly available and identified onco-exaptation events. We found that the events identified in tumor tissues could be validated in cell lines, and L1PA2 contributed the most events, which may be associated with the global hypomethylation of L1PA2 in bladder cancer, which could be reflected in the results of pathway enrichment analysis, where methylation-related pathways exhibited alterations. Furthermore, differential expression analysis revealed that the existence of L1PA2-SYT1 correlated with the activity of the cell cycle and the ERBB pathway. Recent study has confirmed that primate-specific transposable elements, including L1PA2, contribute to a significant number of open chromatin regions in decidual stromal cells. These regions contribute to the binding motif of the progesterone receptor and mediate regulatory effects on pathways like the cell cycle and ERBB signaling.[33] Furthermore, we also identified an onco-exaptation event related to SALL4 (AluJb-SALL4). This event was also identified in the pan-cancer study, although the TSS is predicted to be contributed by a different TE subfamily (MLT1J). The function of SALL4 in bladder cancer remains unexplored, making this event an intriguing candidate for further investigation.[21]

SYT1 protein is a member of the synaptotagmin family and is a membrane transport protein consisting of an N terminus, a variable link domain, a single transmembrane domain, and two C2 domains (C2A, C2B), with C2 domains serving as Ca^{2+}-binding sites.[34] Overexpression of SYT1 was found to promote the invasion and metastasis ability of colon cancer cell lines *in vitro*. We found that L1PA2-SYT1 was the most frequent event in bladder cancer cell lines and was associated with poor prognosis in patients, and *in vitro* overexpression of L1PA2-SYT1 enhances the invasion capability of bladder cancer cell line, suggesting the potential tumor-driving role and clinical application potential of L1PA2-SYT1.

Limitations of the study

In this study, we observed widespread onco-exaptation event in all bladder cancer cell lines, highlighting the significance of these events in promoting bladder cancers. Although initial exploration unraveled epigenetic mechanisms in regulating such events, further evidence is required to elucidate the direct upstream regulator. Besides, multiple interesting events remained unexplored, like the potential effect of L1PA2-XCL1 after losing coding potential, and the association between AluJb-SALL4 identified in this study and MLT1J-SALL4 identified in TCGA-BLCA project.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108482.

ACKNOWLEDGMENTS

This study was funded by the National Key R&D Program of China (SQ2022YFA1300089), the National Science Fund for Excellent Young Scholars (82022055), the National Natural Science Foundation of China (82272950, 81802515, 81801854, 82172871, and 81972391), the Qihang program of Second Military Medical University (2021), the Discipline Development Plan of Changhai Hospital (2019YXK041), the Science and Technology Commission of Shanghai Municipality (20Y11904800 and 22140903700), and the Shanghai Municipal Health Commission (2022YQ010).

AUTHOR CONTRIBUTIONS

Conceptualization, Z.W., Y.Y., and M.W.; methodology, Z.W. and S.Z.; software, Z.W.; validation, Y.Y., M.W., and X.Y.; formal analysis, Z.W.; investigation, Y.Y., M.W., and Q.C.; resources, Y.Y., M.W., and W.H.; data curation, M.W., Q.C., and Y.W.; writing – original draft, Z.W., Y.Y., and M.W.; writing – review and editing, S.Z. and C.X.; visualization, Z.W., Y.Y., and M.W.; supervision, J.L.; project administration, S.Z. and C.X.; funding acquisition, J.L., S.Z., and C.X.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 26, 2023
Revised: October 17, 2023
Accepted: November 15, 2023
Published: November 17, 2023

REFERENCES

1. Bannert, N., and Kurth, R. (2004). Retroelements and the human genome: new perspectives on an old relation. Proc. Natl. Acad. Sci. USA *101*, 14572–14579.
2. Wells, J.N., and Feschotte, C. (2020). A Field Guide to Eukaryotic Transposable Elements. Annu. Rev. Genet. *54*, 539–561.
3. Hur, K., Cejas, P., Feliu, J., Moreno-Rubio, J., Burgos, E., Boland, C.R., and Goel, A. (2014). Hypomethylation of long interspersed nuclear element-1 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis. Gut *63*, 635–646.
4. Daskalos, A., Nikolaidis, G., Xinarianos, G., Savvari, P., Cassidy, A., Zakopoulou, R., Kotsinas, A., Gorgoulis, V., Field, J.K., and Liloglou, T. (2009). Hypomethylation of retrotransposable elements correlates with genomic instability in non-small cell lung cancer. Int. J. Cancer *124*, 81–87.
5. Wolff, E.M., Byun, H.M., Han, H.F., Sharma, S., Nichols, P.W., Siegmund, K.D., Yang, A.S., Jones, P.A., and Liang, G. (2010). Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet. *6*, e1000917.
6. Lee, E., Iskow, R., Yang, L., Gokcumen, O., Haseley, P., Luquette, L.J., 3rd, Lohr, J.G., Harris, C.C., Ding, L., Wilson, R.K., et al. (2012). Landscape of somatic retrotransposition in human cancers. Science *337*, 967–971.

