TEITbase: a database for transposable element (TE)-initiated transcripts in human cancers

Yun Zhang[1,#], Jianqi She[2,3,#], Xueyan Hu[2], Yueqi Jin[1], Changyu Tao[4], Minghao Du[2,*], Ence Yang[1,2,*]

[1] Department of Medical Bioinformatics, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China

[2] Department of Microbiology & Infectious Disease Center, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China

[3] Department of Radiation Medicine, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China

[4] Department of Human Anatomy, Histology and Embryology, School of Basic Medical Sciences, Peking University Health Science Center, Beijing 100191, China

[#] Contributed equally

[*] Correspondence: yangence@pku.edu.cn (Ence Yang), duminghao@pku.edu.cn (Minghao Du)

Abstract

Transposable elements (TEs) are abundant and play a crucial regulatory role in the human genome. Functioning as alternative promoters, TEs can be reactivated to produce TE-initiated transcripts. In cancer, TE-initiated transcription may upregulate oncogene expression and generate novel tumor-specific antigens, which could serve as potential targets for immunotherapy. However, there remains a lack of comprehensive databases that systematically investigate TE-initiated transcription in cancer. To address this gap, we developed a deep learning-based method and identified 38,995 TE-initiated transcripts across 33 cancer types from The Cancer Genome Atlas. Among these, 6,203 were tumor-specific expressed and strongly associated with the tumorigenesis. Using these annotations, we created TEITbase (http://teitbase.medbioinfo.org/), a user-friendly database that provides researchers with tools to conduct various downstream analyses and investigations for TE-initiated transcripts, including 546 novel onco-exaptation events. The establishment of TEITbase provides new insights into transcriptional reprogramming in cancers and enables further investigation into the potential roles of TE-initiated transcripts in cancer diagnostics and therapeutic strategies.

Keywords: transposable element, database, cancer, deep learning

Introduction

Transposable elements (TEs) are mobile genetic elements that are widespread in eukaryotic genomes (1). Approximately 45% of the human genome is derived from various classes of TEs, such as LINEs (long interspersed nuclear elements), SINEs (short interspersed nuclear elements), LTRs (long terminal repeats), and DNA transposons (2,3). In humans, only a small proportion of TEs, including active human-specific L1 (L1HS) elements, remain capable of transposition and inducing genetic changes (4,5). The majority of TEs in the human genome are incomplete and no longer transposable (6). However, TE-derived sequences retain significant regulatory potential, acting as promoters, enhancers, splicing sites, and more, playing crucial roles in both physiological and pathological processes (6–9).

As potential threats to human genome, most TEs have been kept inactive in somatic cells by mutational events and epigenetic repression (6,10). However, accumulating evidence suggests that certain TEs are transcriptionally reactivated in human cancers as a result of epigenetic dysregulation (11,12). Acting as alternative promoters, TEs can enhance their own transcription or regulate the expression of neighboring genes, leading to the production of TE-initiated transcripts with TE-derived transcription start sites (TSSs) (13). In cancer, TE-initiated transcription can upregulate the expression of oncogenes, a phenomenon known as onco-exaptation (14–16). Notable examples include Alujb-*LIN28B*, L1-*FGGY* and HERVH-*CALB1* (15,17,18). Furthermore, TE-initiated transcription may produce novel protein isoforms and tumor-specific antigens, which could serve as potential targets for immunotherapy (19,20). For instance, the tumor-specific membrane protein GABRG2, transcribed from the L1PA2 promoter, can generate aberrant epitopes on the extracellular surface of cancer cells (12).

62 Moreover, endogenous retrovirus (ERV) envelope glycoproteins have been identified as a
63 dominant anti-tumor antibody target and ERV-reactive antibodies exhibit anti-tumor activity
64 that extends survival in mouse models (21).

65 Given the recognized role of TEs in tumor progression, several databases, including
66 ERVcancer and CancerHERVdb, have been developed to investigate TE expression in human
67 cancers (22,23). However, these databases only provide TE expression data at the TE family
68 level, losing locus information and failing to distinguish between TE-initiated transcription and
69 TE exonization. Recently, Gu et al. established TE-TSS, a database compiling 2,681 RNA
70 sequencing datasets that identifies 5,768 human TE-derived TSSs and 2,797 mouse TE-derived
71 TSSs (24). Nevertheless, due to the repetitive nature of TEs and the limitations of short-read
72 RNA-seq, TE-TSS cannot provide accurate annotations of TE-initiated transcripts. Deep
73 learning techniques, which can automatically extract features from DNA sequence information,
74 have been widely applied to predict TSSs (25,26). By integrating RNA-seq data with DNA
75 sequence information, convolutional neural networks could accurately predict TSSs for each
76 assembled transcript, thereby enabling the identification of TE-initiated transcripts. However,
77 there remains a lack of databases that specifically annotate TE-initiated transcripts in human
78 cancers.

79 In this work, we developed the TEITbase (http://teitbase.medbioinfo.org/), an integrated
80 database that systematically studies the TE-initiated transcription across 33 cancer types from
81 The Cancer Genome Atlas (TCGA), offering data at both the TE family and transcript levels.
82 To identify TE-initiated transcripts, we first developed a deep learning-based method to predict
83 transcription start sites (TSSs) for each assembled transcript, utilizing both DNA sequences and

84 short-read RNA sequencing data. Subsequently, we applied this method to 10,079 tumor
85 samples from TCGA, and identified 38,995 TE-initiated transcripts. These annotations enable
86 TEITbase to provide researchers with the tools to conduct various downstream analyses, such
87 as promoter activity of TE families, tumor-specific TE-initiated transcripts, differential
88 expression analysis, stage analysis, and survival analysis. We believe that TEITbase will serve
89 as a valuable resource for further understanding the role of TE-initiated transcription in
90 epigenetically dysregulated tumor tissues and will be instrumental in investigating novel tumor-
91 specific antigens that could potentially be therapeutically targeted.
92

93 # Results

94 **A deep learning-based method for the identification of TE-initiated transcript**

95 Due to the repetitive nature of TEs and the limitations of short-read RNA-seq, identifying TE-
96 derived TSSs based on RNA-seq coverage is not sufficiently accurate. To overcome the
97 challenge, we developed a deep learning-based method that integrates both DNA sequence and
98 RNA-seq coverage information. Using one-hot encoding and convolutional neural network, we
99 predicted TSSs for each assembled transcript and overlapped them with TE annotations to
100 identify TE-initiated transcripts (Fig. 1A).

101 To train our model, we downloaded CapTrap-seq datasets along with matched short-read
102 RNA-seq datasets (27). Unlike CAGE and RAMPAGE, CapTrap-seq combines the cap-
103 trapping strategy with long-read sequencing technologies, enabling the identification of both
104 TSSs and full-length transcripts. Using these CapTrap-seq datasets for training, our model can
105 predict TSSs at the transcript level. For example, *PRKACB* has alternative promoters in brain
106 tissue. While RAMPAGE identifies only two TSS signals, it cannot assign them to specific
107 transcripts. In contrast, both CapTrap-seq and our model can identify the TSS for each transcript
108 (Fig. 1B).

