Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/6/43/eabc3020/DC1

Supplementary Materials for

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

Jumpei Ito, Izumi Kimura, Andrew Soper, Alexandre Coudray, Yoshio Koyanagi, Hirofumi Nakaoka, Ituro Inoue, Priscilla Turelli, Didier Trono, Kei Sato*

*Corresponding author. Email: ksato@ims.u-tokyo.ac.jp

Published 21 October 2020, *Sci. Adv.* **6**, eabc3020 (2020)
DOI: 10.1126/sciadv.abc3020

The PDF file includes:

Figs. S1 to S10

Other Supplementary Material for this manuscript includes the following:

(available at advances.sciencemag.org/cgi/content/full/6/43/eabc3020/DC1)

Tables S1 to S12

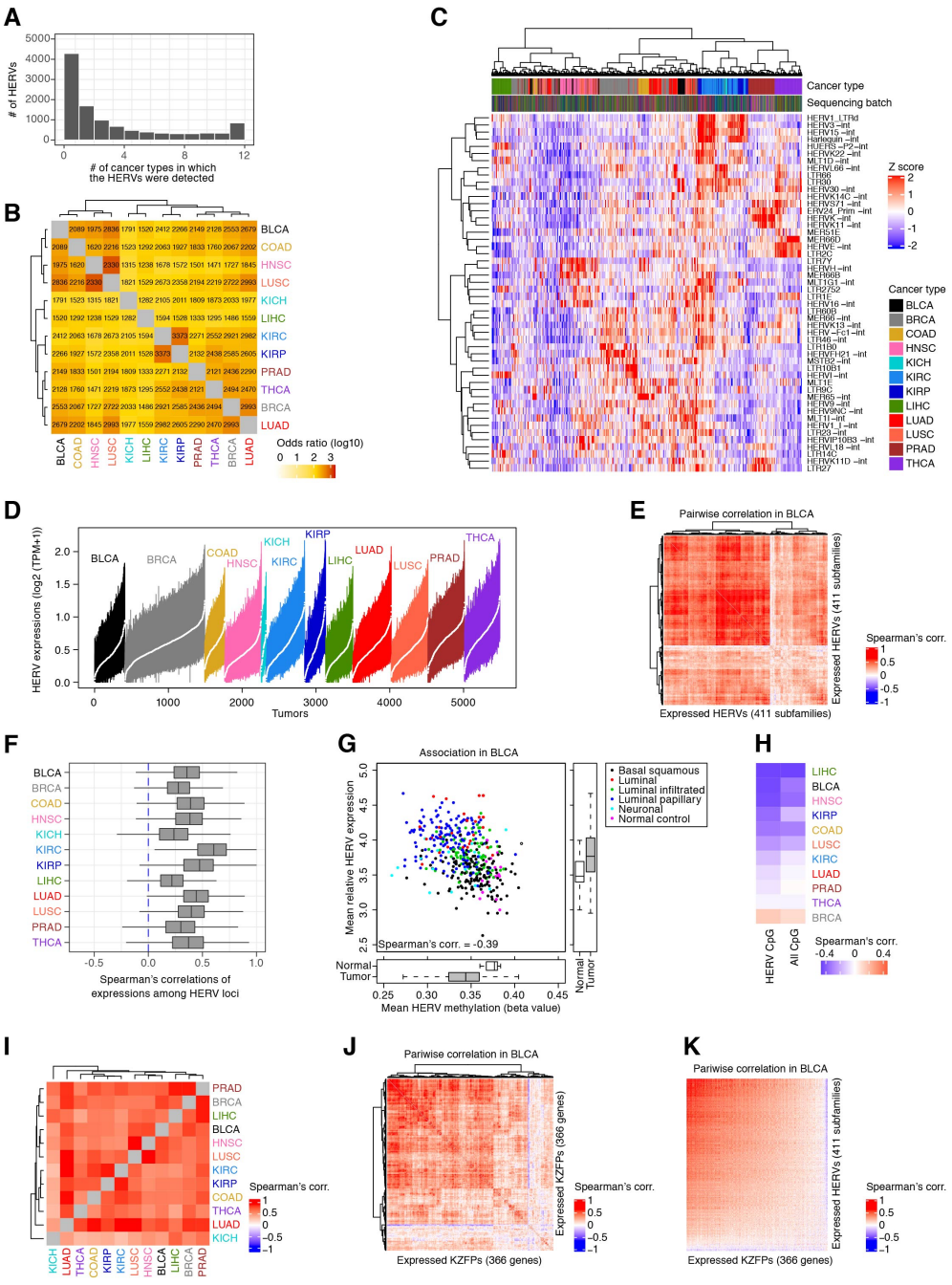

fig. S1 Characterization of HERV expression in 12 types of solid cancers.

(**A**) Histogram showing the numbers of cancer types in which respective expressed HERV loci were detected.

(**B**) Commonality of the sets of expressed HERVs among cancer types. In the heatmap, the color indicates the degree of overlap [log10 (odds ratio)], and the

numerical characters indicate the number of expressed HERV loci that were commonly detected in a pair of cancer types.
(**C**) Hierarchical clustering analysis of TCGA tumors based on the expression profile of HERV subfamilies. The 50 most variably expressed HERV subfamilies were included in the analysis. Information on 161 sequencing batches is indicated.
(**D**) Boxplot showing the distribution of the expression levels of the respective HERV loci in each tumor. The Y-axis indicates the expression levels of the respective HERV loci [log2 (transcripts per million (TPM) + 1)]. A colored line and a white dot indicate the interquartile range and the median value, respectively. Tumors were ordered according to the cancer type and the median expression value.
(**E**) Coexpression of distinct HERV subfamilies in BLCA tumors. The results for all expressed HERV subfamilies are shown.
(**F**) Distribution of the pairwise expressional correlations among HERV loci.
(**G**) Association between the mean HERV expression level and the mean DNA methylation level (beta value) of the CpG sites that are on or proximal [<1 kilo base pairs (kb)] to the expressed HERVs. The result for BLCA is shown, and data for tumors and tumor-adjacent normal tissues are included. The dots are colored according to the sample type (i.e., normal tissue or respective tumor subtypes).
(**H**) Associations of HERV expression levels and DNA methylation levels in respective types of cancers. The results for the HERV-proximal (<1 kb) CpG sites and all CpG sites are shown.
(**I**) Similarities of the gene expression changes upon HERV activation among 12 cancer types.
(**J**) Coexpression of KZFP genes in BLCA tumors. The results for all expressed KZFPs are shown.
(**K**) Coexpression of HERVs and KZFP genes in BLCA tumors. The results for all expressed HERVs and KZFPs are shown.

