

1 **Title (94 characters):**

2 Endogenous retroviruses drive KRAB zinc-finger family protein expression for  
3 tumor suppression

4

5 **Authors:**

6 Jumpei Ito<sup>1†</sup>, Izumi Kimura<sup>1†</sup>, Andrew Soper<sup>2</sup>, Alexandre Coudray<sup>3</sup>, Yoshio  
7 Koyanagi<sup>2</sup>, Hirofumi Nakaoka<sup>4</sup>, Ituro Inoue<sup>4</sup>, Priscilla Turelli<sup>3</sup>, Didier Trono<sup>3</sup>, and  
8 Kei Sato<sup>1,5\*</sup>

9

10 **Affiliations:**

11 <sup>1</sup> Division of Systems Virology, Department of Infectious Disease Control,  
12 International Research Center for Infectious Diseases, Institute of Medical  
13 Science, The University of Tokyo, Minato-ku, Tokyo 1088639, Japan

14 <sup>2</sup> Laboratory of Systems Virology, Institute for Frontier Life and Medical  
15 Sciences, Kyoto University, Kyoto 6068507, Japan

16 <sup>3</sup> School of Life Sciences, Ecole Polytechnique Federale de Lausanne (EPFL),  
17 1015 Lausanne, Switzerland

18 <sup>4</sup> Division of Human Genetics, National Institute of Genetics, Mishima, Shizuoka  
19 4118540, Japan.

20 <sup>5</sup> CREST, Japan Science and Technology Agency, Kawaguchi, Saitama  
21 3320012, Japan

22

23 <sup>†</sup> These authors contributed equally.

24 \* Correspondence: ksato@ims.u-tokyo.ac.jp (K.S.)

25

26 **Conflict of interest:** The authors declare that no competing interests exist.

27 **Short title:** Tumor suppression by ERVs and KZFPs

28 **Keywords:** gene regulatory network; pan-cancer analysis; TCGA; endogenous  
29 retrovirus; KRAB zinc-finger protein

30 **Abstract (150 words)**

31 Numerous genes are aberrantly expressed in tumors, but its cause remains  
32 unclear. Human endogenous retroviruses (HERVs) are repetitive elements in the  
33 genome and have a potential to work as enhancers modulating adjacent genes.  
34 Since numerous HERVs are activated epigenetically in tumors, their activation  
35 could alter gene expression globally in tumors and change the tumor  
36 characteristics. Here, we show the HERV activation in tumors is associated with  
37 the upregulation of hundreds of transcriptional suppressors, Krüppel-associated  
38 box domain-containing zinc-finger family proteins (KZFPs). KZFP genes are  
39 preferentially encoded nearby the activated HERVs in tumors and  
40 transcriptionally regulated by the adjacent HERVs. Increased HERV and KZFP  
41 expression in tumors was associated with better disease conditions. Many KZFPs  
42 could suppress the progressive characteristics of cancer cells by downregulating  
43 genes related to the cell cycle and cell-matrix adhesion. Our data suggest that  
44 HERV activation in tumors drives the concerted expression of KZFP genes for  
45 tumor suppression.

46 **Introduction**

47 Aberrant gene expression is a hallmark of cancers. Gene expression statuses in  
48 tumors are highly diverse among patients and are associated with the  
49 phenotypes of tumors such as proliferation, invasion/metastasis capacity, and  
50 therapeutic response as well as the clinical outcome of patients<sup>1</sup>. Particularly,  
51 many genes that are aberrantly expressed in tumors and associated with cancer  
52 progression have been identified<sup>2</sup>; however, the abnormality of the gene  
53 regulatory network underlying the aberrant expression of these genes in tumors  
54 is poorly understood<sup>3-5</sup>.

55 Decades of research have highlighted the significance of regulatory  
56 sequences derived from human endogenous retroviruses (HERVs) in the  
57 modulation of human gene expression<sup>6</sup>. HERVs are a type of transposable  
58 element (TE) that originates from ancient retroviral infection in host germ cells<sup>7</sup>.  
59 There are several hundred types of HERVs in the human genome, constituting  
60 8% of the genome<sup>8</sup>. Unlike other TEs, HERVs possess long terminal repeat (LTR)  
61 sequences that particularly densely contain transcriptional regulatory  
62 elements<sup>9,10</sup> and function as viral promoters<sup>7</sup>. In addition, HERV LTRs have the  
63 potential to function as promoters or enhancers of adjacent genes<sup>6</sup>. While most  
64 HERVs are epigenetically silenced in normal tissues, some HERVs function as  
65 part of the host gene regulatory network and play crucial roles in diverse  
66 biological events<sup>6,11-16</sup>. For instance, HERVs harboring STAT1- and IRF1-binding  
67 sites are essential for the interferon inducibility of genes related to the innate  
68 immune response<sup>17</sup>.

69 The expression of HERVs in normal tissues is controlled by epigenetic  
70 mechanisms such as DNA methylation and repressive histone modifications<sup>18,19</sup>;  
71 in contrast, HERV expression is highly elevated in various types of cancers<sup>20-24</sup>.  
72 Since the elevation of HERV expression in tumors is presumably caused by  
73 epigenetic reactivation, the expressed HERVs could upregulate the expression  
74 of adjacent genes. Therefore, it is possible that the derepression of numerous  
75 HERVs in tumors globally alters host gene expression and changes the  
76 characteristics of cancers<sup>25,26</sup>. To test this hypothesis, we investigated the multi-  
77 omics dataset of tumors provided by The Cancer Genome Atlas (TCGA)<sup>27</sup> and  
78 assessed the effects of HERV activation on host gene expression. We found that  
79 genome-wide HERV activation in tumors is associated with the upregulation of  
80 potent transcriptional suppressor genes, Krüppel-associated box (KRAB)  
81 domain-containing zinc-finger family protein (KZFP) genes<sup>28</sup>, which are

82 preferentially located in the vicinity of activated HERVs. Although KZFPs are  
83 widely known as transcriptional silencers against TEs, including HERVs<sup>28</sup>, our  
84 data highlight that the expression of KZFP genes is induced by the adjacent  
85 HERVs in tumors, leading to global gene expression alterations and phenotypic  
86 changes.

87 **Results**

88 **Characterization of expressed HERVs across 12 types of solid tumors**

89 We investigated the tumor RNA-sequencing (RNA-Seq) data of 5,470 patients  
90 provided by TCGA (**Data S1**). Only RNA-Seq reads that were uniquely mapped  
91 to the human genome were analyzed. A total of 11,011 loci of expressed HERVs  
92 were identified across twelve types of solid tumors (**Fig. 1A and Data S2**). While  
93 some HERVs were detected in only specific types of cancers, the majority of the  
94 expressed HERVs were detected in multiple types of cancers, and the sets of the  
95 expressed HERV loci were highly similar among all cancer types (**Figs. S1A and**  
96 **S1B**). In nine out of the twelve types of cancers, the overall expression levels of  
97 HERVs were increased compared to that in the adjacent normal tissues (**Fig. 1B**),  
98 consistent with previous reports<sup>20-24</sup>. Dimension reduction analysis based on  
99 HERV expression profiles showed that each type of cancer displays a  
100 distinguishable pattern of HERV expression (**Fig. 1C**). Importantly, the expressed  
101 HERVs preferentially overlapped with the nucleosome-free regions (NFRs)  
102 determined by Assay for Transposase-Accessible Chromatin Sequencing  
103 (ATAC-Seq) (**Fig. 1D**), suggesting that expressed HERVs in tumors are  
104 epigenetically active and have the potential to modulate adjacent gene  
105 expression.

106

107 **Transcriptome signatures associated with the global derepression of**  
108 **HERVs in tumors**

109 Although HERV expression levels tended to be elevated in tumors compared to  
110 the corresponding normal tissues (**Fig. 1B**), the genome-wide expression levels  
111 of HERVs in tumors were highly heterogeneous among patients, even within the  
112 same cancer type (**Figs. S1C and 1E**). Notably, such global HERV activation  
113 occurred regardless of the type of HERV (**Figs. 1E, S1D, and S1E**), although the  
114 regulatory sequences of these HERVs were highly diverse<sup>10</sup>. In many types of  
115 cancers, the global expression levels of HERVs were negatively correlated with  
116 the DNA methylation levels of CpG sites that are on or proximal (<1 kb) to the  
117 expressed HERVs (**Figs. S1F and S1G**), suggesting that the epigenetic  
118 derepression of HERVs is a cause of the elevation of HERV expression in tumors.

119 To elucidate the effects of the global derepression of HERVs on host  
120 gene expression in tumors, we investigated the genes whose expression was  
121 associated with HERV derepression in tumors. We assessed the correlation of  
122 the expression level of each gene with the total expression level of HERVs in

123 tumors and subsequently performed gene set enrichment analysis (GSEA)<sup>29</sup>  
124 based on the above correlation scores. We found that the genes showing a  
125 correlation with HERVs were highly similar among distinct types of cancers (**Figs.**  
126 **S2A and S2B**). KZFP genes (i.e., genes possessing the KRAB domain) were  
127 highly upregulated upon the elevation of HERV expression (**Fig. 1F**). Most KZFP  
128 genes were co-expressed with each other (**Fig. S2C**) and with the major groups  
129 of HERVs in tumors (**Fig. S2D**). Additionally, genes related to the cell cycle, cell-  
130 matrix adhesion, and immune response were downregulated upon the  
131 upregulation of HERV and KZFP genes (**Figs. 1F and S3**). We investigated  
132 another RNA-Seq dataset of cancer cell lines provided by the Cancer Cell Line  
133 Encyclopedia (CCLE)<sup>30</sup> and verified that the expression of HERVs was positively  
134 associated with KZFP genes and negatively associated with genes related to the  
135 cell cycle, cell-matrix adhesion, and immune response (**Fig. S4**). These results  
136 suggest that these associations are arise from the expressional changes that  
137 occur in cancer cells themselves.

138

### 139 **Transcriptional activation of KZFP genes by surrounding HERVs**

140 We hypothesized that derepressed HERVs near KZFP genes induce the  
141 expression of these genes, leading to the synchronized expression of HERVs  
142 and KZFP genes in tumors. It is known that KZFP genes form genomic clusters,  
143 particularly on chromosome 19 in the human genome<sup>31</sup>. We found that the  
144 expressed HERVs in tumors were preferentially present in these clusters of KZFP  
145 genes (**Fig. 2A**). The expressed HERVs in tumors and those with transcriptional  
146 regulatory signals (i.e., NFRs or enhancers defined by GeneHancer<sup>32</sup>) were  
147 highly enriched in the vicinity of transcriptional start sites (TSSs) of KZFP genes  
148 (**Figs. 2B and S5A**). Several types of HERV LTRs, such as LTR70, LTR25,  
149 LTR5B, and LTR5Hs, showed particularly strong enrichments around the TSSs  
150 of KZFP genes (**Fig. 2C**).

151 We next investigated the association between the transcriptional  
152 upregulation of KZFP genes and the epigenetic activation of the adjacent HERVs  
153 in tumors. The mean expression level of KZFP genes was associated with the  
154 mean NFR activity of the expressed HERVs around those genes in tumors (**Fig.**  
155 **2D**). Additionally, the mean expression level of KZFP genes in tumors was  
156 negatively correlated with the mean DNA methylation level of the CpG sites that  
157 are on or proximal (<1 kb) to the expressed HERVs around those genes (**Fig.**

158 **S5B**). These findings suggest that the expression of KZFP genes in tumors is  
159 upregulated by the epigenetic derepression of adjacent HERVs.

160 Next, we searched for genes possibly regulated by respective HERV loci  
161 according to the co-expression, NFR-expression, and DNA methylation–  
162 expression relationships as well as the pre-defined enhancer–gene links<sup>32</sup> (**Fig.**  
163 **2E, left**) (**Data S3**). In these four types of predictions, KZFP genes were highly  
164 enriched in the set of genes possibly regulated by HERVs (**Fig. S5C**), supporting  
165 the significance of HERVs in the transcriptional regulation of KZFP genes. Based  
166 on these interactions, we constructed a network representing the regulation of  
167 KZFP genes by HERVs (**Fig. 2E, middle**). We identified several “hub” HERV loci,  
168 which are connected to many KZFP genes in the network and are likely to be  
169 involved in the transcriptional regulation of these genes (**Fig. 2E, right**).

