METHODS

Cell lines and cellular assays

LoVo, HCT116, HCT8, HCT15, DLD1, SW620, and LS123 were cultured in RPMI-1640 + 10% FBS + 1% Pen/Strep (Gibco/Life Technologies). Cell lines were obtained within the institution and the majority original obtained from ATCC. All cells were tested periodically for mycoplasma contamination giving negative results. For shRNA experiments, HCT116 cells were infected with pLKO-based lentiviruses encoding either scrambled shRNA or shRNAs targeting human *TP53* (D1: TCAGACCTATGGAAACTACTT; D2: GTCCAGATGAAGCTCCCAGAA; D3: CACCATCCACTACAACTACAT. Cells were selected 48 hours after infection with puromycin (1 μg/mL) for 7 days.

Bisulfite DNA Methylation Sequencing

Isolated genomic DNA was bisulfite converted using the EZ DNA Methylation-Lightning Kit (Zymo Research) using the manufacturer's protocol. Briefly, 50-100 ng of DNA was bisulfite treated with the Lightning Conversion Reagent and was incubated at the recommended temperatures for the provided durations. The converted DNA was loaded onto a Zymo-Spin IC Column on which it underwent a binding step, desulphonation step, and elution step, each separated by a wash step. 10 uL of 1X Low EDTA TE Buffer (Quality Biological) was used to elute the bisulfite-converted DNA. Illumina library construction was performed using the EpiGnome Methyl-Seq Kit (Epicentre/Illumina) according the manufacturer's specifications with an input of 9 uL of the eluted bisulfite-treated DNA. Library validation and quantification was performed by quantitative PCR using the KAPA SYBR® FAST Universal qPCR Kit (Kapa Biosystems). The individual libraries were pooled at equal concentrations, and the pool concentration was determined using the KAPA SYBR® FAST Universal qPCR Kit. The pool of libraries underwent paired-end sequencing on a MiSeq using a 150-cycle kit.

The paired-end reads were aligned to the human_g1k_v37.fasta reference genome from http://www.1000genomes.org using BSMAP version 2.74 with default settings. Duplicate reads were marked using the MarkDuplicates tool in picard-tools-1.8.4 and were removed. BSMAP's methratio.py script was used to extract methylation ratio data from the mapped reads with the --zero-meth, --remove-duplicate, and --combine-CpG options. CpG-specific methylation data was extracted by taking the rows of the methratio.py output that had a value of "+" in the strand field and had a sequence context that began with "..CG". Each CpG in the CpG methylation data table was annotated with repeat masker information (ucsc_repeatmasker_2014-06-23.txt from the UCSC genome browser (https://genome.ucsc.edu) and its distance from the nearest Gencode annotation (gencode.v19.annotation.gtf from http://www.gencodegenes.org). CpG methylation fractions for specific repeat types were averaged over all repeats of the given type. Distances to the nearest annotated genes were binned in 5,000-base increments from 0 to 100,000 bases, and CpG methylation fractions were averaged by distance bin within each repeat class. All FASTQ files have been uploaded to NCBI GEO GSE90966.

Cell RNA in situ hybridization (RNA-ISH)

ISH was performed according to the QuantiGene ViewRNA ISH Cell Assay (Affymetrix, Santa Clara, CA) user manual on cell cytospin slides. Briefly, cells were permeabilized by detergent solution for 5 min at room temperature and digested with protease (1:2000) for 10 min. Target probe sets were applied and hybridized to the cells by incubating for 3 hrs at 40°C. Probes were used at a dilution of 1:100 Human LINE-1_ORF1 Type 1(Cat # VA1_14038), and 1:50 Human GAPDH Type 6 (Cat # VA6_10337) (Affymetrix). Signal was amplified through the sequential hybridization of Pre Amplifier and Amplifier mixes to the target probe set and target RNA molecules were detected by applying Label Probe Mix. Cells were then counterstained with DAPI (5 mg/ml, Life Technologies) and slides were mounted using ProLong Gold Antifade Reagent (Life Technologies). Type 1 probes were detected in the Cy3 (550 nm) channel and Type 6 probes in the Cy5 channel (650 nm).

LINE-1 and GAPDH quantification was done post staining using the Keyence BIOREVO fluorescent microscope. Each slide was imaged in 10 different locations with DAPI, Cy3 (green), and Cy5 (red) filters. Channels were merged into a single image, which was then processed through the BIOREVO Analyzer Hybrid Cell Count. Individual cells were recognized by DAPI and cell boundaries were hand adjusted with the fine-edit tool. Red and green signal was detected for each cell above a uniform, automated brightness threshold. Integration values for red and green color represent expression of LINE-1 and GAPDH, respectively, in a single cell. LINE-1 signal was normalized relative to GAPDH by dividing the integration value for LINE-1 by the integration value for GAPDH to yield a single data point for each cell. Data was exported and graphed in Prism.

RNA-seq and Methylation analysis of TCGA Colon Cancer Data

Raw Illumina reads were quality filtered as follows. First, ends of the reads were trimmed to remove N's and bases with quality less than 20. After that the quality scores of the remaining bases were sorted and the quality at the 20th percentile was computed. If the quality at the 20th percentile was less than 15, the whole read was discarded. Also, reads shorter than 40 bases after trimming were discarded. If at least one of the reads in the pair failed the quality check and had to be discarded, we discarded the mate as well.

Quality filtered reads were mapped to the human genome (GENCODE annotation, build 38) and to repbase elements (release 20) using STAR aligner. Aligned reads were assigned to genes using the featureCounts function of Rsubread package using the external (GENCODE) annotation. This produced the raw read counts for each gene. Mapping and counting of the reads is done in two stages. First, reads are mapped to the human genome, and the counts are determined using the gencode annotation and the annotation derived from the repeatmasker output. After that the reads which were not assigned to any feature in either GENCODE and repeatmasker annotation were re-aligned to the repeat consensus sequence (repbase). Counts obtained from repeatmasker and repbase were added together.

Gene expression in terms of log2-CPM (counts per million reads) was computed and normalized across samples using the TMM method as implemented in the calcNormFactors function of edgeR package. We used expression of coding genes only for computing the library size and normalization factors.

