PNAS
www.pnas.org

# Supplementary Information for
## Comprehensive mapping of alternative polyadenylation site usage and its dynamics at single cell resolution

Junliang Wang[1,2#], Wei Chen[3#], Wenjun Yue[2], Wenhong Hou[2], Feng Rao[2], Hanbing Zhong[2], Yuanming Qi[1*], Ni Hong[2*], Ting Ni[3*], Wenfei Jin[2*]

[1] School of Life Sciences, Zhengzhou University, Zhengzhou 450001, Henan, China;

[2] Shenzhen Key Laboratory of Gene Regulation and Systems Biology, School of Life Sciences, Southern University of Science and Technology, Shenzhen 518055, China

[3] State Key Laboratory of Genetic Engineering & MOE Key Laboratory of Contemporary Anthropology, Collaborative Innovation Center of Genetics and Development, School of Life Sciences and Shanghai Cancer Center, Fudan University, Shanghai 200438, China.

[#]Equal contribution

*Corresponding author: jinwf@sustech.edu.cn (W.J.), tingni@fudan.edu.cn (T.N.), hongn@mail.sustech.edu.cn (N.H.), qym@zzu.edu.cn (Y.Q.).

**This PDF file includes:**

Supplementary text
Figures S1 to S11
Legends for Datasets S1 to S6

**Other supplementary materials for this manuscript include the following:**

Datasets S1 to S6

**Supplementary Information Text**

## Materials and Methods

**Cell culture and cell synchronization by double thymidine block.** MDA-MB-468, HEK293T and HeLa cell lines were obtained from American Type Culture Collection (ATCC) and cultured in DMEM (Gibco) with 10% FBS (Gibco) and 1% Penstrep in 5% $CO_2$ at 37°C. MEF was prepared following our recent study(1). Synchronization of HeLa cells by double thymidine block was conducted following protocol in *Ma and Poon* (2). In Brief, Hela cells were separated and cultured in two 25 $cm^2$ flasks, for control and cell synchronization, respectively. To synchronize the cells, thymidine (2 mM, Sigma) was added to a flask and the cells were incubated for another 18h, washed 3 times with PBS, and released into thymidine-free complete medium for 9h. Thymidine (2 mM) was then added again and cultured for an additional 15 h. Hela cells in control flask were cultivated in complete medium without adding thymidine.

**scPolyA-seq**. We used Fluidigm C1™ Single-Cell Auto Prep System (Fluidigm, South San Francisco, CA, USA) for scPolyA-seq library preparation. Library construction followed Fluidigm manual and protocol (3), except these steps adapted to scPolyA-seq. In brief, MDA-MB-468, MEF and HeLa cells were collected and counted in Cellometer Mini (Nexcelom Bioscience). Then cells were mixed and re-suspended to 400 cells/μL cell suspensions. Fluidigm high throughput integrated fluidic circuit (HT IFC) with 800 wells was used for cell capture. Cell suspensions of MEF (mixed with 10% synchronized HeLa cells) and MDA-MB-468 (mixed with 10% HeLa cells) were loaded into two independent inlets of IFC (Fig. 1*B*), up to 400 cells could be captured for each cell suspension. HT IFC was checked under white field microscope to record the cell status in each well: empty, single, debris or dual (*SI Appendix*, Fig. S1*A*).

The C1 Single-Cell mRNA Seq HT Reagent Kit was used for cDNA synthesis. We started scPolyA-seq library preparation by cell lysis. Then we added primers containing polyT and cell barcode for labeling the 3' end of transcripts. Full-length cDNA was synthesized by template-switching reverse transcription, following by amplification, and tagmentation with Tn5 transposases. Nextera XT DNA Library Preparation Kit (Ilumina) was applied to construct scPolyA-seq library. DNA fragments at the 3' end of the cDNA were captured by targeted PCR, and sequencing indexes were added for amplification. The quality of library was examined by Qubit 3.0 Fluorometer and Agilent 2100 Bioanalyzer. Libraries separated by columns were pooled together and sequenced on Illumina HiSeq2500 to obtain 150-bp paired-end reads.

To remove potential outliers, we calculated the median absolute deviation (MAD) of numbers of mapped reads and expressed genes of all cells. The cells were filter out if the read counts or number of expressed genes >medians+3×MAD or <medians-3×MAD. Cells were removed if proportion of their mitochondrial RNA were outlier.

**Identification of cell types and cell populations.** MEF and synchronized HeLa need to be separated by computational approach since they were loaded into one inlet of IFC. To separate synchronized HeLa from MEF, reads of each cell were mapped to hg19_mm10 mega reference genome using STAR (4). The numbers of unique mapped reads on human genome and mouse genome were counted, respectively. Scatter plot showed that the reads from one cell either dominantly mapped to human genome or mouse genome, with few exceptions (*SI Appendix*, Fig. S1*D*). A cell was assigned as synchronized HeLa or MEF if reads from human genome or mouse genome contributed >80% total reads.

The cells from the other inlet of IFC are a mixture of MDA-MB-468 and unsynchronized Hela. These cells and synchronized Hela were combined and projected on uniform manifold approximation and projection (UMAP) by Seurat 3.1.5(5, 6). The cells were clustered into two distinct groups, one of which contained the synchronized Hela (*SI Appendix*, Fig. S2 *A* and *B*). In fact, the two distinct groups on UMAP are MDA-MB-468 and Hela. Further analyses showed that Hela was separated into synchronized Hela and unsynchronized Hela (Fig. 1 *F* and *G*). Differential gene analysis was performed by DESeq2(7).

**Inferring cell cycle status and cell cycle associated genes.** Since we found polyA site usage switch was strongly associated with cell cycle (Fig. 4 *C* and *D*), inference of cell cycle status of each cell potentially provides insights into the dynamics of polyA site usage. Each cell was assigned into a cell cycle phase using its cell cycle score according to Macosko *et al.* (4) (*SI Appendix*, Fig. S6*A*). Briefly, cell cycle related genes of each phase(8) were selected and average expression level of these gene sets were calculated. Genes in each set whose correlation with the average expression level of this set across all cells larger than 0.3 were retained. The average of gene set expression level of each phase was deemed as the cell cycle score of this phase. The cell's phase would be assigned with the top phase score. And cells were ordered according to their phase score of each phase.

The fractions of cells in cell cycle phases of synchronized HeLa are significantly different from that of unsynchronized HeLa ($\chi^2$ test, $P = 2.15\times10^{-4}$). Compared with unsynchronized Hela, the number of synchronized Hela cells increased most in S phase while decreased the most in M/G1 phase (*SI Appendix*, Fig. S6*B*). The observation is consistent with previous report that cells were arrested in S phase after double thymidine block (TdR) (2). The expressions of *CDK4* (G1/S phase), *CCNE2* (G1/S phase and S phase), *MKI67* (G2/M phase), *CCNB2* (G2/M phase and M phase) and *CCNB1* (M phase) reached peaks at their specific representing cell cycle phase (*SI Appendix*, Fig. S6*C*), indicating the inferred cell cycle status is reliable.

We identified cell cycle associated genes by inferring the co-expressed genes of each cell cycle phase representing gene (*SI Appendix*, Dataset S4). E.g. The Cyclin B1 (*CCNB1)* co-expressed genes showed M phase specific high expression (*SI Appendix*, Fig. S6*D*). These genes were enriched in cell division ($P = 3.90\times10^{-32}$), mitotic nuclear division ($P = 3.55\times10^{-27}$), spindle organization ($P = 3.89\times10^{-19}$) and establishment of chromosome localization ($P = 1.55\times10^{-12}$) (*SI Appendix*, Fig. S6*E*), which indicates the functions of these genes are M phase specific.