109 Compared to StringTie, which relies solely on RNA-seq coverage to determine TSS, our
110 model significantly enhances the accuracy of TSS prediction ($P = 4.6 \times 10^{-4}$, Fig. 1C, Table S1).
111 In the validation set, our model accurately identified an average of approximately 26.7% of
112 TSSs, with about 84.7% of TSSs exhibiting identification errors within 100 bp, a notable
113 improvement over StringTie's performance of 2.4% and 70.6%, respectively. As expected, our
114 model performed significantly better in predicting highly expressed transcripts (Fig. 1C, P =

115 1.2×10^{-5}). For transcripts with counts per million (CPM) greater than 2, an average of
116 approximately 71.0% of TSSs were accurately identified by our model, with around 99.2% of
117 TSSs having identification errors within 100 bp. This is substantially better than for low-
118 expressed transcripts (CPM < 0.1), which had 9.1% and 72.7% accuracy within 100 bp,
119 respectively. Furthermore, the prediction score of our model was significantly and negatively
120 correlated with TSS prediction errors (Fig. 1D, Spearman's $r = -0.50$, $P < 2.2 \times 10^{-16}$), enabling
121 us to improve TSS identification accuracy by setting an appropriate prediction score threshold.
122 Through the application of the deep learning model and strict criteria, our method
123 significantly outperforms DeeReCT-TSS in identifying TE-TSS (Table S2, $P = 0.029$) (26). In
124 the H1 cell line, our method identified 954 TE-TSS, of which 626 (65.6%) were supported by
125 CapTrap-seq data. In contrast, DeeReCT-TSS identified only 390 TE-TSS, with just 12 (3.1%)
126 supported by CapTrap-seq data (Fig. 1E). When compared to the TE-TSS database (24), our
127 method identified 64 previously annotated TE-TSS and 908 novel TE-TSS, of which 580
128 (63.9%) were supported by CapTrap-seq data (Fig. 1E). To exclude the potential impact of
129 model overfitting on TE-TSS identification, we downloaded short-read RNA-seq data from 18
130 testis tissues as an external validation dataset (28), with RAMPAGE data serving as the gold
131 standard. Our method identified 3,979 TE-TSS, of which 2,909 (73.1%) were supported by
132 RAMPAGE data, while DeeReCT-TSS identified only 286 TE-TSS, with just 107 (37.4%)
133 supported by RAMPAGE data (Fig. 1F). Compared to the TE-TSS database, our method
134 identified 157 previously annotated TE-TSS and 3,822 novel TE-TSS, of which 2,752 (72.0%)
135 were supported by RAMPAGE data (Fig. 1F). These results suggest that we have developed a
136 deep learning-based method that can accurately identify TE-initiated transcripts using multiple

137 biological replicates of short-read RNA-seq data, significantly outperforming existing methods.

138 Due to the relatively accurate annotation of TE-initiated transcripts, our method effectively

139 distinguishes between TE-initiated transcription and TE exonization, enabling the measurement

140 of promoter activity in TE families. To evaluate the quantitative performance at the family level,

141 we downloaded short-read RNA-seq data and corresponding CAGE data (used as the gold

142 standard for quantitative evaluation) from six cancer cell lines (A549, Hela, HepG2, K562,

143 MCF-7, and SK-N-SH) from the ENCODE database (29). Our method demonstrated

144 significantly better consistency with CAGE data compared to existing TE family expression

145 quantification methods, TEtranscripts (Fig. 1G, $P < 2.2 \times 10^{-16}$). Among all TE families, the

146 median Spearman correlation coefficient between the quantitative results from our method and

147 CAGE data across these six cell lines was 0.26, while TEtranscripts had a value of -0.37.

148 Furthermore, our method identified 105 TE families with high accuracy (Spearman's $r \geq 0.7$),

149 whereas TEtranscripts identified only 6. Notably, the age of these 105 high-accuracy TE

150 families (measured by the evolutionary distance to ancestral TEs: Kimura value) was

151 significantly lower than that of other TE families (Fig. 1G, $P = 9.2 \times 10^{-4}$), which may be related

152 to the higher retention of promoter activity in younger TEs.

153

154 **Genome-wide identification of TE-initiated transcripts in pan-cancers**

155 In 10,079 tumor samples across 33 cancer types from the TCGA, we identified 38,995 TE-

156 initiated transcripts (Table S3), of which 34,670 were not annotated in GENCODE.v46 (30).

157 The expression of these TE-initiated transcripts was associated with the activation of 39,703

158 TE loci, including 11,857 SINE elements, 11,758 LTR elements, 10,911 LINE elements, and

159 2,580 DNA transposons (Fig. 2A). These active TEs contributed to the expression of 14,453
160 neighboring genes, including 9,315 protein-coding genes and 4,771 lncRNA genes (Fig. 2B),
161 suggesting that TEs play a widespread role in regulating gene expression in tumor tissues.

162 Tumor-specific expressed transcripts play a critical role in tumorigenesis and represent
163 important sources of tumor-specific antigens, which may provide new targets for cancer
164 immunotherapy. To investigate this, we performed the expression analysis of TE-initiated
165 transcripts in 10,079 tumor samples, 740 tumor-adjacent samples and 10,620 non-testicular
166 normal tissue samples. Tumor-specific scores were calculated, identifying 6,203 tumor-specific
167 TE-initiated transcripts (scores ≥ 8, Table S4). These activated tumor-specific TE-initiated
168 transcripts were associated with the expression of 1,532 adjacent protein-coding genes. As
169 anticipated, enrichment analysis revealed significant overrepresentation in tumorigenesis-
170 related pathways, including regulation of *PTEN* gene transcription and TP53 activity (Fig. 2C,
171 Table S5). Among the 702 oncogenes annotated in the ONGene database (31), we found that
172 TEs upregulated the expression of 333 oncogenes (a process known as onco-exaptation),
173 generating 696 TE-initiated transcripts (Table S6), including previously reported onco-
174 exaptations such as AluJb-*LIN28B*, L1HS-*SYT1*, L1HS-*FGGY*, and L1HS-*MET* (15,18,32).
175 Compared to a recent similar study (15), we identified 150 reported and 546 novel onco-
176 exaptation events, with 263 (48.2%) supported by CAGE data from at least one cancer cell line.
177 Among them, AluJr-*AKT2* was expressed across all cancer types and detected in 619 tumor
178 samples, further highlighting the universality of onco-exaptation events. Furthermore, TE
179 family enrichment analysis revealed significant overrepresentation of young families, including
180 L1HS, L1P1, and HERVH (Fig. 2D, Table S7). These evolutionarily young TE families retain

181 regulatory sequences that are typically suppressed in somatic cells but are reactivated in the
182 epigenetically dysregulated tumor tissues, driving the TE-initiated transcription and potentially
183 promoting tumor progression.

184 To illustrate, we analyzed the 10 TE-initiated transcripts with the highest tumor-specific
185 scores and found that 8 were initiated by the L1HS family (Fig. 2E). These 10 TE-initiated
186 transcripts were expressed in 275 to 1,069 tumor samples but only in 1 or 2 normal tissue
187 samples, indicating their high tumor specificity. Moreover, these activated TEs were critical in
188 regulating the expression of adjacent genes, with L1HS-*CNBD1*, L1HS-*CFAP47*, and L1P2-
189 *PATE2* serving as the primary transcripts for their respective genes. Furthermore, *EYA4* and
190 *NLRP7* have been shown to promote cell proliferation, invasion, and migration in various
191 cancers (33,34), suggesting that the expression of L1HS-*EYA4* and AluSz-*NLRP7* may have
192 significant oncogenic effects. Survival analysis further indicated that the expression of L1HS-
193 *EYA4*, L1HS-*CRPPA*, L1HS-*CFAP47*, and L1P2-*PATE2* was associated with poor prognosis in
194 at least one cancer type (Fig. 2F). In conclusion, these highly tumor-specific TE-initiated
195 transcripts may play a role in tumorigenesis and hold potential as biomarkers for tumor
196 diagnosis and prognosis.

197

198 **Content and usage of TEITbase**

199 Based on the identification of TE-initiated transcripts across various cancer types, we have
200 developed a free, user-friendly website, TEITbase (http://teitbase.medbioinfo.org/), which
201 provides expression data for 1,027 TE families and 38,995 TE-initiated transcripts (Fig. 3A).
202 The website consists of several main pages: Home, Browse, Analyze, Statistics, Download,

203 Help, and Contact. TEITbase offers an intuitive interface and a range of core functionalities,
204 enabling researchers to explore and understand the role of TE-initiated transcription in cancers.
205 The Browse page is divided into two sections: TE Family and TE-initiated Transcripts.
206 Users can query TE family entries by TE family, TE superfamily, and species. The query results
207 display the TE family, Dfam database ID, TE superfamily, TE classification, species, consensus
208 sequence length, Kimura value, and tissue-specificity score (Tau index). Users can also query
209 TE-initiated transcripts using TEITbase database ID, TE family, and gene. The results for these
210 queries include the TEITbase database ID, TE family, TE superfamily, TE classification,
211 chimeric gene, transcript classification, coding types, tumor specificity score, and TSS location.
212 According to the GENCODE v46 annotation, TE-initiated transcripts are classified into three
213 categories: annotated, chimeric (with a novel splice site at the 5' end), and intergenic (without
214 known splice sites). Based on predicted open reading frames (ORFs), TE-initiated transcripts
215 are further categorized into six types: annotated, 5' truncated, 5' chimeric, chimeric truncated,
216 other coding types, and non-coding. The Analyses page allows users to query tumor-specific
217 TE-initiated transcripts, as well as survival-associated, stage-associated, and differentially
218 expressed TE-initiated transcripts (or TE families). Additionally, the Statistics page provides
219 the basic information of TEITbase, such as the number of TE-initiated transcripts. The Help
220 page includes a user guide for the database, while the Download page offers links to download
221 all annotation information, expression matrices, and analysis results.
222 Using the functions of TEITbase, we investigated the dysregulation of the L1HS family in
223 cancer. We found that lung squamous carcinoma (LUSC), esophageal carcinoma (ESCA), and
224 bladder urothelial carcinoma (BLCA) exhibit higher L1HS promoter activity, while normal