7. Ade, C., Roy-Engel, A.M., and Deininger, P.L. (2013). Alu elements: an intrinsic source of human genome instability. Curr. Opin. Virol. *3*, 639–645.
8. De Brakeleer, S., De Grève, J., Lissens, W., and Teugels, E. (2013). Systematic detection of pathogenic alu element insertions in NGS-based diagnostic screens: the BRCA1/BRCA2 example. Hum. Mutat. *34*, 785–791.
9. Halling, K.C., Lazzaro, C.R., Honchel, R., Bufill, J.A., Powell, S.M., Arndt, C.A., and Lindor, N.M. (1999). Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. Hum. Hered. *49*, 97–102.
10. Roulois, D., Loo Yau, H., Singhania, R., Wang, Y., Danesh, A., Shen, S.Y., Han, H., Liang, G., Jones, P.A., Pugh, T.J., et al. (2015). DNA-Demethylating Agents Target Colorectal Cancer Cells by Inducing Viral Mimicry by Endogenous Transcripts. Cell *162*, 961–973.
11. Mehdipour, P., Marhon, S.A., Ettayebi, I., Chakravarthy, A., Hosseini, A., Wang, Y., de Castro, F.A., Loo Yau, H., Ishak, C., Abelson, S., et al. (2020). Epigenetic therapy induces transcription of inverted SINEs and ADAR1 dependency. Nature *588*, 169–173.
12. Lev-Maor, G., Sorek, R., Shomron, N., and Ast, G. (2003). The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. Science *300*, 1288–1291.
13. Rebollo, R., Romanish, M.T., and Mager, D.L. (2012). Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu. Rev. Genet. *46*, 21–42.
14. Fitzpatrick, T., and Huang, S. (2012). 3′-UTR-located inverted Alu repeats facilitate mRNA translational repression and stress granule accumulation. Nucleus *3*, 359–369.
15. Babaian, A., and Mager, D.L. (2016). Endogenous retroviral promoter exaptation in human cancer. Mob. DNA *7*, 24.
16. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F.A., and Wheeler, T.J. (2016). The Dfam database of repetitive DNA families. Nucleic Acids Res. *44*, D81–D89.
17. Lamprecht, B., Walter, K., Kreher, S., Kumar, R., Hummel, M., Lenze, D., Köchert, K., Bouhlel, M.A., Richter, J., Soler, E., et al. (2010). Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat. Med. *16*, 571–579. 1p following 579.
18. Criscione, S.W., Theodosakis, N., Micevic, G., Cornish, T.C., Burns, K.H., Neretti, N., and Rodić, N. (2016). Genome-wide characterization of human L1 antisense promoter-driven transcripts. BMC Genom. *17*, 463.
19. Weber, B., Kimhi, S., Howard, G., Eden, A., and Lyko, F. (2010). Demethylation of a LINE-1 antisense promoter in the cMet locus impairs Met signalling through induction of illegitimate transcription. Oncogene *29*, 5775–5784.
20. Moqtaderi, Z., Wang, J., Raha, D., White, R.J., Snyder, M., Weng, Z., and Struhl, K. (2010). Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nat. Struct. Mol. Biol. *17*, 635–640.
21. Jang, H.S., Shah, N.M., Du, A.Y., Dailey, Z.Z., Pehrsson, E.C., Godoy, P.M., Zhang, D., Li, D., Xing, X., Kim, S., et al. (2019). Transposable elements drive widespread expression of oncogenes in human cancers. Nat. Genet. *51*, 611–617.
22. Klijn, C., Durinck, S., Stawiski, E.W., Haverty, P.M., Jiang, Z., Liu, H., Degenhardt, J., Mayba, O., Gnad, F., Liu, J., et al. (2015). A comprehensive transcriptional portrait of human cancer cell lines. Nat. Biotechnol. *33*, 306–312.
23. Shah, N.M., Jang, H.J., Liang, Y., Maeng, J.H., Tzeng, S.C., Wu, A., Basri, N.L., Qu, X., Fan, C., Li, A., et al. (2023). Pan-cancer analysis identifies tumor-specific antigens derived from transposable elements. Nat. Genet. *55*, 631–639.
24. Zuiverloon, T.C.M., de Jong, F.C., Costello, J.C., and Theodorescu, D. (2018). Systematic Review: Characteristics and Preclinical Uses of Bladder Cancer Cell Lines. Bladder Cancer *4*, 169–183.
25. Jiang, J.C., Rothnagel, J.A., and Upton, K.R. (2021). Integrated transcription factor profiling with transcriptome analysis identifies L1PA2 transposons as global regulatory modulators in a breast cancer model. Sci. Rep. *11*, 8083.
26. Lu, H., Hao, L., Yang, H., Chen, J., and Liu, J. (2019). miRNA-34a suppresses colon carcinoma proliferation and induces cell apoptosis by targeting SYT1. Int. J. Clin. Exp. Pathol. *12*, 2887–2897.
27. Su, S.F., de Castro Abreu, A.L., Chihara, Y., Tsai, Y., Andreu-Vieyra, C., Daneshmand, S., Skinner, E.C., Jones, P.A., Siegmund, K.D., and Liang, G. (2014). A panel of three markers hyper- and hypomethylated in urine sediments accurately predicts bladder cancer recurrence. Clin. Cancer Res. *20*, 1978–1989.