**Identification of genes showing expression change or polyA site usage switch during cell cycle.** We obtained the RNA-seq data of synchronized MCF-7 at G1, S and M phase from Liu *et al*. (9). We calculated the significantly differentially expressed genes between any two cell cycle phases using *exactTest* function in edgeR package(10). Mapped reads within 100nt upstream of each polyA site in RNA-seq were counted to quantify the polyA site usage. Based on the number of reads on each ploy A site, we calculated the genes showing significantly polyA site usage switch using fisher's exact test.

**Deletion of polyA site by CRISPR/Cas9.** In order to check the effects of polyA site usage on cell cycle, we deleted the distal polyA site of *MSL1*, *SCCPDH*, *ZC3HAV1* and *NUP160* by CRISPR/Cas9. In brief, gene-targeted guide RNAs (*MSL1*: 5'-TAAATATATGAACAACCAAT-3'; *SCCPDH*: 5'-AAAGGTTGTCAATCGAATGGTGG -3'; *ZC3HAV1*: 5'-TCAAATTAAGCTTGAAGAAC-3'; *NUP160*: 5'-TAACAATGTGAGTTTCTCTG-3') were synthesized by Sangon Biotech and was cloned into pLKO.1 vector, which was used to package lentiviral particles in HEK293T cells cultured in 6-cm plates. Lentiviruses in the medium were collected by centrifugation at 24 hours and 48 hours, and resuspended in culture medium at a 1:1 ratio to infect MDA-MB-468 cells for 48 hours. Puromycin at 2 ug/ml was applied to select successful transfected cells.

**Cell cycle analysis by flow cytometry**. Transfected MDA-MB-468 cells were cultured in complete medium supplemented with thymidine (2 mM, Sigma) for 18 hours, washed twice with PBS and released into thymidine-free complete medium for 9 h. Thymidine (2 mM) was added for an additional 15 hours. Cells were washed with PBS and cultured in complete medium, and collected at different time points (0 hour，3 hours，6 hours，9 hours，12 hours and 18 hours). The cell cycle status was assessed by flow cytometry using Cell Cycle and Apoptosis Analysis Kit (Beyotime) according to the manufacturer. Briefly, cells were washed and fixed with 70% EtOH for 2 hours at -20 °C, stained with propidium iodide (PI), followed by Flow cytometry analysis using BD FACSAria SORP and analyzed by Flowjo.

**SI References**


1. W. Chen *et al.*, Single-Cell Transcriptome Analysis Reveals Six Subpopulations Reflecting Distinct Cellular Fates in Senescent Mouse Embryonic Fibroblasts. *Front Genet* **11**, 867 (2020).
2. H. T. Ma, R. Y. Poon, Synchronization of HeLa Cells. *Methods Mol Biol* **1524**, 189-201 (2017).
3. D. M. DeLaughter, The Use of the Fluidigm C1 for RNA Expression Analyses of Single Cells. *Curr Protoc Mol Biol* **122**, e55 (2018).
4. E. Z. Macosko *et al.*, Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
5. T. Stuart *et al.*, Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902 e1821 (2019).
6. B. Zhou, W. Jin, Visualization of Single Cell RNA-Seq Data Using t-SNE in R. *Methods Mol Biol* **2117**, 159-167 (2020).

7. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
8. M. L. Whitfield *et al.*, Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* **13**, 1977-2000 (2002).
9. Y. Liu *et al.*, Transcriptional landscape of the human cell cycle. *Proc Natl Acad Sci U S A* **114**, 3473-3478 (2017).
10. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139-140 (2010).

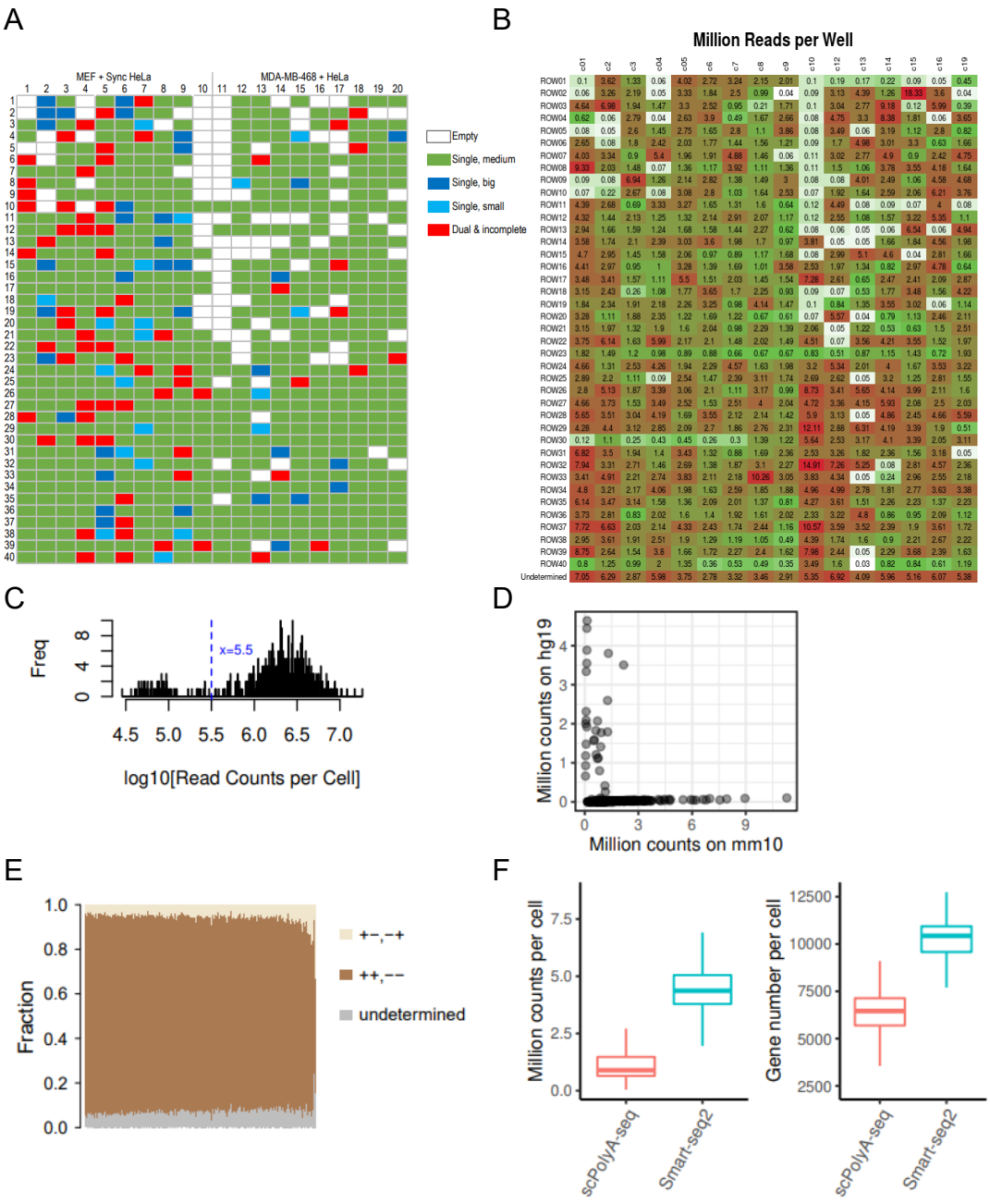


**Fig. S1. Quality control for scPolyA-seq data.** (A) Status of chip captured cells in each well under white-field microscope. (B) Heatmap of million reads per well after high throughout sequencing. (C) Histogram of counts per cell in log10 scale to determine the threshold for empty cells. (D) Uniquely mapped reads to hg19 and mm10 based on reads mapped to hg19_mm10 mega reference. One dot denotes a cell. (E) scPolyA-seq is a strand-specific library determine by RSeQC infer-experiment.py. Each column represents read fraction in a cell. (F) Detected counts and gene numbers of MEF cells between scPolyA-seq and Smart-seq2(Chen et al. 2020 Front Genet).