225 tissues show lower activity (Fig. 3B). Differential expression analysis revealed significant
226 activation of the L1HS promoter in 16 out of 21 cancer types, with the most pronounced
227 increase in LUSC (Fig. 3C, $P < 2.2 \times 10^{-16}$). In LUSC, L1HS promoter activity in stage II
228 patients was significantly higher than in stage I patients (Fig. 3D, $P = 0.038$). Survival analysis
229 indicated that in 5 out of 33 cancer types, including liver hepatocellular carcinoma (LIHC),
230 abnormal activation of the L1HS promoter was associated with poorer tumor prognosis.
231 Conversely, activation of the L1HS promoter is significantly correlated with better tumor
232 prognosis in LUSC (Fig. 3E), adrenocortical carcinoma (ACC), and mesothelioma (MESO).
233 To explore how L1HS promoter activation improves prognosis, we conducted a correlation
234 analysis between L1HS promoter activity and gene expression, followed by enrichment
235 analysis of genes positively correlated with its expression. We discovered that in LUSC,
236 activation of the L1HS promoter was linked to apoptosis-regulating signaling pathways (Fig.
237 3F), suggesting that active transposition of L1HS may disrupt tumor genome stability, leading
238 to cancer cell apoptosis (35).

239 By utilizing the 'Tumor-specific transcripts' function of TEITbase, we investigated the
240 role of TE_11880 (L1HS-*IRF1*) in cancers. TE_11880 is initiated by L1HS with a tumor-
241 specific score of 32.5 and is chimeric with *IRF1*, a transcriptional regulator and tumor
242 suppressor that activates genes involved in antitumor immunity (36). These findings suggest
243 that the activation of TE promoters in cancer tissues may also confer benefits to the host. Further
244 details about TE_11880, including its sequence and other relevant information, can be accessed
245 in the TEITbase database. TE_11880 is silent in most normal tissue samples but activated in
246 several cancer types (Fig. 3G, H). We selected specific cancer types to explore the differential

247 expression, stage, and survival associations of TE_11880. Our differential expression analysis
248 revealed that TE_11880 expression was primarily restricted to LUSC and stomach
249 adenocarcinoma (STAD), with no expression in adjacent normal tissues (Fig. 3I). In STAD,
250 TE_11880 expression in stage III patients was significantly higher than in both stage I and stage
251 II patients (Fig. 3J). Additionally, TE_11880 expression is associated with a favorable prognosis
252 in colon adenocarcinoma (COAD), despite not reaching statistical significance ($P = 0.069$) (Fig.
253 3K), suggesting its potential prognostic value.
254

255 ## Discussion

256 The activation of TE promoters can significantly impact the transcriptomic and proteomic
257 profiles of cancer (20). Recent studies suggest that TE-initiated transcription may be linked to
258 tumorigenesis and could generate tumor-specific antigens, offering potential targets for
259 immunotherapies (12,15). However, the widespread use of short-read RNA sequencing
260 technologies poses challenges due to the loss of TSS information. To address this, we developed
261 a deep learning-based approach that predicts TSSs and accurately identifies TE-initiated
262 transcripts, enabling the creation of a specialized database, TEITbase. TEITbase offers
263 abundant data resources and various analytical tools, allowing users to explore the roles of TE-
264 initiated transcripts in cancer progression. Importantly, our findings have identified over 6000
265 tumor-specific TE-initiated transcripts, with the L1HS family being notably enriched. As the
266 youngest L1 family in humans, the L1HS family retains transposable ability and may play a
267 significant role in tumorigenesis (35).

268 Due to its powerful model fitting capabilities, deep learning has become widely applied to
269 predicting key features from DNA sequences such as gene expression, RNA editing, TSS,
270 alternative splicing, and polyadenylation (25,37–40). Several CNN-based tools have been
271 developed for the accurate prediction of promoters and TSS, including Puffin, TSSPlant,
272 TransPrise, and DeeReCT-TSS (25,26,41,42). Among them, DeeReCT-TSS accurately
273 identifies active TSS across the genome by integrating DNA sequences and RNA-seq coverage
274 information. However, DeeReCT-TSS uses CAGE-seq data to train the model, which loses
275 relevant transcript structure information and cannot assign identified active TSS to specific
276 transcripts. Besides, reads derived from TE sequences often produce ambiguous alignments,

leading to errors in RNA-seq coverage estimation and reducing the accuracy of the identification of TE-derived TSS (43). To address this, we downloaded CapTrap-seq data to train the model (27), enabling the prediction of TSS for each transcript, and applied strict criteria to identify TE-initiated transcript. Our method significantly outperforms DeeReCT-TSS in identifying TE-derived TSS, providing a new tool to investigate TE-initiated transcription.

In summary, TEITbase is a user-friendly database that provides valuable insights and tools for investigating TE-initiated transcription in tumorigenesis, as well as exploring the potential roles of these transcripts in cancer diagnostics and therapeutic strategies. While it already offers a comprehensive data collection, there are still opportunities for optimization and further development. Future versions will incorporate TE exonization, which results from noncanonical splice junctions between exons and TEs, serving as a source of tumor-specific antigens (8). With ongoing advancements in full-length long-read RNA-seq technologies (16), TEITbase will integrate publicly available long-read RNA-seq data to capture the full length of TE-initiated and TE-exonization transcripts. These updates are expected to further enhance TEITbase's utility and broaden its applicability across various research fields.

293 # Methods

294 **Data collection**

295 We downloaded short-read RNA-seq data from 10,819 samples across 33 cancer types from the
296 The Cancer Genome Atlas (TCGA, phs000178.v11.p8) (44). Additionally, we obtained RNA-
297 seq data from 10,620 samples across 45 human body sites from the GTEx project (by dbGaP,
298 study accession: phs000424.v8.p2) (45). Short-read RNA-seq and CapTrap-seq datasets were
299 downloaded from the ENCODE project and the ArrayExpress repository (E-MTAB-13063)
300 (27,29). Ten samples of these samples were used as the training set (H1 cell line and heart), and
301 three samples were used as the validation set (endodermal cells and brain). Short-read RNA-
302 seq and RAMPAGE datasets of testis were download from the Gene Expression Omnibus
303 database (GSE135791) and the ENCODE project (28). Short-read RNA-seq and CAGE
304 datasets of six cancer cell lines were download from the ENCODE project. GENCODE Version
305 46 was download from https://www.gencodegenes.org/human/ and used as the transcript
306 reference (30). TE annotations and consensus sequences of TE family were downloaded from
307 the Dfam database (https://www.dfam.org/releases/Dfam_3.8/) (46).

308

309 **Deep learning model for TSS prediction**

310 CapTrap-seq and matched short-read RNA-seq datasets were used to train the model. For each
311 3' end of the first exon, the 10 kb DNA sequence on either side, along with the corresponding
312 RNA-seq coverage (including both unique-mapping and multi-mapping reads), was extracted
313 and used as input to the model. The corresponding TSS signal from the CapTrap-seq data served
314 as the model's label. Our model is based on the architecture of the Puffin-D model, a

convolutional neural network specifically designed for TSS prediction using DNA sequences (25). To enable TSS prediction for each sample, we adjusted the model to extract information from both the DNA sequence and RNA-seq coverage. To address the imbalance between positive and negative samples during model training, we employed the KL divergence loss function:

$$L = \sum_{i} target_i^* [ln(target_i^* + \epsilon) - ln(pred_i^* + \epsilon)]$$

The *target* represents the true TSS signal at position *i*, *pred* represents the TSS signal predicted by the model, and ϵ is a constant set to e^{-10}.

