**EXPERIMENTAL MODEL AND SUBJECT DETAILS**

**QUANTIFICATION AND STATISTICAL ANALYSIS**

**PolyA-seq Data Processing**

Six cell lines HCT116, DKO, PrEC, LNCap, DU145, and DICER were prepared for a polyA-sequencing library. Each cell lines were split into 4 replicates and sequenced in more than one flowcells. The 100-bp single-end reads in the FASTQ files are demultiplexed, allowing a maximum of 2 mismatches in the sample barcodes. Total of 24 FASTQ files was generated. Both multiplexing barcode and read tags were stripped off. The former was used for sample identification and the latter was appended to a header of each read in the FASTQ file for a PCR duplicate read removal. Reads with highly enriched A’s from the 3’ end (the number of 'A's is >=9 in a 10 bp-long window) are considered for determining polyA sites.

After removing barcodes and trimming 3’end of the reads, we aligned the reads to hg19 with Bowtie2 (version 2.3.3.1) in a single-end read mode. No additional parameters were applied.

Reads aligned with higher confidence (i.e., MAPQ>=20) were only retained. Any read mapped to either PhiX control or decoy of chromosome contigs are excluded. Taking into account both the read tag mentioned in the pre-processing step and 5' end alignment position, we recognized whether the reads are PCR duplicated or not. All duplicated reads were collapsed. Some reads can be potentially originated from a highly-enriched A’s region in the reference genome sequence. Therefore, we examined the mapping coordinate of each read 3’ end. If the number of 'A' in the reference genome sequence is >=7 out of 8 bp window, we exclude the reads from further analysis.

**Recruiting Putative PolyA Sites**

Based on the 24 polyA-Seq BAM files after the quality control, we determined poly(A) site positions, or equivalently named ‘processing region’ (PR). All 3’ end mapping positions are counted. If two positions are within 10 bp apart from each other), we group them together and both are contributed from the same isoform. A unique poly(A) ID is assigned to a group of 3’ ends when it is supported by more than 10 reads.

Gene assignments to poly(A) sites were based on the Ensembl build 87 gene models on GRCh37 and genes annotated in assembled chromosomes (chr1-22, chrX, chrY, and chrM) are only considered. We retained only the genes matched by Ensembl annotation curated by UCSC genome database [<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4937336/> and Supplementary Table 2]. Biotypes of genes annotated by Ensembl include "protein\_coding", "processed\_transcript", "lincRNA", "antisense", "polymorphic\_pseudogene", "sense\_overlapping", "sense\_intronic", and "3prime\_overlapping\_ncrna". Not only known poly(A) sites but also we kept a *de novo* poly(A) site appeared in either intronic regions or beyond (up to 5,000 bp) the known 3’ UTR end position based on the existing gene model (Supplementary Table 2).

**Differential PolyA Site Usage Testing**

Differential poly(A)-site usages between two cell lines HCT116 and DKO are tested. The read counts at each PR are normalized across all replicates. A significance testing is conducted with DEXSeq 1.24.4 with the default parameter. P-value is adjusted by Benjamini-Hochberg false discovery rate at 0.05. In order to obtain high confidence differential poly(A) sites, multiple quality controls are applied. PR sites were only considered when the difference between two PR usage fractions is ≥ 0.1 and the adjusted P-value is < 0.0001. Also, we retained only PR’s where the log2 fold change of PR usage fraction in absolute value is > 1.5 and at least one group sample’s mean PR usage fraction is ≥ 0.05. A total of 718 significant PRs are observed across 546 genes (Table1 and Supplementary Table 3).

**Genomic function enrichment in regions between putative polyA sites**

Among 546 genes discovered in the differential testing, we focused on 412 APA genes where the pA site is shifted toward the 5’-upstream. Shortly, let us denote the genes by ‘proximal’ APA genes. We interrogated the genomic regions of shifted PR from HCT116 to DKO at each proximal APA gene to see if it is often associated with transcription factors in common. We randomly chose a pair of PRs in any gene having more than one PR under 10,000 permutations and compared those with the 412 APA regions. ENCODE regulatory of transcription factor binding sites (including ChIP-Seq binding sites for colon cancer cell lines in a previous study [<https://www.ncbi.nlm.nih.gov/pubmed/26257180>]) are annotated each testing region. Both Z scores and a frequency of binding proteins are calculated (Figure 1C and Supplementary table 1f), showing that the APA regions have a higher binding affinity of a certain protein than the others.