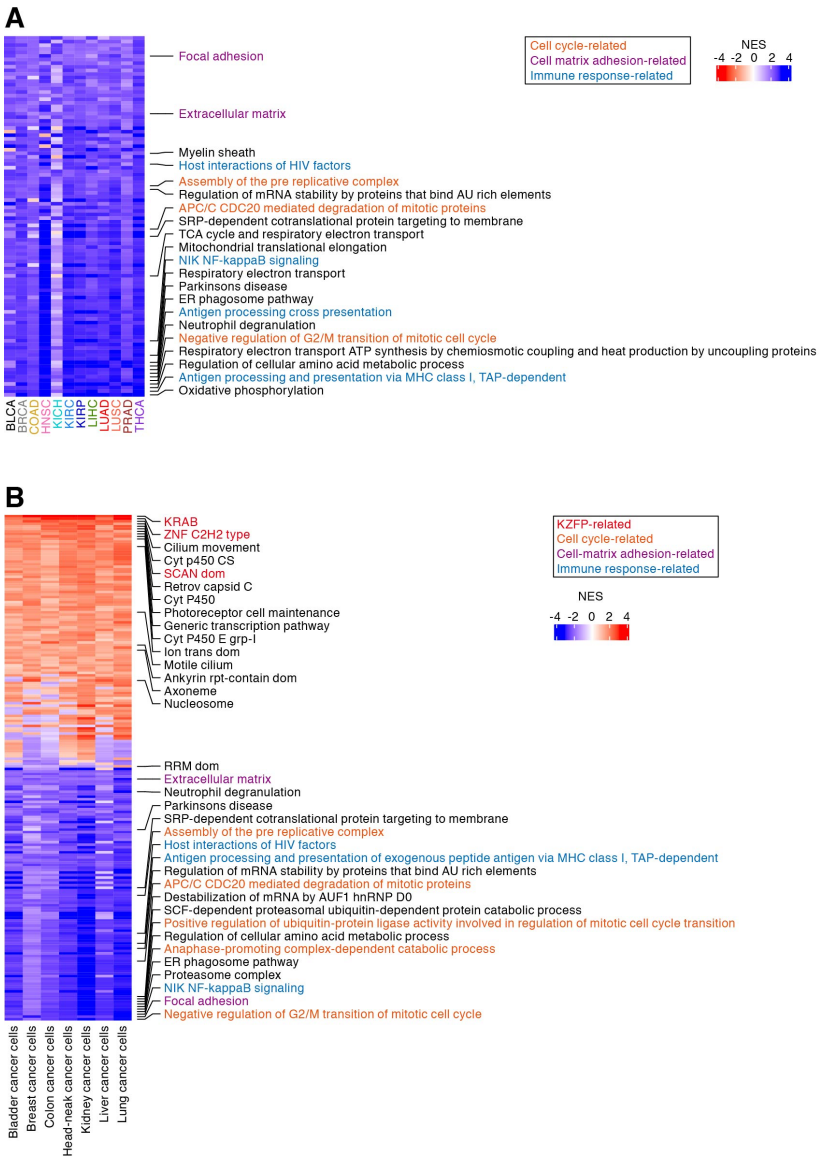

fig. S2 Gene expression signatures associated with increased expression of HERVs and KZFPs

(**A**) Results of GSEA summarizing genes whose expression levels were correlated with the overall expression levels of KZFP genes in the TCGA dataset. The GSVA score (*49*) of the KZFP genes was used to denote their overall expression level. The 100 highest-scored gene sets with negative correlations are shown. Of these, the top 10 gene sets and the gene sets indicated in **Fig. 1F** are annotated.

(**B**) Results of GSEA summarizing genes whose expression levels were correlated with the global expression levels of HERVs in the CCLE dataset. The

100 highest-scored gene sets with positive or negative correlations are shown. Of these, the top 10 gene sets and the gene sets indicated in **Fig. 1F** are annotated.