28. Feng, Y., Yang, Z., and Xu, X. (2022). c-Met: A Promising Therapeutic Target in Bladder Cancer. Cancer Manag. Res. *14*, 2379–2388.
29. Tao, Y., Li, X., Zhang, Y., He, L., Lu, Q., Wang, Y., Pan, L., Wang, Z., Feng, C., Xie, Y., et al. (2022). TP53-related signature for predicting prognosis and tumor microenvironment characteristics in bladder cancer: A multi-omics study. Front. Genet. *13*, 1057302.
30. Choi, S.H., Worswick, S., Byun, H.M., Shear, T., Soussa, J.C., Wolff, E.M., Douer, D., Garcia-Manero, G., Liang, G., and Yang, A.S. (2009). Changes in DNA methylation of tandem DNA repeats are different from interspersed repeats in cancer. Int. J. Cancer *125*, 723–729.
31. Florl, A.R., Franke, K.H., Niederacher, D., Gerharz, C.D., Seifert, H.H., and Schulz, W.A. (2000). DNA methylation and the mechanisms of CDKN2A inactivation in transitional cell carcinoma of the urinary bladder. Lab. Invest. *80*, 1513–1522.
32. Jiang, J.C., Rothnagel, J.A., and Upton, K.R. (2021). Widespread Exaptation of L1 Transposons for Transcription Factor Binding in Breast Cancer. Int. J. Mol. Sci. *22*, 5625.
33. Mika, K., and Lynch, V.J. (2022). Transposable Elements Continuously Remodel the Regulatory Landscape, Transcriptome, and Function of Decidual Stromal Cells. Genome Biol. Evol. *14*, evac164.
34. Suo, H., Xiao, N., and Wang, K. (2022). Potential roles of synaptotagmin family members in cancers: Recent advances and prospects. Front. Med. *9*, 968081.
35. Earl, J., Rico, D., Carrillo-de-Santa-Pau, E., Rodríguez-Santiago, B., Méndez-Pertuz, M., Auer, H., Gómez, G., Grossman, H.B., Pisano, D.G., Schulz, W.A., et al. (2015). The UBC-40 Urothelial Bladder Cancer cell line index: a genomic resource for functional studies. BMC Genom. *16*, 403.
36. Mohammad, T.A., Tsai, Y.S., Ameer, S., Chen, H.I.H., Chiu, Y.C., and Chen, Y. (2019). CeL-ID: cell line identification using RNA-seq data. BMC Genom. *20*, 81.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Rabbit monoclonal to Synaptotagmin-1	Cell Signaling Technology	Cat#14558; RRID:AB_2798510
Rabbit monoclonal to GAPDH	Cell Signaling Technology	Cat#2118; RRID:AB_561053
Mouse Anti-Rabbit IgG (Light-Chain Specific) (D4W3E) mAb (HRP Conjugate)	Cell Signaling Technology	Cat#93702; RRID:AB_2800208
Bacterial and virus strains		
DH5alpha	Vazyme, Jiangsu, China	Cat#C502
Chemicals, peptides, and recombinant proteins		
Ultra GelRed (10,000 ×)	Vazyme, Jiangsu, China	Cat#GR501
VeZol Reagent	Vazyme, Jiangsu, China	Cat#R411
ReverTra Ace -a- High Efficient Reverse Transcription Kit	Toyobo, Osaka, Japan	Cat#FSK-101
KOD -Plus- Ver.2	Toyobo, Osaka, Japan	Cat#KOD-211
Gel Extraction Kit	Cwbio, Jiangsu, China	Cat#CW2302
HiScript-TS 5′/3′ RACE Kit	Vazyme, Jiangsu, China	Cat#RA101
FastPure Cell/Tissue Total RNA Isolation Kit V2	Vazyme, Jiangsu, China	Cat#RC112
HiScript III All-in-one RT SuperMix Perfect for qPCR	Vazyme, Jiangsu, China	Cat#R333
Taq Pro Universal SYBR qPCR Master Mix	Vazyme, Jiangsu, China	Cat#Q712
FastPure Cell/Tissue DNA Isolation Mini Kit	Vazyme, Jiangsu, China	Cat#DC102
Ligation Sequencing Kit	Oxford Nanopore Technologies	Cat#SQK-LSK110
The PCR Barcoding Expansion 1-96	Oxford Nanopore Technologies	EXP-PCB096
LongAmp® *Taq* DNA Polymerase	New England BioLabs, Massachusetts, US	Cat#M0323S
AMPure XP Beads	Beckman Coulter Life Sciences	Cat#A63880
pCE2 TA/blunt-Zero vector	Vazyme, Jiangsu, China	Cat#C601
EcoRI	New England BioLabs, Massachusetts, US	Cat#R3101S
BamHI	New England BioLabs, Massachusetts, US	Cat#R3136S
ClonExpress II One Step Cloning Kit	Vazyme, Jiangsu, China	Cat#C112
Ampicillin (100mg/ml,1000X)	Beyotime, Shanghai, China	Cat#ST008
EndoFree Mini Plasmid Kit II	Tiangen Biotech, Beijing, China	Cat#DP118
PEI 40K Transfection Reagent	Servicebio, Hubei, China	Cat#G1802
Pierce BCA Protein Assay Kit	Thermo Scientific	Cat#23225
10× ice-bath free fast transfer buffer	Servicebio, Hubei, China	Cat#G2154
Clarity Western ECL Substrate	Biorad	Cat#1705061
Deposited data		
SRP343250	Sequence Read Archive (SRA)	SRA: SRP343250
SRP095491	Sequence Read Archive (SRA)	SRA: SRP095491
SRP265359	Sequence Read Archive (SRA)	SRA: SRP265359
SRP343258	Sequence Read Archive (SRA)	SRA: SRP343258
SRP186687	Sequence Read Archive (SRA)	SRA: SRP186687
SRP103878	Sequence Read Archive (SRA)	SRA: SRP103878
PRJNA891747	Sequence Read Archive (SRA)	SRA: PRJNA891747