The model utilized the Adam optimizer for parameter optimization, with a learning rate of 0.005, a batch size of 64, and a maximum of 10 training epochs. Model construction and training were carried out using PyTorch. The distance between the predicted and actual TSS was used to evaluate the model's performance. The example of an alternative promoter was visualized on WashU Epigenome Browser (47).

Pipeline for the detection of TE initiated transcripts

First, RNA-seq data were aligned to the reference human genome (GRCh38) using STAR v2.7.10b (48). Sambamba was employed to remove PCR duplicates, and deepTools was used to extract RNA-seq coverage (49,50). StringTie v2.1.7 was used to assemble the reads into full-length transcripts (stringtie –m 200 –c 1 -t) (51). The TSS for each assembled transcript was then predicted using the trained deep learning model. These assembled transcripts across different samples were subsequently merged with a custom script. Briefly, the first exons of multi-exon transcripts were extracted. Following this, first exons overlapping with other

337 internal exons were excluded, as they may have resulted from RNA degradation. Then, for each

338 first exon, the TSS was corrected according to the score calculated as follows (52):

339 $$S_t = s_t + \sum_{k \in T} \frac{\mathrm{w} - d_k}{\mathrm{w} * s_k}$$

340 The *S* is the weight score for position *t*; *s* represents the number of samples supporting the

341 position as a TSS of the first exon; *w* is the window size used to calculate the weight score; *T*

342 is the set of all positions within a distance from *t* that is less than the window size; and *d* is the

343 distance between positions *t* and *k*.

344 Corrected TSSs overlapping with TEs were retained, and the full-length transcripts of TE-

345 initiated transcripts were merged using a reference-guided approach across samples. The count

346 of junction reads supporting the first exon and the expression of TE-initiated transcripts at the

347 transcript level were then calculated using StringTie. TE-initiated transcripts with at least one

348 junction read supporting the first exon were considered expressed in the sample. The count of

349 junction reads supporting the first exon of TE-initiated transcripts from the same TE family was

350 summed, and the total read count mapped to the GRCh38 genome was used for normalization

351 to obtain TE promoter activity at the family level.

352

353 **Computational validation of TE-derived TSSs and TE promoter activity**

354 To evaluate the precision of TE-derived TSSs identified by our pipeline, we used CapTrap-seq

355 or RAMPAGE data as the gold standard. A TE-derived TSS with supporting reads within 100

356 bp was considered positive. Plots of the CapTrap-seq and RAMPAGE signals surrounding the

357 TSSs were generated using deepTools. To assess the precision of TE promoter activity

358 calculated by our pipeline, the Spearman correlation with CAGE data was calculated.

359

360 **Annotation and expression analysis of TE-initiated transcripts**

361 Identified TE-initiated transcripts were annotated with Dfam database by using BEDTools
362 v2.31.1 (46,53). Based on the relationship with reference transcripts, TE-initiated transcripts
363 were classified into three categories: annotated, chimeric, and intergenic. Coding potential of
364 TE-initiated transcripts was calculated by CPC2 (54). For coding TE-initiated transcripts, we
365 subsequently compared predicted CDS with annotated CDS using BLAST 2.14.1 (55). Then
366 the coding TE-initiated transcripts were annotated into five classes: annotated, 5' chimeric
367 normal, 5' chimeric truncated, 5' truncated and novel. Tumor-specific scores for each TE-
368 initiated transcript were calculated as follows:

369 $$S = \frac{(T_e + 1)/T_t}{(N_e + 1)/N_t}$$

370 The S represents the tumor-specific score for the TE-initiated transcript. T_e represents the
371 number of tumor samples expressing the TE-initiated transcript, N_e represents the number of
372 tumor-adjacent or normal tissue samples expressing the TE-initiated transcript (excluding
373 testicular tissue), and T_t and N_t represent the total number of tumor and normal samples,
374 respectively. A TE-initiated transcript with a tumor-specific score greater than eight was
375 considered tumor-specific expressed.

376 The reactome pathway enrichment analysis of gene-chimeric TE-initiated transcripts was
377 conducted by the ReactomePA package (56). The TE family enrichment analysis was conducted
378 using a two-tailed Fisher's exact test. All P-values were corrected using the Benjamini-
379 Hochberg (BH) procedure.

380

Clinically associated TE families or TE-initiated transcripts

The Wilcoxon rank-sum test was used to assess statistical differences in the expression levels of TE families and TE-initiated transcripts between tumor and adjacent normal tissues. The Kruskal-Wallis test was applied to evaluate statistical differences in the expression levels of TE families and TE-initiated transcripts across patients with different tumor stages. The survdiff function from the survival package was used to perform the Log-Rank test, comparing differences between two survival curves, and calculating the hazard ratio (HR) and *P*-value. All *P*-values were adjusted using the Benjamini-Hochberg (BH) procedure.

Construction of the TEITbase website

The backend application and frontend interface of the TEITbase website were developed using the Django framework, with Gunicorn serving as the WSGI HTTP server to execute the Django code. NGINX was employed as a reverse proxy server to forward client requests to Gunicorn. To ensure system stability, Supervisor was used to monitor the operation of the Gunicorn and Django processes. The frontend interface was designed and implemented using HTML, JavaScript, and CSS. Additionally, annotations for TE families and TE initiated transcripts were securely stored in an SQLite3 database.

Declarations

Ethics approval and consent to participate

Not applicable.

Availability of data and materials

Details about data analyzed in this study were included in the Methods section. The modified pipeline, together with the deep learning model for TSS prediction developed in this study are available in (https://github.com/yunzhang1998/Deep-TEIRI).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by grant from the Young Scientists Fund of the National Natural Science Foundation of China (32400534), the China Postdoctoral Science Foundation (2024M760136), and the Ministry of Science and Technology of China (2021ZD0203203).

Author contributions

E.Y. conceived and designed the study. Y.Z., J.Q.S., X.Y.H., Y.Q.J., C.Y.T., and M.H.D. analyzed the data. Y.Z., J.Q.S., M.H.D., and E.Y. interpreted the data and wrote the manuscript. All authors approved the final version of the manuscript.

Acknowledgments

We would like to express our gratitude to the TCGA and GTEx committees for their valuable contributions. We also acknowledge the ENCODE Consortium and the ENCODE production laboratories.

Additional files

Additional file 1: Table S1-7. Supplementary tables (Table S1-S7, XLSX).

References

1. Britten RJ, Kohne DE. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science*. 1968;161(3841):529-540. doi:10.1126/science.161.3841.529

2. Kapitonov VV, Jurka J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat Rev Genet*. 2008;9(5):411-412; author reply 414. doi:10.1038/nrg2165-c1

3. Wells JN, Feschotte C. A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet*. 2020;54:539. doi:10.1146/annurev-genet-040620-022145

4. Beck CR, Collier P, Macfarlane C, et al. LINE-1 retrotransposition activity in human genomes. *Cell*. 2010;141(7):1159-1170. doi:10.1016/j.cell.2010.05.021

5. Brouha B, Schustak J, Badge RM, et al. Hot L1s account for the bulk of retrotransposition in the human population. *Proc Natl Acad Sci U S A*. 2003;100(9):5280-5285. doi:10.1073/pnas.0831042100

6. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. *Nat Rev Genet*. 2016;18(2):71. doi:10.1038/nrg.2016.139

7. Bourque G, Leong B, Vega VB, et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res*. 2008;18(11):1752-1762. doi:10.1101/gr.080663.108

8. Merlotti A, Sadacca B, Arribas YA, et al. Noncanonical splicing junctions between exons and transposable elements represent a source of immunogenic recurrent neo-antigens in patients with lung cancer. *Sci Immunol*. 2023;8(80):eabm6359. doi:10.1126/sciimmunol.abm6359

9. Miao B, Fu S, Lyu C, Gontarz P, Wang T, Zhang B. Tissue-specific usage of transposable element-derived promoters in mouse development. *Genome Biol*. 2020;21(1):255. doi:10.1186/s13059-020-02164-3

10. Deniz Ö, Frost JM, Branco MR. Regulation of transposable elements by DNA modifications. *Nat Rev Genet*. 2019;20(7):417-431. doi:10.1038/s41576-019-0106-6