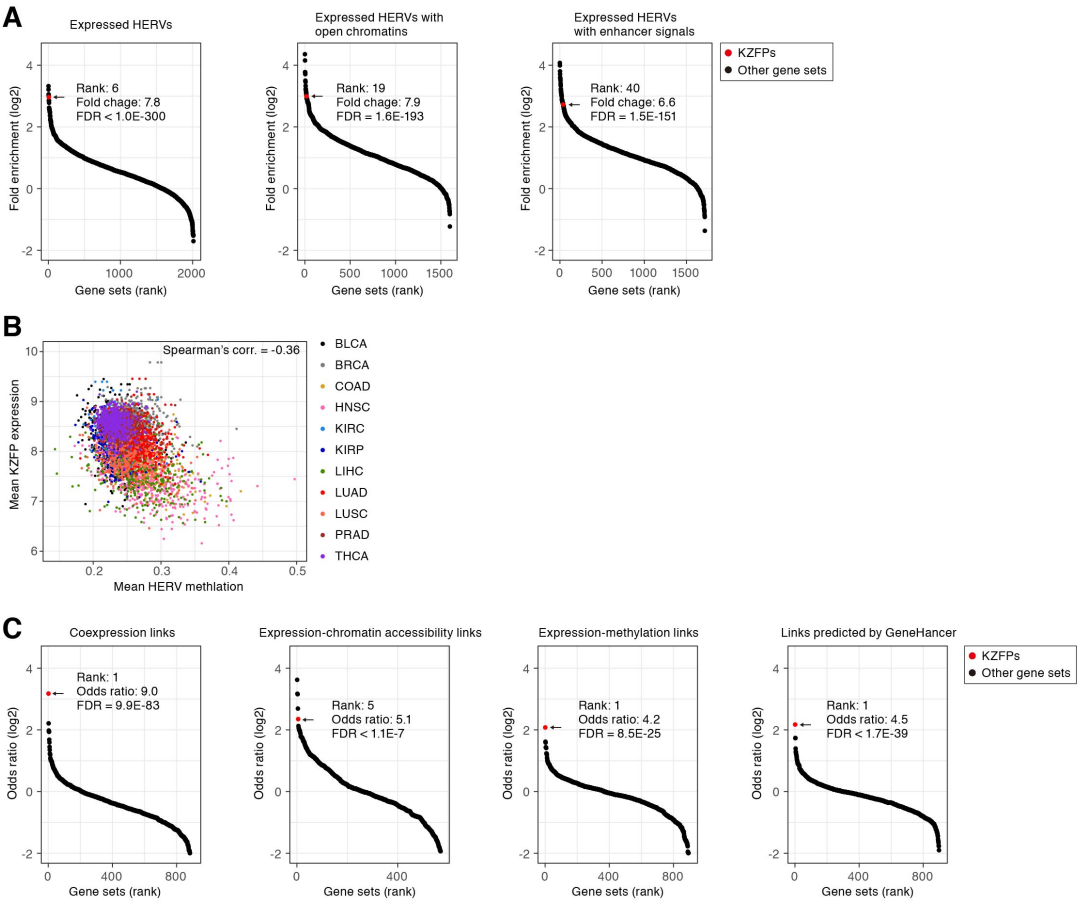

fig. S3 Transcriptional modulation of KZFP genes by adjacent HERVs in tumors.

(**A**) Gene Ontology (GO) enrichment analyses to identify sets of genes that are preferentially present in the vicinity (<10 kb) of the expressed HERVs. The gene sets are ranked according to the fold enrichment scores, and the gene set "KZFP family" is highlighted. Only gene sets with ≥5 hits of HERVs are shown.

(**B**) Association between the mean methylation level of CpG sites that are on or proximal (<1 kb) to the expressed HERVs in the vicinity (<50 kb) of KZFP genes and the mean expression levels of those genes.

(**C**) GO enrichment analyses to identify sets of genes that are preferentially related to the expressed HERVs in the predicted gene regulatory network. Only gene sets with ≥1 hit of HERVs are shown.

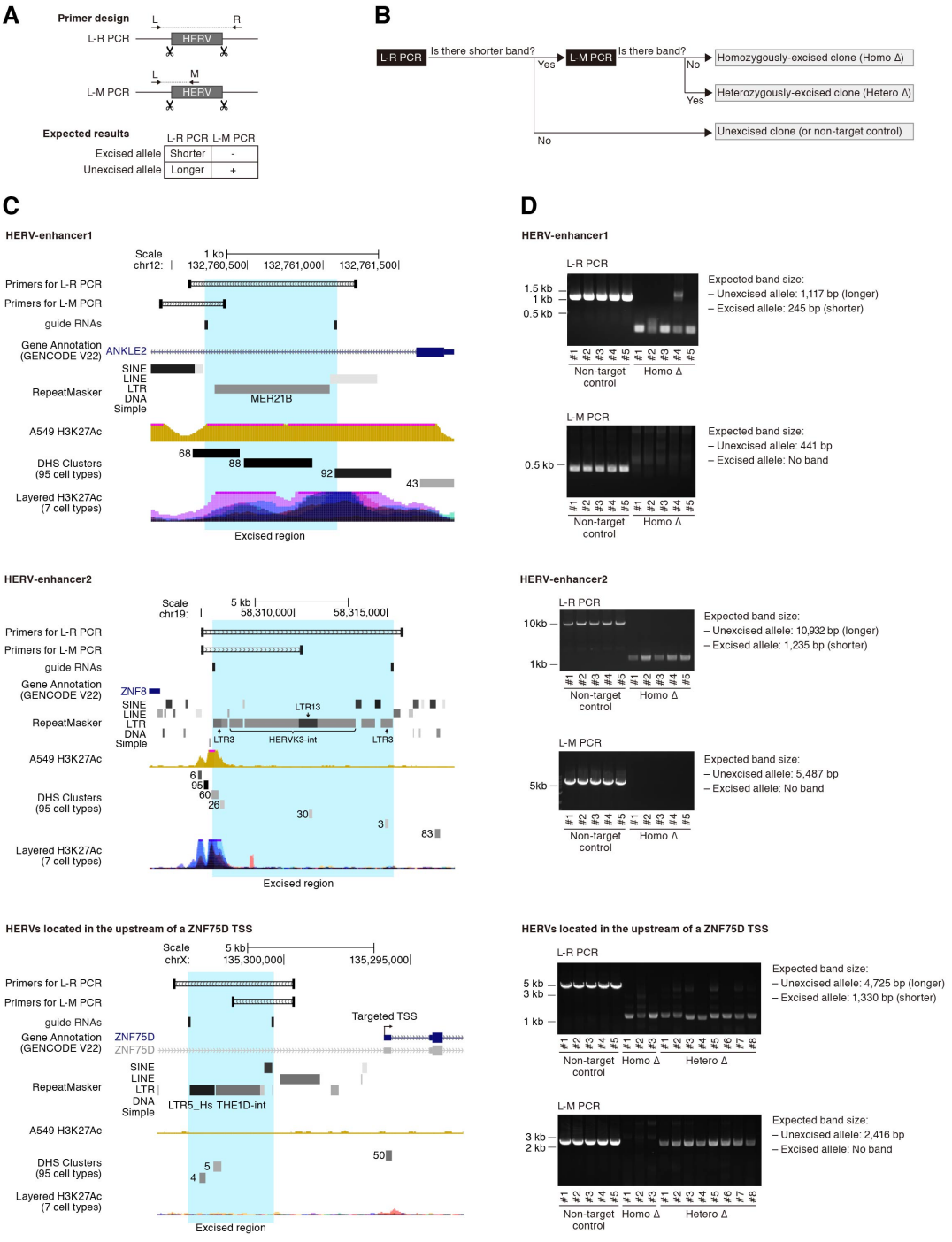

fig. S4 Establishment of HERV-excised A549 cell clones using the CRISPR-Cas9 system.