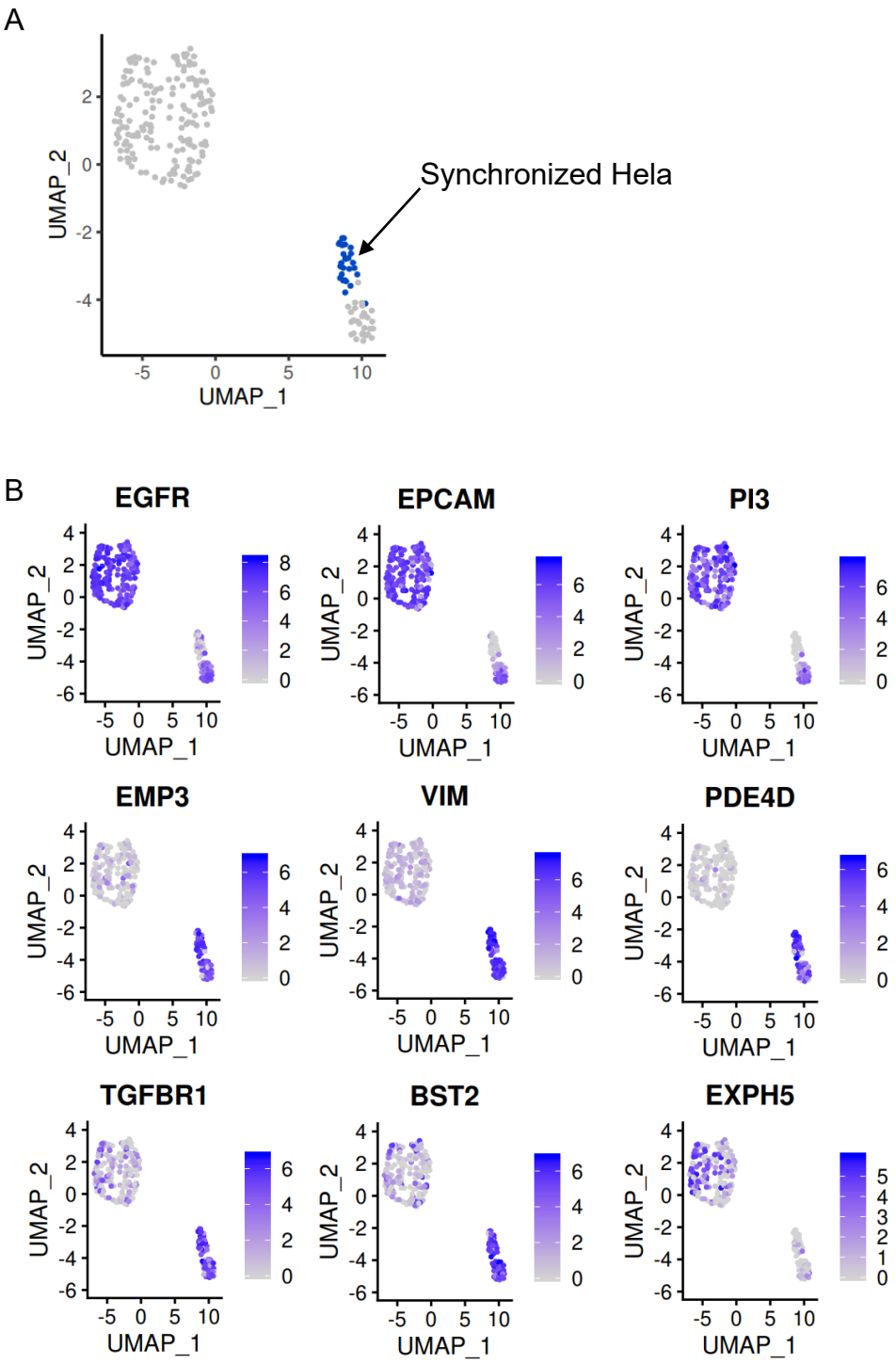


**Fig. S2. Cell clustering and cell type specific genes.** (A) UMAP plot of cells from human cell lines, with 2 sub-clusters showing up, while the top half of cluster 1 contained synchronized HeLa cells. (B) The expression levels of cell type specific genes.