We have used the 364 colorectal non-FFPE samples from TCGA (COAD and READ) which were sequenced using paired-end reads. We have computed Pearson correlation between expression of HERV-K and LINE-1 as well as between expression of HERV-H and LINE-1. We have assessed the statistical significance of the observed correlation coefficient using the t-test as implemented in the cor.test function in R.
Heatmap shows expression of the various LINE-1 and HERV repeat elements. Expression was transformed to the Z-score, samples and genes were reordered according to the dendrogram produced by the hierarchical clustering.

We have identified the 354 TCGA COAD and READ tumors which had paired-end RNA-Seq data as well as matched Illumina HumanMethylation450 methylation data. We have computed the total expression of LINE-1 elements from the sum of the read counts for all LINE-1 elements. We have computed the representative genome methylation for each sample as the median beta-value across all the probes in the array. We have sorted the samples according to the computed total expression of LINE-1 elements and split them into the three equal groups (low, median, high LINE-1). We have computed the correlation between the total LINE-1 expression and the representative methylation beta-value using the Pearson correlation.
The contingency table for LINE-1 level and MSI status was calculated using the Chi-square test for significance. Statistical analyses (t-test for the representative methylation and for significance of the correlation) were performed using R.

References:

[STAR] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner, Bioinformatics 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25

[Rsubread] Liao Y, Smyth GK and Shi W (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41(10):e108.

[edgeR] Robinson MD, McCarthy DJ and Smyth GK (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.
Bioinformatics 26, 139-140

Robinson, MD, and Oshlack, A (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology 11, R25.

Human Colon Cancer Specimens

Human tumor tissues were obtained from the Massachusetts General Hospital according to an IRB-approved protocols 2012P000039 and 2014P000731. This cohort of resected human primary colorectal cancers from the Massachusetts General Hospital spanned over 2001-2013. In this cohort, MSI+ instability status was assessed by immunohistochemistry loss for MLH1, PMS2, MSH6 and MSH2. BRAF mutational status was determined by PCR based genotyping as part of the clinical work up.

Manual RNA-ISH on FFPE tissue sections

FFPE tissue sections were stained by ViewRNA ISH tissue assay (Affymetrix, Santa Clara, CA) for one-plex assay (LINE-1). Briefly, baked tissue sections were subjected to Histoclear deparaffinization, followed by ethanol dehydration. To unmask the RNA targets, dewaxed sections were incubated in 500mL of 1X pretreatment buffer at 90 to 95°C for 10 minutes and digested with 1:100 dilution protease at 40°C for 10 minutes, followed by fixation with 10% neutral buffered formalin at room temperature for 5 minutes.

For one-plex assay, unmasked tissue sections were subsequently hybridized with 1:100 dilution of the Human LINE-1_ORF1 Type 1 (Cat # VA1_14038) for 3 hours at 40°C, followed by a series of post-hybridization washes. Signal amplification was achieved by a series of sequential hybridizations and washes as described in the user manual. The specific conditions were as follows: pre-AMP: 25 minutes at 40°C; AMP: 15 minutes at 40°C; hybridization with labeled probe: 1:1000 dilution for 15 minutes at 40°C; signal detection with fast-red substrate: 30 minutes at 40°C.

The intensity of LINE-1 and HERV-H ISH signal in tumor cells was compared to the adjacent stromal cells. High repeat RNA was defined as tumor cell signal greater than stromal signal. Low repeat RNA was defined as tumor cells signal less than or equal to that seen in stromal cells

Automated RNA-ISH and IHC on FFPE tissue sections

Monoclonal antibodies for T-reg marker FOXP3 and RNA-ISH probe for HERV-H were sequentially applied on a single histologic slide, and contrasting chromogens were used to visualize the antibody: brown (diaminobenzidine [DAB]) for FOXP3 and RNA: red (fast red) for HERV-H. FFPE sections were mounted on coated slides and were placed in an oven for 60 minutes at 60°C. The sequential double-staining protocol was performed using the Leica Bond Rx automated immunostainer. Deparaffinization (View RNA Dewax1 protocol) and on-board antigen retrieval were performed for 20 minutes at approximately 100°C with HIER2 reagent, which is an EDTA-based proprietary Leica solution (pH 8.0–8.5). Monoclonal FOXP3 antibody (eBioscience Clone, Cat # 150D/E4) was diluted as 1:50 in Leica antibody diluent solution. For ISH part View- RNA eZL Detection Kit (Affymetrix) was used on the Bond RX immunohistochemistry and ISH Staining System with BDZ 6.0 software (Leica Biosystems). The Bond RX user-selectable settings for part 2 were as follows: ViewRNA eZ-L Detection 1-plex (Red) protocol; ViewRNA Dewax1 Preparation protocol; ViewRNA Enzyme 2 (10); ViewRNA Probe Hybridization 3hrs . With these settings, the RNA unmasking conditions for the FFPE tissue consisted: 10-minute incubation with Proteinase K from the Bond Enzyme Pretreatment Kit at 1:1000 dilution (Leica Biosystems). HERV-H (Cat # DVF1-19702) Ez probes were diluted as 1:40 in ViewRNA Probe Diluent (Affymetrix). Post run, slides were rinsed with water, air dried for 30 minutes at room temperature and mounted using Dako Ultramount (Dako, Carpinteria,CA), and visualized using a standard bright-field microscope. Punted dote like red color hybridization signals in the cell cytoplasm defined as positive signals for HERV-H and brown nuclear reactivity was considered as positive for FOXP3.

Additional details on the automation are provided at: http://www.affymetrix.com/estore/browse/level_three_category_and_children.jsp?category=cat640026&categoryIdClicked=cat640026&expand=true&parent=cat640022.