(**A**) Design of PCR primers to check HERV excision. Two types of PCRs were designed: L-R PCR and L-M PCR. In L-R PCR, shorter and longer bands are amplified from HERV-excised and -unexcised alleles, respectively. In L-M PCR,

a single band is amplified from the HERV-unexcised allele (but not from the HERV-excised allele).
(**B**) Scheme for screening HERV-excised cell clones using L-R and L-M PCRs.
(**C**) UCSC genome browser views of the target HERVs used in the excision experiments. The views for HERV-enhancer1 (top), HERV-enhancer2 (middle), and HERVs located upstream of a *ZNF75D* TSS (bottom) are shown.
(**D**) PCR results of HERV-excised clones or nontarget control clones. The results for HERV-enhancer1 (top), HERV-enhancer2 (middle), and HERVs located upstream of a *ZNF75D* TSS (bottom) are shown.

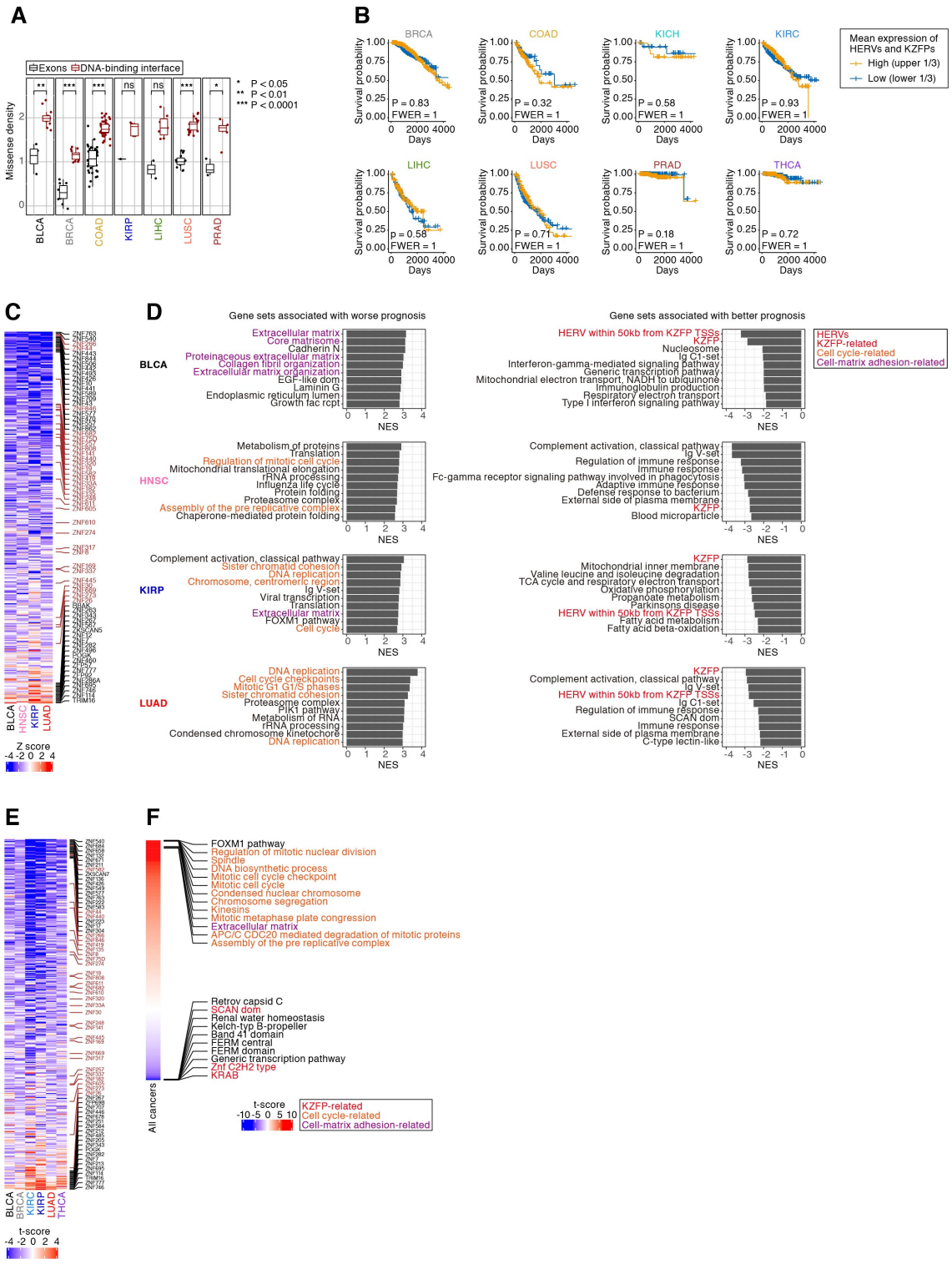

fig. S5 Association of HERV/KZFP expression status in tumors and cancer disease conditions

(**A**) Accumulation of somatic missense mutations in the DNA-binding amino acid residues of KZFP genes. The mutation density (counts per Mb per patient) of KZFP genes was compared between the DNA-binding amino acid residues (red) and the whole coding regions (black). The results for KZFP genes with ≥1 mutations are shown. *P* values were calculated by the two-sided Wilcoxon rank sum test.