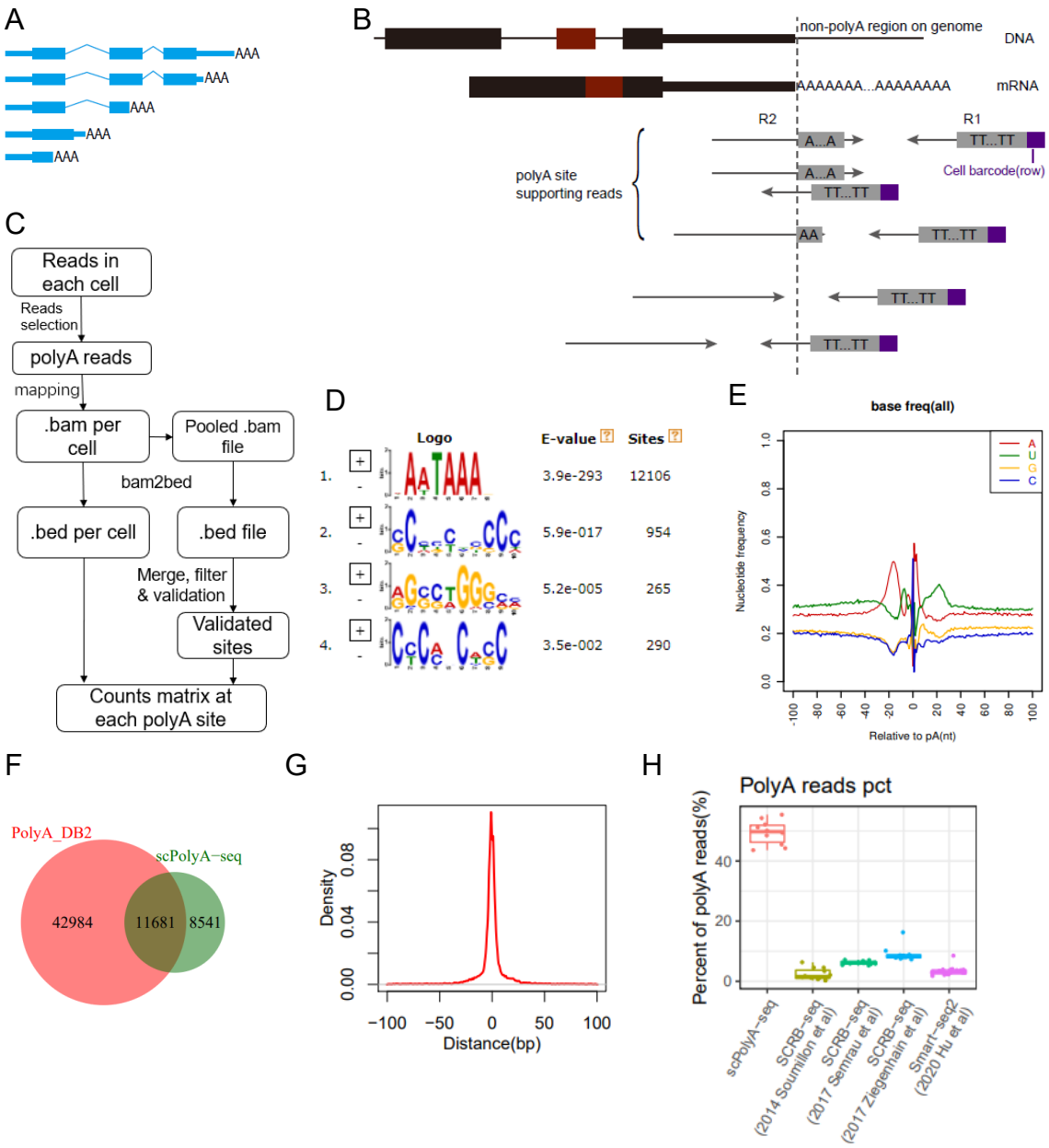


**Fig. S3. Identification of polyA sites and their features.** (A) Scheme of a gene with multiple polyA sites, either coding functional protein or not. (B) Scheme of polyA site supporting reads that used for identification of polyA site. (C) The pipeline for *de novo* polyA site identification in this study. (D) The most significantly enriched motifs near polyA sites (between -60bp and 0bp) inferred by MEME. (E) Base content frequency of 100bp upstream and downstream around each polyA site. (F) Venn diagram showed the relationship between polyA sites identified in scPolyA-seq and polyA sites in PolyA_DB2. Two polyA sites are treated as the same one if their distance <12nt. (G) Distance distribution between polyA site discovered by scPolyA-seq and the nearest polyA site in PolyA_DB2. (H) The percentage of polyA-site supporting (PASS) reads in scPolyA-seq, SCRB-seq and Smart-seq2.

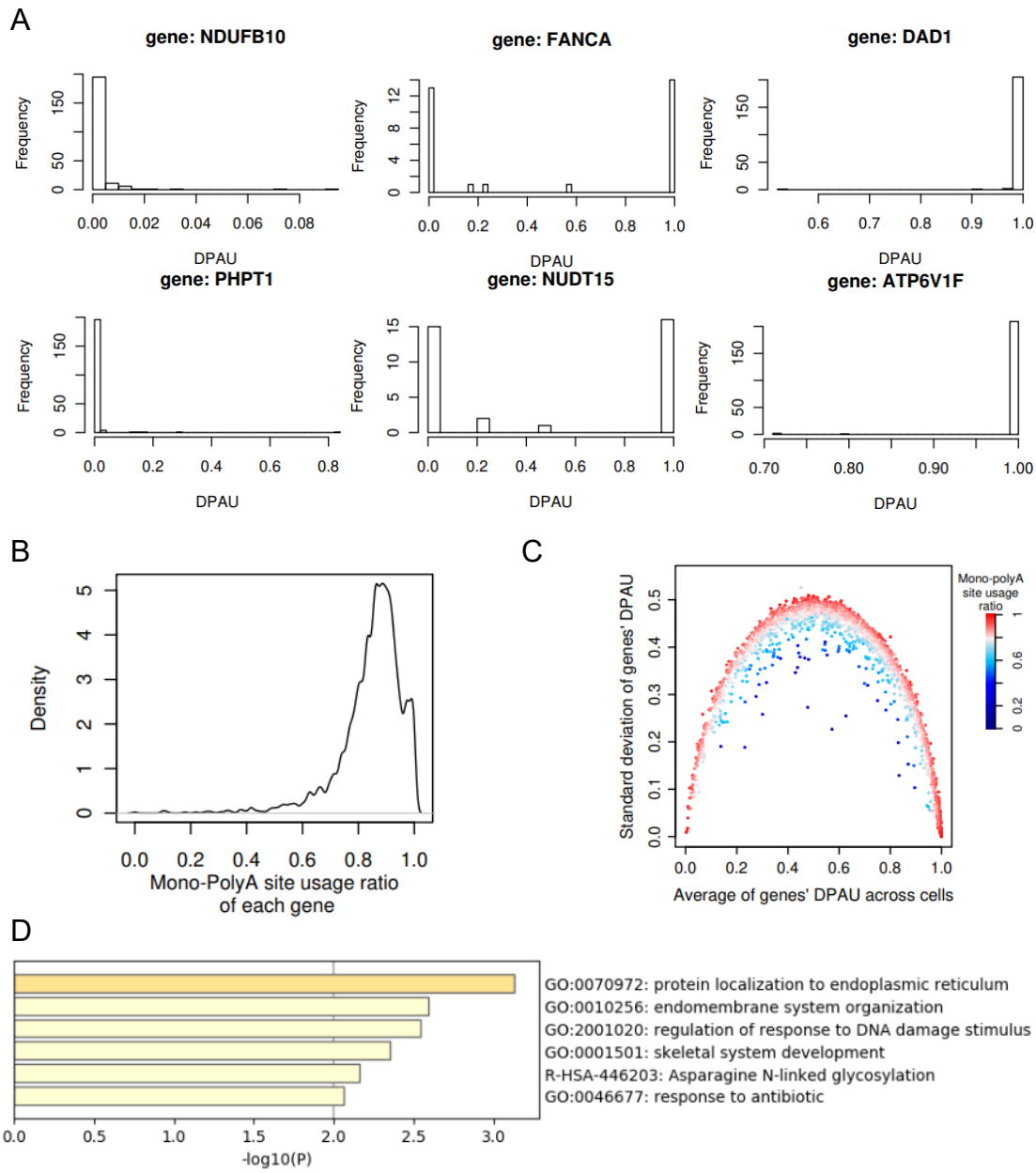


**Fig. S4. Mono-polyA site usage are prevail for genes.** (A) Histograms of DPAU of NDUFB10, PHPT1, FANCA, NUDT15, DAD1 and ATP6V1F. (B) Distribution of genes' mono-polyA site usage ratio, defined as the ratio of cells whose reads mainly at proximal or distal sites (>= 95% at either site) for each gene, shows most genes tend to favor one site in a single cell. (C) Blue points represent genes with low mono-polyA site usage ratio, which means they tend to use both sites of a gene's two poly(A) sites in one single cell, demonstrated lower variation of DPAU compared with the same level of average DPAU. (D) GO term of genes which tend to use both sites in one cell (mono-polyA site usage ratio<0.5, n=31 genes).