PrimeFlow^{TM} RNA assay

ISH was performed according to PrimeFlow^{TM} RNA assay (Affymetrix, Santa Clara, CA) user manual. Briefly, 1–5 x 10^6 cells were aliquoted in Flow Cytometry Staining Buffer, fixed in PrimeFlow RNA Fixation Buffer 1 for 30 minutes at 2–8°C, permeabilize with 1X PrimeFlow RNA Permeabilization Buffer followed by fixation with 1X PrimeFlow RNA Fixation Buffer 2 for 60 minutes at room temperature. Human LINE-1_ORF1 Type 1(Cat # VA1_14038), 1/20 dilution and Human GAPDH Type6 (Cat # VA6_10337), 1/20 dilution target Probe were hybridized for 2 hours at 40°C followed by series of washing. Target probe hybridization was followed by signal amplification steps as: PreAmp hybridization for 1.5 hours at 40°C, Amp hybridization for 1.5 hours at 40°C and Label Probe hybridization for 1 hour at 40°C. Samples were then analyzed on a flow cytometer. fcs files were then analyzed using FlowJo software. LINE-1 and GAPDH signal expression value was extracted per single cell with the software. LINE-1 signal was normalized relative to GAPDH by dividing the value for LINE-1 by the value for GAPDH to yield a single data point for each cell. Data was exported and graphed in Prism.

LINE-1 quantitative reverse transcriptase PCR (qRT-PCR)

RNA was extracted from cells using the Trizol reagent. DNase treatment was conducted using the Ambion DNA-free kit. Quantitative RT-PCR was conducted on the DNase treated RNA using the KAPA SYBR-fast onestep qRT-PCR kit according to manufacturer recommendations. All qPCR assays were conducted in triplicate.

Primers used were:
Human *TP53* (For: 5'-AACCCACAGCTGCACAG-3'; Rev: 5'-CCTTCCCAGAAAACCTACCAG-3')
Human *GAPDH* (For: GGAGCGAGATCCCTCCAAAAT; Rev: GGCTGTTGTCATACTTCTCATGG)
Human *LINE-1* 5' UTR (For: GAACAGCTCCGGTCTACAGC, Rev: TCACCCCTTTCTTTGACTCG)
Human *LINE-1* 3' UTR (For: TGATGAGTTCATATCCTTTGTAGGG, Rev: GATATTCCCCTTCCTGTGTCC).

CellTiter-Glo Cell Viability Assay: CRC cell lines were cultured in 96-well clear bottom black polystyrene plates for 24 hours. At 24 hours, cells were treated with 0.078, 0.312, 0.625,1.25, 2.5,5, 10, and 20 µM of AZA (Sigma-Aldrich). At 72 hours post treatment, CellTiter-Glo cell viability reagent (Promega, Madison, WI, USA; Cat # G7571) was added to each well in 1:1 ratio with media. Then, the assay was performed according to the manufacturer's instructions.

Luminescence was measured on SpectraMax M5 at an integration time of 1000 milliseconds and IC50 was calculated for each cell line with GraphPad Prism software.