170 To experimentally address the significance of HERVs in the  
171 transcriptional modulation of KZFP genes in cancer cells, we performed CRISPR-  
172 Cas9 excision of a hub HERV locus (HERV-enhancer1; **Fig. 2E, right**) in human  
173 lung adenocarcinoma (LUAD) (A549) cells (**Figs. 2F and S6**). We particularly  
174 selected this HERV locus because it displayed active histone marks in A549 cells  
175 (**Fig. 2F**). We demonstrated that the homozygous excision of this HERV  
176 decreased the expression of adjacent genes, including many KZFP genes (**Fig.**  
177 **2G**).

178

### 179 **Biological relevance of the expression status of KZFPs and HERVs to 180 cancer progression**

181 Since KZFPs are potent transcriptional suppressors<sup>28</sup>, it is possible that the  
182 synchronized induction of many KZFPs in tumors would alter gene expression  
183 globally and change the characteristics of tumors. We found that somatic  
184 mutations accumulated particularly in the DNA-binding interfaces of KZFPs in  
185 tumors (**Fig. 3A**), suggesting that the aberration of the DNA-binding activity of  
186 KZFPs is associated with tumor progression. We therefore investigated the  
187 associations of the expression of KZFPs and HERVs with the clinical outcomes  
188 of cancer patients and found the following marked associations: in four (bladder  
189 carcinoma (BLCA), head and neck squamous cell carcinoma (HNSC), kidney  
190 renal papillary cell carcinoma (KIRP), and LUAD) out of twelve types of cancers,  
191 patients with high expression levels of KZFPs and HERVs in the tumors tended  
192 to show a better prognosis than those with low expression levels (**Figs. 3B and**  
193 **S7**). Furthermore, we examined the association of the expression levels of

194 respective genes and HERV loci with cancer prognosis and found that KZFP  
195 genes and HERVs tended to show a stronger association with better prognosis  
196 than the other genes (**Figs. 3C, 3D, and S8**). Conversely, genes related to the  
197 cell cycle and cell-matrix adhesion tended to show a stronger association with a  
198 worse prognosis (**Fig. S8B**). We further examined the association of the overall  
199 expression level of KZFPs and cancer stage, which reflects the degree of  
200 invasion and metastasis of tumors. The overall expression level of KZFPs  
201 decreased as the cancer stage progressed in multiple types of cancers (**Figs. 3E,**  
202 **3F, and S9**). Conversely, genes related to the cell cycle and cell-matrix adhesion  
203 increased as the cancer stage progressed (**Fig. S9B**).

204

205 **Gene expression and phenotypic changes induced by the overexpression**  
206 **of KZFP genes in LUAD cells**

207 The analysis of the chromatin immunoprecipitation-sequencing (ChIP-Seq)  
208 dataset of KZFPs (Imbeault et al.<sup>33</sup>) showed that many KZFPs preferentially  
209 bound to genes related to the cell cycle and cancer-associated signaling  
210 pathways, such as TGF-related pathways (TGF- $\beta$ , BMP, SMAD2/3 pathways)  
211 and Wnt pathway (**Fig. S10**). These pathways are critical for the regulation of  
212 cell-matrix adhesion and are associated with cell migration/invasion and  
213 proliferation in cancers<sup>34,35</sup>. Notably, the expression levels of the genes related to  
214 the cell cycle and cell-matrix adhesion were negatively correlated with those of  
215 KZFP genes in tumors (**Fig. S3**) and associated with worse disease conditions  
216 (**Figs. S8B and S9B**), suggesting that KZFPs can modulate cancer phenotypes  
217 by altering the expression of these genes.

218 To assess the effects of elevated KZFP expression on cancer cells, we  
219 established a panel of A549 LUAD cells overexpressing 30 types of KZFPs  
220 (referred to as A549/KZFP cells) (**Fig. S11**) and subsequently investigated the  
221 phenotypic and gene expression changes caused by these KZFPs. Most of the  
222 tested KZFPs induced apoptosis (**Fig. 4A**), while many of the KZFPs suppressed  
223 cell growth, migration, and invasion (**Figs. 4B–D**). In total, the expression of 2,368  
224 genes was altered by the overexpression of any of the tested KZFPs (**Fig. 4E**).  
225 Of note, the genes related to the cell cycle and cell-matrix adhesion were  
226 significantly downregulated by the overexpression of many types of KZFPs (**Fig.**  
227 **S12A**). Although the phenotypic and gene expression alterations caused by  
228 KZFPs were relatively similar among all types of A549/KZFP cells (**Figs. 4E and**  
229 **S12B**), these alterations were clearly associated (**Figs. S12C and S12D**),

230 suggesting that the phenotypic changes in A549/KZFP cells were caused by  
231 alterations in gene expression. Overall, we demonstrated that a substantial  
232 fraction of KZFPs could suppress the phenotypes associated with cancer  
233 progression by altering gene expression in LUAD cells.

234 To identify the target genes of KZFPs that are likely to be critical for  
235 cancer progression, we developed a scoring system for genes according to their  
236 expressional negative correlation with KZFPs, the association of their expression  
237 with worse clinical conditions, and their expressional suppression in A549/KZFP  
238 cells as well as considering the frequency of KZFP binding (**Fig. S13**). In this  
239 system, the high-scored genes included a substantial number of genes related to  
240 the cell cycle and cell-matrix adhesion (**Figs. 4F and S13D**). In particular, many  
241 genes related to cytoskeletal regulation (i.e., *ACTG1*, *GIT1*, *PFN1*, *RAC1*, and  
242 *RRAS*) that are critical for cell-matrix adhesion and modulate cell  
243 migration/invasion and proliferation<sup>36</sup> were identified as targets of KZFPs.  
244 Additionally, a serine-threonine kinase gene (*AURKB*) and ubiquitin-proteasome  
245 pathway genes (*UBC*, *RPS27A*, *PSMB4*, and *PSMA7*) that are critical for cell  
246 cycle regulation<sup>37,38</sup> were identified.

247

248 **Transcriptional modulation of cancer phenotype-associated KZFP genes  
249 by the adjacent HERVs in LUAD cells**

250 *ZNF75D* was capable of altering all four investigated cancer phenotypes (**Figs.**  
251 **4A–E**). In the region approximately 5 kb upstream of a TSS of *ZNF75D*, two  
252 HERV integrants (LTR5\_Hs and THE1D-int) were present (**Fig. 4G**). THE1D-int  
253 was co-expressed with *ZNF75D* in LUAD tumors (**Fig. 4H**). A luciferase reporter  
254 assay showed that these two HERV elements exhibit enhancer activity in A549  
255 cells (**Figs. 4I and S14**) regardless of their orientation (**Fig. S14E**). To test the  
256 significance of these HERVs on the transcriptional modulation of *ZNF75D*, we  
257 excised these two HERVs using the CRISPR-Cas9 system in A549 cells (**Fig.**  
258 **S6**) and demonstrated that the deletion of these HERVs decreased *ZNF75D*  
259 expression in an allelic number-dependent manner (**Fig. 4J**). These results  
260 suggest that these HERVs are involved in the transcriptional modulation of the  
261 *ZNF75D* in LUAD cells. Moreover, for 12 out of the 30 KZFP genes tested above,  
262 we investigated the transcriptional modulation potential of the adjacent HERVs  
263 by performing a luciferase reporter assay. HERVs in the vicinity of 7 KZFP genes  
264 (*ZNF141*, *ZNF248*, *ZNF30*, *ZNF320*, *ZNF44*, *ZNF611*, and *ZNF846*) enhanced  
265 the promoter activities of these genes in A549 cells (**Figs. S14F and S14G**).

266 These results support the significance of HERVs in the transcriptional regulation  
267 of these KZFP genes in cancer cells.

268

269 **Discussion**

270 In the present study, we found that the global activation of HERVs occurred in a  
271 substantial fraction of tumors (**Figs. 1E and S1C**). Although the ultimate cause  
272 of HERV activation in tumors is still unclear, the attenuation of the epigenetic  
273 silencing of HERVs would be related to HERV activation (**Figs. S1F and S1G**).  
274 HERV activation was associated with the synchronized induction of KZFP gene  
275 expression (**Figs. 1F and S2C**). Further analyses including *in vitro* experiments  
276 suggest that KZFPs are transcriptionally regulated by the adjacent HERVs (**Figs.**  
277 **2 and 4G–4J**). Notably, the coordinated induction of KZFP expression was  
278 clearly associated with better disease conditions in multiple types of cancers  
279 (**Figs. 3B–F**). A substantial fraction of KZFPs could suppress the phenotypes  
280 related to tumor progression by altering gene expression in cultured cells (**Figs.**  
281 **4A–4E**). These findings suggest that a repertoire of KZFPs cooperatively exerts  
282 suppressive effects on tumor progression. Collectively, we highlight the presence  
283 of tumor heterogeneity driven by the gene regulatory network comprising HERVs  
284 and KZFPs — the activation of HERVs in tumors induces the expression of  
285 adjacent KZFP genes, leading to the suppression of the progressive  
286 characteristics of cancers by altering gene expression.

287 Although our data highlight the significance of HERVs in the  
288 transcriptional regulation of KZFPs (**Fig. 2**), it is widely considered that one of the  
289 primary functions of KZFPs is the silencing of the disordered expression of TEs,  
290 including HERVs<sup>28</sup>. Such seemingly paradoxical findings suggest the presence  
291 of a transcriptional negative feedback loop between HERVs and KZFPs — once  
292 HERVs are derepressed globally, the regulatory activities of HERVs around  
293 KZFP genes are reactivated simultaneously, resulting in the induction of KZFP  
294 expression. In other words, KZFP genes seem to utilize HERVs as their  
295 regulatory sequences to detect the global derepression of HERVs. A previous  
296 report proposed the possibility that such negative feedback functions when the  
297 embryonic genome activation occurs to silence the activation of TEs including  
298 HERVs effectively<sup>39</sup>. This feedback system, at least regarding the induction of  
299 KZFP expression by HERVs, seems to work also in cancer cells and cause  
300 aberrant gene expression in tumors.

301 **Materials and Methods**

302

303 **Ethical approval**

304 The utilization of the TCGA multi-omics dataset was authorized by the National  
305 Cancer Institute (NCI) data access committee through the Database of  
306 Genotypes and Phenotypes (dbGaP; <http://dbgap.ncbi.nlm.nih.gov>) for the  
307 following projects: “Systematic identification of reactivated human endogenous  
308 retroviruses in cancers (#15126)”, “Effects of the genome-wide activation of  
309 human endogenous retroviruses on gene expression and cancer phenotypes  
310 (#18470)”, and “Screening of subclinical viral infections in healthy human tissues  
311 (#19481)”.

312

313 **Construction of the gene-HERV transcript model for RNA-Seq analysis**

314 For the gene transcript model, GENCODE version 22 (for GRCh38/hg38)  
315 obtained from the GENCODE website (<http://www.gencodegenes.org/>) was used.  
316 For the HERV transcript model, the RepeatMasker output file (15-Jan-2014; for  
317 GRCh38/hg38) obtained from the UCSC genome browser  
318 (<http://genome.ucsc.edu/>) was used. From the gene model, transcripts with the  
319 flag “retained intron” were excluded. From the HERV model, HERV loci with low  
320 reliability scores (i.e., Smith-Waterman score < 2,500) were excluded.  
321 Additionally, the regions of HERV loci overlapping with the gene transcripts were  
322 also excluded. A gene-HERV transcript model was generated by concatenating  
323 the gene and HERV models. This model includes 60,483 protein-coding/non-  
324 coding genes in addition to 138,124 HERV loci, which occupy 3.4% of the  
325 genome.

326

327 **RNA-Seq data analysis of the TCGA dataset**

328 Poly A-enriched RNA-Seq (mRNA-Seq) data provided by TCGA were analyzed.  
329 Of the RNA-Seq data, we analyzed only the data produced by pair-ended  
330 sequencing with a read length of 48–50 bp. The BAM-formatted read alignment  
331 file (for GRCh38/hg38) of the RNA-Seq data was downloaded from the Genomic  
332 Data Commons (GDC) data portal site (<http://portal.gdc.cancer.gov/>) using the  
333 GDC Data Transfer Tool (<http://gdc.cancer.gov/access-data/gdc-data-transfer-tool/>). To measure expression levels of HERVs and genes, RNA-Seq fragments  
335 mapped on HERVs and the exons of genes were counted using Subread  
336 featureCounts<sup>40</sup> with the BAM file and the gene-HERV transcript model. The

337 option “fracOverlap” was set at 0.25. The RNA-Seq fragments assigned to  
338 multiple features were not counted.

