**EXPERIMENTAL MODEL AND SUBJECT DETAILS**

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**Poly(A)-seq Data Processing**

Poly(A)-seq libraries for six cell lines (HCT116, DKO, PrEC, LNCap, DU145, and DICER) were sequenced in one lane. The 100 bp single-end reads in the FASTQ files were demultiplexed. A maximum of 2 mismatches were allowed in the sample barcodes. 24 FASTQ files were generated. Multiplexing barcodes and random hexamer sequences were removed and used for sample identification and PCR duplicate read removal respectively. Reads with ≥9 ‘As’ in a 10 bp window were used to determine poly(A) sites.

After removing barcodes and trimming the enriched ‘As’ at the 3′ end, the reads were aligned to hg19 using Bowtie2 (version 2.3.3.1) in a single-end read mode. No additional parameters were applied.

Reads with MAPQ ≥20 were retained. Reads mapped to either PhiX control or decoy of chromosome contigs were excluded. Reads with identical random hexamer sequence were considered PCR duplicates. All duplicated reads were collapsed. Reads with ≥7 ‘As’ in an 8 bp window in the reference genome were excluded from further analysis.

**Recruiting Putative Poly(A) sites**

Based on the 24 poly(A)-seq BAM files after the quality control, we determined poly(A) site (pAS) positions, or equivalently named ‘processing region’ (PR). All 3′ end mapping positions were counted. Positions within 10 bp of each other were grouped together and assigned to a single PR. A unique poly(A) ID was assigned to any group of 3′ ends with ≥10 reads

pASs were assigned to genes based on Ensembl 87 (e87) gene models on GRCh37. Only genes annotated in assembled chromosomes (chr1-22, chrX, chrY, and chrM) were considered. We retained only the genes matched by Ensembl annotation curated by UCSC genome database(cite Bhasin and Ting 2016) [Table S2]. Biotypes of genes annotated by Ensembl include "protein\_coding", "processed\_transcript", "lincRNA", "antisense", "polymorphic\_pseudogene", "sense\_overlapping", "sense\_intronic", and "3prime\_overlapping\_ncrna". We also considered *de novo* pASs that were identified in intronic regions or up to 5,000 bp beyond the known 3′UTR.

**Differential pAS Usage Testing**

The read counts at each PR were normalized across all replicates. Significance testing was conducted with DEXSeq 1.24.4(cite Anders, Reyes and Huber, 2012 and Reyes et, 2013) using the default parameter. P-value was adjusted by Benjamini-Hochberg false discovery rate (FDR=0.05). PR sites were only considered when 1) the difference between two PR usage fractions was ≥ 0.1, 2) the FDR adjusted P-value for differential usage testing was < 0.0001, 3) the log2 fold change of PR usage fraction in absolute value was > 1.5, and 4) at least one group sample’s mean PR usage fraction was ≥ 0.05. (Table1 and Table S3).

**Transcription factor binding site profiles between putative pASs**

Using 412 genes that preferentially use the proximal pAS in DKO, we interrogated the genomic regions between shifting pASs in HCT116 and DKO to identify enriched binding of ENCODE transcription factors. We randomly sampled 10,000 pairs of PRs in any gene having more than one PR without replacement per each APA gene and compared those with the 412 APA regions. ENCODE(cite Encode) regulatory of transcription factor binding sites (including ChIP-Seq binding sites for colon cancer cell lines in a previous study(cite Maurano et al, 2015) are annotated each testing region. Both Z scores and binding frequency were calculated (Figure 1C and Table S1F).

**RNA-Seq read Data Processing**

HCT116 and DKO RNA-seq FASTQ files contain 101 bp paired-end reads. The FASTQ files were converted to hg19 BAM using the two-pass mode of STAR (v2.5.4b) with the default parameter for paired-end reads. For gene expression quantification, the BAM files were converted to a read count matrix by RSubread::featureCounts()(cite Liao, Smyth and Shi, 2019) to quantify based on the e87 gene model with the options: ountMultiMappingReads=FALSE, allowMultiOverlap=FALSE, isPairedEnd=T, strandSpecific=2, minMQS=5, ignoreDup=FALSE**,** and fraction=FALSE. Log2-RPKM (read per kilobase per million mapped reads) values adjusted by gene length were computed. Gene expression of 17 known cleavage factors (*CPSF1, CPSF2, CPSF3, CPSF3L, CPSF4, CPSF6, CPSF7, NUDT21, CSTF1, CSTF2, CSTF2T, CSTF3, FIP1L1, CLP1, PCF11, WDR33*, and *SYMPK*) was compared between HCT116 and DKO (Figure S1D).