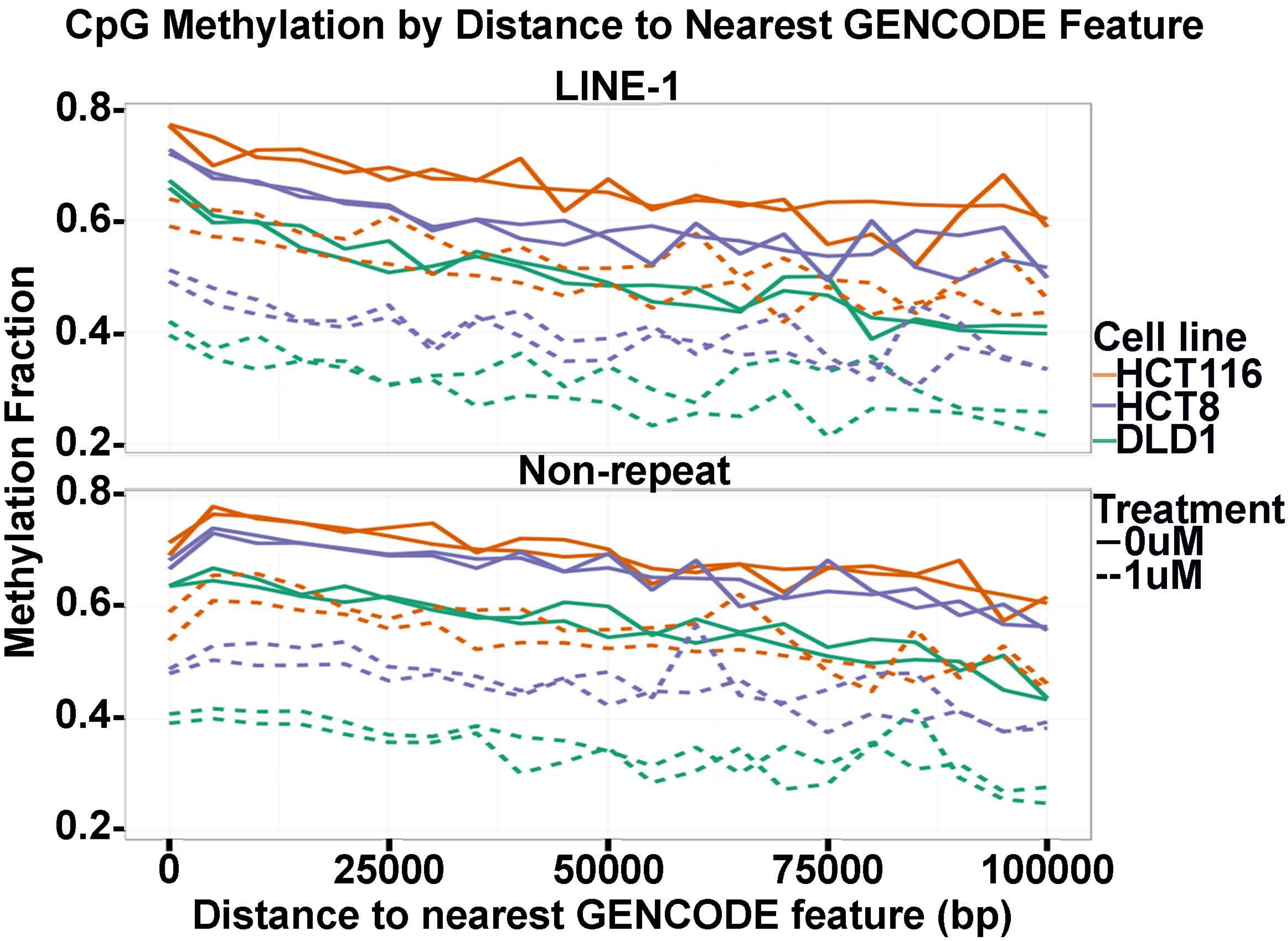
Annexin V

CRC cell lines were cultured in 10 cm plate for 24 hours. Cells were treated with 1uM of AZA for 48 hours and harvested by dissociating with Trypsin, washed with 1 X PBS, then once in 1X Binding Buffer. Cell suspensions were stained with Annexin V/FITC and Propidium iodide (PI) according to the manufacturer's instructions (eBioscience, Affymetrix, San Diego, CA USA). Briefly, cells were suspended in 100 µl 1 X binding buffer at a concentration of 1-5 × 10^6 combined with 5 µl Annexin V/FITC. After 15 minutes of incubation in the dark at room temperature, cells were washed with 2 mL 1X Binding Buffer and resuspend in 200 µL of 1X Binding Buffer. 5 µL of Propidium Iodide was added to the suspension and analyzed by flow cytometry. Cells that were Annexin V-negative and PI-negative were considered viable cells. Cells positive for Annexin V only were considered early apoptotic, and cells positive for Annexin V and PI were considered necrotic or late apoptotic. All samples were prepared in triplicate.

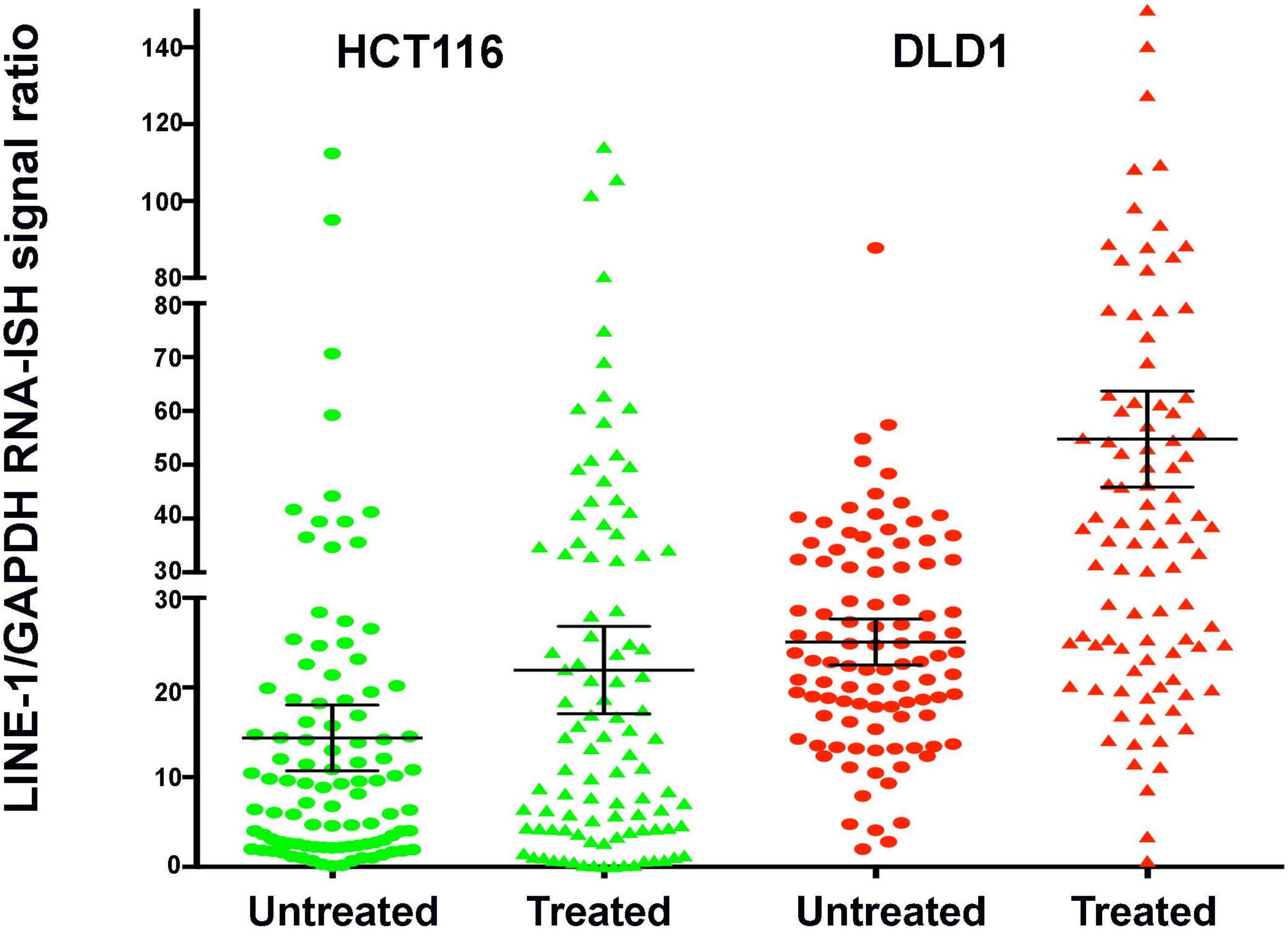
Statistics

All ordinal data (Flow ISH and FOXP3 IHC) p values were calculated based on a 2-tailed *t* test with GraphPad Prism 7 or Microsoft Excel software. All sets of data met normal distribution and displayed homogenous variance. Values of $P<0.05$ were considered significant. The sample size of each experiment is specified in figure legends.