457 11. Kong Y, Rose CM, Cass AA, et al. Transposable element expression in tumors is
458 associated with immune infiltration and increased antigenicity. *Nat Commun*.
459 2019;10(1):5228. doi:10.1038/s41467-019-13035-2

460 12. Rodriguez-Martin B, Alvarez EG, Baez-Ortega A, et al. Pan-cancer analysis of whole
461 genomes identifies driver rearrangements promoted by LINE-1 retrotransposition. *Nat*
462 *Genet*. 2020;52(3):306-319. doi:10.1038/s41588-019-0562-0

463 13. Modzelewski AJ, Shao W, Chen J, et al. A mouse-specific retrotransposon drives a
464 conserved Cdk2ap1 isoform essential for development. *Cell*. 2021;184(22):5541-
465 5558.e22. doi:10.1016/j.cell.2021.09.021

466 14. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. *Mob*
467 *DNA*. 2016;7:24. doi:10.1186/s13100-016-0080-x

468 15. Jang HS, Shah NM, Du AY, et al. Transposable elements drive widespread expression of
469 oncogenes in human cancers. *Nat Genet*. 2019;51(4):611-617. doi:10.1038/s41588-019-
470 0373-3

471 16. Shi Q, Li X, Liu Y, Chen Z, He X. FLIBase: a comprehensive repository of full-length
472 isoforms across human cancers and tissues. *Nucleic Acids Res*. 2023;52(D1):D124-D133.
473 doi:10.1093/nar/gkad745

474 17. Attig J, Pape J, Doglio L, et al. Human endogenous retrovirus onco-exaptation counters
475 cancer cell senescence through calbindin. *J Clin Invest*. 2023;133(14):e164397.
476 doi:10.1172/JCI164397

477 18. Sun Z, Zhang R, Zhang X, et al. LINE-1 promotes tumorigenicity and exacerbates tumor
478 progression via stimulating metabolism reprogramming in non-small cell lung cancer.
479 *Mol Cancer*. 2022;21:147. doi:10.1186/s12943-022-01618-5

480 19. Laumont CM, Vincent K, Hesnard L, et al. Noncoding regions are the main source of
481 targetable tumor-specific antigens. *Sci Transl Med*. 2018;10(470):eaau5516.
482 doi:10.1126/scitranslmed.aau5516

483 20. Liang Y, Qu X, Shah NM, Wang T. Towards targeting transposable elements for cancer
484 therapy. *Nat Rev Cancer*. 2024;24(2):123-140. doi:10.1038/s41568-023-00653-8

485 21. Ng KW, Boumelha J, Enfield KSS, et al. Antibodies against endogenous retroviruses
486 promote lung cancer immunotherapy. *Nature*. 2023;616(7957):563-573.
487 doi:10.1038/s41586-023-05771-9

488 22. Lei X, Mao S, Li Y, et al. ERVcancer: a web resource designed for querying activation of
489 human endogenous retroviruses across major cancer types. *J Genet Genomics Yi Chuan*
490 *Xue Bao*. Published online September 10, 2024:S1673-8527(24)00241-8.
491 doi:10.1016/j.jgg.2024.09.004

492 23. Stricker E, Peckham-Gregory EC, Scheurer ME. CancerHERVdb: Human Endogenous
493 Retrovirus (HERV) Expression Database for Human Cancer Accelerates Studies of the
494 Retrovirome and Predictions for HERV-Based Therapies. *J Virol*. 2023;97(6):e0005923.
495 doi:10.1128/jvi.00059-23

496 24. Gu X, Wang M, Zhang XO. TE-TSS: an integrated data resource of human and mouse
497 transposable element (TE)-derived transcription start site (TSS). *Nucleic Acids Res*.
498 2023;52(D1):D322. doi:10.1093/nar/gkad1048

499 25. Dudnyk K, Cai D, Shi C, Xu J, Zhou J. Sequence basis of transcription initiation in human
500 genome. *Science*. 2024;384(6694):eadj0116. doi:10.1126/science.adj0116

501 26. Zhou J, Zhang B, Li H, et al. Annotating TSSs in Multiple Cell Types Based on DNA
502 Sequence and RNA-seq Data via DeeReCT-TSS. *Genomics Proteomics Bioinformatics*.
503 2022;20(5):959-973. doi:10.1016/j.gpb.2022.11.010

504 27. Carbonell-Sala S, Perteghella T, Lagarde J, et al. CapTrap-seq: a platform-agnostic and
505 quantitative approach for high-fidelity full-length RNA sequencing. *Nat Commun*.
506 2024;15(1):5278. doi:10.1038/s41467-024-49523-3

507 28. Özata DM, Yu T, Mou H, et al. Evolutionarily conserved pachytene piRNA loci are highly
508 divergent among modern humans. *Nat Ecol Evol*. 2020;4(1):156-168.
509 doi:10.1038/s41559-019-1065-1

510 29. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the
511 human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247

512 30. Harrow J, Frankish A, Gonzalez JM, et al. GENCODE: the reference human genome
513 annotation for The ENCODE Project. *Genome Res*. 2012;22(9):1760-1774.
514 doi:10.1101/gr.135350.111

515 31. Liu Y, Sun J, Zhao M. ONGene: A literature-based database for human oncogenes. *J*
516 *Genet Genomics*. 2017;44(2):119-121. doi:10.1016/j.jgg.2016.12.004

517 32. Hur K, Cejas P, Feliu J, et al. Hypomethylation of long interspersed nuclear element-1
518 (LINE-1) leads to activation of proto-oncogenes in human colorectal cancer metastasis.
519 *Gut*. 2014;63(4):635-646. doi:10.1136/gutjnl-2012-304219

520 33. de la Peña Avalos B, Tropée R, Duijf PHG, Dray E. EYA4 promotes breast cancer
521 progression and metastasis through its role in replication stress avoidance. *Mol Cancer*.
522 2023;22(1):158. doi:10.1186/s12943-023-01861-4

523 34. Li B, Qi ZP, He DL, et al. NLRP7 deubiquitination by USP10 promotes tumor progression
524 and tumor-associated macrophage polarization in colorectal cancer. *J Exp Clin Cancer*
525 *Res CR*. 2021;40(1):126. doi:10.1186/s13046-021-01920-y

526 35. Gu Z, Liu Y, Zhang Y, et al. Silencing of LINE-1 retrotransposons is a selective

527 dependency of myeloid leukemia. *Nat Genet*. 2021;53(5):672-682. doi:10.1038/s41588-
528 021-00829-8

529 36. Ghislat G, Cheema AS, Baudoin E, et al. NF-κB-dependent IRF1 activation programs
530 cDC1 dendritic cells to drive antitumor immunity. *Sci Immunol*. 2021;6(61):eabg3570.
531 doi:10.1126/sciimmunol.abg3570

532 37. Avsec Ž, Agarwal V, Visentin D, et al. Effective gene expression prediction from sequence
533 by integrating long-range interactions. *Nat Methods*. 2021;18(10):1196-1203.
534 doi:10.1038/s41592-021-01252-x

535 38. Bogard N, Linder J, Rosenberg AB, Seelig G. A deep neural network for predicting and
536 engineering alternative polyadenylation. *Cell*. 2019;178(1):91-106.
537 doi:10.1016/j.cell.2019.04.046

538 39. Fu ZC, Gao BQ, Nan F, Ma XK, Yang L. DEMINING: A deep learning model embedded
539 framework to distinguish RNA editing from DNA mutations in RNA sequencing data.
540 *Genome Biol*. 2024;25(1):258. doi:10.1186/s13059-024-03397-2

541 40. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, et al. Predicting splicing
542 from primary sequence with deep learning. *Cell*. 2019;176(3):535-548.
543 doi:10.1016/j.cell.2018.12.015

544 41. Pachganov S, Murtazalieva K, Zarubin A, Sokolov D, Chartier DR, Tatarinova TV.
545 TransPrise: a novel machine learning approach for eukaryotic promoter prediction. *PeerJ*.
546 2019;7:e7990. doi:10.7717/peerj.7990

547 42. Shahmuradov IA, Umarov RK, Solovyev VV. TSSPlant: a new tool for prediction of plant
548 Pol II promoters. *Nucleic Acids Res*. 2017;45(8):e65. doi:10.1093/nar/gkw1353