(**B**) Kaplan–Meier survival plots of cancer patients with high or low expression levels of HERVs and KZFPs. Cancer patients were stratified according to the mean value of the gene set-wise expression scores [GSVA scores (*49*)] between KZFPs and HERVs. Statistical significance was evaluated by the two-sided log-rank test. The results for BRCA, COAD, KICH, KIRC, LIHC, LUSC, PRAD, and THCA tumors are shown (results for the others are shown in **Fig. 3A**).

(**C**) Associations of the expression of respective KZFP genes with cancer prognosis. The Z score in the Cox proportional hazards model is shown. The top 20 KZFP genes with respect to the association with better or worse prognoses are annotated. In addition, the KZFP genes used in the overexpression experiments (**Fig. 4**) are annotated and highlighted.

(**D**) The high-scored gene sets from GSEA based on the Z scores in the Cox proportional hazards model. Redundant gene sets were removed from the results.

(**E**) Associations of the expression of respective KZFP genes with cancer progression. The associations of the expression levels of respective genes with cancer progression were evaluated using single linear regression analysis. Positive and negative t-scores indicate the tendencies of increased and decreased expression, respectively, of the genes along with cancer progression. The top 20 KZFP genes with respect to the positive and negative associations are annotated. In addition, the KZFP genes used in the overexpression experiments (**Fig. 4**) are annotated and highlighted.

(**F**) The high-scored gene sets in the linear regression analysis (shown in **Fig. 3E**). The top 10 gene sets and the gene sets indicated in **Fig. 1F** are annotated.

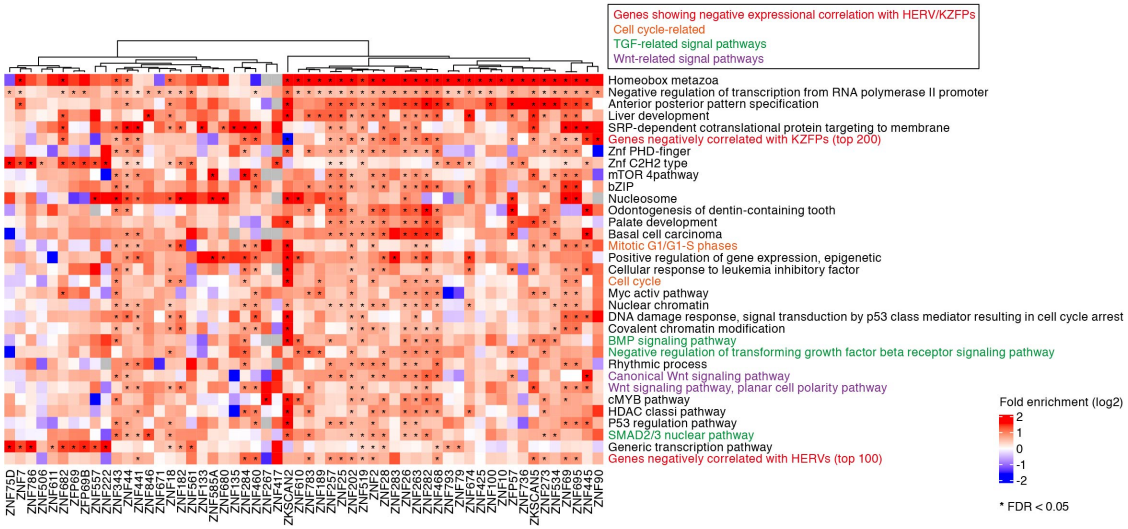

fig. S6 GO enrichment analysis to identify sets of genes that are preferentially bound by respective KZFPs.

This analysis was based on a publicly available ChIP-Seq dataset of KZFPs [Imbeault et al. (*33*)]. In the heatmap, the color indicates the log2-transformed fold enrichment. An asterisk denotes a significant enrichment (fold enrichment > 1.5; FDR < 0.05 in binomial test). The heatmap includes 1) gene sets that were significant in >12 KZFPs; and 2) KZFPs in which ≥1 gene set above was significant.