339 To control the quality of the RNA-Seq data used in the present study, we  
340 checked the proportion of non-assigned RNA-Seq fragments (i.e., the fragments  
341 that were uniquely mapped on the reference genome but not on HERVs or exons  
342 of genes) in each sequence library. For this proportion, outlier libraries were  
343 detected recursively using the Smirnov-Grubbs test (the threshold was set at  
344 0.05). These outlier libraries were excluded from the downstream analyses. The  
345 final RNA-Seq data used in this study are summarized in **Data S4**.

346 The expression count matrices of the RNA-Seq data were separately  
347 prepared for the datasets of the respective types of cancers. In addition, the  
348 expression matrix including all tumor data was also prepared. Furthermore, the  
349 expression matrix, including the data from the tumors and corresponding normal  
350 adjacent tissues, was also prepared for each type of cancer. Genes and HERVs  
351 with low expression levels were removed from the expression matrix as follows.  
352 The counts per million (CPM) value of each gene and HERV locus were  
353 calculated in the respective RNA-Seq libraries. Subsequently, genes and HERVs  
354 were discarded from the expression matrix if the 90th percentile of CPM values  
355 was less than 0.2.

356 In each type of cancer, the expressed HERVs in tumors, which are  
357 HERVs included in the expression matrix of the corresponding types of cancers,  
358 were determined.

359 The total expression level of the HERVs was normalized as CPM. The  
360 expression levels of genes and HERV loci were normalized using variance-  
361 stabilizing transformation (VST) implemented in DESeq2 (version 1.18.1)<sup>41</sup>. This  
362 VST-normalized expression level was used unless otherwise noted.  
363

#### 364 **RNA-Seq data analysis of the CCLE dataset**

365 The BAM-formatted read alignment file (for GRCh37/hg19) of the mRNA-Seq  
366 data was downloaded from the GDC data portal site  
367 (<http://portal.gdc.cancer.gov/>) using the GDC Data Transfer Tool  
368 (<http://gdc.cancer.gov/access-data/gdc-data-transfer-tool/>). The RNA-Seq data  
369 of CCLE used in this study are summarized in **Data S5**. Since the gene-HERV  
370 transcript model prepared above is for GRCh38/hg38, the genomic coordinates  
371 of the gene-HERV transcript model were converted to those in GRCh37/hg19  
372 using UCSC liftOver

373 ([http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86\\_64/liftOver](http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver)). The option  
374 “minMatch” was set at 0.95. The generation of the expression count matrix,  
375 filtering of genes and HERVs with low expression levels, and normalization of the  
376 expression data were performed using the same procedures as those in the  
377 above section (“**RNA-Seq data analysis of the TCGA dataset**”).

378

### 379 **RNA-Seq analysis of A549/KZFP cells**

380 The RNA-Seq sample information is summarized in **Data S6**. Low quality  
381 sequences in RNA-Seq fragments were trimmed using Trimmomatic (version  
382 0.36)<sup>42</sup> with the option “SLIDINGWINDOW:4:20”. RNA-Seq fragments were  
383 mapped to the human reference genome (GRCh38/hg38) using STAR (ver.  
384 2.5.3a)<sup>43</sup> with the gene-HERV transcript model. STAR was run using the same  
385 options and parameters as those used in the GDC mRNA Analysis Pipeline  
386 ([https://docs.gdc.cancer.gov/Data/Bioinformatics\\_Pipelines/Expression\\_mRNA\\_Pipeline](https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline)). The generation of the expression count matrix, filtering of genes and  
387 HERVs with low expression levels, and normalization of the expression data were  
388 performed using the same procedures as those in the above section (“**RNA-Seq**  
389 **data analysis of the TCGA dataset**”).

390

### 392 **Dimension reduction analysis of HERV expression profiles using t- 393 distributed stochastic neighbor embedding (t-SNE)**

394 The expression matrix including all tumor data was used in this analysis. The  
395 expression levels of the 1000 most highly expressed HERVs were used in the  
396 analysis. t-SNE analysis was performed using the “Rtsne” R package. For the  
397 analysis, the first 10 principle components of the HERV expression profiles were  
398 used, and the parameter “perplexity” was set at 70.

399

### 400 **ATAC-Seq data analysis**

401 The ATAC-Seq data of tumors and normal adjacent tissues provided by TCGA  
402 (TCGA-ATAC\_PanCan\_Log2Norm\_Counts.rds) was downloaded from the GDC  
403 website (<https://gdc.cancer.gov/about-data/publications/ATACseq-AWG>). This  
404 file contains the normalized read count matrix comprising all ATAC-Seq samples  
405 (n=796) and ATAC-Seq peaks (NFRs) (n=562,709) analyzed in the previous  
406 study<sup>4</sup>. In the respective types of cancers, the upper 1/4 of NFRs with respect to  
407 the mean value were regarded as the NFRs that are active in the corresponding  
408 cancer types.

409 To calculate the fold enrichment of the overlaps between the expressed  
410 HERVs in tumors and NFRs, randomization-based enrichment analysis was  
411 performed as follows: genomic regions of NFRs were randomized using bedtools  
412 “shuffle”<sup>44</sup> and subsequently, the number of NFRs on the expressed HERVs was  
413 counted. This process was repeated 1,000 times, and the mean value of the  
414 counts in the randomized datasets was regarded as the random expectation  
415 value. The fold enrichment was calculated by dividing the observed count by the  
416 random expectation value.

417

#### 418 **DNA methylation data analysis**

419 The DNA methylation data (produced by the methylation microarray  
420 HumanMethylation450 (Illumina)) of tumors and normal tissue controls were  
421 downloaded from the GDC data portal (<http://portal.gdc.cancer.gov/>) using the  
422 GDC Data Transfer Tool (<http://gdc.cancer.gov/access-data/gdc-data-transfer-tool/>). These data describe the methylation level (beta value; proportion of  
423 methylated CpGs at a CpG site) of each probe in the array. Probes overlapping  
424 with single nucleotide polymorphisms (SNPs) with >0.05 minor allele frequency  
425 were excluded from the analysis using the function “rmSNPandCH” implemented  
426 in the “DMRcate” library<sup>45</sup> in R. The CpG sites that were on or proximal (<1 kb)  
427 to HERVs were extracted using the “slop” and “intersect” functions in bedtools<sup>44</sup>.  
428 DNA methylation data used in this study is summarized in **Data S7**.

429

#### 430 **Preparation of gene sets for enrichment analyses**

431 As sources of gene sets, “GO biological process”, “GO cellular component”,  
432 “MSigDB canonical pathway”, and “InterPro” were used. The gene sets in these  
433 sources were concatenated and used. “InterPro” is the collection of gene sets  
434 according to protein families or domains and includes the gene set “KRAB”,  
435 representing the KZFP family genes. “GO biological process” and “GO cellular  
436 component” were obtained from Gene Ontology (GO) consortium  
437 (<http://geneontology.org/>; GO validation date: 08/30/2017); “canonical pathway”  
438 was from MSigDB (<http://software.broadinstitute.org/gsea/msigdb>; version 6.1);  
439 and “InterPro” was from BioMart on the Ensembl website  
440 (<https://www.ensembl.org>; on 2/13/2018).

441 In addition, we defined the gene sets “HERVs” and “HERVs around KZFP  
442 genes”. The “HERV” gene set includes all expressed HERVs in tumors, while  
443 “HERVs around KZFP genes” includes the HERVs present in the genomic

445 regions within 50 kb from the TSSs of KZFP genes expressed in tumors. These  
446 gene sets were used in **Figs. 3D, 3F, S8B, and S9B** in addition to the pre-defined  
447 gene sets.

448 Furthermore, we defined gene sets according to the expressional  
449 negative correlation with HERVs or KZFP genes as follows. In the respective  
450 tumor datasets of TCGA, Spearman's correlations between the expression levels  
451 of respective genes and the total expression level of HERVs were calculated, and  
452 genes were ranked according to their median value in the datasets. The top 100,  
453 200, and 500 genes with respect to their negative expressional correlation with  
454 HERVs were used as gene sets. Using the same procedures as above, the top  
455 100, 200, and 500 genes with respect to the negative expressional correlation  
456 with KZFP genes were extracted and used as gene sets. As the representative  
457 value of KZFP expression, the gene set-wise expression score (Gene Set  
458 Variation Analysis (GSVA) score<sup>46</sup>) of the KZFP genes was used. The GSVA  
459 score is described in the following section (“**Calculation of the gene set-wise**  
460 **expression score using GSVA**”). These gene sets were used in **Fig. S10** in  
461 addition to the pre-defined gene sets.

462

### 463 **Calculation of the gene set-wise expression score using GSVA**

464 The VST-normalized expression matrix was converted to the gene set-wise  
465 expression score matrix using GSVA<sup>46</sup> with the gene sets prepared above. The  
466 option “minimum size of gene set” was set at 20.

467

### 468 **GSEA**

469 To perform GSEA<sup>29</sup>, the R package “fgsea”<sup>47</sup>, a fast implementation of GSEA,  
470 was used. The parameters of “number of permutations” and “minimum size of  
471 gene set” were set at 10,000 and 50, respectively. In the analyses of **Figs. 1F**  
472 **and S4**, the Spearman's correlations between the expression levels of respective  
473 genes and the total expression level of HERVs were used as statistical scores.  
474 In the analysis of **Fig. S3**, the Spearman's correlations between the expression  
475 levels of respective genes and the GSVA score of the KZFP genes were used. In  
476 the analyses of **Figs. 3D and S8B**, Z scores in Cox proportional hazards  
477 regression were used (the Z score is described in the “**Survival analysis of the**  
478 **cancer patients**” section). In the analysis of **Fig. S12A**, Wald statistics of the  
479 respective genes in the differential expression analysis were used (the Wald  
480 statistic is described in the “**Differential expression analysis**” section).

481

482 **Summarizing the results of GSEA and GO enrichment analysis by removing  
483 redundant gene sets**

484 Since the gene members of some gene sets highly overlapped with each other,  
485 redundant gene sets were removed from the results of the enrichment analyses  
486 as follows. Gene sets were ranked according to the score of interest (e.g., the  
487 mean value of normalized enrichment score (NES)). If the gene members of a  
488 certain gene set were highly overlapped with those of the upper-ranked gene sets,  
489 the gene set was removed from the result. As a statistic of the overlap, the  
490 Szymkiewicz–Simpson coefficient was used, and two gene sets were regarded  
491 as highly overlapped if the coefficient was greater than 0.7. This gene set filtering  
492 was applied to the analyses shown in **Figs. 1F, S8B, S9B, S10, and S12A**, which  
493 show only the top-ranked gene sets.

494

495 **GO enrichment analysis to identify gene sets that are preferentially present  
496 in the vicinity of the expressed HERVs**

497 Randomization-based GO enrichment analysis was performed as follows. Only  
498 genes whose expression levels were detected in the TCGA tumor datasets were  
499 used. Regions of interest were defined as the regions within 50 kb from the TSSs  
500 of the gene members of a certain gene set. The genomic regions of HERVs were  
501 randomized using the “shuffle” function of bedtools<sup>44</sup>, and subsequently, the  
502 number of HERVs in the region of interest was counted. This process was  
503 repeated 1,000 times, and the mean value of the counts in the randomized  
504 datasets was regarded as the random expectation value. The fold enrichment  
505 was calculated by dividing the observed count by the random expectation value.

506 Additionally, we calculated the fold enrichments of HERVs in the regions  
507 within 10, 100, and 500 kb and 1 mb from the TSSs of the KZFP genes using the  
508 same procedures as above.