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
PRJNA552055	Sequence Read Archive (SRA)	SRA: PRJNA552055
ONT-seq data for J82, HT1376, and RT4	Genome Sequence Archive (GSA)-human; This paper	GSA: HRA005343
Original code	This paper	Zenodo: https://doi.org/10.5281/zenodo.8417976
Experimental models: Cell lines		
J82	the Cell Bank of the Type Culture Collection of the Chinese Academy of Sciences (Shanghai, China)	Cat#TCHu218; RRID:CVCL_0359
HT1376	the Cell Bank of the Type Culture Collection of the Chinese Academy of Sciences (Shanghai, China)	RRID: CVCL_1292
RT4	the Cell Bank of the Type Culture Collection of the Chinese Academy of Sciences (Shanghai, China)	Cat#TCHu226; RRID: CVCL_0036
Oligonucleotides		
All primers used in this paper	See Table S1	N/A
Software and algorithms		
R version 4.1.2	R project	https://www.r-project.org/
fastp version 0.23.2	Github	https://github.com/OpenGene/fastp
STAR version 2.7.10b	Github	https://github.com/alexdobin/STAR
Bismark version 0.24.0	Github	https://github.com/FelixKrueger/Bismark
Methylkit version 1.24.0	Bioconductor	https://bioconductor.org/packages/release/bioc/html/methylKit.html
tximport version 1.26.1	Bioconductor	https://bioconductor.org/packages/release/bioc/html/tximport.html
Varscan2 version 2.3.9	Github	https://github.com/dkoboldt/varscan
vcftools version 0.1.16	Github	https://github.com/vcftools/vcftools
StringTie version 2.2.1	Github	https://github.com/gpertea/stringtie
TEPROF2 version 0.1	Github	https://github.com/twlab/TEProf2Paper
ggsashimi version 1.1.5	Github	https://github.com/guigolab/ggsashimi
sangeranalyseR version 1.8.0	Bioconductor	https://bioconductor.org/packages/release/bioc/html/sangeranalyseR.html
ggmsa version 1.3.4	Bioconductor	https://bioconductor.org/packages/release/bioc/html/ggmsa.html
Salmon version 1.10.0	Github	https://github.com/COMBINE-lab/salmon
TEtranscript version 2.2.3	Github	https://github.com/mhammell-laboratory/TEtranscripts
DESeq2 version 1.38.3	Bioconductor	https://bioconductor.org/packages/release/bioc/html/DESeq2.html
minimap2 version 2.26(r1175)	Github	https://github.com/lh3/minimap2
survival version 3.5-0	CRAN	https://cran.r-project.org/web/packages/survival/
Integrative Genomics Viewer (IGV) version 2.16.0	Broad Institute	https://software.broadinstitute.org/software/igv/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Other		
The Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13)	GENCODE project	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/GRCh38.primary_assembly.genome.fa.gz
Comprehensive gene annotation (Gencode V42)	GENCODE project	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.annotation.gtf.gz
the Catalogue Of Somatic Mutations In Cancer (COSMIC) v97 database	the COSMIC database	https://cancer.sanger.ac.uk/cosmic
dbSNP build 150	National Center for Biotechnology Information (NCBI)	https://www.ncbi.nlm.nih.gov/snp/
repeatmasker track	UCSC Genome Browser	https://genome.ucsc.edu/