**ChIP-Seq Data Processing**

ChIP-Seq data was generated for 5 proteins (CTCF, SMC1, RAD21, Pol2Ser2, and Pol2Ser5) and H3K27ac in HCT116 and DKO. One input control was included for each cell line. The FASTQ files contain 75 bp single-end reads. The primary analysis, including alignments and peak callings, was done by ENCODE Transcription Factor and Histone ChIP-Seq processing pipeline(cite Landt et al, 2012) (https://github.com/kundajelab/chipseq\_pipeline). The following parameters were used for the software with the default parameters, suggested for unreplicated experiments:

"--se --species hg19 --peak-caller macs2 --blacklist ${blacklist\_file} --fastq1 ${chipseq\_fastq\_file} --ctl\_fastq1 ${input\_control\_fastq\_file} --type $type".

We specified "TF" to "--type" for all transcript factor samples whereas "histone" for the samples tagged by H3K27Ac, Pol2Ser2, and Pol2Ser5. The ENCODE pipeline collapsed duplicated reads and relied on uniquely mapped reads (e.g., MAPQ score 37). All peaks passed reproducibility tests for peak detection statistics (e.g., overlapping peaks and irreproducible discovery rate).

For MBD-seq data, the FASTQ files were converted to BAM files via Bowtie2 (v.2.3.4.1) with default parameters for a single-end read. All mapped reads were utilized for further analysis.

**Differential ChIP-seq binding site analysis**

Peaks of binding sites for the ChIP-seq data are available from the ENCODE pipeline output. The peak calls were done by MACS2(cite Zhang et al, 2008). For each protein, peaks (\*.narrowPeak) between HCT116 and DKO were compared directly via MANorm v1.1.4(cite Shao et al, 2012) (<https://github.com/shao-lab/MAnorm>) under the program option: manorm --p2 HCT116.narrowPeak --p1 DKO.narrowPeak --r2 HCT116.tagAlign.bed --r1 DKO.tagAlign.bed -w $width -m 1 -p 0.01 --name2 HCT116 --name1 DKO. A window length of 500 bp (CTCF, SMC1, and RAD21) or 1,000 bp (H3K27Ac, Pol2Ser5, and Pol2Ser2) was used to calculate the read density. Both fold change (M-Value) and normalized binding depth are reported. All the regions reported were kept for further analysis.

Similarly, MBD-seq peaks were called by MACS2 with the run option: “-f BAM -g hs -B --broad”. Peaks from each sample group (i.e., HCT116 and DKO) were simply concatenated. The mean depth and the genomic region of the peaks were used for a differential binding analysis using DESeq2 (v1.18.1) under default parameters. Differential binding sites with a log2 fold change ≥1.0 and mean depth >3 before normalization in either sample group were kept for further analysis.

**Integrative clustering analysis on multiple differential binding sites**

We compiled a total of 8 genomic tracks with binding sites called by either 6 differential ChIP-seq MANorm or 2 MBD-seq DESeq2 analysis. Using the 412 candidate genes that preferentially use the proximal pAS in DKO, we focused on genomic regions where at least one binding site overlaps the region between the two shifting PRs +/- 500 bp. Firstly, we divided the binding sites into smaller segments such that the mean depth is equivalent within each ChIP-seq and thus the differential binding sites can be summarized with either log2 fold change or mean read depth in 8 dimensional feature vector. .