509

510 **Prediction of genes regulated by HERVs**

511 The regulatory interactions between HERV loci and genes were predicted  
512 according to the following information: co-expression between HERVs and genes,  
513 positive correlations between HERV NFR activities and gene expression,  
514 negative correlations between HERV DNA methylation and gene expression, and  
515 pre-defined links between the regulatory sequences on HERVs and genes. The  
516 co-expression interaction was used in only pairs of HERVs and genes within 50

517 kb of each other, while the NFR-expression, methylation-expression, and pre-  
518 defined interactions were used in only pairs of HERVs and genes within 500 kb  
519 of each other. A co-expression interaction was defined if the expression of the  
520 HERV and gene were positively correlated (Spearman's correlation > 0.4) in any  
521 type of cancer in TCGA. A methylation-expression interaction was defined if the  
522 DNA methylation level of the CpG site that is on or proximal (<1 kb) to a HERV  
523 and the expression of the gene were negatively correlated (Spearman's  
524 correlation < -0.3) in any type of cancer or in the pan-cancer dataset in TCGA.  
525 As the source of NFR-expression interactions, the interactions defined in a  
526 previous study<sup>4</sup> were used. As the source of pre-defined regulatory interactions,  
527 the interactions recorded in GeneHancer version 4.7 obtained from GeneLoc  
528 database (<https://genecards.weizmann.ac.il/geneloc/index.shtml>) were used.  
529

### 530 **Mutation analysis**

531 To define DNA-binding amino acids of KZFP genes, we first determined the  
532 precise genomic positions of KRAB and C2H2 zinc-finger domains as follows.  
533 For both of KRAB and C2H2 zinc-finger domains, Hidden Markov Model (HMM)  
534 profiles were generated using hmmbuild from HMMER2 [<http://hmmer.org/>].  
535 Multiple sequence alignments used to build the HMM profiles were generated  
536 from the seed sequences downloaded from Pfam  
537 [<https://academic.oup.com/nar/article/44/D1/D279/2503120>]. Next, the human  
538 reference genome (GRCh37/hg19) was scanned using hmmpfam from HMMER2  
539 with the built HMM profiles. The both strands of chromosomes translated in 3  
540 reading frames were scanned. KZFP genes were collected if a KRAB domain had  
541 ≥2 downstream C2H2 zinc-fingers found on the same strand within 40kb, which  
542 corresponds to the maximum length from the first base of KRAB domain to the  
543 last base of zinc-finger domain. Detected KZFP genes were then annotated  
544 according to the Ensembl annotation (version 92; for GRCh37/hg19). Finally, the  
545 DNA-binding amino acid positions were inferred from C2H2 zinc-fingers  
546 annotated above, taking position 4th, 6th, 7th, and 10th (also called -1, +2, +3,  
547 and +6 positions) after the second cysteine of the C2H2. Only zinc-finger with a  
548 canonical C2H2 structure and associated with a KRAB domain was taken into  
549 account.

550 Processed mutation data were obtained from International Cancer  
551 Genome Consortium (ICGC) (release 27) (<https://icgc.org/>). Then we measured  
552 the somatic missense mutation density (counts per mb per patient) of KZFP

553 genes in the DNA-binding amino acids and the whole exonic regions of the  
554 canonical transcript.

555

### 556 **Survival analysis of the cancer patients**

557 The overall survival rate of the cancer patients was used for survival analyses  
558 with the R package “survival”. The survival curve of the patients was estimated  
559 by the Kaplan–Meier method, and statistical significance was evaluated by the  
560 two-sided log-rank test. With respect to the expression level of interest, the upper  
561 and lower third of patients were regarded as patients with higher and lower  
562 expression statuses, respectively. In **Figs. S7B and S7C**, the patients were  
563 stratified according to the GSVA expression scores of HERVs and KZFPs in  
564 tumors, respectively. In **Figs. 3B and S7A**, the patients were stratified according  
565 to the mean value of the GSVA scores of HERVs and KZFPs in tumors.

566 To examine the association of the expression level of each gene and  
567 HERV locus with the prognosis of cancer patients, Cox proportional hazards  
568 regression analysis was performed with adjustment for the effects of sex and race  
569 of the patients. In addition to HERVs, genes that were included in any of the gene  
570 sets prepared above were used.

571

### 572 **Association analysis of gene expression and cancer progression**

573 Prostate adenocarcinoma (PRAD) tumors were excluded from the analysis since  
574 information on cancer stage for most PRAD patients was not available. In the  
575 analysis, cancer stage was regarded as an interval scale. For each type of cancer,  
576 the association between the expression of each gene and the progression of the  
577 cancer stage was evaluated by single linear regression. Similarly, the association  
578 between the GSVA score of each gene set and the progression of cancer stage  
579 for each type of cancer was evaluated using the same procedure. To evaluate  
580 the pan-cancer association of the GSVA score of each gene set and the  
581 progression of cancer stage, multiple linear regression analysis with adjustment  
582 for the effects of cancer type was performed.

583

### 584 **Analysis of a publicly available ChIP-Seq dataset of KZFPs**

585 This analysis was based on a publicly available ChIP-Seq dataset of KZFPs in  
586 HEK293T cells presented in a previous study (Imbeault et al.<sup>33</sup>; GEO accession  
587 #: GSE78099). Information on pre-defined ChIP-Seq peaks (GSE78099\_RAW.tar)  
588 was downloaded from the Gene Expression Omnibus (GEO) database

589 (<https://www.ncbi.nlm.nih.gov/geo/>). Since these ChIP-Seq peaks (referred to as  
590 transcription factor binding sites; TFBSs) are for GRCh37/hg19, the genomic  
591 coordinates of these TFBSs were converted to those in GRCh38/hg38 using  
592 UCSC liftOver ([http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86\\_64/liftOver](http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/liftOver)).  
593 The option “minMatch” was set at 0.95. If multiple technical replicates of ChIP-  
594 Seq are available for one KZFP, the replicate files were merged using the  
595 bedtools “merge”<sup>44</sup> with the options “-c 5 -o mean”. KZFPs were removed from  
596 the downstream analyses if the total number of TFBSs was less than 500. If  
597 >10,000 TFBSs were available for one KZFP, only the high-scored 10,000 TFBSs  
598 were used for the analyses.

599 To identify sets of genes that are preferentially targeted by a certain KZFP,  
600 genomic region enrichment analysis (GREAT)<sup>48</sup> was performed as follows. Only  
601 genes whose expression levels were detected in the TCGA tumor datasets were  
602 used. Regions of interest were defined as the regions within 10 kb from the TSSs  
603 of the gene members of a certain gene set. Regions of background were defined  
604 as the regions within 10 kb from the TSSs of genes belonging to any of the gene  
605 sets. The lengths of the regions of interest and regions of background were  
606 calculated and referred to as  $L_i$  and  $L_b$ , respectively. In the regions of interest and  
607 regions of background, the numbers of TFBSs were counted (referred to as  
608 counts of interest ( $C_i$ ) and counts of background ( $C_b$ ), respectively). The fold  
609 enrichment value was calculated by dividing  $C_i/C_b$  by  $L_i/L_b$ , and the statistical  
610 significance was evaluated using a binomial test.

611

## 612 **Differential expression analysis**

613 Differential expression analysis was performed using DESeq2 (version 1.18.1)<sup>41</sup>  
614 in R. Genes that were included in any of the gene sets prepared above were used.  
615 A549/KZFP cells and empty vector-transduced cells was compared (**Fig. 4E**).  
616 Additionally, comparison was conducted between A549 cells in which HERV-  
617 enhancer1 were excised versus the non-target control cells (**Fig. 2G**). Statistical  
618 significance was evaluated by the Wald test with false discovery rate (FDR)  
619 correction using the Benjamini-Hochberg (BH) method.

620

## 621 **Scoring system of genes for predicting the targets of KZFPs critical for 622 cancer progression**

623 The scheme is summarized in **Fig. S13A**. For each gene, the following scores  
624 were defined. The TCGA expressional correlation score was defined as the

625 Spearman's correlation between the expression of each gene and GSVA score  
626 of KZFPs in the TCGA dataset (the median value among all cancer types was  
627 used). The CCLE expressional correlation score was also defined using the same  
628 procedure but on the CCLE dataset. The prognosis score was defined as the Z  
629 score representing the association of each gene with the prognosis of cancer  
630 patients (the mean value among BLCA, HNSC, KIRP, and LUAD tumors was  
631 used). This Z score was described in the above section "**Survival analysis of**  
632 **the cancer patients**". The progression score was defined as the t-score  
633 representing the association of each gene with cancer progression (the mean  
634 value among BLCA, BRCA, KIRC, KIRP, LUAD, and thyroid carcinoma (THCA)  
635 tumors was used). This t-score was described in the above section "**Association**  
636 **analysis of gene expression and cancer progression**". The suppression score  
637 was defined as the mean value of the Wald scores in the differential expression  
638 analysis among the A549/KZFP cells. This Wald score was described in the  
639 above section "**RNA-Seq analysis of A549/KZFP cells**". Regarding the TCGA  
640 and CCLE correlation scores and suppression scores, the signs of the scores  
641 were inverted. All scores were standardized as Z scores and subsequently  
642 quantile-normalized. Genes were extracted if the minimum score was greater  
643 than 0.5 and the median score was greater than 1. Of the extracted genes, genes  
644 targeted by  $\geq 10$  KZFPs were further extracted and regarded as the target genes  
645 of KZFPs critical for cancer progression. A gene was regarded as the target of a  
646 certain KZFP if the KZFP bound to the regions within 10 kb from the TSSs of the  
647 gene. In this analysis, only TSSs of "principal transcripts" (Principals 1–3) defined  
648 by APPRIS<sup>49</sup> were used. If  $>1,000$  genes were assigned to a certain KZFP as its  
649 targets, only the top 1,000 genes having high-scored TFBSs were used.  
650

## 651 **Data visualization**

652 All visualizations were performed in R. Graphs were plotted using the "ggplot2"  
653 package or the pre-implemented function "plot" unless otherwise noted.  
654 Heatmaps were drawn using the "ComplexHeatmap" package<sup>50</sup>. Networks were  
655 plotted using the "igraph" package. Kaplan–Meier plots were drawn using the  
656 "ggsurvplot" function in the "survminer" package.  
657

## 658 **Cell culture**

659 HEK293T cells (CRL-11268; ATCC, Manassas, VA) were cultured in Dulbecco's  
660 modified Eagle's medium (Sigma-Aldrich, St .Louis, MO; #D6046) with 10% fetal

661 bovine serum (FBS; Sigma-Aldrich #172012-500ML) and 1% penicillin  
662 streptomycin (Sigma-Aldrich #P4333-100ML). A549 cells (CCL-185; ATCC) were  
663 cultured in Ham's F-12K (Kaighn's) medium (Thermo Fisher Scientific, Waltham,  
664 MA; #21127022) with 10% FBS (guaranteed doxycycline free; Thermo Fisher  
665 Scientific; #2023-03) and 1% penicillin streptomycin. A459/KZFP cells were  
666 cultured in F-12K medium with 1.0  $\mu$ g/ml puromycin (Invivogen, San Diego, CA;  
667 #ant-pr-1). An A549 cell line stably expressing Cas9 (A549/Cas9 cells) was  
668 cultured in F-12K medium with 10% FBS (guaranteed doxycycline free; Thermo  
669 Fisher Scientific; #2023-03) and 5.0  $\mu$ g/ml blasticidin (Invivogen #ant-bl-1). All  
670 cells were cultured in 5% CO<sub>2</sub> at 37°C.

671

## 672 **Establishment of a panel of A549/KZFP cells**

673 A549 cells were selected as the parental cells since the expression levels of  
674 KZFPs (and HERVs) were relatively low in this cell line (**Fig. S11A**). We selected  
675 30 types of KZFP genes satisfying the following criteria: 1) showing a positive  
676 correlation (Spearman's correlation > 0.3) between its expression and the total  
677 expression of HERVs in >2 types of cancers; 2) possessing expressed HERVs  
678 within the vicinity (<20 kb) of its TSSs in tumors; 3) showing a positive correlation  
679 (Spearman's correlation > 0.3) between its expression and the expression of  
680 HERV loci in the vicinity (<20 kb) of its TSSs in >2 types of cancers; 4) having  
681 available ChIP-Seq data presented by a previous study (Imbeault et al.<sup>33</sup>).  
682 Information of the selected KZFP genes is summarized in **Data S8**.