RESOURCE AVAILABILITY

Lead contact

Further information and requests should be directed to the lead contact, Chuanliang Xu (chuanliang_xu@126.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

3rd generation long-read transcriptome RNA-seq data have been deposited at GSA-Human and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.

All original code has been deposited at Zenodo and is publicly available as of the date of publication. DOIs are listed in the key resources table.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell lines

The RT4, HT1376, and J82 cell lines were purchased from the Cell Bank of the Type Culture Collection of the Chinese Academy of Sciences (Shanghai, China) in December 2022. HT1376 and J82 were cultured in MEM medium (Gibco, C11095500BT), and RT4 was cultured in RPMI-1640 medium (Gibco, C11875500BT). All culture media were supplemented with 10% fetal bovine serum (Gibco, 10099-141) and 1% penicillin/streptomycin (Gibco, 15140122) and cells were grown at 37°C with 5% CO2. RT4 and J82 come from male patients, and J82 comes from a female patient. All cel lines have been authenticated through short tandem repeat (STR) profiling and have been tested to be free from mycoplasma contamination.

METHOD DETAILS

Genome build and genomic locations

The Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13) of the human genome was used. Only the reference chromosomes were analyzed, while all genetic elements that mapped to scaffolds, assembly patches and alternate loci (haplotypes) were excluded from this study. All genomic locations were described using hg38 coordinates.

The collection and processing of datasets

We searched for transcriptome sequencing (RNA-seq) data for bladder cancer cell lines in the SRA and ENA datasets using the query "((cell line AND "biomol rna"[Properties] AND "platform illumina"[Properties])) AND ((((bladder) OR urothelia) OR urothelium) AND "biomol rna"[Properties] AND "platform illumina"[Properties])". Cell lines were selected by reviewing the sample processing methods in the corresponding literature and datasets, and those treated with DMSO or empty vectors were excluded, leaving only untreated cell lines. Cell line names were standardized according to the Cellosaurus database (https://www.cellosaurus.org/). TCCSUPB from the CCLE database was renamed TCCSUP, and BV from the CCLE database was renamed 253JBV. 253JP and 253J were collectively referred to as 253J. Whole genome

sequencing data for bladder cancer cell lines were searched for in the CCLE project (SRP186687). Cell line annotations were referenced from Tahlita et al. 2018[24] and Julie et al. 2015,[35] which reviewed the grading, staging, and molecular subtypes of 127 bladder cancer cell lines and the genomic instability (GIN) of 40 bladder cancer cell lines, respectively. The 22 cases of paired bladder tumors and adjacent normal samples were downloaded from the SRA database through project ID PRJNA552055. The MIBC data was downloaded from the SRA database through project ID PRJNA891747, while the corresponding follow-up data for the patients were obtained from the original publication27. The publicly available raw fastq data were downloaded using Aspera and quality-controlled using fastp v0.23.228, with adapters and low-quality sequences removed. Detailed quality control metrics of all cell line RNA-seq datasets were provided in Table S2. STAR v2.7.10b was used for 2-pass alignment based on the hg38 genome build and Gencode V42 transcript annotation (gencodegenes.org). For novel splice junction (SJ) events identified in all samples during 1-pass mapping (stored in the SJ.out file), canonical and intron motifs found having at least 15 unique alignments in 5 samples were retained. These novel junctions were then input along with the original SJ.out files in the 2-pass mapping for each sample. The mRRBS methylation data of bladder cancer cell lines in the CCLE project were downloaded. Detailed quality control metrics of all cell line datasets were provided in Table S3. After quality control, the mRRBS data were aligned using Bismark v0.24.029, and the methylation status of CpG sites were analyzed using the R Methylkit v1.24.0 package30, with the "mincov" parameter set to 5.