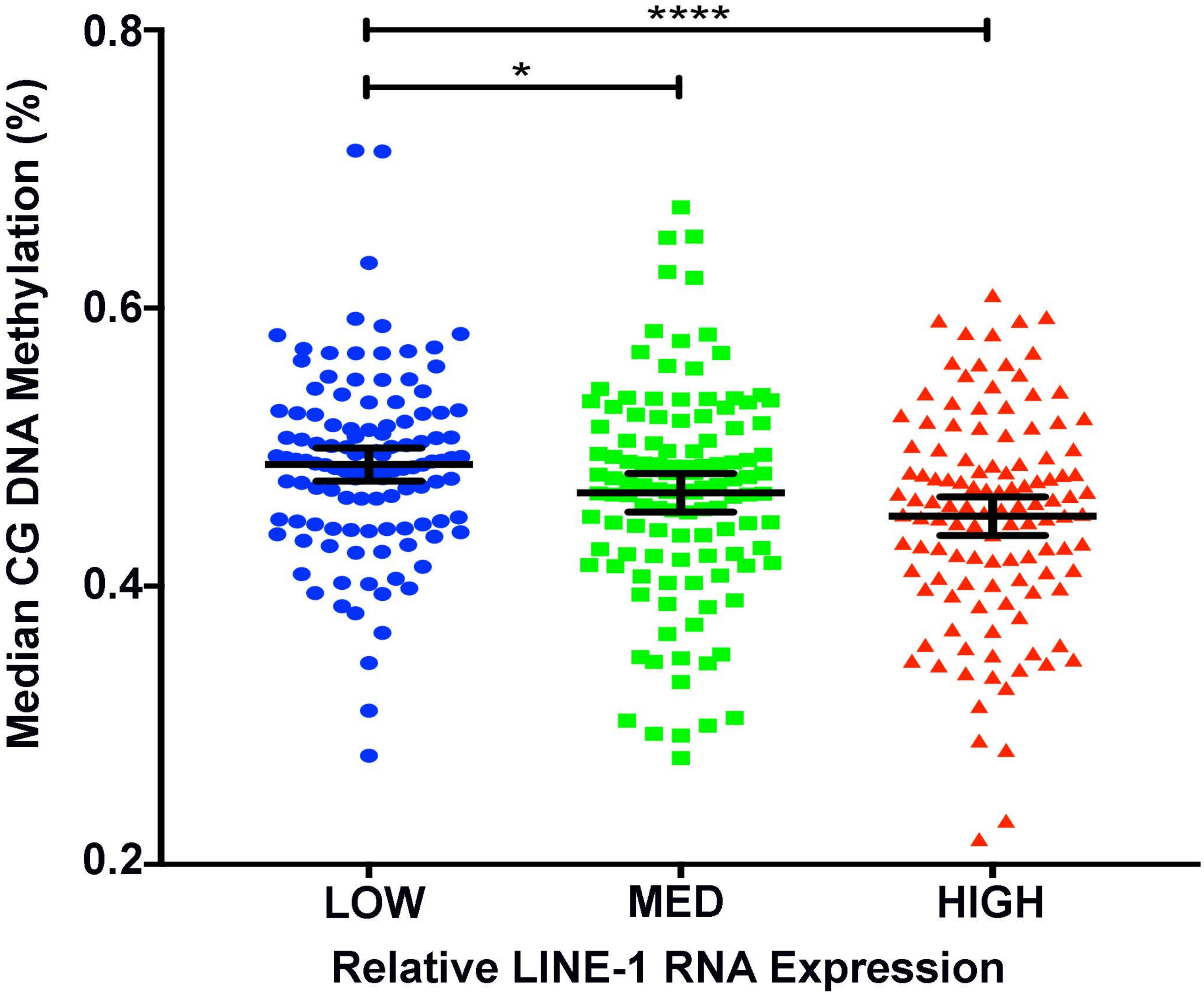
For clinical data, Chi square test was used to compare categorical data and independent t-test for comparing continuous variables. All analyses were performed using IBM SPSS- version 21. We used a two-sided significance level of 0.05 for all analyses.



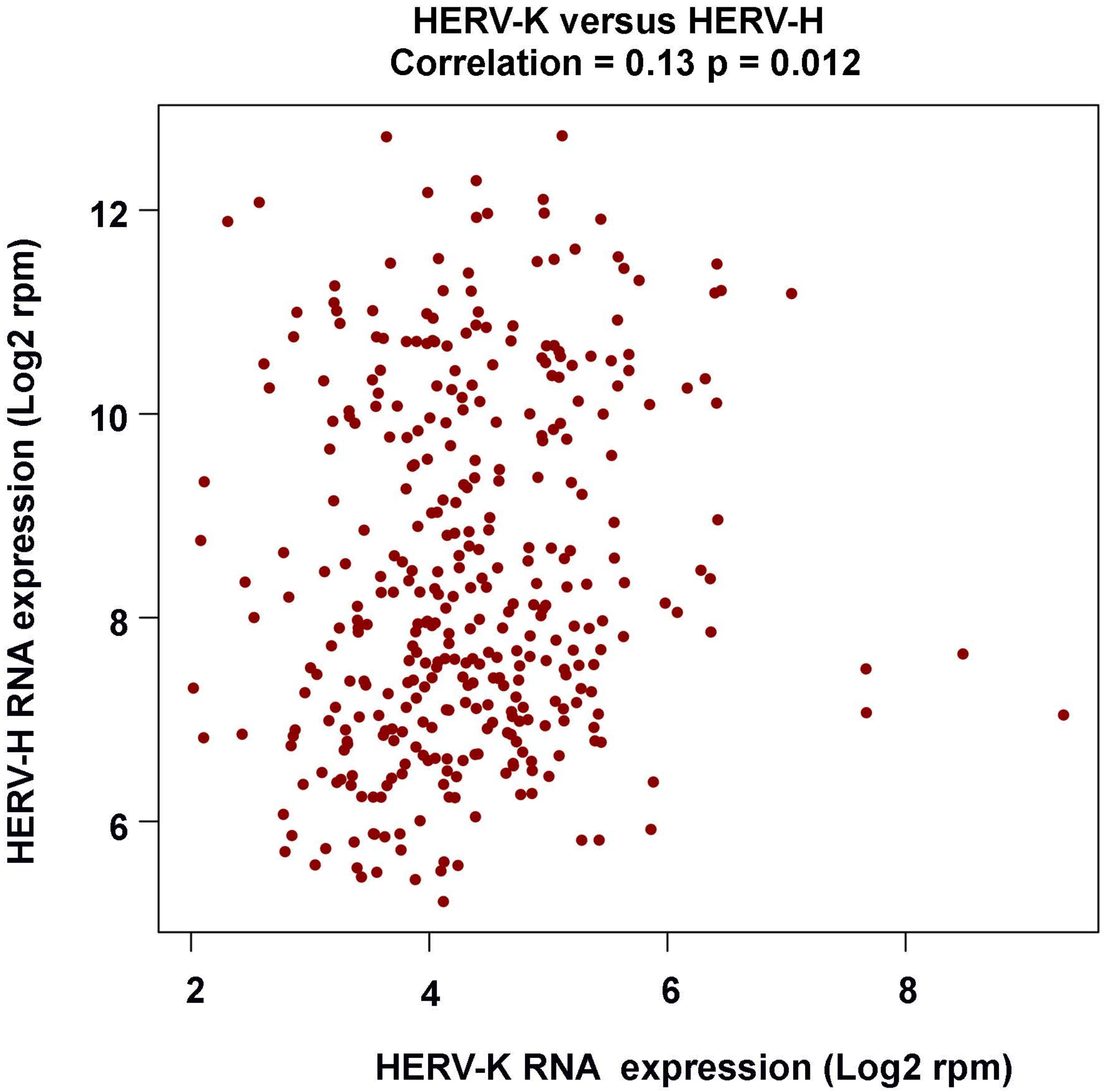
Supplemental Figure 1: LINE-1 and non-repeat CpG methylation fraction in relation to nearest GENCODE feature in cell lines with or without AZA treatment.