683 To prepare the lentiviral vectors expressing x3 HA-tagged KZFPs,  
684 HEK293T cells were co-transfected with 12  $\mu$ g of pCAG-HIVgp (RDB04394,  
685 kindly provided by Dr. Hiroyuki Miyoshi), 10  $\mu$ g of pCMV-VSV-G-RSV-Rev  
686 (RDB04393, kindly provided by Dr. Hiroyuki Miyoshi), and 17  $\mu$ g of pEXPpSIN-  
687 TRE-GW ZNF-3xHA<sup>33</sup> by the calcium phosphate method. The pEXPpSIN-TRE-  
688 GW ZNF-3xHA plasmids encode respective HA-tagged KZFP proteins. After 12  
689 hours of transfection, the culture medium was changed to fresh F-12K medium.  
690 After 48 hours of transfection, the culture supernatant including lentivector  
691 particles was collected. A549 cells were infected with these particles at a  
692 multiplicity of infection (MOI) of 0.1. After 2 days of infection, the cells were  
693 selected with puromycin (1  $\mu$ g/ml) for 7 days. Three days before the start of the  
694 experiments, doxycycline (1.0  $\mu$ g/ml) was added to induce the expression of  
695 KZFP. The expression of KZFP was verified by western blotting with the HA-  
696 specific antibody (Roche, Basel, Switzerland; #12013819001). Empty vector-

697 transduced A549 cells (referred to as negative control cells (NC cells)) were  
698 established according to the procedures described above.

699

700 **Apoptosis detection assay**

701 A549/KZFP cells and NC cells were stained with Annexin V conjugated to Alexa  
702 Fluor™ 647 (Invitrogen Carlsbad, CA; #S32357). After staining, the number of  
703 Annexin V-positive cells was counted by a FACSCalibur system (BD Biosciences,  
704 San Jose, CA), and the rate of apoptotic cells was calculated. A single set of  
705 triplicate experiments was performed, and the mean and standard error of the  
706 mean (SEM) values are shown in **Fig. 4A**. Statistical tests were performed by  
707 two-sided Student's t-test with a threshold of 0.05.

708

709 **Cell growth assay**

710 A549/KZFP cells and NC cells were seeded at  $1.0 \times 10^5$  cells/well in 6-well plates  
711 (Thermo Fisher Scientific). After 72 hours of seeding, the number of cells was  
712 counted manually under a microscope, and the growth rate of the cells was  
713 calculated. Single-replicate experiments were performed at least 7 times  
714 independently, and the mean and SEM values are shown in **Fig. 4B**. Statistical  
715 tests were performed by two-sided Student's t-test with a threshold of 0.05.

716

717 **Cell scratch assay (wound-healing assay<sup>51</sup>)**

718 A549/KZFP cells and NC cells were seeded in 12-well plates (Thermo Fisher  
719 Scientific) and cultured until >90% confluence. A single straight wound was  
720 formed in each well by scratching with a sterile 1,000  $\mu$ l pipette tip. The cells were  
721 washed with phosphate-buffered saline (PBS), and 2 ml of F-12K medium was  
722 added. Images were taken under a microscope immediately after the scratch and  
723 again after 24 hours. Using ImageJ<sup>52</sup> with in-house scripts, the area (pixels) in  
724 which cells migrated for 24 hours was calculated. Triplicate experiments were  
725 performed independently twice. Regarding the mean and SEM, the average  
726 values between the two sets of experiments are shown in **Fig. 4C**. Statistical tests  
727 were performed by two-sided Student's t-test in each set of experiments with a  
728 threshold of 0.05. Only if a significant difference was observed in both sets of  
729 experiments, the comparison was considered significant.

730

731 **Cell invasion assay**

732 An invasion assay was performed using a 96-well Transwell plate (8.0- $\mu$ m pore  
733 size) (Corning, Corning, NY #3374) with Corning Matrigel Basement Membrane  
734 Matrix (Corning #354234). The Matrigel matrix was diluted 50-fold with serum-  
735 free F-12K medium. To coat the Transwell insert plate, 30  $\mu$ l of Matrigel matrix  
736 was dispensed into the insert plate. After 2 hours of incubation, 20  $\mu$ l of the  
737 supernatant was removed from the coated Transwell plate. Subsequently,  
738 A549/KZFP cells and NC cells were seeded at  $5.0 \times 10^4$  cells/well in the insert  
739 plate. The insert plate was filled with serum-free F-12K medium, while the  
740 reservoir plate was filled with F-12K medium with 10% FBS. After incubation at  
741 37°C for 48 hours, the cells that had invaded the Matrigel and migrated to the  
742 opposite side of the insert plate were washed with PBS, stripped with Trypsin-  
743 EDTA, and stained with calcein AM (Invitrogen #C3100MP). To evaluate the  
744 degree of cell invasion, the fluorescence intensity of the cells was measured  
745 using a 2030 ARVO X multi-label counter (PerkinElmer, Waltham, MA). The  
746 relative fluorescence intensity was calculated as  $(Fl_i - Fl_b) / (Fl_c - Fl_b)$ , where  $Fl_i$   
747 denotes the fluorescence intensity of the A549/KZFP cells of interest,  $Fl_b$  denotes  
748 the intensity of the blank, and  $Fl_c$  denotes the intensity of NC cells. Triplicate  
749 experiments were performed independently twice. Regarding the mean and SEM,  
750 the average values between the two sets of experiments are shown in **Fig. 4D**.  
751 Statistical tests were performed by two-sided Student's t-test in each set of  
752 experiments with a threshold of 0.05. Only if a significant difference was observed  
753 in both sets of experiments, the comparison was considered significant.  
754

#### 755 **Construction of plasmids for the luciferase reporter assay**

756 Genomic DNA from the human peripheral blood lymphocytes of a healthy donor  
757 was used as the DNA source. A luciferase reporter vector, pGL3-basic (Promega,  
758 Madison, WI), was used. Using nested PCR, the genomic region indicated by the  
759 arrow in **Figs. S14A–14B** was cloned into pGL3-basic.

760 Information on the plasmids and primers prepared in this section is  
761 summarized in **Data S9 and S10**, respectively.

762

#### 763 **Luciferase reporter assay to assess the promoter activity of genes**

764 A549 cells were seeded at  $1.0 \times 10^5$  cells/well in 12-well plates (Thermo Fisher  
765 Scientific). After 24 hours of seeding, the luciferase reporter plasmid was  
766 transfected using polyethylenimine transfection. To fairly compare the reporter  
767 activities of the two plasmids with different sequence lengths, 1  $\mu$ g of the longer

768 plasmid and the same molar of the shorter plasmid were used for the transfection.  
769 After 12 hours of transfection, the culture medium was changed to fresh F-12K  
770 medium. After 48 hours of transfection, the luminescence intensity of the  
771 transfected cells was measured using a 2030 ARVO X multi-label counter  
772 (PerkinElmer) or a GloMax® Explorer Multimode Microplate Reader 3500  
773 (Promega) with a BrillianStar-LT assay system (Toyo-b-net, Tokyo, Japan; #307-  
774 15373 BLT100). A single set of triplicate experiments was performed, and the  
775 mean and SEM values are shown in **Figs. 4I and S14E–G**. Statistical tests were  
776 performed by two-sided Student's t-test with a threshold of 0.05.  
777

#### 778 **Establishment of HERV-excised cells**

779 First, an A549 cell line stably expressing Cas9 (referred to as A549/Cas9 cells)  
780 was established as follows. To prepare the lentiviral vectors expressing Cas9,  
781 HEK293T cells were co-transfected with 12 µg of pCAG-HIVgp, 10 µg of pCMV-  
782 VSV-G-RSV-Rev, and 17 µg of plentiCas9-Blast (Addgene, Watertown, MA;  
783 #52962) by the calcium phosphate method. After 12 hours of transfection, the  
784 culture medium was changed to fresh F-12K medium. After 48 hours of  
785 transfection, the culture supernatant including lentivector particles was collected.  
786 A549 cells were infected with these particles at an MOI of 0.1. After 2 days of  
787 infection, the cells were selected with blasticidin (5 µg/ml) for 7 days. After  
788 selection, single cell clones were obtained through the limiting dilution method.  
789 By screening the expression level of Cas9 among the candidate clones,  
790 A549/Cas9 cells were established.

791 To excise the target HERV, a pair of guide RNAs (gRNAs) were designed  
792 in the upstream and downstream regions of the HERV using the web applications  
793 of sgRNA designer<sup>53</sup> (<http://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design>) or CRISPOR<sup>54</sup> (<http://crispor.tefor.net>). The gRNA  
794 information is summarized in **Data S11**. The gRNA was cloned into a gRNA  
795 expression plasmid, lentiGuide-Puro (Addgene #52963). A pair of gRNA  
796 expression plasmids was co-transfected into the A549/Cas9 cells by  
797 electroporation using the NEON Transfection System (ThermoFisher) (1200 V;  
798 30 ms; 2 times pulse; 1.0 x 10<sup>5</sup> cells; and 500 ng of each plasmid). After  
799 transfection, the cells were selected with 1 µg/ml puromycin for 3 days. After  
800 selection, single cell clones were obtained through the limiting dilution method.  
801 Of these candidate clones, the clones in which homozygous or heterozygous  
802 excision of the target HERV occurred were screened using PCR (**Fig. S6**).  
803

804 Regarding homozygous clones, the PCR fragments were checked through  
805 molecular cloning into a TOPO vector (Invitrogen #450245) followed by Sanger's  
806 sequencing.

807

### 808 **qRT-PCR**

809 Total RNA was extracted from cells by the QIAamp RNA Blood Mini Kit (QIAGEN,  
810 Hilden, Germany; # 52304) and subsequently treated with DNase I,  
811 Amplification Grade (Invitrogen #18068015). cDNA was synthesized by reverse  
812 transcription of the total RNA using SuperScript III reverse transcriptase (Life  
813 technologies #18080044) with Oligo(dT)12-18 Primer (Invitrogen #18418012).  
814 qRT-PCR was performed on the cDNA using a CFX Connect Real-Time PCR  
815 Detection System (Bio-Rad, Richmond, CA; #1855201J1) with a TaqMan® Gene  
816 Expression Assay kit (Thermo Fisher Scientific). The primer and TaqMan probe  
817 information are listed in **Data S12**. GAPDH was used as an internal control.

818

### 819 **Preparation of RNA-Seq samples and sequencing**

820 Cells were seeded at  $1.0 \times 10^6$  cells in 100 mm dishes (Thermo Fisher Scientific  
821 EasYDish #150466). After 48 hours of seeding, the cells were harvested and  
822 stored at  $-80^{\circ}\text{C}$ . Total RNA was extracted from the cells by the QIAamp RNA  
823 Blood Mini Kit (QIAGEN #52304) and subsequently treated with RNase-Free  
824 DNase Set (QIAGEN #79254).

825 Quality checks, library construction, and sequencing were performed by  
826 Novogene (<https://en.novogene.com>). Pair-end 150-bp read length sequencing  
827 was performed on an Illumina NovaSeq 6000 system.

828

### 829 **Code availability**

830 Computer codes used in the present study will be available in the GitHub  
831 repository ([https://github.com/TheSatoLab/HERV\\_Pan-cancer\\_analysis](https://github.com/TheSatoLab/HERV_Pan-cancer_analysis)).

832

### 833 **Data availability**

834 RNA-seq data reported in this paper will be available in GEO  
835 (<https://www.ncbi.nlm.nih.gov/geo/>; GSE141803).

836

837 **References:**

838 1 Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis  
839 project. *Nat Genet* **45**, 1113-1120, doi:10.1038/ng.2764 (2013).

840 2 Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome.  
841 *Science* **357**, doi:10.1126/science.aan2507 (2017).

842 3 Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in  
843 Cancer. *Cell* **168**, 629-643, doi:10.1016/j.cell.2016.12.013 (2017).

844 4 Corces, M. R. *et al.* The chromatin accessibility landscape of primary  
845 human cancers. *Science* **362**, doi:10.1126/science.aav1898 (2018).

846 5 Chen, H. *et al.* A Pan-Cancer Analysis of Enhancer Expression in Nearly  
847 9000 Patient Samples. *Cell* **173**, 386-399.e312,  
848 doi:10.1016/j.cell.2018.03.027 (2018).

849 6 Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of  
850 transposable elements: from conflicts to benefits. *Nat Rev Genet* **18**, 71-  
851 86, doi:10.1038/nrg.2016.139 (2017).