**RNA-Seq read Data Processing**

HCT116 and DKO RNA-seq FASTQ files contain 101 bp paired-end reads. The FASTQ files are converted to hg19 BAM files by Bowtie2 (version 2.3.3.1). Default parameter for paired-end reads was used except one additional parameter ‘--very-sensitive-local’. Then, for the gene expression quantification, the BAM files are converted to a read count matrix by RSubread::featureCounts() to quantify based on the Ensembl build 87 gene model with the options: ountMultiMappingReads=FALSE, allowMultiOverlap=FALSE, and fraction=FALSE. Log2-RPKM (read per kilobase per million mapped reads) values adjusted by gene length are computed. In total, 17 known cleavage factors (*CPSF1, CPSF2, CPSF3, CPSF3L, CPSF4, CPSF6, CPSF7, NUDT21, CSTF1, CSTF2, CSTF2T, CSTF3, FIP1L1, CLP1, PCF11, WDR33*, and *SYMPK*) are chosen to compare the gene expression levels between HCT116 and DKO (Supplementary Figure 1D).

**ChIP-Seq Data Processing**

Twelve ChIP-seq data (two cell lines, HCT116 and DKO, by six proteins, CTCF, SMC1, RAD21, H3K27Ac, Pol2Ser2, and Pol2Ser5) are generated, along with two input controls one for each cell line. The FASTQ files contains 75 bp single-end reads. The primary analysis including alignments and peak callings is done by ENCODE Transcription Factor and Histone ChIP-Seq processing pipeline (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 performs a quality control by collapsing duplicated reads and relying on uniquely mapped reads (e.g., MAPQ score 37). Reproducibility tests of peak detection statistics (e.g., overlapping peaks and irreproducible discovery rate) are all passed. No further quality control is performed.

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

**Differential ChIP-seq binding site analysis**

Peaks of binding sites for the twelve ChIP-seq data are available from the ENCODE pipeline output. The peak calls are done by MACS2. Between two cell lines, we conduct a differential binding analysis. For each protein, peaks (\*.narrowPeak) between two binding peaks are compared directly via MANorm v1.1.4 (<https://github.com/shao-lab/MAnorm>, <https://genomebiology.biomedcentral.com/track/pdf/10.1186/gb-2012-13-3-r16>) 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. In the option (-w) for a window length to calculate the read density, 500 bp is set for CTCF, SMC1, and RAD21 but 1,000 bp is set for H3K27Ac, Pol2Ser5, and Pol2Ser2.

MANorm takes narrow peaks and reads coordinates supporting the peaks (e.g., summit region surrounding by 500 bp extension for a transcription factor ChIP-seq) in BED file format. Then, MAnorm collects peaks commonly bound between two samples. Based on an assumption that the fold change (M) is independent of the average read density (A), MAnorm generates an MA linear transformation regression fitting to the plot data. Consequently, the same transformation can be applied to all peaks in the two samples (i.e., peaks shared by both and private members of peaks). Finally, a P-value associated with each peak was calculated to quantify the significance of differential binding at this locus using a Bayesian model developed by Audic and Claverie [<https://genome.cshlp.org/content/7/10/986.long>]. All the regions reported are kept for further analysis.

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

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

We compiled the total of 8 genomic tracks with binding sites called by either 6 differential ChIP-seq MANorm or 2 MBD-seq DESeq2 analysis. But, we paid attention to the genomic region where at least one binding site is overlapped by the 412 proximal APA shifted PR regions which we define as a genomic region between two PR locations flanked by 500 bp and the polyA usage is significantly different and the pA site in DKO is shifted toward to 5’ upstream. Then, we considered to generate a feature by count matrix for a clustering analysis. At first, we split the overlapped genomic regions into smaller segments such that each smaller segment can be summarized by 8 figures of either log2 fold change or mean read depth in a vector form and a genomic coordinate in bp (chromosome:start ~end).