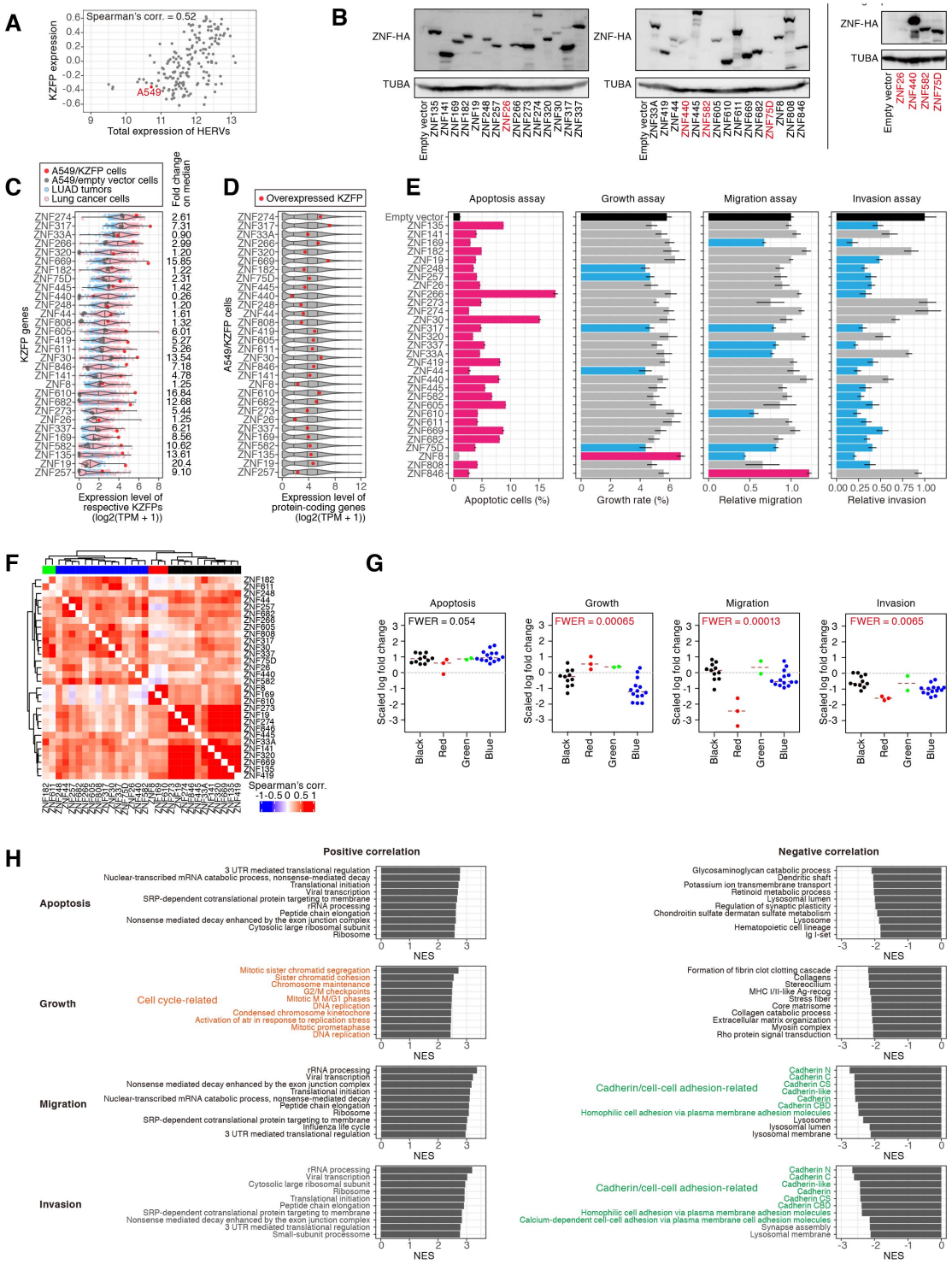

fig. S7 Phenotypic and gene expression changes caused by the overexpression of KZFPs in lung adenocarcinoma cells.

(**A**) Expression levels of HERVs and KZFPs in lung cancer cells in the CCLE dataset. The X-axis indicates the total expression levels of HERVs (log2-

transformed CPM), and the Y-axis indicates the overall expression levels of KZFP genes (GSVA score). A dot corresponding to A549 cells is highlighted.
(**B**) Western blotting to confirm the exogenous expression of KZFP proteins in A549/KZFP cells using an anti-HA antibody. Since the target bands in several A549/KZFP cells were relatively faint (indicated as red in the left panel), the results with long exposure are also shown for these cells (in the right panel).
(**C**) Expression level of overexpressed KZFPs in A549/KZFP cells (red), A549/empty vector cells (gray), lung adenocarcinoma (LUAD) tumors (blue), and lung cancer cell lines (pink). The median value is indicated. In the right of the panel, the fold change value (A549/KZFP cells vs. the median) is indicated.
(**D**) Expression level of all protein-coding genes expressed in A549/KZFP cells. The value for the overexpressed KZFP is indicated as a red dot. Quantiles are shown.
(**E**) Examinations of the phenotypic changes in a panel of A549/KZFP cells. The results of the apoptosis assay, growth assay, migration assay, and invasion assay are shown. The black bar indicates the result of the empty vector-transduced cells. Red or blue bars indicate the result of the cells in which the value significantly increased or decreased, respectively, compared to that in the empty vector-transduced cells (P value < 0.05). The error bar indicates the standard error of the mean (SEM).
(**F**) Similarity of the gene expression alterations induced by KZFPs among the A549/KZFP cells. Spearman's correlations of the fold changes of gene expression were calculated among the A549/KZFP cells. The clusters indicated on the upper side of the heatmap are the same as those in **Fig. 4A**.
(**G**) Phenotypic differences of the cells among the gene expression-based clusters. Statistical significance was evaluated by one-way ANOVA.
(**H**) Genes whose expression levels were associated with the measured phenotypes in A549/KZFP cells. For each phenotype, Spearman's correlations of the expression levels of respective genes and the phenotype score were calculated, and GSEA was subsequently performed according to those correlation scores. Regarding the positive and negative correlations, the results for the top 10 gene sets are shown.