852 7 Coffin, J. M. *Retroviruses*. (Cold Spring Harbor Laboratory Press, 2002).

853 8 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome.  
854 *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

855 9 Sundaram, V. *et al.* Widespread contribution of transposable elements to  
856 the innovation of gene regulatory networks. *Genome Res* **24**, 1963-1976,  
857 doi:10.1101/gr.168872.113 (2014).

858 10 Ito, J. *et al.* Systematic identification and characterization of regulatory  
859 elements derived from human endogenous retroviruses. *PLoS Genet* **13**,  
860 e1006883, doi:10.1371/journal.pgen.1006883 (2017).

861 11 Kunarso, G. *et al.* Transposable elements have rewired the core regulatory  
862 network of human embryonic stem cells. *Nat Genet* **42**, 631-634,  
863 doi:10.1038/ng.600 (2010).

864 12 Pi, W. *et al.* Long-range function of an intergenic retrotransposon. *Proc  
865 Natl Acad Sci U S A* **107**, 12992-12997, doi:10.1073/pnas.1004139107  
866 (2010).

867 13 Emera, D. *et al.* Convergent evolution of endometrial prolactin expression  
868 in primates, mice, and elephants through the independent recruitment of  
869 transposable elements. *Mol Biol Evol* **29**, 239-247,  
870 doi:10.1093/molbev/msr189 (2012).

871 14 Wang, J. *et al.* Primate-specific endogenous retrovirus-driven transcription  
872 defines naive-like stem cells. *Nature* **516**, 405-409,  
873 doi:10.1038/nature13804 (2014).

874 15 Ferreira, L. M. *et al.* A distant trophoblast-specific enhancer controls HLA-  
875 G expression at the maternal-fetal interface. *Proc Natl Acad Sci U S A* **113**,  
876 5364-5369, doi:10.1073/pnas.1602886113 (2016).

877 16 Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons  
878 demarcate topologically associating domains in human pluripotent stem  
879 cells. *Nat Genet* **51**, 1380-1388, doi:10.1038/s41588-019-0479-7 (2019).

880 17 Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory evolution of innate  
881 immunity through co-option of endogenous retroviruses. *Science* **351**,  
882 1083-1087, doi:10.1126/science.aad5497 (2016).

883 18 Slotkin, R. K. & Martienssen, R. Transposable elements and the  
884 epigenetic regulation of the genome. *Nat Rev Genet* **8**, 272-285,  
885 doi:10.1038/nrg2072 (2007).

886 19 Deniz, O., Frost, J. M. & Branco, M. R. Regulation of transposable  
887 elements by DNA modifications. *Nat Rev Genet* **20**, 417-431,  
888 doi:10.1038/s41576-019-0106-6 (2019).

889 20 Rooney, M. S., Shukla, S. A., Wu, C. J., Getz, G. & Hacohen, N. Molecular  
890 and genetic properties of tumors associated with local immune cytolytic  
891 activity. *Cell* **160**, 48-61, doi:10.1016/j.cell.2014.12.033 (2015).

892 21 Smith, C. C. *et al.* Endogenous retroviral signatures predict  
893 immunotherapy response in clear cell renal cell carcinoma. *J Clin Invest*  
894 **128**, 4804-4820, doi:10.1172/jci121476 (2018).

895 22 Solovyov, A. *et al.* Global Cancer Transcriptome Quantifies Repeat  
896 Element Polarization between Immunotherapy Responsive and T Cell  
897 Suppressive Classes. *Cell Rep* **23**, 512-521,  
898 doi:10.1016/j.celrep.2018.03.042 (2018).

899 23 Panda, A. *et al.* Endogenous retrovirus expression is associated with  
900 response to immune checkpoint blockade in clear cell renal cell carcinoma.  
901 *JCI Insight* **3**, doi:10.1172/jci.insight.121522 (2018).

902 24 Attig, J. *et al.* LTR retroelement expansion of the human cancer  
903 transcriptome and immunopeptidome revealed by de novo transcript  
904 assembly. *Genome Res* **29**, 1578-1590, doi:10.1101/gr.248922.119  
905 (2019).

906 25 Babaian, A. & Mager, D. L. Endogenous retroviral promoter exaptation in  
907 human cancer. *Mob DNA* **7**, 24, doi:10.1186/s13100-016-0080-x (2016).

908 26 Jang, H. S. *et al.* Transposable elements drive widespread expression of  
909 oncogenes in human cancers. *Nat Genet* **51**, 611-617,  
910 doi:10.1038/s41588-019-0373-3 (2019).

911 27 Hutter, C. & Zenklusen, J. C. The Cancer Genome Atlas: Creating Lasting  
912 Value beyond Its Data. *Cell* **173**, 283-285, doi:10.1016/j.cell.2018.03.042  
913 (2018).

914 28 Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins.  
915 *Development* **144**, 2719-2729, doi:10.1242/dev.132605 (2017).

916 29 Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based  
917 approach for interpreting genome-wide expression profiles. *Proc Natl  
918 Acad Sci U S A* **102**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).

919 30 Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line  
920 Encyclopedia. *Nature* **569**, 503-508, doi:10.1038/s41586-019-1186-3  
921 (2019).

922 31 Huntley, S. *et al.* A comprehensive catalog of human KRAB-associated  
923 zinc finger genes: insights into the evolutionary history of a large family of  
924 transcriptional repressors. *Genome Res* **16**, 669-677,  
925 doi:10.1101/gr.4842106 (2006).

926 32 Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers  
927 and target genes in GeneCards. *Database (Oxford)* **2017**,  
928 doi:10.1093/database/bax028 (2017).

929 33 Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins  
930 contribute to the evolution of gene regulatory networks. *Nature* **543**, 550-  
931 554, doi:10.1038/nature21683 (2017).

932 34 Horbelt, D., Denkis, A. & Knaus, P. A portrait of Transforming Growth  
933 Factor beta superfamily signalling: Background matters. *Int J Biochem Cell  
934 Biol* **44**, 469-474, doi:10.1016/j.biocel.2011.12.013 (2012).

935 35 Wang, Y. Wnt/Planar cell polarity signaling: a new paradigm for cancer  
936 therapy. *Mol Cancer Ther* **8**, 2103-2109, doi:10.1158/1535-7163.Mct-09-  
937 0282 (2009).

938 36 Hall, A. The cytoskeleton and cancer. *Cancer Metastasis Rev* **28**, 5-14,  
939 doi:10.1007/s10555-008-9166-3 (2009).

940 37 Dar, A. A., Goff, L. W., Majid, S., Berlin, J. & El-Rifai, W. Aurora kinase  
941 inhibitors--rising stars in cancer therapeutics? *Mol Cancer Ther* **9**, 268-278,  
942 doi:10.1158/1535-7163.Mct-09-0765 (2010).

943 38 Nakayama, K. I. & Nakayama, K. Ubiquitin ligases: cell-cycle control and  
944 cancer. *Nat Rev Cancer* **6**, 369-381, doi:10.1038/nrc1881 (2006).

945 39 Pontis, J. *et al.* Hominoid-Specific Transposable Elements and KZFPs  
946 Facilitate Human Embryonic Genome Activation and Control Transcription  
947 in Naive Human ESCs. *Cell Stem Cell* **24**, 724-735.e725,  
948 doi:10.1016/j.stem.2019.03.012 (2019).

949 40 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general  
950 purpose program for assigning sequence reads to genomic features.  
951 *Bioinformatics* **30**, 923-930, doi:10.1093/bioinformatics/btt656 (2014).

952 41 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change  
953 and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550,  
954 doi:10.1186/s13059-014-0550-8 (2014).

955 42 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for  
956 Illumina sequence data. *Bioinformatics* **30**, 2114-2120,  
957 doi:10.1093/bioinformatics/btu170 (2014).

958 43 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*  
959 **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2013).

960 44 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for  
961 comparing genomic features. *Bioinformatics* **26**, 841-842,  
962 doi:10.1093/bioinformatics/btq033 (2010).

963 45 Peters, T. J. *et al.* De novo identification of differentially methylated regions  
964 in the human genome. *Epigenetics Chromatin* **8**, 6, doi:10.1186/1756-  
965 8935-8-6 (2015).

966 46 Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation  
967 analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7,  
968 doi:10.1186/1471-2105-14-7 (2013).

969 47 Sergushichev, A. A. Fast gene set enrichment analysis. *bioRxiv*,  
970 doi:10.1101/060012 (2016).

971 48 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-  
972 regulatory regions. *Nat Biotechnol* **28**, 495-501, doi:10.1038/nbt.1630  
973 (2010).

974 49 Rodriguez, J. M. *et al.* APPRIS: annotation of principal and alternative  
975 splice isoforms. *Nucleic Acids Res* **41**, D110-117,  
976 doi:10.1093/nar/gks1058 (2013).

977 50 Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and  
978 correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-  
979 2849, doi:10.1093/bioinformatics/btw313 (2016).

980 51 Liang, C. C., Park, A. Y. & Guan, J. L. In vitro scratch assay: a convenient  
981 and inexpensive method for analysis of cell migration in vitro. *Nat Protoc*  
982 **2**, 329-333, doi:10.1038/nprot.2007.30 (2007).

983 52 Schneider, C. A., Rasband, W. S. & Eliceiri, K. W. NIH Image to ImageJ:  
984 25 years of image analysis. *Nat Methods* **9**, 671-675 (2012).

985 53 Sanson, K. R. *et al.* Optimized libraries for CRISPR-Cas9 genetic screens  
986 with multiple modalities. *Nat Commun* **9**, 5416, doi:10.1038/s41467-018-  
987 07901-8 (2018).

988 54 Concorde, J. P. & Haeussler, M. CRISPOR: intuitive guide selection for  
989 CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids*  
990 *Res* **46**, W242-w245, doi:10.1093/nar/gky354 (2018).

991 55 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks  
992 with increased coverage, supporting functional discovery in genome-wide  
993 experimental datasets. *Nucleic Acids Res* **47**, D607-d613,  
994 doi:10.1093/nar/gky1131 (2019).

995

996 **Acknowledgments**

997 We would like to thank Naoko Misawa (Institute for Frontier Life and Medical  
998 Sciences, Kyoto University, Japan) and Kyoichiro Nomura (Yamaguchi University,  
999 Japan) for technical supports; Julien Pontis (School of Life Sciences, Ecole  
1000 Polytechnique Federale de Lausanne (EPFL), Switzerland), Shohei Kojima and  
1001 Junna Kawasaki (Institute for Frontier Life and Medical Sciences, Kyoto  
1002 University, Japan), and Shiro Yamada (Tokai University, Japan) for thoughtful  
1003 comments. The super-computing resource, SHIROKANE, was provided by  
1004 Human Genome Center, The Institute of Medical Science, The University of  
1005 Tokyo, Japan. The results shown here are in part based upon data generated by  
1006 the TCGA Research Network (<https://www.cancer.gov/tcga>). **Funding:** This  
1007 study was supported in part by AMED J-PRIDE JP19fm0208006 (to K.S.); AMED  
1008 Research Program JP19fk0410014 (to Y.K. and K.S.) and JP19fk0410019 (to  
1009 K.S.); JST CREST (to K.S.); JSPS KAKENHI Scientific Research B JP18H02662  
1010 (to K.S.); JSPS KAKENHI Scientific Research on Innovative Areas JP16H06429  
1011 (to K.S.), JP16K21723 (to K.S.), JP17H05813 (to K.S.), and JP19H04826 (to  
1012 K.S.); JSPS Research Fellow PD JP19J01713 (to J.I.) and DC1 JP19J20488 (to  
1013 I.K.); JSPS Core-to-Core program (A. Advanced Research Networks) (to Y.K.  
1014 and K.S.); Joint Usage/Research Center program of Institute for Frontier Life and  
1015 Medical Sciences, Kyoto University (to K.S.); Takeda Science Foundation (to  
1016 K.S.); ONO Medical Research Foundation (to K.S.); Ichiro Kanehara Foundation  
1017 (to K.S.); Lotte Foundation (to K.S.); and Mochida Memorial Foundation for  
1018 Medical and Pharmaceutical Research (to K.S.).