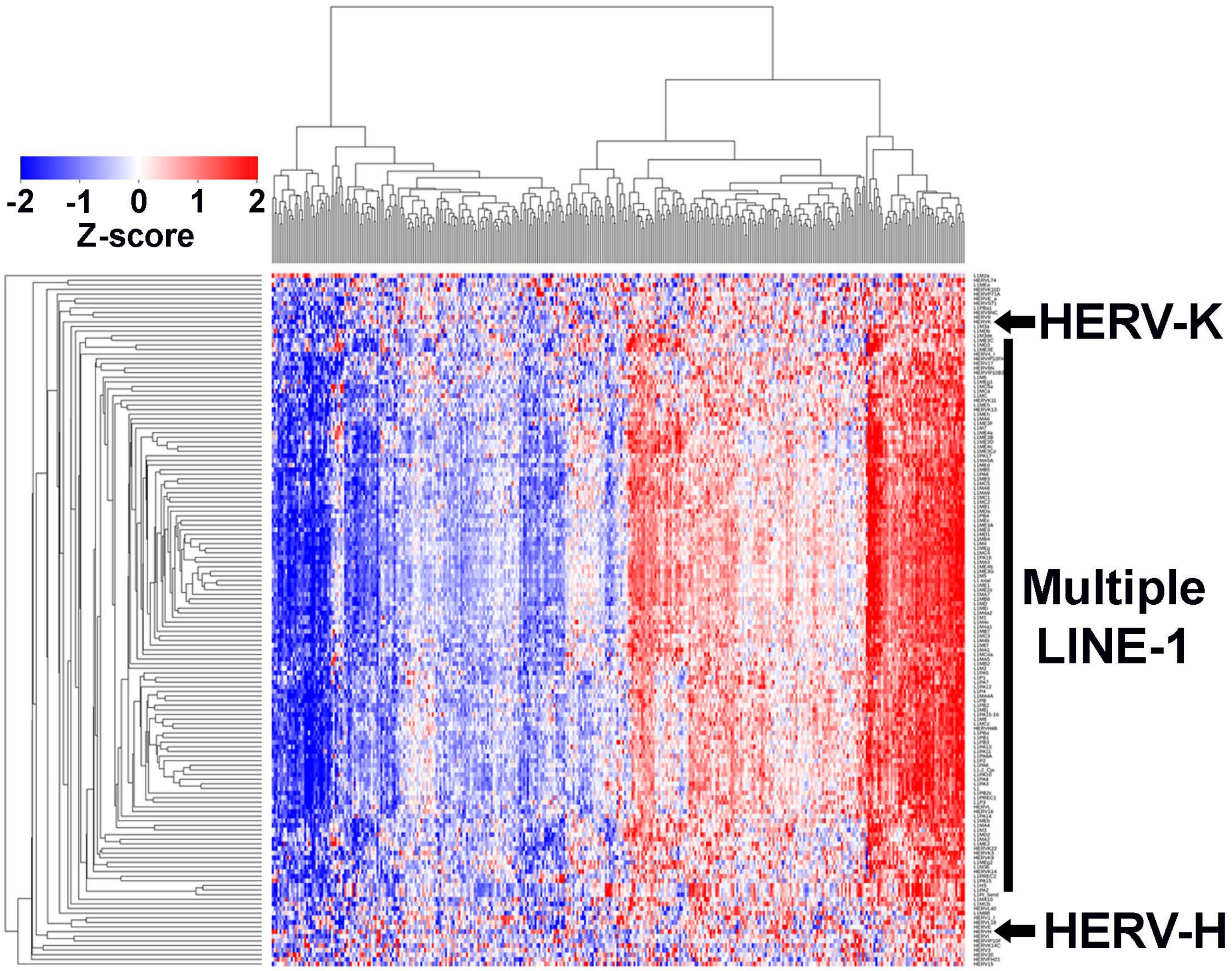
Supplemental Figure 2: LINE-1/GAPDH RNA-ISH quantitation of HCT116 and DLD1 cell lines untreated and treated with 1 μM AZA (Bar = mean and error bar = 95% CI). The experiment was repeated three times and at least 100 cells were counted per experiment.