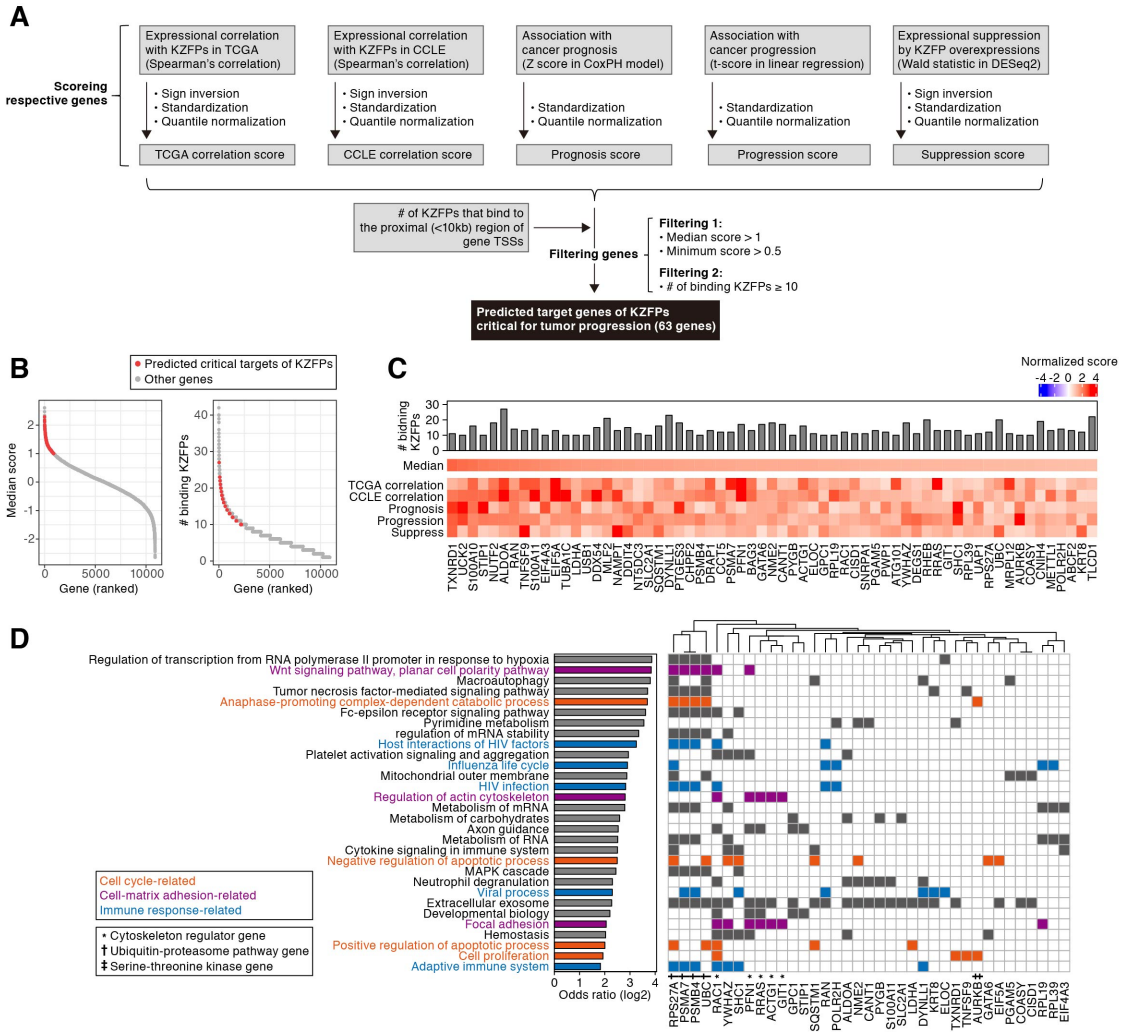

fig. S8 Identification of genes that are likely to be targeted by KZFPs and critical for cancer progression.

(**A**) Scheme for extracting the target genes of KZFPs critical for tumor progression. For each gene, the 1) TCGA (i.e., primary tumors) correlation score, 2) CCLE (i.e., cancer cell lines) correlation score, 3) prognosis score, 4) progression score, and 5) suppression score were calculated. Genes with a median score >1 and a minimum score >0.5 were extracted. Subsequently, genes targeted by ≥10 KZFPs were extracted. Finally, 63 genes were extracted. Details are described in the **"Scoring system of genes for predicting the targets of KZFPs critical for cancer progression"** subsection in the **Materials and Methods** section.

(**B**) Distribution of the median score (left) and number of binding KZFPs (right) of the respective genes.

(**C**) The number of binding KZFPs (upper), the median score (middle) and respective scores (lower) of the 63 genes.
(**D**) GO enrichment analysis summarizing the target genes of KZFPs critical for cancer progression. Of the significant gene sets (i.e., FDR < 0.1 in Fisher's exact test; number of hits ≥ 5), the top 30 gene sets according to the odds ratios are shown. Log2-transformed odds ratios (left) and gene memberships (right) are indicated. Cytoskeleton regulator genes (*), ubiquitin-proteasome pathway genes (†), and serine-threonine kinase genes (‡) are annotated under the heatmap.

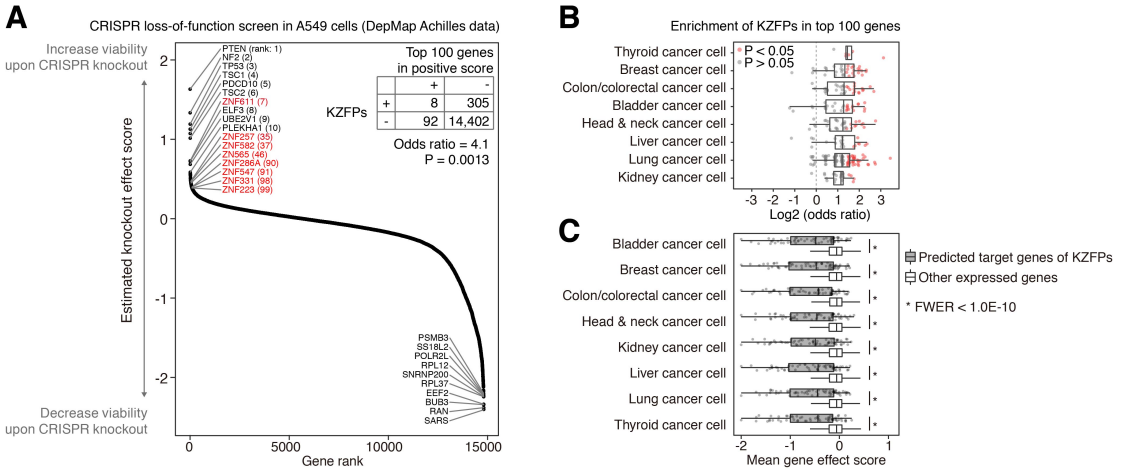

fig. S9 Analysis of the CRISPR knockout screening dataset provided by Cancer Dependency Map (DepMap) Achilles project.