1019

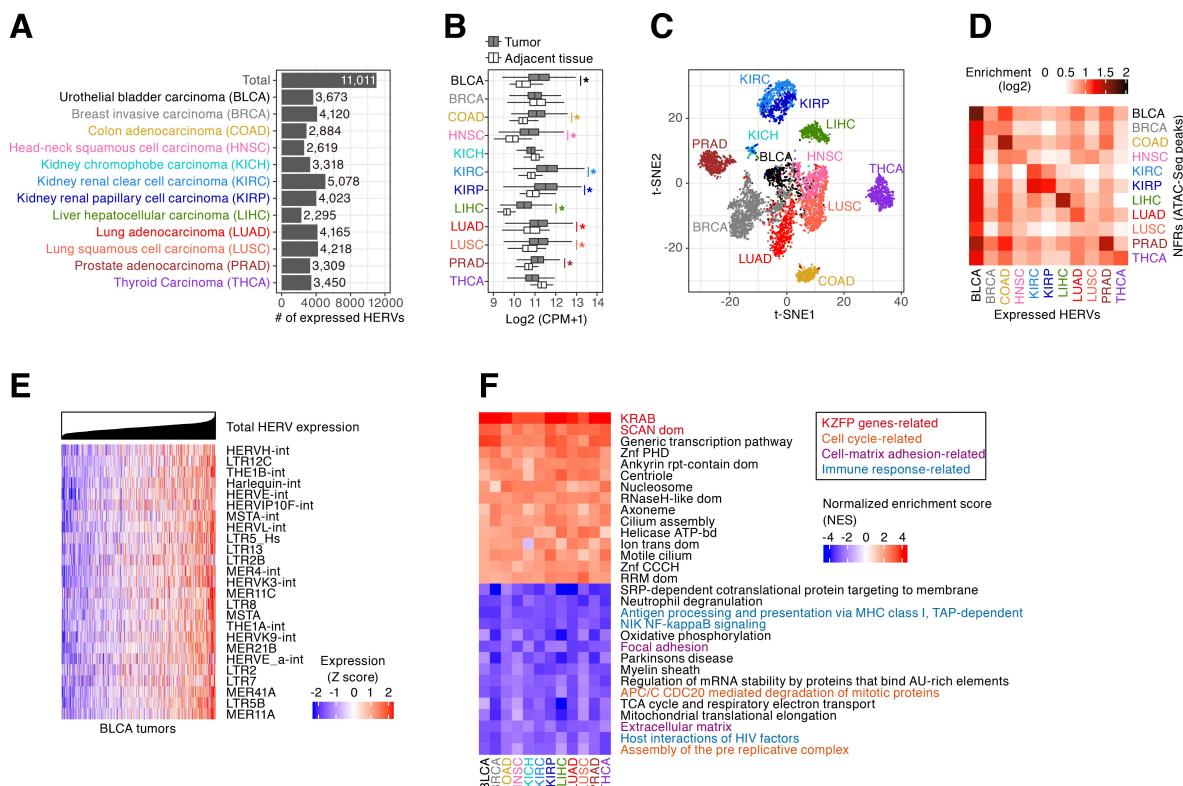
1020 **Author contributions:**

1021 J.I. conceived the study; J.I. and A.C. mainly performed bioinformatics analyses;  
1022 I.K., H.N., I.I., P.T., and D.T. supported bioinformatics analyses; I.K mainly  
1023 performed experimental analyses; A.S. and Y.K. supported experimental  
1024 analyses; Y.K., P.T., and D.T. provided reagents; J.I., I.K., and K.S. prepared the  
1025 figures; J.I., I.K., and K.S. wrote the initial draft of the manuscript; all authors  
1026 contributed to data interpretation, designed the research, revised the paper, and  
1027 approved the final manuscript.

1028

1029 **Declaration of Interests:**

1030 The authors declare that they have no competing interests.



1031

**Fig. 1 Landscape of HERV expression in 12 types of solid cancers.**

1032 A) Numbers of the expressed HERV loci identified in respective types of cancers.

1033 B) Total expression levels of HERVs (log2 (counts per million (CPM) + 1)) in

1034 cancers and adjacent normal tissues. An asterisk denotes a significant increase

1035 in the values in tumors compared to that in normal tissues (Bonferroni-corrected

1036  $P$  value  $< 0.05$  in two-sided Wilcoxon rank sum test).

1037 C) t-SNE plot representing the expression patterns of HERVs among tumor

1038 samples. Dots indicate tumor sample data. The expression levels of the 1000

1039 most highly expressed HERVs were used in the analysis.

1040 D) Fold enrichments of the overlaps between expressed HERV loci and

1041 nucleosome-free regions (NFRs; i.e., ATAC-Seq peaks) identified in respective

1042 types of cancers. The enrichment value was calculated based on the random

1043 expectation.

1044 E) Expression levels of the respective HERV groups in BLCA tumors. Normalized

1045 expression levels (Z scores) of the 25 most highly expressed HERV groups are

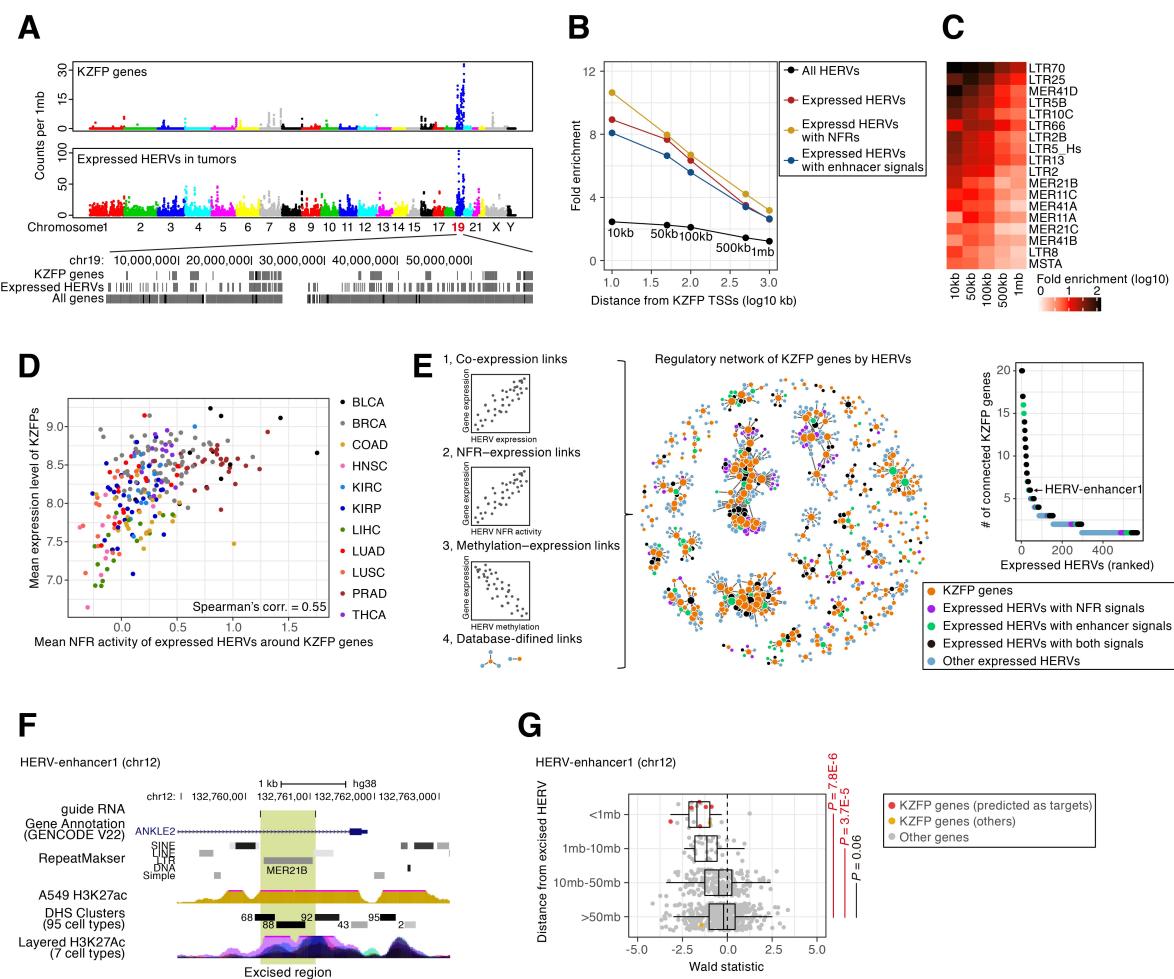
1046 shown. Tumors were ordered according to the total value.

1047 F) Gene set enrichment analysis (GSEA)<sup>29</sup> summarizing genes whose

1048 expression levels were correlated with the global expression levels of HERVs.

1049 Spearman's correlation scores between the expression levels of respective

1051 genes and the total expression level of HERVs were calculated, and GSEA was  
1052 subsequently performed based on those scores. For the positive (red) and  
1053 negative (blue) correlations, the high-scored 15 gene sets (regarding the mean  
1054 value among cancer types) are shown. Redundant gene sets were removed from  
1055 the results.

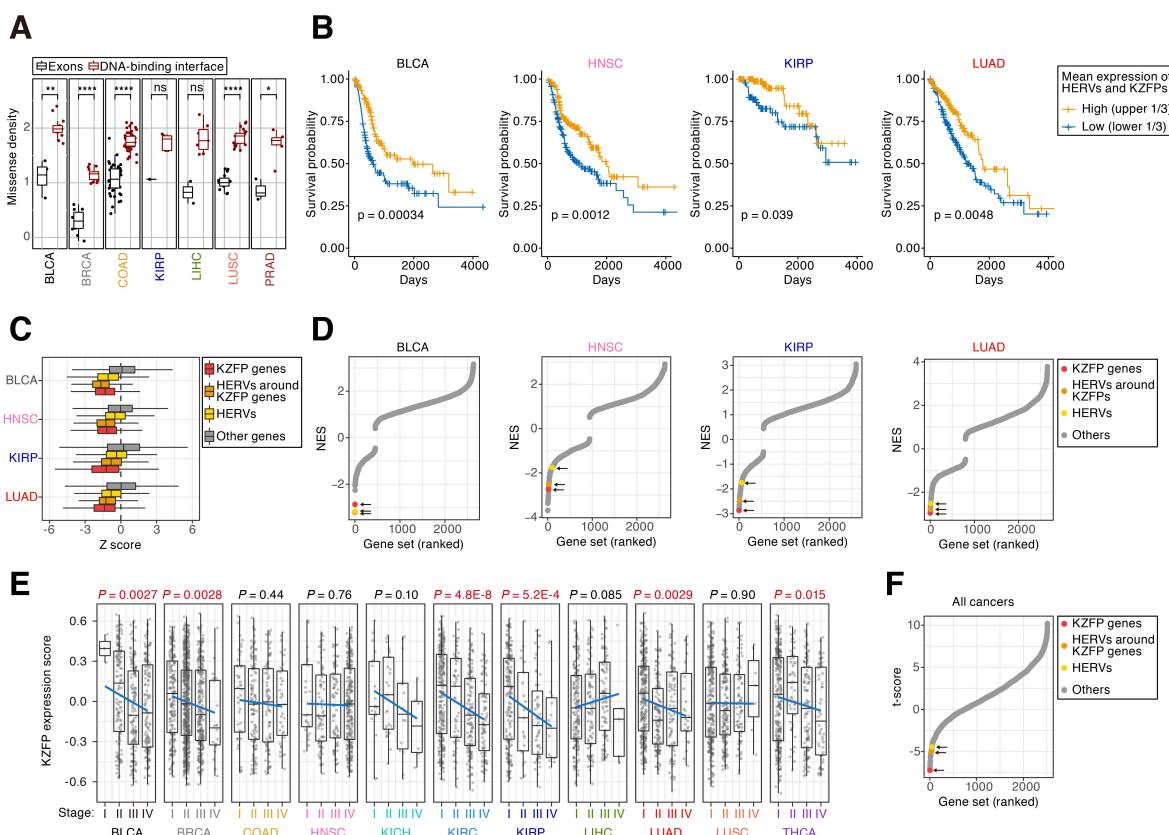


1056

## 1057 Fig. 2 Transcriptional activation of KZFP genes by the adjacent HERVs.

1058 A) Genomic positions of KZFP genes and the expressed HERVs in tumors. Top)  
 1059 The genomic densities of KZFP genes and the expressed HERVs (counts per 1  
 1060 megabase pairs (mb)). Bottom) Genomic locations of KZFP genes, the  
 1061 expressed HERVs, and all genes on chromosome 19.  
 1062 B) Enrichments of the expressed HERVs in tumors around the transcription start  
 1063 sites (TSSs) of KZFP genes. Fold enrichments of the four categories of HERVs  
 1064 (all HERVs, expressed HERVs, expressed HERVs with NFRs, and expressed  
 1065 HERVs overlapped with the enhancers defined by GeneHancer<sup>32</sup>) in the regions  
 1066 within 10, 50, 100, and 500 kb and 1 mb from the TSSs of KZFP genes are shown.  
 1067 The enrichment value was calculated based on the random expectation.  
 1068 C) Fold enrichments of respective groups of expressed HERVs (LTRs) in the  
 1069 vicinity of the TSSs of KZFP genes. LTR groups that were significantly (FDR <  
 1070 0.05) enriched within 50 kb from the TSSs are shown.