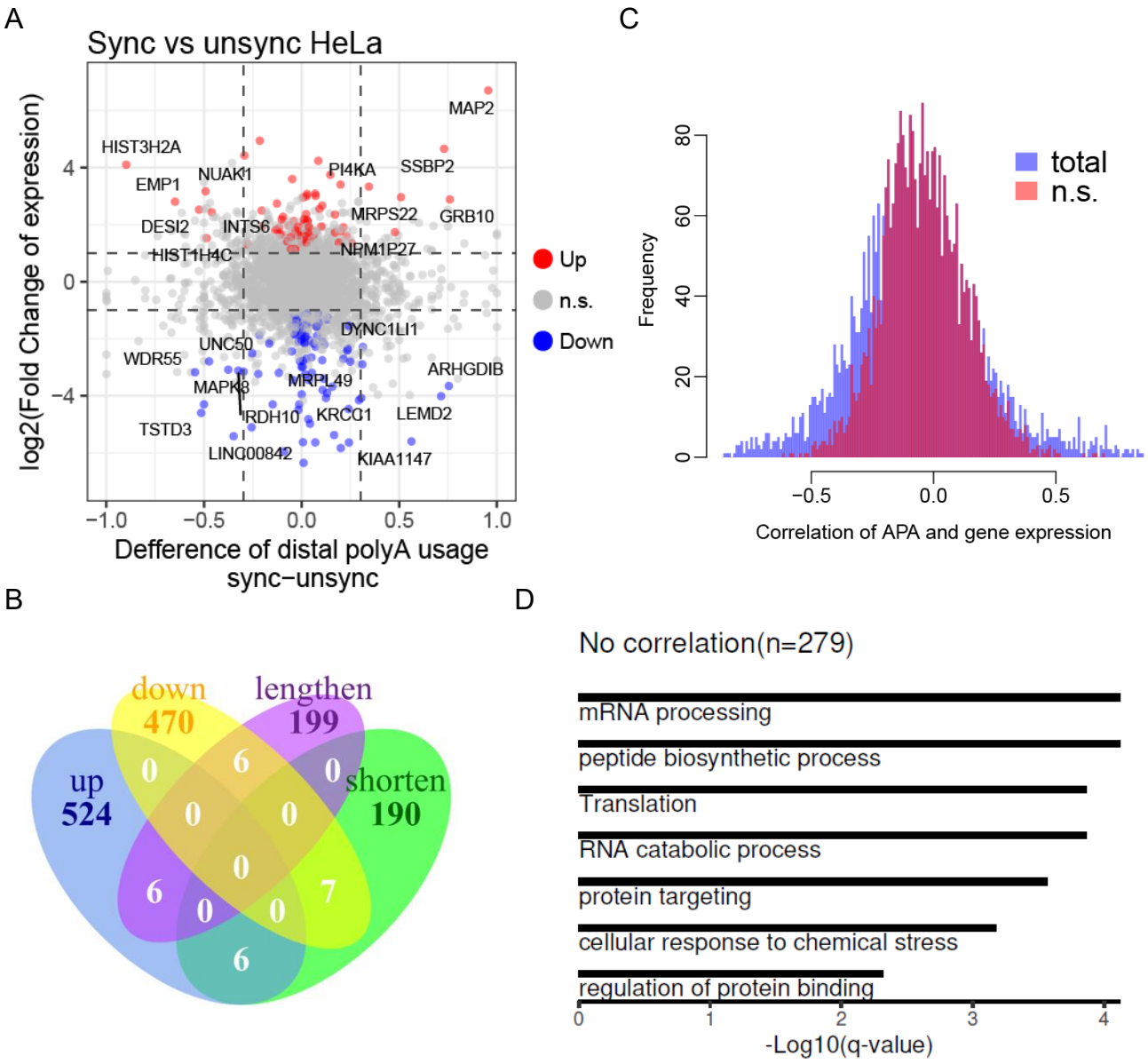


**Fig. S5. Correlation between gene expression level and APA level.** Scatterplot (A) and Venn plot (B) showed little overlaps between significant changed genes at expression level and APA level in HeLa cell before and after synchronization. (C) Histogram of Spearman correlation between gDPAU and expression from Fig. 3A, while total means all genes and n.s. denote the p-value of gDPAU and expression correlation >=0.05. (D) GO enrichment analyses of genes showing no correlation between gDPAU and expression (|Spearman correlation|<0.02).

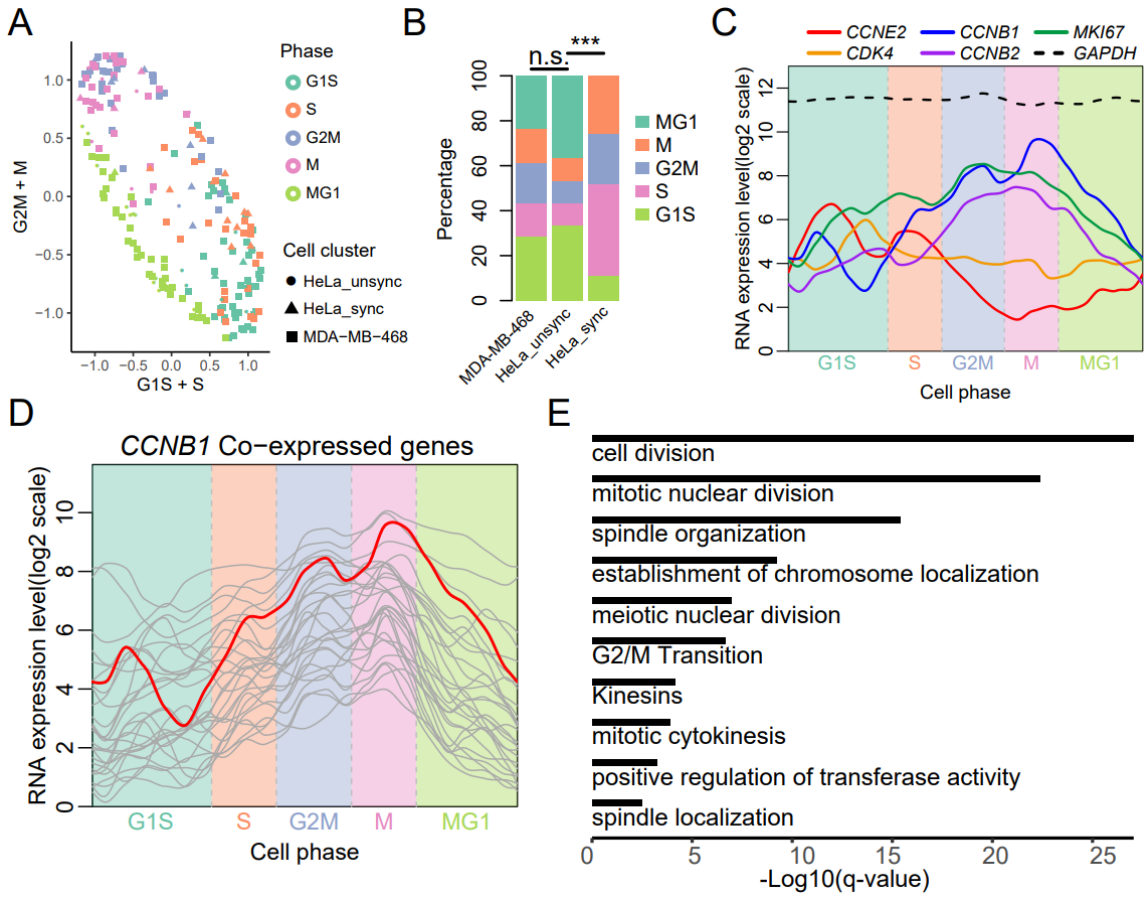


**Fig. S6. Cell cycle status and its representative genes inferred by scPolA-seq data.** (A) Inferred cell cycle status of all cells in the 3 cell subpopulations, namely MDA-MB-468, unsynchronized HeLa and synchronized HeLa. (B) Bar plot of cell cycle status for the 3 cell subpopulations. The cell cycle status of synchronized HeLa is significantly different from that of unsynchronized Hela ($\chi 2$ test, $P = 2.15\times10^{-4}$), with more cell in S phase in synchronized HeLa. (C) The expressions of cell cycle phase representing genes *CDK4* (G1/S phase), *CCNE2* (G1/S phase and S phase), *MKI67* (G2/M phase), *CCNB2* (G2/M phase and M phase) and *CCNB1* (M phase). Lines are smoothed using LOESS, with parameter span as 0.1. Using GAPDH as control. (D) Genes highly co-expressed with *CCNB1*, with Pearson correlation >0.4 and adjusted P<0.05 (n=29 genes). (E) GO enrichment analyses of genes co-expressed with *CCNB1*.