(**A**) Gene ranking according to the estimated knockout effect score [CERES score (*39*)] in A549 cells. The negative and positive values respectively represent the increase and decrease of cancer cell viability upon the CRISPR gRNA library transduction. This score is adjusted so that the median knockout effects of predefined essential and nonessential genes are represented as minus one and zero values, respectively. The top 10 genes in both sides and the KZFP genes in the top 100 on the positive side (indicated in red) are indicated. The numbers in parentheses on the positive side denote the rank of the gene indicated. Statistical enrichment of KZFP genes in the top 100 genes in the positive score was evaluated by the two-sided Fisher's exact test (inset).

(**B**) Enrichment of KZFP genes in the top 100 genes in the positive score in respective cancer cell lines.

(**C**) Comparison of the distributions of the gene effect scores between the predicted KZFP target genes (shown in **Fig. 4D**) and the other expressed genes. The mean value of the scores in respective cancer types is shown. Statistical significance was evaluated by the two-sided Wilcoxon rank sum test.

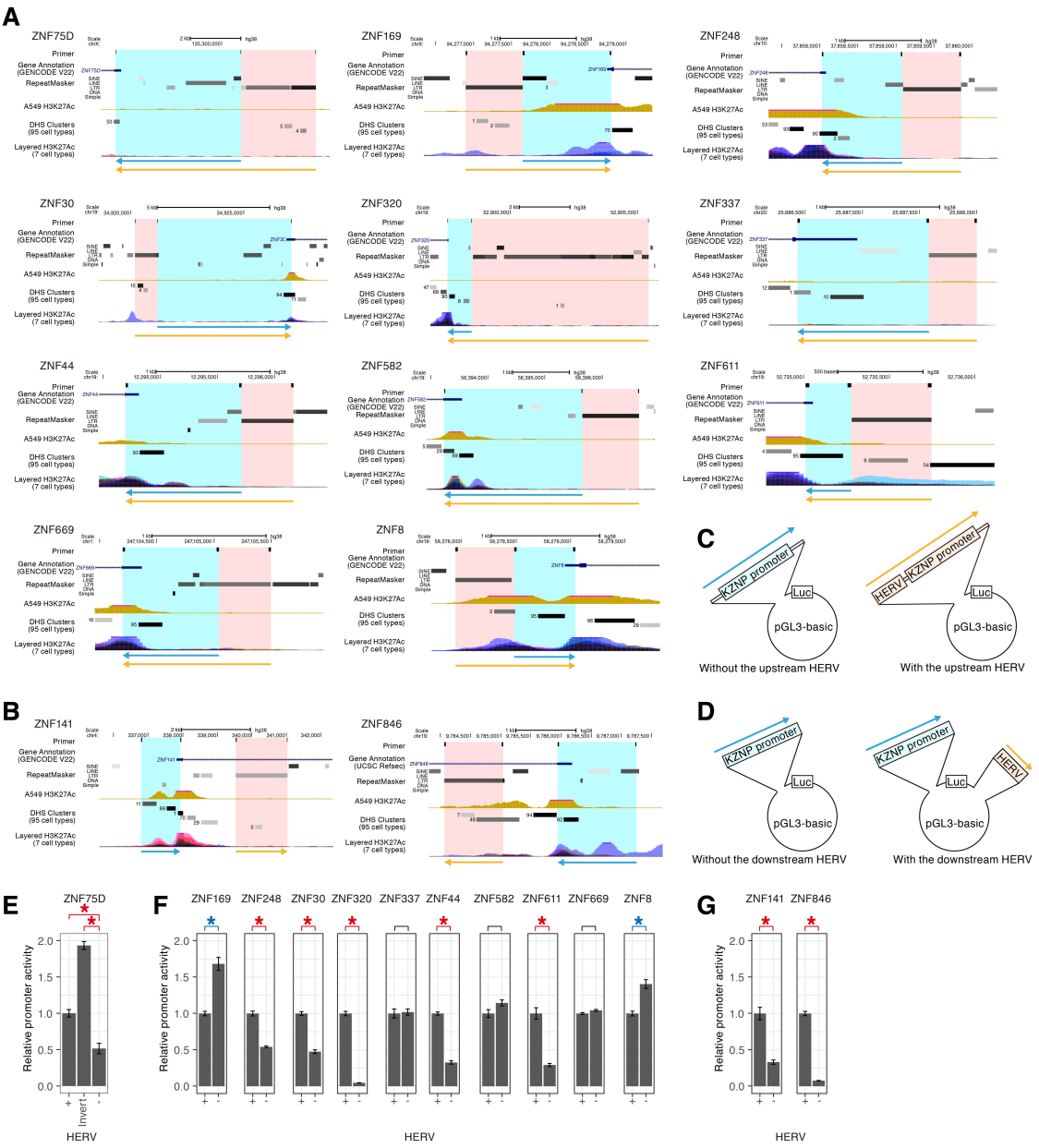

fig. S10 Luciferase reporter assay to assess the effects of the surrounding HERVs on the promoter activities of KZFP genes.

(**A**) and (**B**) UCSC genome browser views of the target HERVs and KZFP promoters used in the luciferase reporter assay. The panel for a HERV in the upstream region of the KZFP promoter is shown in (**A**), while that for a HERV in the downstream region is shown in (**B**). The genomic region inserted into the reporter plasmid is indicated by the arrow. In (**A**), the orange or blue arrows indicate genomic fragments with or without a HERV sequence, respectively. In (**A**), the orange or blue arrows indicate the HERV or KZFP promoter, respectively.

(**C**) and (**D**) Schematics of the reporter plasmids. A schematic for a HERV in the upstream region of the KZFP promoter is shown in (**C**), while that for a HERV in the downstream region is shown in (**D**).
(**E**) Assessment of the directional effect of the HERVs on the promoter activity of *ZNF75D*.
(**F**) Assessment of the effect of the upstream HERVs on the promoter activity of the KZFP gene.
(**G**) Assessment of the effect of the downstream HERVs on the promoter activity of the KZFP gene.