1071 D) Association between the mean expression level of KZFPs and the mean NFR  
1072 activity of the expressed HERVs in the vicinity (<50 kb) of KZFP genes in tumors.  
1073 E) Prediction of the genes regulated by the expressed HERVs. Left) Schematics  
1074 of the prediction of the regulatory relationships. The prediction was based on the  
1075 following information: 1) co-expression interactions, 2) NFR–expression  
1076 interactions, 3) DNA methylation–expression anti-correlation interactions, and 4)  
1077 interactions predicted by GeneHancer<sup>32</sup>. The co-expression interaction was used  
1078 in only pairs of HERV and KZFP genes within 50 kb of each other, while the other  
1079 interactions were used in only pairs within 500 kb of each other. Middle)  
1080 Integrated network representing the predicted regulations of KZFP genes by  
1081 HERVs. Right) Numbers of connected KZFPs of the respective HERV nodes in  
1082 the network. The HERVs were ranked according to connectivity. The target  
1083 HERV for the CRISPR-Cas9 excision experiment is denoted.  
1084 F) UCSC genome browser view of the target HERV (HERV-enhancer1).  
1085 G) Effect of the excision of HERV on the expression of the adjacent genes in lung  
1086 adenocarcinoma (A549) cells. The cells in which the target HERV was  
1087 homozygously excised (5 clones) and the non-target control cells (5 clones) were  
1088 compared. The X-axis indicates the Wald statistic, in which the positive and  
1089 negative values indicate the up- and downregulation, respectively, of the gene  
1090 expression compared to that in the non-target control cells. Genes were stratified  
1091 according to the distance from the excised HERV, and the distributions of Wald  
1092 statistics were compared between the indicated categories. *P* values were  
1093 calculated by two-sided Student's t-test.



1094

1095 **Fig. 3 Association of the expression status of KZFPs and HERVs in tumors**  
1096 **with cancer prognosis and progression.**

1097 A) Accumulation of somatic missense mutations in the DNA-binding amino acid  
1098 residues of KZFP genes. The mutation density (counts per mb per patient) of  
1099 KZFP genes was compared between the DNA-binding amino acid residues (red)  
1100 and the whole exonic regions (black). Results for KZFP genes with  $\geq 1$  mutations  
1101 are shown.  $P$  values were calculated by two-sided Wilcoxon rank sum test. \*,  $P$   
1102  $< 0.05$ ; \*\*,  $P < 0.01$ ; \*\*\*\*,  $P < 0.0001$ .

1103 B) Kaplan–Meier survival plots of cancer patients with high or low expression  
1104 levels of HERVs and KZFPs. The results for BLCA, HNSC, KIRP, and LUAD  
1105 tumors are shown (results for the other cancer types are shown in **Fig. S7A**). The  
1106 stratification of the patients was according to the mean value of the gene set-wise  
1107 expression scores (GSVA scores<sup>46</sup>) between KZFPs and HERVs. The results for  
1108 the stratifications according to the GSVA scores of HERVs and KZFPs are shown  
1109 in **Figs. S7B and S7C**, respectively. The  $P$  value was calculated by the two-sided  
1110 log-rank test.

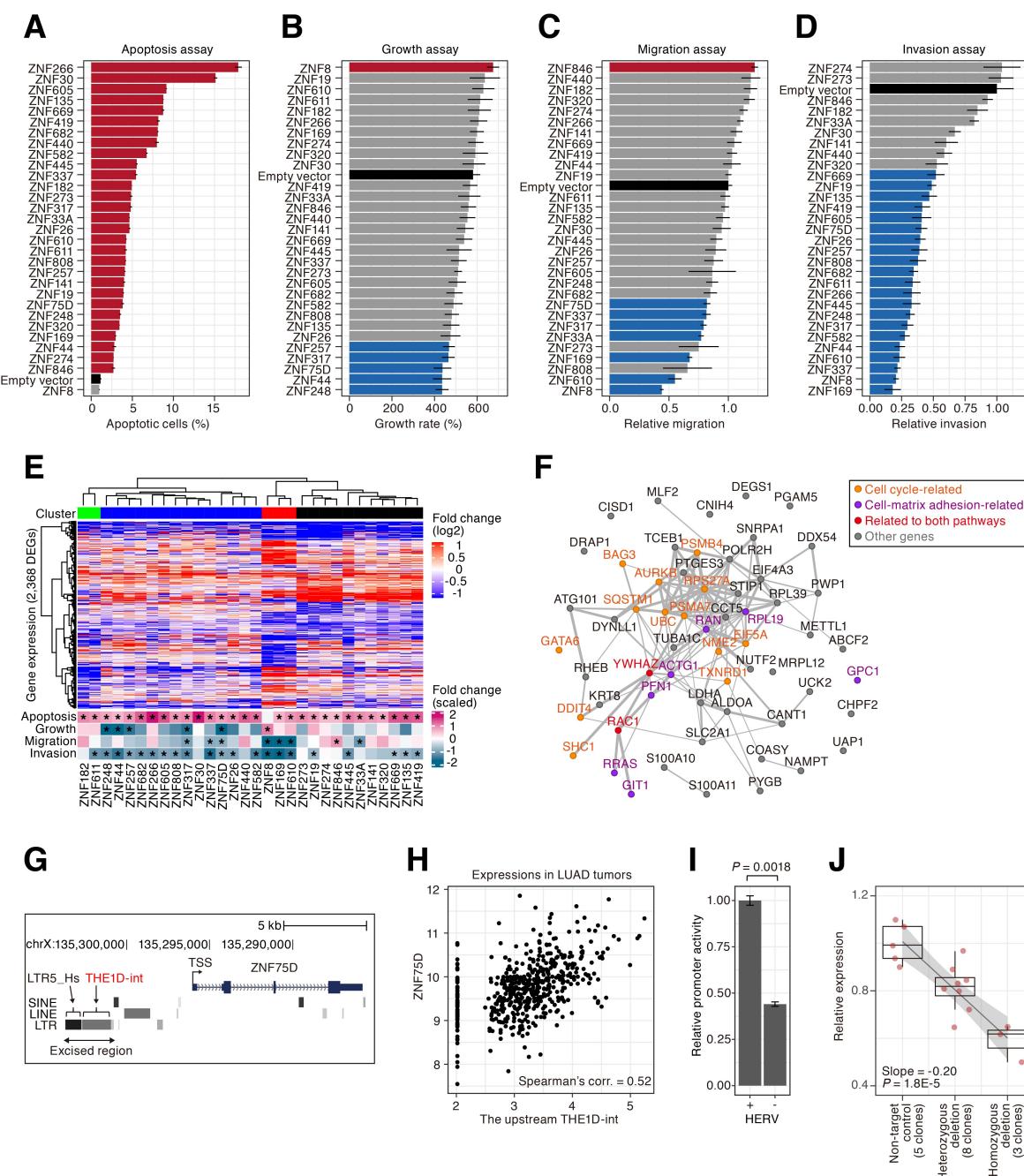
1111 C) Associations of respective genes and HERVs with the prognosis of cancer  
1112 patients. The association was evaluated as the Z score in the Cox proportional

1113 hazards model, and the distributions of the Z score were compared among  
1114 KZFPs, HERVs, HERVs around KZFPs (within 50 kb), and the other genes.  
1115 Positive and negative Z scores indicate the associations with worse or better  
1116 prognoses, respectively.

1117 D) Results of GSEA based on the Z scores in the Cox proportional hazards model.  
1118 Positive and negative NES values indicate the associations with worse or better  
1119 prognoses, respectively. Gene sets were ranked according to the NES value, and  
1120 the gene sets of interest are highlighted. The high-scored gene sets are shown  
1121 in **Fig. S8B**.

1122 E) Overall expression levels of KZFPs in respective cancer stages. The Y-axis  
1123 indicates the GSVA score of KZFPs. The *P* value was calculated by single linear  
1124 regression.

1125 F) Associations of the expression levels of respective gene sets with cancer  
1126 progression. For each gene set, multiple linear regression analysis was  
1127 performed with adjustment for cancer type-specific effects. Positive and negative  
1128 t-scores indicate the tendencies of increase and decrease, respectively, in the  
1129 GSVA scores along with cancer progression. Gene sets were ranked according  
1130 to the t-score, and the gene sets of interest are highlighted. The high-scored gene  
1131 sets are shown in **Fig. S9B**.



1132

1133 **Fig. 4 Phenotypic and gene expression changes caused by the**  
 1134 **overexpression of KZFPs in lung adenocarcinoma cells.**

1135 A–D) Examinations of phenotypic changes in a panel of lung adenocarcinoma  
 1136 (A549) cells overexpressing 30 types of KZFPs (referred to as A549/KZFP cells).  
 1137 These 30 KZFPs satisfy the following criteria: 1) showing a positive correlation  
 1138 with the total expression of HERVs in tumors; 2) possessing expressed HERVs  
 1139 in the vicinity of its TSSs in tumors; and 3) having available ChIP-Seq data

1140 presented by a previous study (Imbeault et al.<sup>33</sup>). The results for the apoptosis  
1141 assay (A), growth assay (B), migration assay (C), and invasion assay (D) are  
1142 shown. The black bar indicates the result of the empty vector-transduced cells.  
1143 Red or blue bars indicate the result of the cells in which the value significantly  
1144 increased or decreased, respectively, compared to that in the empty vector-  
1145 transduced cells in two-sided Student's t-test ( $P$  value < 0.05). The error bar  
1146 indicates the standard error of the mean (SEM).

1147 E) Phenotypic and gene expression changes in A549/KZFP cells. Upper)  
1148 Heatmap showing the gene expression alterations of 2,368 differentially  
1149 expressed genes (DEGs) identified in any of A549/KZFP cells compared to the  
1150 empty vector-transduced cells. Gene expression-based clusters are indicated at  
1151 the top of the heatmap. Lower) Heatmap summarizing the results of the  
1152 experiments shown in A–D). For visualization, the values were log2-transformed  
1153 and subsequently scaled (i.e., the standard deviation was adjusted at 1). An  
1154 asterisk denotes a significant change in the value.

1155 F) Possible target genes of KZFPs critical for cancer progression. The details are  
1156 described in **Fig. S13**. Edge indicates protein–protein interactions defined by the  
1157 Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) (version  
1158 11.0)<sup>55</sup>. The edge width represents the reliability score of the interaction.

1159 G) Schematic view of the *ZNF75D* gene locus. The region excised by CRISPR-  
1160 Cas9 is indicated by the arrow.

1161 H) Expressional correlation between *ZNF75D* and the upstream THE1D-int in  
1162 LUAD tumors.

1163 I) Effect of the HERV integrants on the promoter activity of *ZNF75D*. The effect  
1164 was assessed by a luciferase reporter assay in A549 cells. A pair of the reporter  
1165 plasmids harboring the *ZNF75D* promoters with and without these HERVs were  
1166 constructed, and subsequently, the promoter activities were compared. Error  
1167 bars indicate the SEM.  $P$  values were calculated by two-sided Student's t-test.

1168 J) Effect of the CRISPR-Cas9 excision of these HERVs on the expression of  
1169 *ZNF75D* in A549 cells. The mRNA expression level of *ZNF75D* in each clone of  
1170 cells was measured by qRT-PCR.  $P$  values were calculated using linear  
1171 regression.