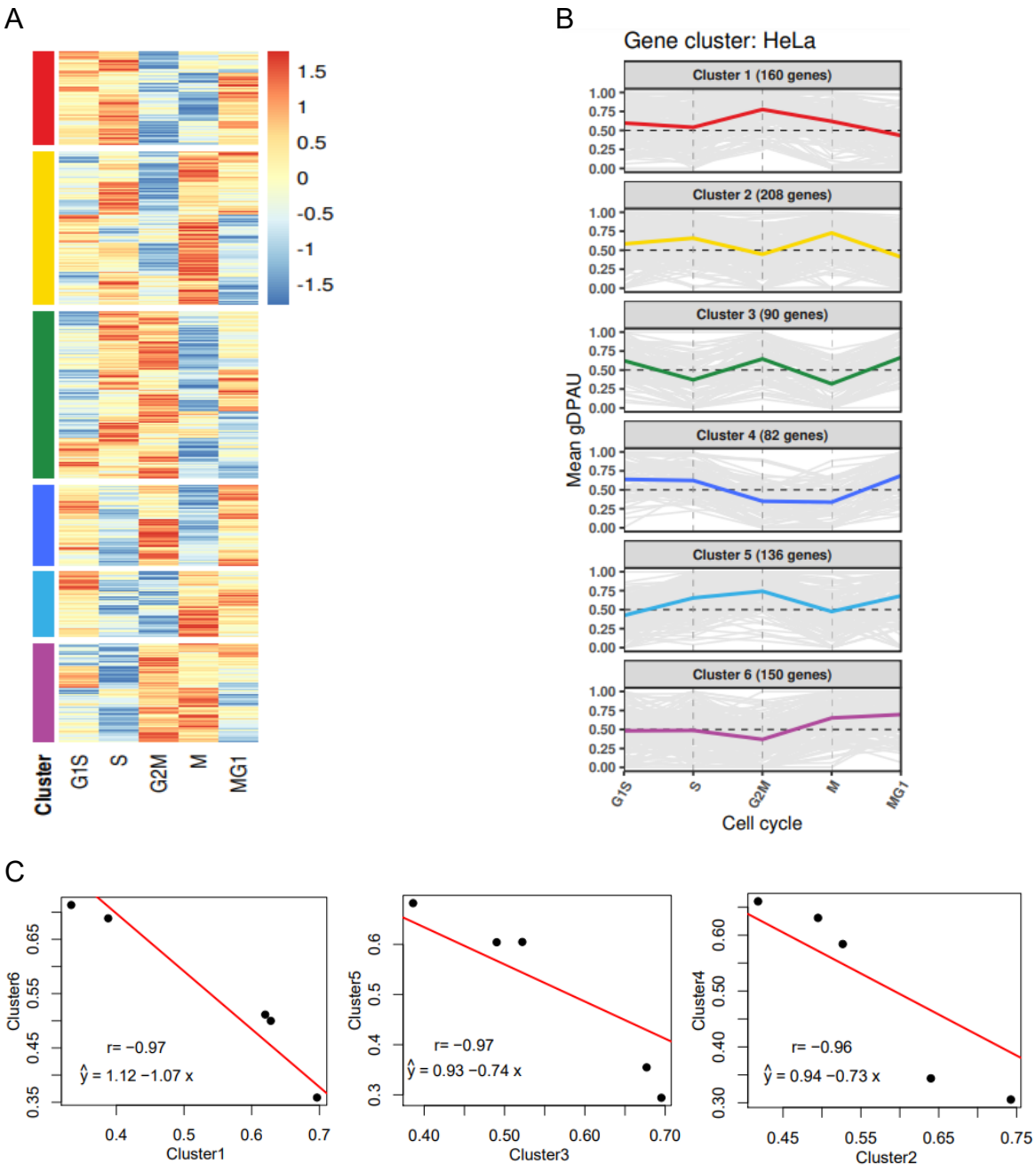


**Fig. S7. gDPAU of genes in each gene cluster along cell cycle.** (A) Heatmap of genome-wide APA dynamics along cell cycle phases in breast cancer cell line MDA-MB-468. (B) Dynamics of gDPAU of each gene at each phase for the 6 gene clusters in HeLa cells. (C) Correlation between negatively correlated cluster-pairs at APA level.

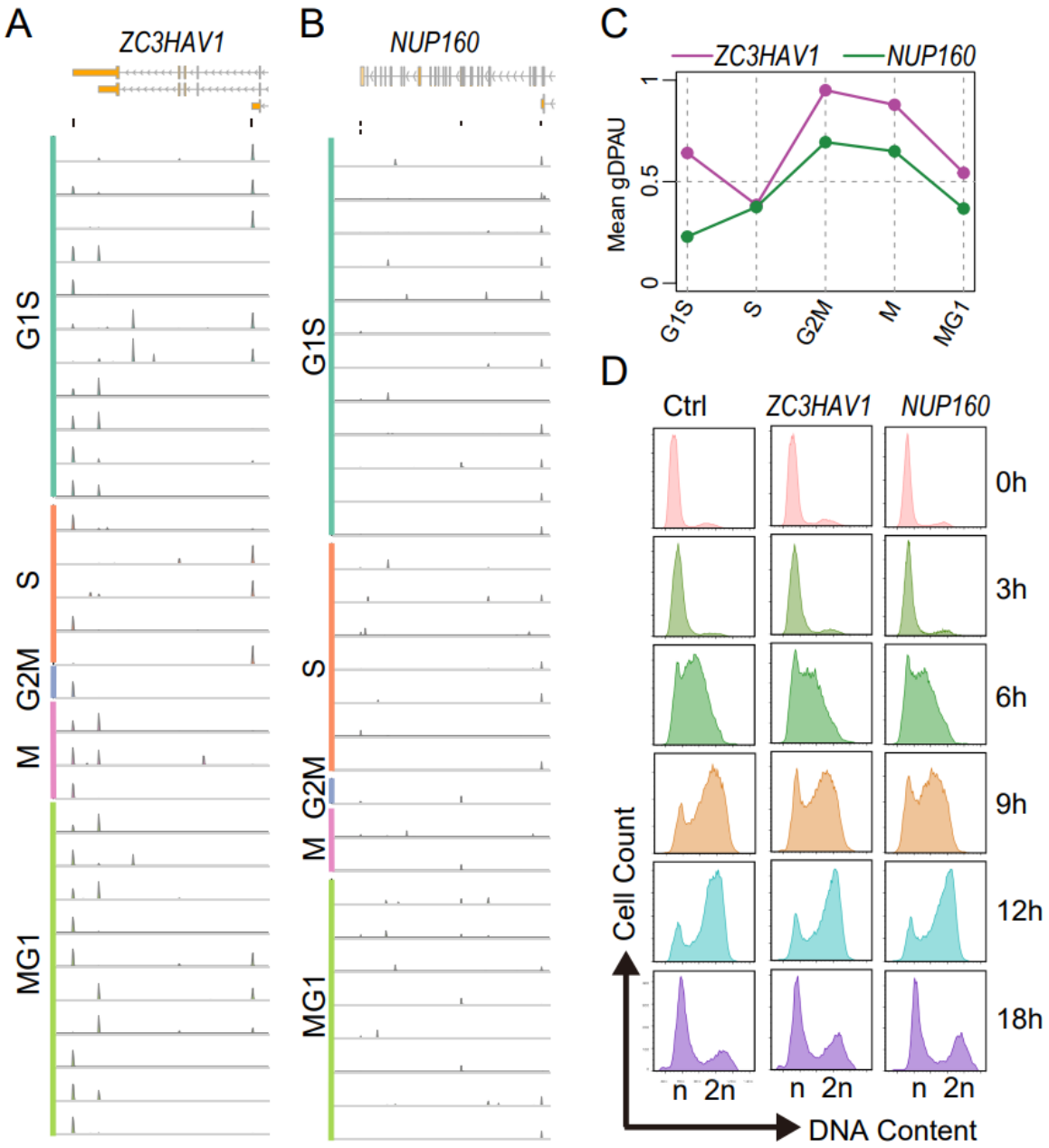


**Fig. S8. Change of cell cycle after deletion of distal polyA site of *ZC3HAV1* and *NUP160*.** (A, B) IGV screenshots of reads density around polyA site at single cell level showed the dynamics of polyA site usage in *ZC3HAV1* (A) and *NUP160* (B) during cell cycle. The top track and second track show gene structure at 3'-end and position of polyA sites of the gene, respectively, with following tracks showing reads density of each cell. (C) The mean gDPAU of *ZC3HAV1* and *NUP160* at each cell phase, indicating dynamics of polyA usage along cell cycle. (D) Cell cycles of distal polyA-site deleted MDA-MB-468 cells in *ZC3HAV1* and *NUP160* were slower than that of wild type MDA-MB-468 cells (Ctrl).