Mutation calling

To exclude cell line contamination and identity confusion, the identity of the cell lines was identified using both transcriptome-based phenotype and mutation-based methods. The fast quantification was performed using tximport v1.26.131, and the top 5000 variable genes were used to plot the principal component analysis (PCA) to observe if the same cell lines exhibit cross-dataset similarity. The mutation-based method was adapted from Cel-ID.[36] Varscan2 v2.3.933 was used to identify mutations in RNA-seq data and WGS data. For mutation calling in RNA-seq data, vcftools34 was used to retain mutations with DP>10, FREQ>25, only locating in the exon regions, and have been recorded in COSMIC v97 database (https://cancer.sanger.ac.uk/cosmic). Finally, the distance between the data was described using the 1-Pearson correlation coefficient of the allele fraction (AF) and the correlation heatmap was plotted. For germline mutation calling in WGS data, only germline mutations curated in dbSNP build 151 were called.

Identification and quantification of onco-exaptation events

Onco-exaptation events were identified using TEPROF2 (https://doi.org/10.5281/zenodo.7670515). Detailed methodology has been described before23. In brief, StringTie v2.2.1 was used to perform *de novo* assembly on the STAR 2-pass alignment results35. Based on the GTF assembly, TEPROF2 searched for onco-exaptation events in assembled transcripts that originated from TE sequences (annotated in Repeatmasker as LTR, LINE, SINE, and DNA) and were spliced into known exons of transcripts annotated as "appris principal" in Gencode V42. To remove false positive TE-exaptation structures, multiple heuristic filters were applied to candidates, including: 1. First exon length < 99th percentile of known first exon length in Gencode V42, and the first exon contains a retained intron; 2. TE has at least 10 unique alignments, and at least one pair-end fragment crosses the TE-gene splicing site; 3. Transcript is present in two or more samples; 4. Suspected exonization of TE is excluded (more than 15% of reads end within TE rather than start within TE); The cutoff values for all filters in the software were kept as default. The class, family, and subfamily of all events, the genes spliced into by TE were recorded. StringTie ballgown module implemented within the TEPROF2 pipeline was used to quantify the TPM values of onco-exaptation events and the total expression of the corresponding genes. The upstream intron read count of splice targets was obtained using StringTie ballgown. If an event had sufficient expression (>1 TPM) and reads crossing the splicing site in a sample, it was identified as "exist" in the sample. The novel SJ from TE to splice target was visualized using ggsashimi v1.1.536. The distribution of onco-exaptation events compared to the relative presence of each class of TEs in the human genome was quantified with Fisher enrichment score using the formula (ad) / (bc), where 'a' is the number of onco-exaptation events originating from a specific TE class, 'b' is the total number of onco-exaptation events, 'c' is the number of copies of the specific TE class in the entire genome, and 'd' is the number of copies of all TE classes in the entire genome.

Quantification of gene expression

Gene expression was quantified with Salmon version 1.10.0. The library type was detected automatically by the software. Salmon quantifies and outputs TPM abundance of transcripts annotated in the Gencode V42. The raw count matrix at gene-level was generated by tximport using the TPM abundance and length information of each transcript. The raw count matrix at gene-level was used for principal component analysis (PCA) and differential expression analysis.

Quantification of TE expression

TE expression at subfamily level was quantified with TEtranscript version 2.2.3 using the 'TE count' module. We used the default parameter expect for the '–stranded' parameter, which was set according to whether the RNA-seq library was first-stranded or unstranded (Table S2). TEtranscripts quantifies and outputs raw counts of genes annotated in the Gencode V42 and TE subfamilies annotated in Repeatmasker as LTR, LINE, SINE, and DNA simultaneously. The raw count matrix at gene- and TE-level were used for principal component analysis (PCA).

PCA

The raw count matrix, generated by either Salmon or TEtranscripts, was normalized using the variance stabilizing transformation (vst) algorithm, and was adjusted for dataset-derived batch effect using the removeBatchEffect function from the limma package. The top5000 most variable features were selected based on the standard deviation of the normalized, batch-corrected expression matrix and used for calculating the principal components. The 1st and 2nd principal components were displayed on the scatter plot.