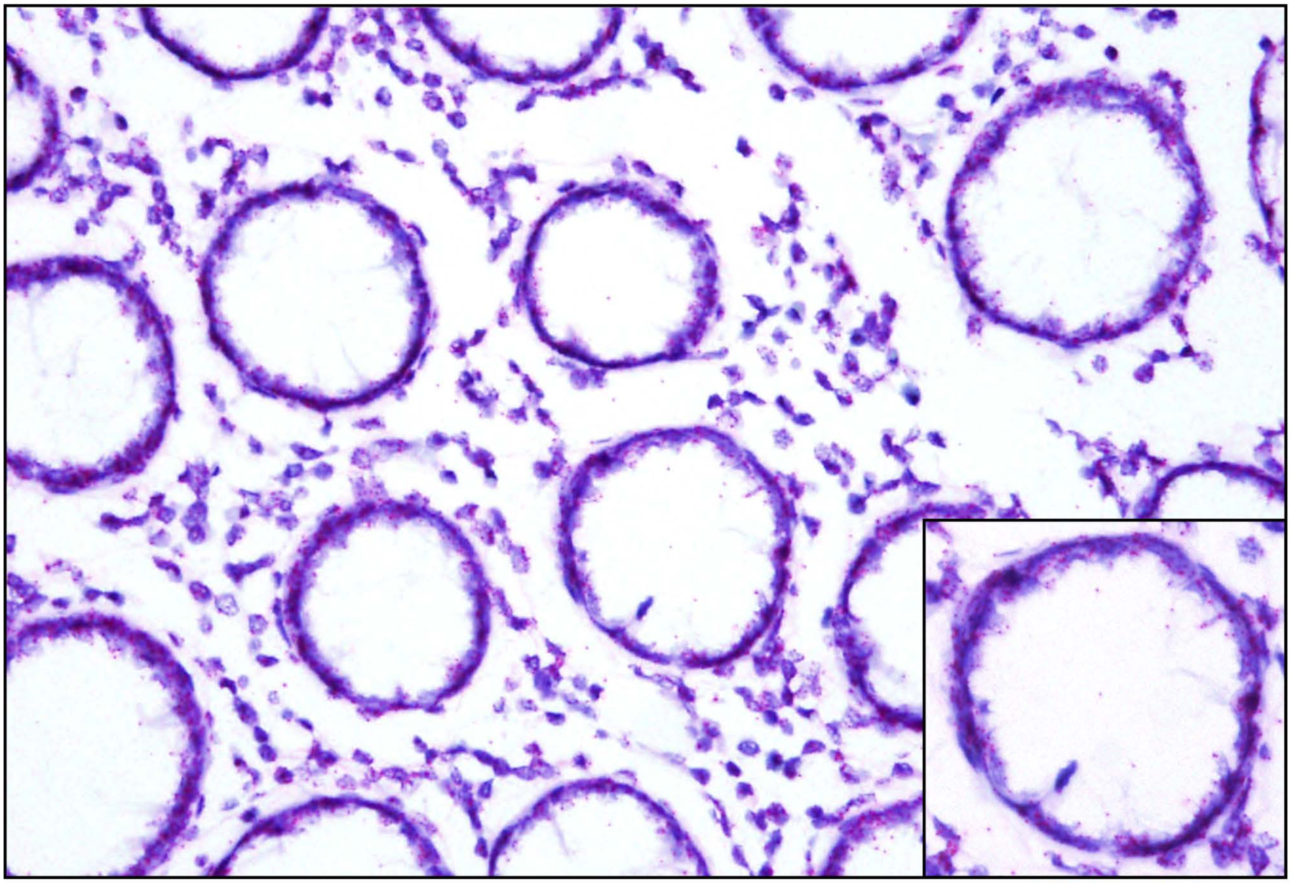
Supplemental Figure 3: .TCGA primary colon cancer tumors analysis with matched CpG methylation and mRNA-seq data showing significant differnces of median global CpG methylation (y-axis) with LINE-1 RNA expression. Samples separated in tertiles of LOW, MED, and HIGH LINE-1 RNA expression. (Bar = mean and error bar = 95% CI).P value calculated with unpaired 2-tailed t-test.



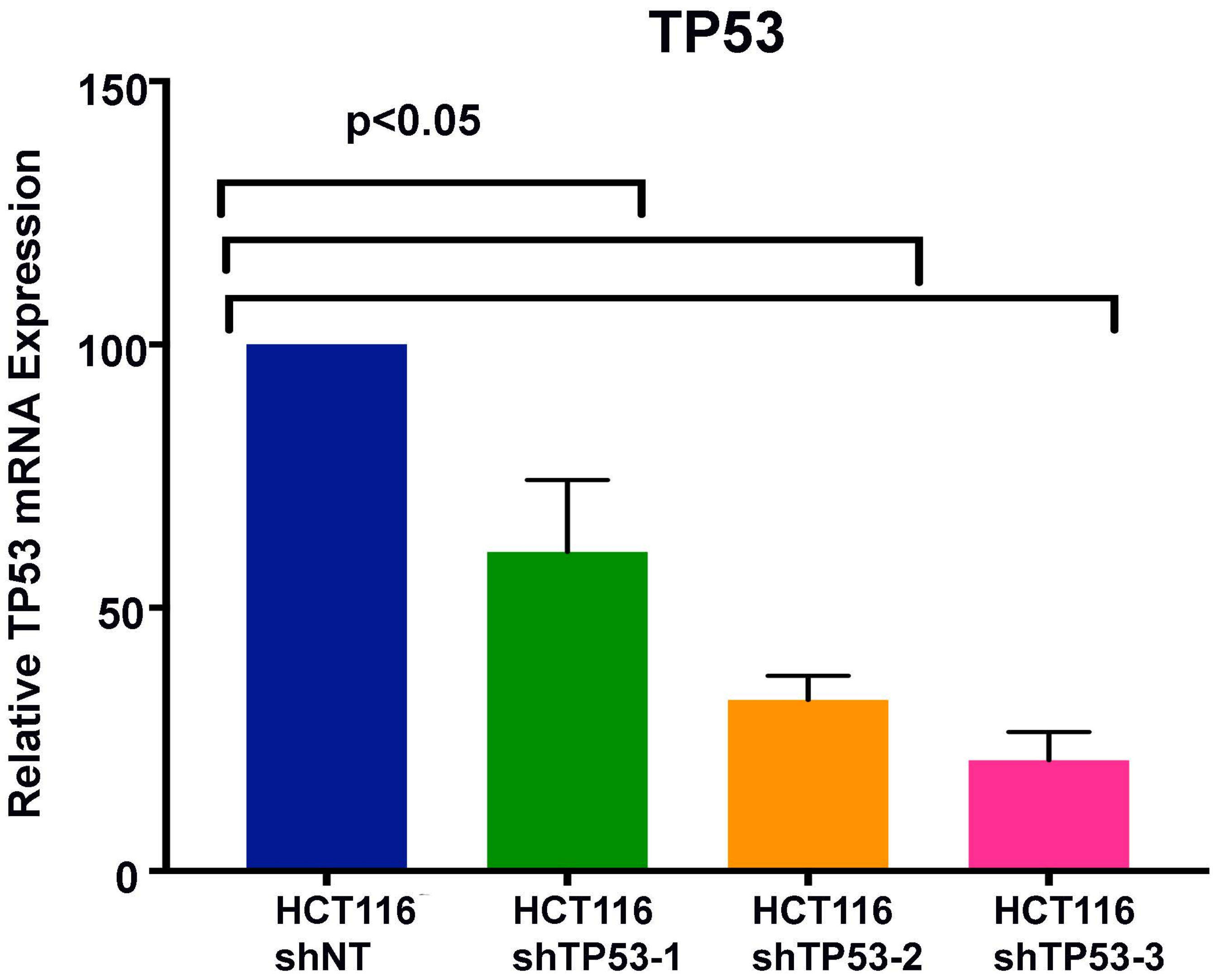
Supplemental Figure 4: Expression of HERV-H vs HERV-K in primary colon cancer RNA-seq from TCGA. Although there is a significant correlation between these HERV elements, it is much less significant than between HERV-K and LINE-1.This indicates that HERV-H and HERV-K expression is distinct from each other. Pearson correlation shown.