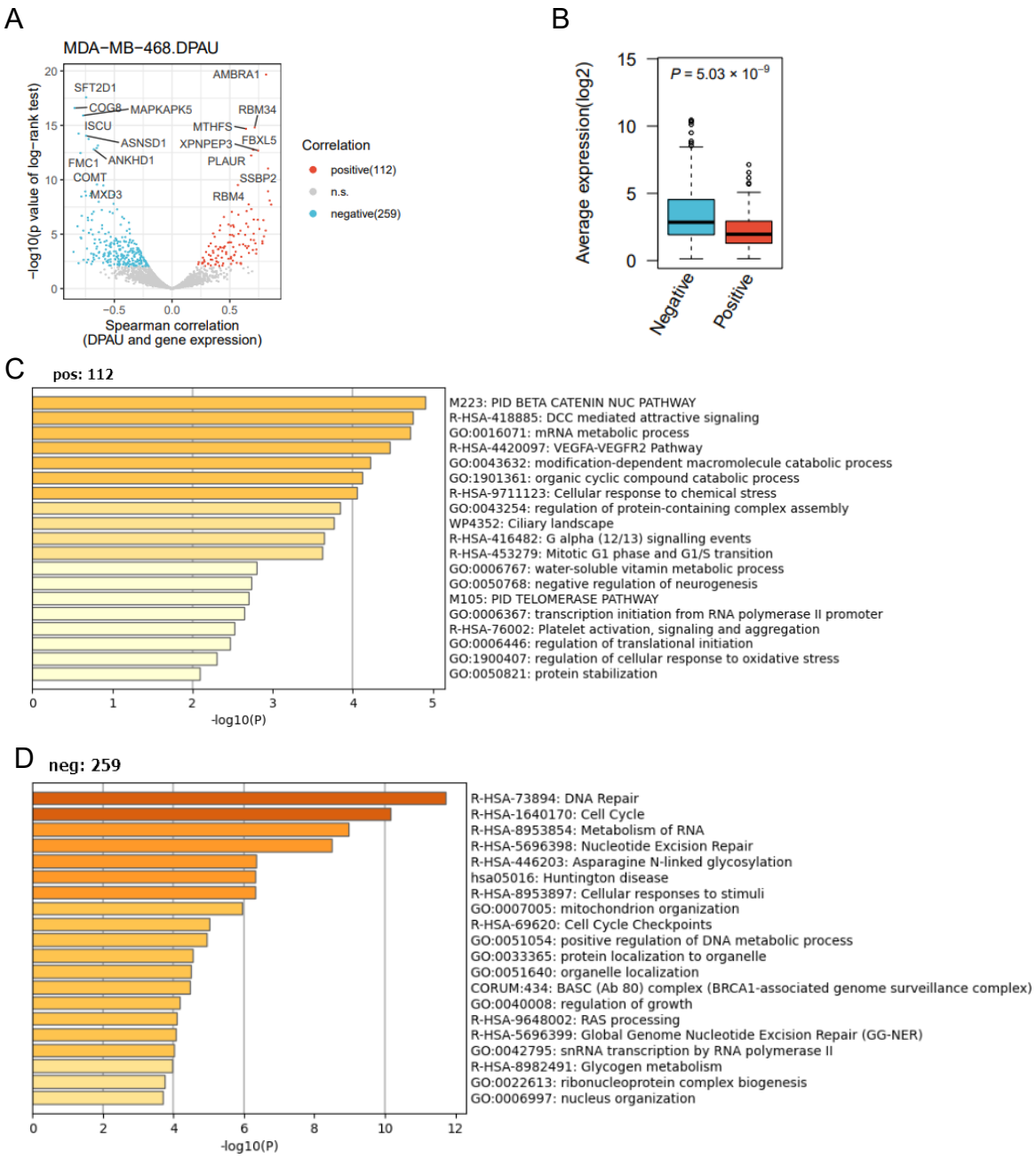


**Fig. S9. Relationship between polyA site usage and expression level using genes with 2 polyA sites.** (A) Identification of genes showing significant correlation between distal polyA usage (DPAU) and expression level, with each dot representing a gene. (B) Genes showing negative correlation between polyA site usage and expression level have higher expression level than those showing positive correlation. (C-D) GO enrichment of genes showing positive correlation (C) and negative correlation (D) between polyA site usage and expression level.