549 43. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression.
550 *Nat Rev Genet*. 2020;21(12):721-736. doi:10.1038/s41576-020-0251-y

551 44. Cancer Genome Atlas Research Network. Comprehensive genomic characterization
552 defines human glioblastoma genes and core pathways. *Nature*. 2008;455(7216):1061-
553 1068. doi:10.1038/nature07385

554 45. GTEx Consortium, Laboratory, Data Analysis &Coordinating Center (LDACC)—
555 Analysis Working Group, Statistical Methods groups—Analysis Working Group, et al.
556 Genetic effects on gene expression across human tissues. *Nature*. 2017;550(7675):204-
557 213. doi:10.1038/nature24277

558 46. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of
559 transposable element families, sequence models, and genome annotations. *Mob DNA*.
560 2021;12(1):2. doi:10.1186/s13100-020-00230-y

561 47. Zhou X, Wang T. Using the Wash U Epigenome Browser to examine genome-wide

562 sequencing data. *Curr Protoc Bioinforma*. 2012;Chapter 10:10.10.1-10.10.14.
563 doi:10.1002/0471250953.bi1010s40

564 48. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner.
565 *Bioinformatics*. 2012;29(1):15. doi:10.1093/bioinformatics/bts635

566 49. Ramírez F, Ryan DP, Grüning B, et al. deepTools2: a next generation web server for deep-
567 sequencing data analysis. *Nucleic Acids Res*. 2016;44(W1):W160-165.
568 doi:10.1093/nar/gkw257

569 50. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS
570 alignment formats. *Bioinformatics*. 2015;31(12):2032-2034.
571 doi:10.1093/bioinformatics/btv098

572 51. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome
573 assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*.
574 2019;20(1):278. doi:10.1186/s13059-019-1910-1

575 52. Tang AD, Soulette CM, van Baren MJ, et al. Full-length transcript characterization of
576 SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained
577 introns. *Nat Commun*. 2020;11(1):1438. doi:10.1038/s41467-020-15171-6

578 53. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
579 features. *Bioinforma Oxf Engl*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033

580 54. Kang YJ, Yang DC, Kong L, et al. CPC2: a fast and accurate coding potential calculator
581 based on sequence intrinsic features. *Nucleic Acids Res*. 2017;45(W1):W12-W16.
582 doi:10.1093/nar/gkx428

583 55. Camacho C, Coulouris G, Avagyan V, et al. BLAST+: architecture and applications. *BMC*
584 *Bioinformatics*. 2009;10:421. doi:10.1186/1471-2105-10-421

585 56. Yu G, He QY. ReactomePA: an R/Bioconductor package for reactome pathway analysis
586 and visualization. *Mol Biosyst*. 2016;12(2):477-479. doi:10.1039/c5mb00663e

587

588

589 # Figure legends

590 **Fig. 1 Detection of TE-initiated transcripts**

591 (**A**) Pipeline designed to identify TE-initiated transcripts by using a deep learning model. (**B**)

592 RAMPAGE and CapTrap-seq signals of the *PRKACB* gene, along with TSS predictions of our

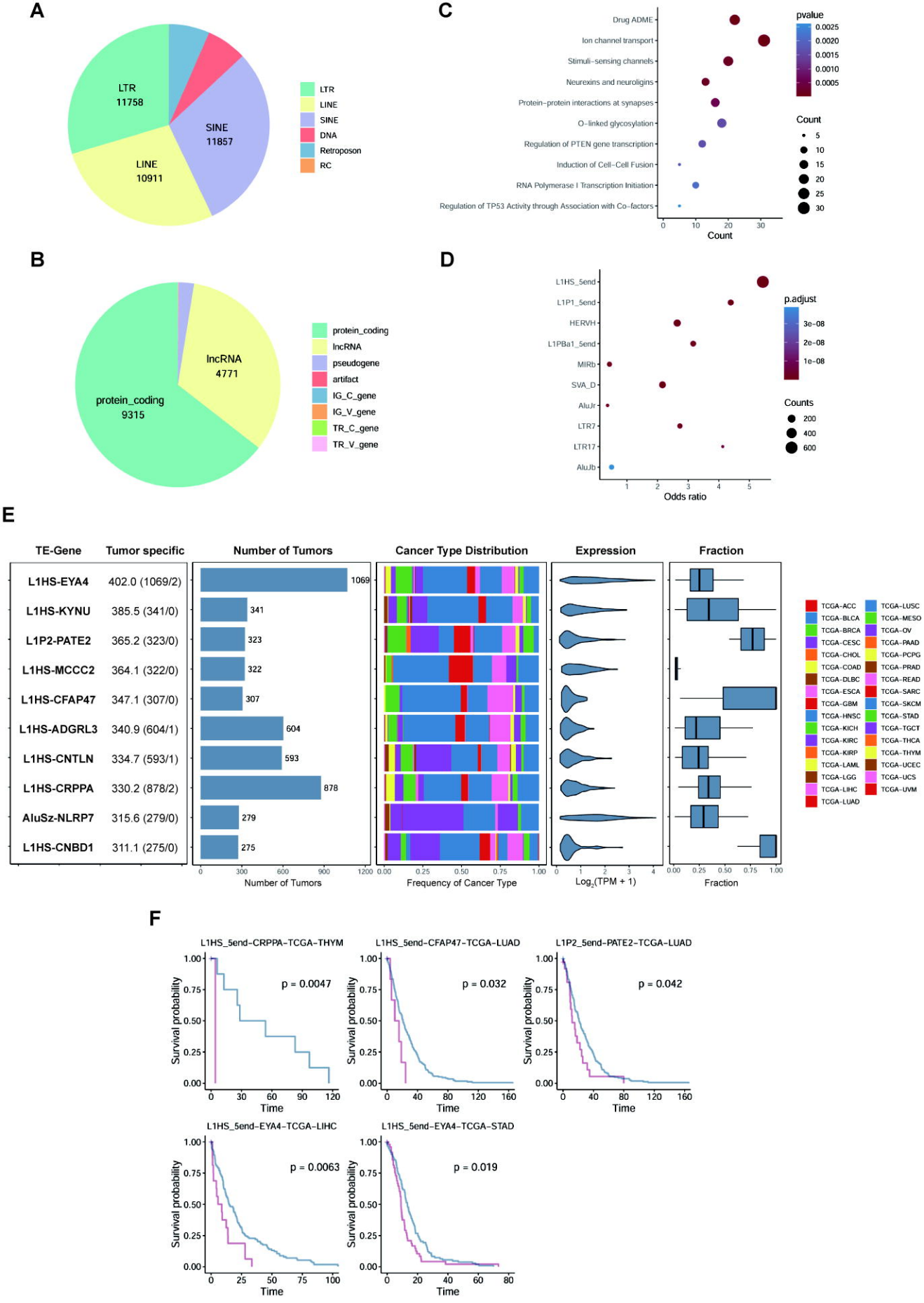
593 method. (**C**) The distances to the true TSSs of predicted TSSs from our method (left) and
594 Stringtie (right). (**D**) Prediction score and distance of predicted TSSs to the true TSSs. (**E**)
595 CapTrap-seq and CAGE signals around the TE-derived TSSs identified by our method or
596 DeeReCT-TSS (left) and those identified in this work, TE-TSS database, or both (right) in H1
597 cell lines. (**F**) RAMPAGE signals around the TE-derived TSSs identified by our method or
598 DeeReCT-TSS (left) and those identified in this work, TE-TSS database, or both (right) in testis
599 tissues. (**G**) Spearman correlation coefficient between CAGE data and the quantitative results
600 from our method or TEtranscripts (left). The Kimura value of high-accuracy TE families and
601 others (right).
602
603 **Fig. 2 Tumor-specific TE-initiated transcription promotes tumorigenesis**
604 (**A**) The distribution of the activated TE loci across the TE classes. (**B**) The distribution of the
605 neighboring genes across the gene type. (**C**) The reactome pathway enrichment analysis of
606 neighboring genes associated with the tumor-specific TE-initiated transcripts. (**D**) The TE
607 family enrichment analysis of tumor-specific TE-initiated transcripts. (**E**) The 10 TE-initiated
608 transcripts with the highest tumor-specific scores are presented. The left-most panel shows the
609 TE-gene along with the tumor-specific score. The next two panels display the number of tumor
610 samples in which each TE-initiated transcript is expressed and the distribution across cancer
611 types. The final two panels show the expression levels of the TE-initiated transcript and the
612 percentage of total oncogene expression contributed by TE-initiated transcription. (**F**) Kaplan–
613 Meier plots demonstrating overall survival associated with the expression of TE-initiated
614 transcripts.