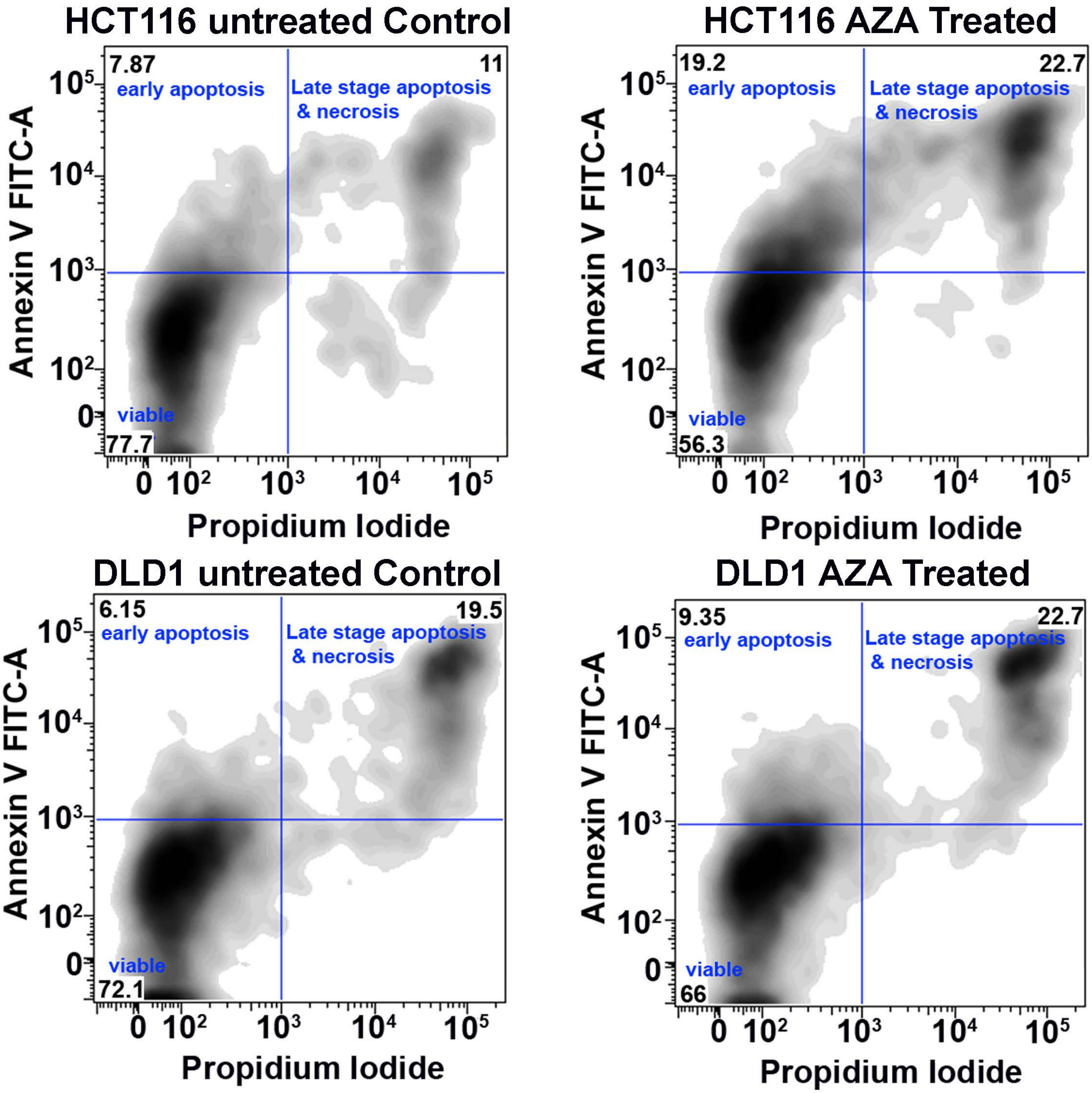
Supplemental Figure 5: Unsupervised clustering of HERV and LINE-1 subclasses across TCGA colon cancer mRNA-seq data demonstrating different co-expression patterns with the majority of LINE-1 subclasses and HERV-K behaving similarly across samples, while HERV-H and other HERV repeats cluster separately (bottom). Heatmap colors represent Z-score values as shown in legend (upper left).



Supplemental Figure 6: LINE-1 RNA-ISH on FFPE normal colon revealed LINE-1 RNA signal (red) consistent with the known baseline expression in normal tissues (200X magnification).Inset - 400X magnification of a single gland



Supplemental Figure 7: Suppression of *TP53* with shRNA confirmed with qRT PCR.Relative expession to HCT116 shNT control shown.(Bar = mean and error bar = SD). Experiment shown was done with n=3 technical replicates. P = Unpaired 2-tailed t test



Supplemental Figure 8: Flow cytometry plots of Annexin V/Propidium Iodide apoptosis quantitation in HCT116 and DLD1 cell lines untreated and treated with AZA at 1 μM total concentration. Experiment was repeated three times.