Cell culture

The RT4, HT1376, and J82 cell lines were purchased from the Cell Bank of the Type Culture Collection of the Chinese Academy of Sciences (Shanghai, China) in December 2022. HT1376 and J82 were cultured in MEM medium (Gibco, C11095500BT), and RT4 was cultured in RPMI-1640 medium (Gibco, C11875500BT). All culture media were supplemented with 10% fetal bovine serum (Gibco, 10099-141) and 1% penicillin/streptomycin (Gibco, 15140122) and cells were grown at 37°C with 5% CO2. The cell lines were authenticated using STR profiling.

PCR and qPCR analysis

Total RNA was isolated from RT4, HT1376, and J82 cells using RNA extraction reagent (R401, Vazyme, Jiangsu, China) and subjected to reverse transcription-polymerase chain reaction (RT-PCR) for complementary DNA (cDNA) synthesis (FSK-101, Toyobo, Osaka, Japan) according to manufacturer's instructions. Polymerase chain reaction (PCR) (KOD-211, Toyobo, Osaka, Japan) was performed to amplify DNA sequence specific to L1PA2-SYT1 using four pairs of primers spanning L1PA2-SYT1 junction (Table S4A). The PCR product was then subjected to electrophoresis on a 2% agarose gel containing 1× GelRed DNA Dye (GR501, Vazyme, Nanjing, China). DNA band was visualized and cut under ultraviolet light for DNA purification using a gel extraction kit (CW2302, Cwbio, Jiangsu, China). The purified DNA was Sanger-sequenced on the ABI 3730XL platform (Applied Biosystem, California, USA). The sequencing data were processed and visualized using the R packages sangeranalyseR v1.8.037 and ggmsa v 1.3.438, respectively. To quantify the expression of L1PA2-SYT1, total RNA isolated from J82, RT4, and HT1376 cell lines was subjected to reverse transcription for qPCR (R333, Vazyme, Jiangsu, China). The expression of L1PA2-SYT1 relative to GAPDH was detected with SYBR (Q712, Vazyme, Jiangsu, China) using the ΔΔCt method (QuantStudio 6Pro/6 Flex, ThermoFisher Scientific) (Table S4B). Forward primers were designed for exon1 and reverse primers for exon2-10 to validate the full-length of predicted L1PA2-SYT1 by PCR (Table S4C). To validate L1PA2-SYT1 being an alternative splicing event other than genomic fusion event, genomic RNA was also isolated from J82, RT4, and HT1376 cell lines (DC102, Vazyme, Jiangsu, China) and subjected to PCR electrophoresis (data not shown).

3rd generation long-read transcriptome sequencing

1ug total RNA was prepared for cDNA libraries using cDNA-PCR Sequencing Kit (SQK-LSK110+EXP-PCB096) protocol provided by Oxford Nanopore Technologies (ONT). Briefly, the template switching activity of reverse transcriptase enrich for full-length cDNAs and add defined PCR adapters directly to both ends of the first-strand cDNA. And following cDNA PCR for 14 circles with LongAmp Tag (New England BioLabs, Massachusetts, US). The PCR products were then subjected to ONT adaptor ligation using T4 DNA ligase (New England BioLabs, Massachusetts, US). Agencourt XP beads was used for DNA purification according to ONT protocol. The final cDNA libraries were added to FLO-MIN109 flowcells and run on PromethION platform at Biomarker Technology Company (Beijing, China). The data has been deposited in Genome Sequence Archive (GSA)-Human with the accession number HRA005343.

5'RACE analysis

5'RACE gene-specific primer (GSP) was designed to target the 3rd exon of the predicted L1PA2-SYT1 (Table S4D). The PCR product was amplified with 5'/3' RACE reagent kit (Cat#RA101, Vazyme), enriched with electrophoresis, extracted with gel extraction kit (Cat#CW2302, Cwbio), connected with the pCE2 TA/blunt-Zero vector (Cat#C601, Vazyme), cloned into DH5α competent cell (Cat#C502, Vazyme), and Sanger- sequenced on ABI 3730XL platform using the M13R primer (5'-CAGGAAACAGCTATGAC-3'). The sequenced reads were mapped using minimap2 version 2.26-r1175 and visualized using Integrated Genomics Viewer (IGV).