Secondly, we filtered out spurious short segment region of 10 bp or shorter (equivalently, 95% of the regions are retained). A total of 8,013 distinct genomic regions become available for 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 tags and 8,013 sample objects for genomic segments, respectively.

Thirdly, a non-negative matrix factorization (NMF) method clusters the feature count matrix. A general consensus clustering runs from the number of clusters K=2 to K=15 in 80% of resampling without replacement from the results of 500 NMF clusterings. The K=9 was found optimal, based on the delta value curves and the consistency of cluster membership. The distance between two clusters are 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 each segment genomic coordinates for visualization. Total of 63 read depth samples (9 clusters x 7 proteins) is grouped by a cluster membership and a protein. Subsequently, each group is split into two subgroups; one for HCT116 and the other for DKO. An R function geom\_density from ggplot2 is used to visualize two density distributions of the mean depth in log2 value for each group. In order to quantify the difference between two distributions, we measured the Kullback-Leibler (K-L) Divergence in bits and performed χ2 test (Supplementary Table file 5).

**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, BLCA, BRCA, COAD, ESCA, KIRC, LUAD, LUSC, PRAD, SKCM, STAD, and UCEC in TCGA database. An average of 480 patient samples per cancer type was analyzed.

**Predicting polyA usage in TCGA RNA-seq data based on our colon cancer cell line model**

A 3’ polyA-seq is not available in the TCGA database. Thus, the polyA isoform usage at each 546 APA gene was inferred from the RNA-Seq BAM files. Under the assumption that when two pA sites exist, the read depth in a region shared by two isoforms is higher than a region uniquely transcribed by either of one, we manually define at least one pair of genomic regions for each of the 546 APA gene observed from the polyA-seq data from our colon cancer cell line model (Supplementary Table 3). Incorporating with a gene model (e.g., Ensembl GRCh37.87) and the polyA site positions appeared in the 546 APA genes, we prepared a total of 651 pairs of genomic regions (Supplementary Table X) in a proximity of the pA sites in a way that, from matched RNA-seq BAM files to our polyA-seq data, the mean depth ratio between the pair of genomic regions for testing is proportional to the polyA usage rate. Each pair has two genomic regions: 1) 5’ upstream regions where both proximal and distal isoform commonly share and 2) 3’ downstream region solely used by only one polyA isoform, respectively.

For RNA-seq quality control, we sorted genes by the FPKM-UQ (Fragments Per Kilobase of transcript per Million mapped reads upper quartile) value already available from NCI (National Cancer Institute) Genomic Data Commons (GDC) and considered the genes confidently expressed. We retained the genes whose FPKM-UQ value is within 8,000 from the top for further analysis. Relative polyA usage rate was directly estimated from the ratio of the read depth between the pair of the genomic region in TCGA mRNA-seq BAM files. Cases in which the mean depth in the downstream (unique) region is greater than the one in common were excluded from further analysis.

**Normalization of DNA methylation beta-values**

Among average 485,000 CpG sites reported in the methylation array, approximately, 400,000 probes have valid beta (i.e., methylation fraction) values. In order to correct the technical issues caused by the Illumina type 2 probe, we used the beta mixture quantile dilation (BMIQ) method [<https://academic.oup.com/bioinformatics/article/29/2/189/204142>] to obtain a normalized methylation value at each CpG site. The strategy involves the application of a three-state beta-mixture model to assign probes to methylation states, subsequent transformation of probabilities into quantiles and finally a methylation-dependent dilation transformation to preserve the monotonicity and continuity of the data.

**Correlation between predicted polyA usage ratio and CG methylation rate**

The samples where both beta values and polyA usage rate are available are considered. A normalized methylation rate is defined at each CG site whereas a predicted polyA is defined at a gene (or pair of genomic regions defined in the Section, “**Predicting polyA usage in TCGA RNA-seq data based on our colon cancer cell line model”**) which the CG site is belonged to. Over the multiple TCGA samples, we generated a scatter plot between methylation rate on the x-axis and a polyA usage rate on the y-axis. Pearson correlation (adjusted by Benjamini-Hochberg FDR 𝛼=0.05) is computed to quantify dependence or relationship between two variables.

[Reference]

[1]<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3546795/>

[2]<http://isaim2018.cs.virginia.edu/papers/ISAIM2018_Deebani_Kachouie.pdf>