Secondly, we filtered out spurious short segment regions ≤10 bp, leaving 8,013 distinct genomic regions for the clustering analysis. A matrix of a distinct genomic segment by ChIP-seq tags are filled with a normalized log2 fold change (M-Value from MANorm output) for CTCF, SMC1, RAD21, H3K27Ac, Pol2Ser2, and Pol2Ser5 and raw mean depths from two cell lines MBD-seq, respectively. A feature count matrix represents 8 features for ChIP-seq/MBD-seq binding sites by 8,013 sample objects for genomic segments.

Thirdly, a standard consensus clustering method(cite Monti et al, 2003) was applied based on a non-negative matrix factorization (NMF) method(cite Berry et al, 2007). NMF was run on the feature count matrix 500 times and 80% of the clustering results were randomly sampled for consensus membership analysis. The job was repeated with the number of clusters K=2 to K=15. K=9 was found to be optimal, based on the delta value curves and the consistency of cluster membership. A unique cluster ID was assigned to each genomic segment in the feature count matrix. The distance between two clusters was measured as the average distance between each point in one cluster to every point in the other cluster.

Fourthly, from ChIP-seq BAM files, we retrieved mean read depths (e.g., read depth normalized by MANorm and raw mean depth for MBD-seq) corresponding to a coordinate of each genomic segment in the feature count matrix for visualization. Finally, we split the mean depths into 63 groups from 9 clusters by 6 proteins (i.e., CTCF, SMC1, RAD21, Pol2Ser2, Pol2Ser5, and MBD) and H3K27Ac. Subsequently, each group is split into two subgroups; one for HCT116 and the other for DKO. An R function ggplot2::geom\_density() is used to visualize two density distributions of the mean depth in log2 value for each group. The difference between two distributions are quantified by Mann–Whitney-Wilcoxon nonparametric test(cite reference below) since the distribution is not necessary to follow normal distribution (Table S5).

**The Cancer Genome Atlas (TCGA) Dataset**

As of July 2018, we downloaded matched mRNA-seq BAM files and Infinium HumanMethylation 450k BeadChip array data for 5,284 patients across 11 cancer types, Bladder Urothelial Carcinoma (BLCA), Breast invasive carcinoma (BRCA), Colon adenocarcinoma (COAD), Esophageal carcinoma (ESCA), Kidney renal clear cell carcinoma (KIRC), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Pancreatic adenocarcinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), and Uterine Corpus Endometrial Carcinoma (UCEC) in TCGA database. An average of 480 patient samples per cancer type was analyzed.

**Predicting poly(A) usage in TCGA RNA-seq data based on our colon cancer cell line model**

Poly(A) isoform usage at the 546 APA candidate genes was inferred from the RNA-Seq BAM files. We manually defined at least one pair of common (shared by both proximal and distal isoforms) and unique (unique to either the proximal or distal isoform) genomic regions for each of the 546 APA candidate genes (Table S3). The mean depth ratio between common and unique genomic regions was used to determine relative pAS usage.

Genes were sorted by the FPKM-UQ (fragments per kilobase of transcript per million mapped reads upper quartile)(cite Shahriyari, 2017) values available from the National Cancer Institute Genomic Data Commons. and considered the genes confidently expressed. The top 8,000 genes underwent further analysis. Cases in which the mean depth in the unique region is greater than the common region were excluded from further analysis.

**Normalization of DNA methylation β-values**

Among 485,000 CpG sites reported in the methylation array, approximately 400,000 probes have valid β (i.e., methylation fraction) values. The β mixture quantile dilation method(cite Teschendorff et al, 2012) was used to obtain a normalized methylation value at each CpG site.

**Correlation between predicted poly(A) usage ratio and CG methylation rate**

A normalized methylation rate was defined at each CG site. Predicted pAS usage was defined at a gene containing the CG site +/-5,000 bp. Samples lacking either a β value or pAS usage data were excluded. We generated a scatter plot between methylation rateand pAS usage for each CG site. Pearson correlation (adjusted by Benjamini-Hochberg FDR 𝛼=0.05) was computed to quantify dependence or relationship between two variables.

reference above: [ref:Myles Hollander and Douglas A. Wolfe (1973). Nonparametric Statistical Methods. New York: John Wiley & Sons. Pages 27--33 (one-sample), 68--75 (two-sample). Or second edition (1999)]