Overexpression of L1PA2-SYT1 in J82 cell line

The predicted L1PA2-SYT1 sequence was amplified and cloned into pLVX-IRES-Puro plasmids. Specifically, primers were designed according to the predicted L1PA2-SYT1 sequence. PCR (KOD-211, Toyobo, Osaka, Japan) reaction was performed using RT4 cell line cDNA as the template to synthesize the L1PA2-SYT1 sequence. Double digestion of pLVX-IRES-Puro plasmids was performed using restriction enzymes EcoRI and BamHI (R3101S & R3136S, New England BioLabs, Massachusetts, US) according to the manufacturer's instructions. Predicted L1PA2-SYT1 PCR product and linearized vectors were subjected to electrophoresis on 1% agarose gel and purified in the same manner as the above-mentioned method. Linearized vector and the PCR products was subjected to recombination using recombinase (C112, Vazyme, Nanjing, China). The recombinant plasmid was transformed into DH5α competent E. coli cells (C502, Vazyme, Nanjing, China) and plated on LB plate containing 100μg/ml ampicillin (ST008, Beyotime, Shanghai, China). Single colonies were picked from the plate using pipette tip and transferred into LB broth containing 100μg/ml ampicillin. Plasmid DNA was extracted from the expanded culture using extraction kit (DP118, Tiangen Biotech, Beijing, China) and Sanger-sequenced on ABI 3730XL platform. After validation of recombinant pLVX-IRES- L1PA2-SYT1-Puro plasmid, it was co-transfected into 293T cell line along with packaging plasmids pMD2.G and psPAX2 using PEI 40K reagent (G1802,

Servicebio, Hubei, China) for lentivirus packaging. Medium containing lentiviral particles was harvested and transferred to J82 cell line for lentivirus infection. Puromycin-containing medium was added to J82 cell 48h after lentivirus infection for selection. Expression of L1PA2-SYT1 in J82 cell line was validated using PCR and Western blotting.

Invasion assays

Invasion assays were conducted in Transwell chamber with Matrigel (Corning). Tumor cells containing 4×10E4 were seeded inside Transwell inserts with 200μl culture medium without FBS in triple. As a chemoattractant, 600μl culture media containing 10% FBS was placed in the lower chamber. After 24h, 4% formaldehyde fixed the lower surface of the filters, then the filters were stained with 0.1% crystal violet solution.

Western blot assay

The cells were lysed in buffer containing 50 mM Tris-HCl (pH 7.4), 150 mM NaCl, 1 mM EDTA-2Na2, 0.25% Sodium deoxycholate, 1% NP-40, and 1X protease inhibitor. After centrifugation at 12,000rpm for 5min at 4°C, the supernatants were recovered and total protein concentrations were measured using a bicinchoninic acid assay (Pierce BCA Protein Assay Kit, Thermo Scientific). For immunoblot analysis, extracted proteins were separated on 10% PAGE Gel (10% PAGE Gel Fast Preparation Kit, sEpizyme) under 60V for 30min and 110V for 60min and transferred onto PVDF membranes with 1× ice-bath free fast transfer buffer (10× ice-bath free fast transfer buffer, Servicebio) under 400mA for 30min. The membranes were blocked with 1x protein-free rapid blocking buffer (Servicebio) for 30min and incubated overnight at 4°C with a synaptotagmin-1 rabbit monoclonal antibody (mAb) (1:3000, cat#14588, Cell Signaling Technology) and an GAPDH rabbit mAb (1:5000, Cat#2118, Cell Signaling Technology) as a loading control. Primary antibodies were detected with horseradish peroxidase (HRP)-conjugated species-specific secondary antibody (1:3000, Cat#D4W3E, Cell Signaling Technology) and ECL Substrate (Clarity Western ECL Substrate, Cat#1705061, Biorad).

Survival analysis

The survival analysis was performed with survival version 3.5-0. The separation of survival probability between the exists and not exists group was visualized using the Kaplan-Meier curve and tested by log-rank test. Multivariable Cox regression model was visualized using the forest plot and tested by Wald test.

Differential gene analysis and enrichment analysis

Differential analysis was performed between the L1PA2-SYT1 exist group and not exists group. Gene-level counts were quantified using Salmon and differential analysis was performed using DESeq2. Genes curated in the Gencode V42 annotations were considered in differential expression analysis. The gne expression matrix was filtered to include genes with TPM > 1 in at least one sample. Other parameters of DESeq2 remain default. Genes with $|log2FC| \geq 1$ and FDR < 0.05 were defined as differentially expressed genes. For the up-regulated genes in the exist group, GO enrichment analysis was performed using Metascape (metascape.org/).

QUANTIFICATION AND STATISTICAL ANALYSIS

The statistical details could be found in the figure legends. The significance level was defined as $P < 0.05$. R (version 4.1.2) was utilized for data processing and visualization.