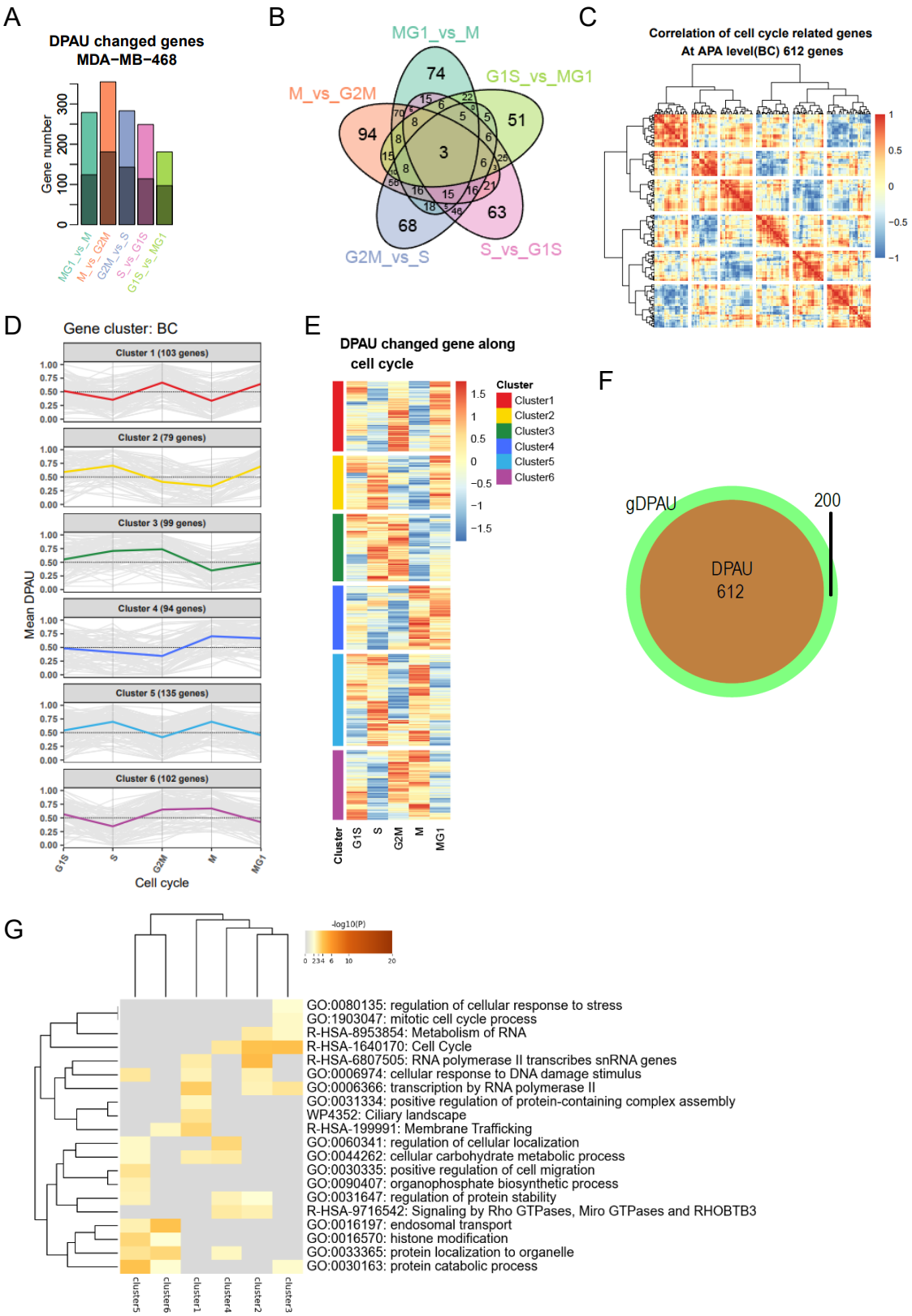


**Fig. S10. Dynamics of polyA site usage during cell cycle using genes with 2 polyA sites.** (A) The number of genes with different polyA site usages between neighboring cell cycle phases. Dark

color and light color denote shortening and lengthening of APA events. (B) Venn diagram of genes with different polyA site usages between neighboring cell phases. (C) Heatmap of correlation of DPAU of each gene across 5 cell phases, resulting genes clustered into 6 gene clusters. (D) Dynamics of DPAU of each gene at each phase for the 6 gene clusters. Each line denotes one gene. Y-axis denotes this gene's average DPAU at each phase. Colored lines are averaged DPAU of each gene cluster. (E) Heatmap of genome-wide APA dynamics along cell cycle phases. (F) The gene sets based on 2 polyA sites is a subset of the genes based on multi-polyA sites. (G) GO enrichment of the six gene clusters.

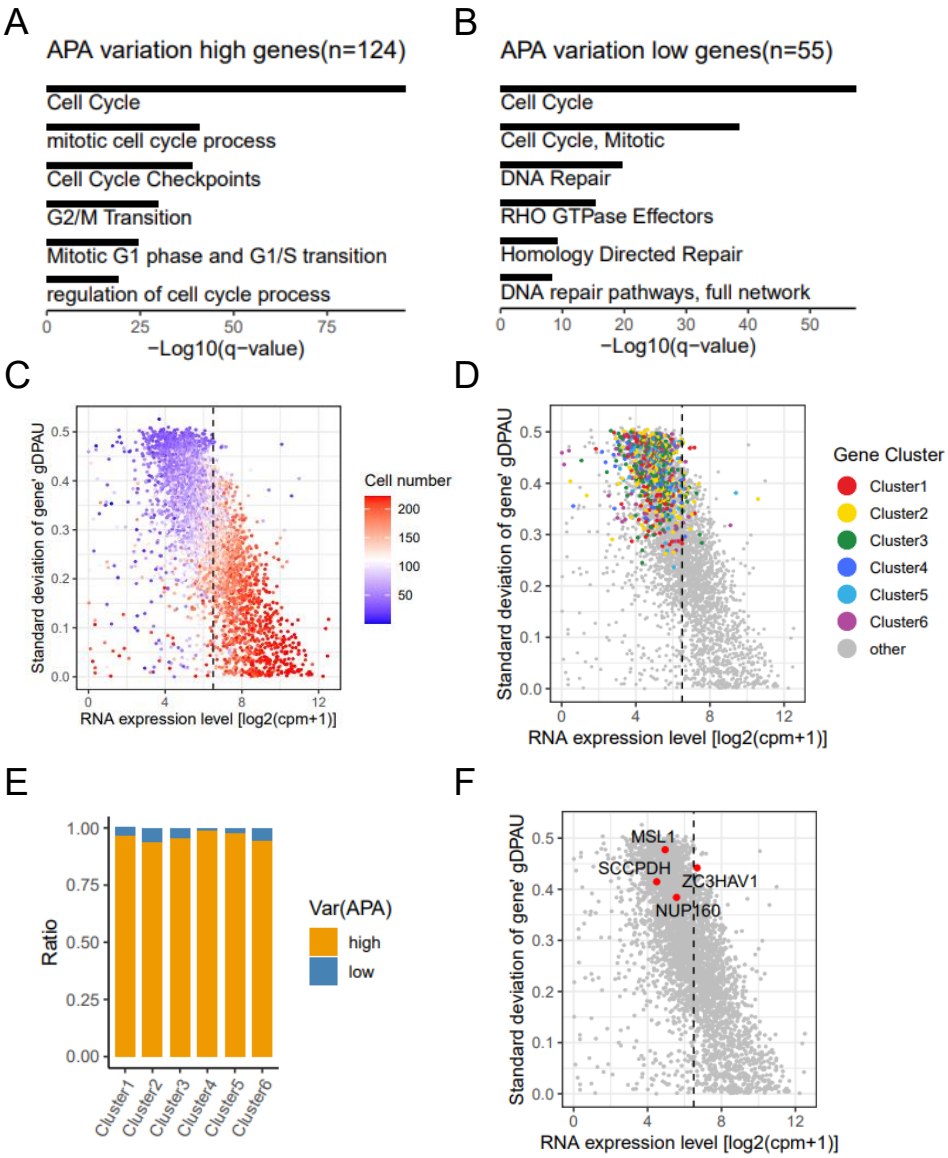


**Fig. S11. 'Cell cycle' genes have different APA behaviors with cell cycle process and regulation subset showing high APA variation and dynamics.** (A) GO analyses of genes with high APA variation and low expression level in GO term 'Cell cycle'. (B) GO analyses of genes with low APA variation and high expression level in GO term 'Cell cycle'. (C) Scatter plot of average expression level and standard variation of gDPAU, with all the genes ≥2 polyA sites. Vertical line separates high and low expressed genes. Dot denotes gene. (D) The 6 gene clusters showing cell cycle dynamics in Fig.5 are high APA variation genes. (E) Bar plot of high APA variation genes and low APA variation genes in each of the 6 gene clusters. (F) The 4 experimentally validated genes in Fig. 6 and Fig. S8 are high APA variation genes.

**Dataset S1 (separate file).** Spearman correlation between gDPAU and gene expression.

**Dataset S2 (separate file).** Significantly differential expressed genes between unsynchronized HeLa and synchronized HeLa by DESeq2.

**Dataset S3 (separate file).** PolyA site usage switched genes between synchronized HeLa and unsynchronized HeLa.

**Dataset S4 (separate file).** Cell cycle associated genes used to infer cell cycle phase in MDA-MB-468 cell line.

**Dataset S5 (separate file).** Genes showing polyA site usage switches between neighboring cell cycle phases in MDA-MB-468 can be grouped into six gene clusters.

**Dataset S6 (separate file).** The six gene clusters inferred using correlation of average gDPAU of each gene across five cell phases in MDA-MB-468 cells.