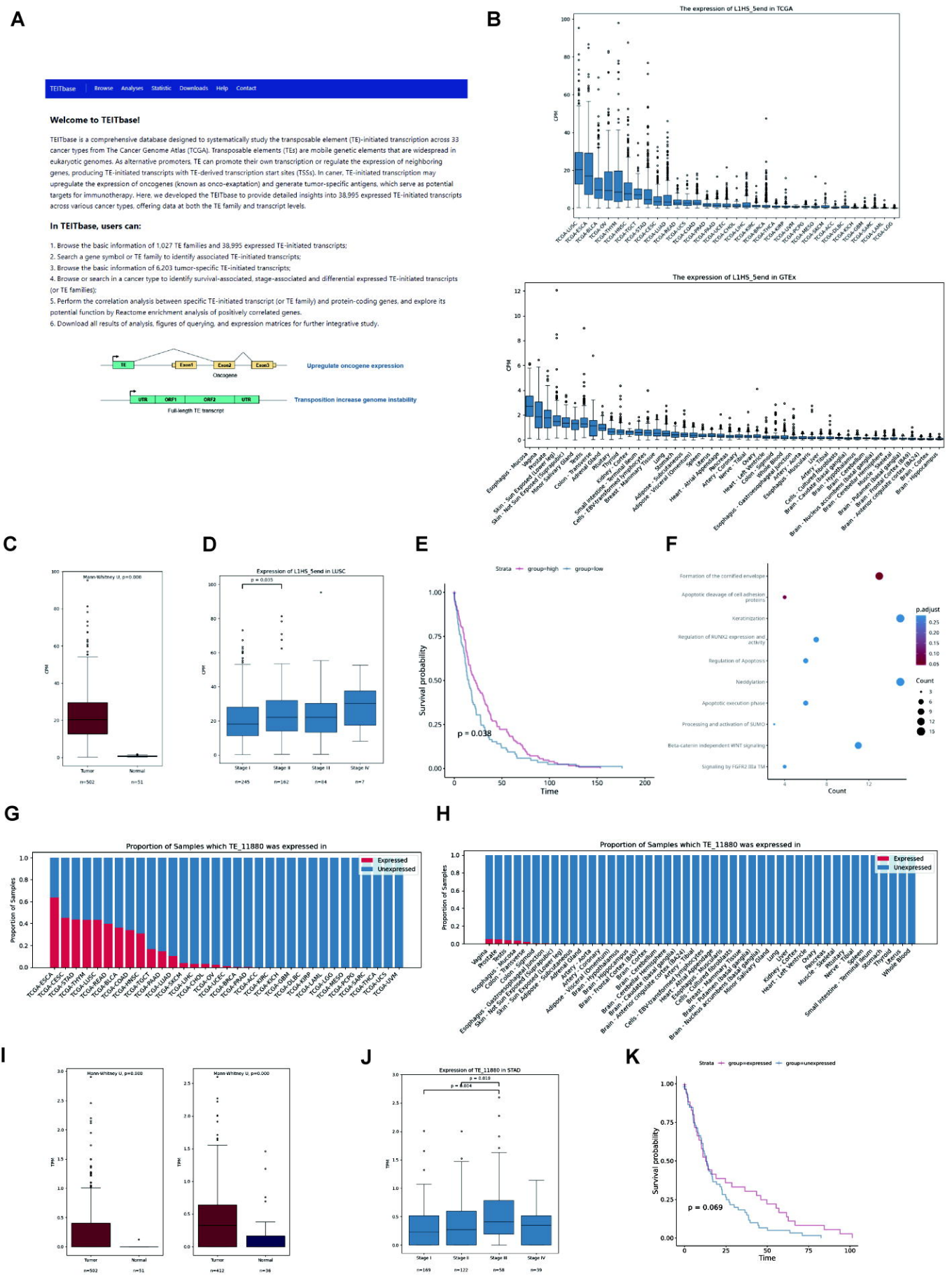
615

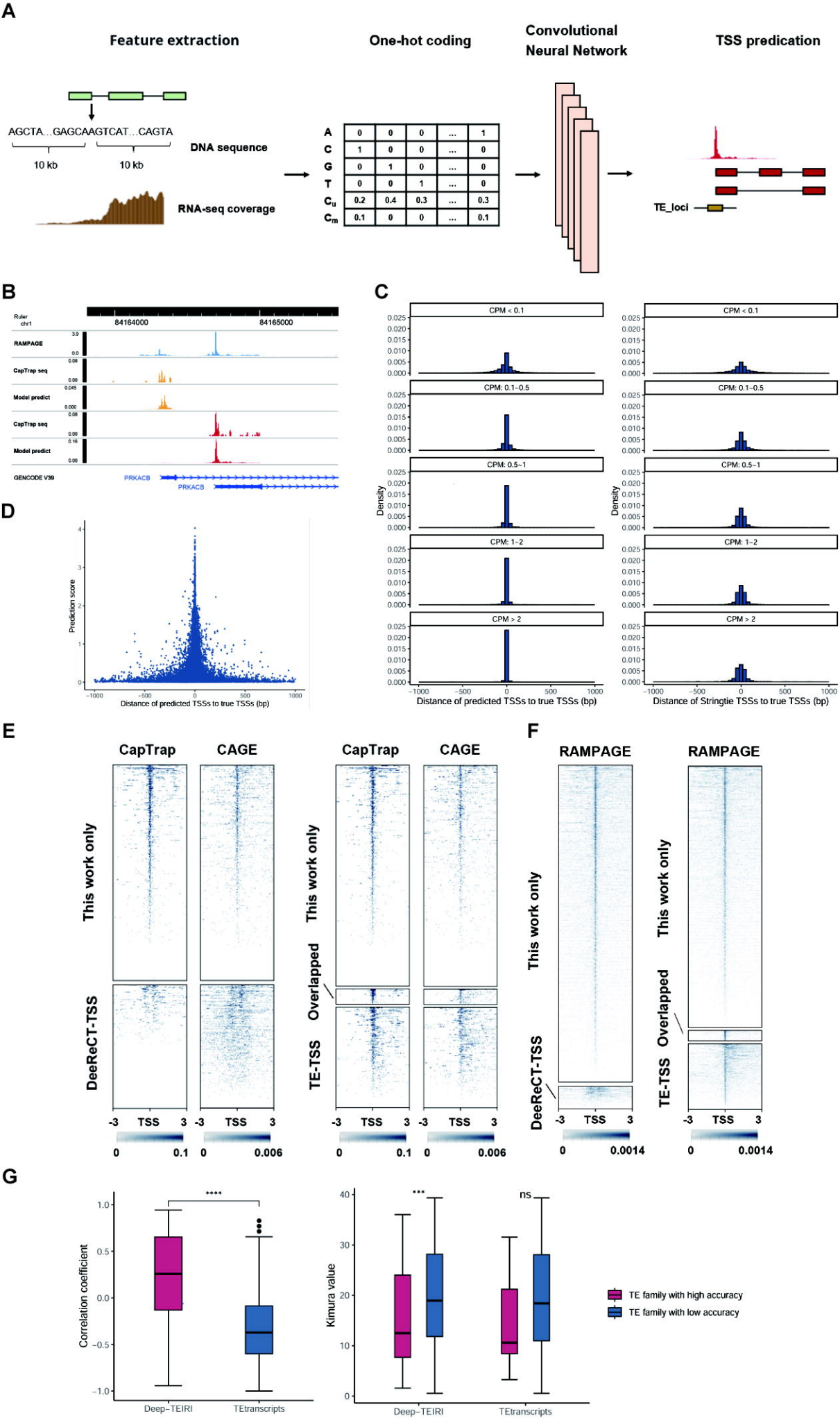
616 **Fig. 3 Applications of TEITbase**

617 (**A**) The homepage of TEITbase. (**B**) Box plot showing the promoter activity of L1HS in TCGA
618 tumors (top) and GTEx normal tissues (bottom). (**C**) Box plot showing the promoter activity of
619 L1HS in LUSC and adjacent normal tissues. (**D**) Box plot illustrating the promoter activity of
620 L1HS across different stages of LUSC. (**E**) Kaplan–Meier plots showing overall survival
621 associated with the promoter activity of L1HS in LUSC. (**F**) The reactome pathway enrichment
622 analysis of genes positively correlated with the promoter activity of L1HS in LUSC. (**G**, **H**)
623 Proportion of TCGA tumor samples (**G**) and GTEx normal tissue samples (**H**) expressing
624 TE_11880. (**I**) Box plot showing the expression level of TE_11880 in LUSC, STAD and
625 adjacent normal tissues. (**J**) Box plot illustrating the expression level of TE_11880 across
626 different stages of STAD. (**K**) Kaplan–Meier plots showing overall survival associated with the
627 expression of TE_11880 in COAD.

