Fig. S1

TCGA-LUAD and TCGA-LUSC raw sequencing data

Pre-processing

Conversion to FASTQ — BioBamBam2
Trimming and QC — Trim Galore
Alignment — Bowtie 2 (-k 100)
Sorting and indexing — Samtools
HERV annotation — Telescope

Raw Count matrix
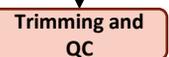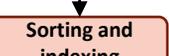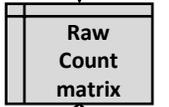
Differential Expression Analysis

HERVs with a mean of TMM < 5 across all samples were removed

Filtering — edgeR

Norm. and Diff. Expr. Analysis — LIMMA
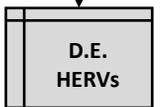
|Log2FC| > 0.58 and an adj.p.value < 0.05

D.E. HERVs

Survival Analysis

Filtering → Normalization → TMM matrix

HERVs with a geometric mean of RPM < 1 across all samples were removed

Survival Analysis — Univariate Cox P-Val

HERV ass. with OS

HERV ass. with RFS
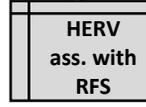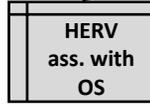
Pathway Analysis

RPM of D.E. HERVs
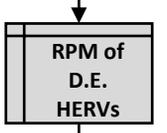
For both HERVs and coding-protein genes identified as differentially expressed (D.E.) in TCGA-LUAD and TCGA-LUSC, we extracted their original raw counts and we scaled them to Read Per Million mapped reads (RPM)

D.E. genes → RPM of D.E. genes → Correlation analysis — Spearman correlation

HERV-gene corr.matrix
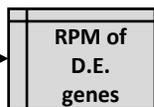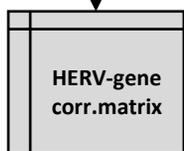
HERV

List of HERV's correlated genes with their ENTREZ IDs and Log2FC values
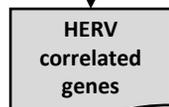
HERV correlated genes — Spearman corr. coef. > 0.2 and adj.P.Value < 0.05

MITHrIL was runned for each HERV's correlated gene list.

Pathway analysis — MITHrIL

HERV-pathway clustering
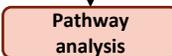
HERV-specific pathways — P.Value < 0.05
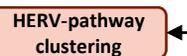
ConsensusClusterPlus ComplexHeamap
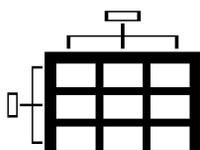
Heatmap with HERV-pathway clustering