628


A
LTR
11758
SINE
11857
LINE
10911
LTR
LINE
SINE
DNA
Retroposon
RC
B
lncRNA
4771
protein_coding
9315
protein_coding
lncRNA
pseudogene
artifact
IG_C_gene
IG_V_gene
TR_C_gene
TR_V_gene
C
Drug ADME
Ion channel transport
Stimuli-sensing channels
Neurexins and neuroligins
Protein-protein interactions at synapses
O-linked glycosylation
Regulation of PTEN gene transcription
Induction of Cell-Cell Fusion
RNA Polymerase I Transcription Initiation
Regulation of TP53 Activity through Association with Co-factors
Count
pvalue
0.0025
0.0020
0.0015
0.0010
0.0005
Count
5
10
15
20
25
30
D
L1HS_5end
L1P1_5end
HERVH
L1PBa1_5end
MIRb
SVA_D
AluJr
LTR7
LTR17
AluJb
Odds ratio
p.adjust
3e-08
2e-08
1e-08
Counts
200
400
600
E
TE-Gene
Tumor specific
Number of Tumors
Cancer Type Distribution
Expression
Fraction
L1HS-EYA4 402.0 (1069/2) 1069
L1HS-KYNU 385.5 (341/0) 341
L1P2-PATE2 365.2 (323/0) 323
L1HS-MCCC2 364.1 (322/0) 322
L1HS-CFAP47 347.1 (307/0) 307
L1HS-ADGRL3 340.9 (604/1) 604
L1HS-CNTLN 334.7 (593/1) 593
L1HS-CRPPA 330.2 (878/2) 878
AluSz-NLRP7 315.6 (279/0) 279
L1HS-CNBD1 311.1 (275/0) 275
Number of Tumors
Frequency of Cancer Type
Log2(TPM + 1)
Fraction
TCGA-ACC
TCGA-BLCA
TCGA-BRCA
TCGA-CESC
TCGA-CHOL
TCGA-COAD
TCGA-DLBC
TCGA-ESCA
TCGA-GBM
TCGA-HNSC
TCGA-KICH
TCGA-KIRC
TCGA-KIRP
TCGA-LAML
TCGA-LGG
TCGA-LIHC
TCGA-LUAD
TCGA-LUSC
TCGA-MESO
TCGA-OV
TCGA-PAAD
TCGA-PCPG
TCGA-PRAD
TCGA-READ
TCGA-SARC
TCGA-SKCM
TCGA-STAD
TCGA-TGCT
TCGA-THCA
TCGA-THYM
TCGA-UCEC
TCGA-UCS
TCGA-UVM
F
L1HS_5end-CRPPA-TCGA-THYM
p = 0.0047
L1HS_5end-CFAP47-TCGA-LUAD
p = 0.032
L1P2_5end-PATE2-TCGA-LUAD
p = 0.042
L1HS_5end-EYA4-TCGA-LIHC
p = 0.0063
L1HS_5end-EYA4-TCGA-STAD
p = 0.019
Survival probability
Time


A
TEITbase
Browse
Analyses
Statistic
Downloads
Help
Contact
Welcome to TEITbase!
TEITbase is a comprehensive database designed to systematically study the transposable element (TE)-initiated transcription across 33 cancer types from The Cancer Genome Atlas (TCGA). Transposable elements (TEs) are mobile genetic elements that are widespread in eukaryotic genomes. As alternative promoters, TE can promote their own transcription or regulate the expression of neighboring genes, producing TE-initiated transcripts with TE-derived transcription start sites (TSSs). In caner, TE-initiated transcription may upregulate the expression of oncogenes (known as onco-exaptation) and generate tumor-specific antigens, which serve as potential targets for immunotherapy. Here, we developed the TEITbase to provide detailed insights into 38,995 expressed TE-initiated transcripts across various cancer types, offering data at both the TE family and transcript levels.
In TEITbase, users can:
1. Browse the basic information of 1,027 TE families and 38,995 expressed TE-initiated transcripts;
2. Search a gene symbol or TE family to identify associated TE-initiated transcripts;
3. Browse the basic information of 6,203 tumor-specific TE-initiated transcripts;
4. Browse or search in a cancer type to identify survival-associated, stage-associated and differential expressed TE-initiated transcripts (or TE families);
5. Perform the correlation analysis between specific TE-initiated transcript (or TE family) and protein-coding genes, and explore its potential function by Reactome enrichment analysis of positively correlated genes.
6. Download all results of analysis, figures of querying, and expression matrices for further integrative study.
TE
Exon1
Exon2
Exon3
Oncogene
Upregulate oncogene expression
UTR
ORF1
ORF2
UTR
Full-length TE transcript
Transposition increase genome instability
B
The expression of L1HS_5end in TCGA
CPM
The expression of L1HS_5end in GTEx
CPM
C
Mann-Whitney U, p=0.000
CPM
Tumor
Normal
n=502
n=51
D
Expression of L1HS_5end in LUSC
p = 0.035
CPM
Stage I
Stage II
Stage III
Stage IV
n=245
n=162
n=84
n=7
E
Strata
group=high
group=low
Survival probability
p = 0.038
Time
F
Formation of the cornified envelope
Apoptotic cleavage of cell adhesion proteins
Keratinization
Regulation of RUNX2 expression and activity
Regulation of Apoptosis
Neddylation
Apoptotic execution phase
Processing and activation of SUMO
Beta-catenin independent WNT signaling
Signaling by FGFR2 IIIa TM
p.adjust
Count
G
Proportion of Samples which TE_11880 was expressed in
Expressed
Unexpressed
Proportion of Samples
H
Proportion of Samples which TE_11880 was expressed in
Expressed
Unexpressed
Proportion of Samples
I
Mann-Whitney U, p=0.000
TPM
Tumor
Normal
n=502
n=51
Mann-Whitney U, p=0.000
TPM
Tumor
Normal
n=412
n=36
J
Expression of TE_11880 in STAD
p = 0.818
p = 0.014
TPM
Stage I
Stage II
Stage III
Stage IV
n=169
n=122
n=58
n=39
K
Strata
group=expressed
group=unexpressed
Survival probability
p = 0.069
Time


A
Feature extraction
One-hot coding
Convolutional Neural Network
TSS predication
AGCTA...GAGCAAGTCAT...CAGTA
10 kb
10 kb
DNA sequence
RNA-seq coverage
A
C
G
T
Cu
Cm
TE_loci
B
Ruler
chr1
84164000
84165000
RAMPAGE
CapTrap seq
Model predict
CapTrap seq
Model predict
GENCODE V39
PRKACB
PRKACB
C
CPM < 0.1
CPM: 0.1~0.5
CPM: 0.5~1
CPM: 1~2
CPM > 2
Density
Distance of predicted TSSs to true TSSs (bp)
Distance of Stringtie TSSs to true TSSs (bp)
D
Prediction score
Distance of predicted TSSs to true TSSs (bp)
E
CapTrap
CAGE
CapTrap
CAGE
This work only
DeeReCT-TSS
This work only
Overlapped
TE-TSS
-3
TSS
3
F
RAMPAGE
RAMPAGE
This work only
DeeReCT-TSS
This work only
Overlapped
TE-TSS
G
Correlation coefficient
Deep-TEIRI
TEtranscripts
Kimura value
Deep-TEIRI
TEtranscripts
TE family with high accuracy
TE family with low accuracy