**APAIQ identifies and quantifies sample-specific alternative polyadenylation from RNA-seq data**

**Methods and Materials**

**Data Preprocessing**

The normal RNA-Seq data are mapped with STAR 2.7.0 with encode standard options and --outWigType bedGraph to generate bedGraph file at the same time. Only the unique mapped reads will be used for further studies. The 3’mRNA-Seq data are also mapped with STAR 2.7.0 with encode standard options after trimming adapter and leading T. Then, the 3’ end position are clustered and mapped to polyADB3[3] and encode PAS database within +-25 bp to get the expression level of PAS. If the adjacent PAS’s distance is smaller than 50bp, the PAS with less expression level will be dropped out. Each PAS should be at least 0.05 RPM in 3’mRNA-Seq and with usage>0.05 (relative expression level of PAS in the same gene). Then, we extract 100bp upstream and 100 bp downstream genomic DNA sequence and RNA-Seq coverage of the cleavage sites from human/mouse genome GRCh38/GRCm38 and RNA-Seq unique mapped bedGraph files. Each PAS has 201bp genomic DNA sequence and 201bp RNA-Seq coverage. D. We also filter out those PAS with upstream RNA-Seq coverage<0.05 RPM. Finally, We got about 20,000 PAS as the positive dataset each cell line during training. For negative dataset, we randomly selected sites in transcriptome as false cleavage sites with upstream RNA-Seq coverage>0.05 and distance to ground truth PAS and other false PAS should be greater than 50bp.

**Model**

Figure ? illustrates the architecture of APAIQ. The genomic DNA sequence data and RNA-Seq coverage data served as inputs separately. The genomic DNA sequence is one-hot encoding to 201x4 matrix and the RNA-Seq coverage data should be normalized to RPM and convert to 201x1 matrix. They firstly go through a convolution layer consisting of 32 filters with kernel size set as 6. It followed by a group normalization layers with group size set as 4. RELU is applied to the normalized results as the activation function. After a max-pooling layer with pooing window setting as 6, features are flattened into 1D dimensional array and fed to the fully connected layer. Then, two features from genomic DNA sequence data and RNA-Seq coverage data are concatenated together and fed into another fully connected layer followed by a softmax activation function to approximate the probability function. Dropout is introduced after max-pooling layer as regularization to reduce over-fitting problems.

**Training and Evaluation**

The transcriptome was split into small blocks with length about 1e6 bp. Specifically, we randomly partition the blocks into five folds and used four of them for training and one for validation. We set the training epoch as 500 and the mini-batch size as 32. Data augmentation was applied during the training, each PAS random shift from -12 to 12 bp of the ground truth position during each epoch. We saved the model that has the highest accuracy on validation set and used this model to scan the whole transcriptome of window length as 201 with sliding step 1. Item with upstream coverage less than 0.05 RPM were filtered and ignored. Each item had a prediction score between 0 and 1. We considered item with prediction score higher than 0.5 as PAS containing and PAS not in the middle of the window but not too far also had prediction score higher than 0.5. Therefore, we collapse continuous position in transcriptome as higher than 0.5 one PAS and predict the location of PAS. This algorithm goes through the genome in two direction (from forward strand to reverse strand and from reverse strand to forward) to correlate the system bias. If a position was predicted as true, the prediction score was added to sum. If false, one point penalty is applied. We find the maximum sum in two direction and if they both are greater than a threshold T, the position of the PAS will be average the two maximum sum position. Even though we collapsed the transcription, we still got too many false positive by directing applying the training model. Therefore, we randomly replaced half of the negative dataset with the false positive PAS to increase the complexity of the training set and trained again. After, four iteration training, the recall the precision converged. We kept the specific model with best F1 score.

**Data availability**

3’mRNA-Seq raw read sequences of four human cell lines (K562, HepG2, SNU398, THLE2) have been submitted to the NCBI Sequence Read Archive (SRA) (<http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>) under accession number SRAXXXXXX. The mouse BL6 fibroblast 3’mRNA-Seq raw read sequences are available on the European Nucleotide Archive (<http://www.ebi.ac.uk/ena>) under the accession number PRJEB15336, run accession from ERR1616672 to ERR1616675. The raw RNA-Seq data that support the finding of the study are available on public database (THLE2: NCBI SRA accession: PRJNA495931, SRA runs: SRR8040781-SRR8040792; SNU398: NCBI SRA accession: PRJNA562266, SRA runs: SRR10022387-SRR10022388; HepG2: ENCODE Library ENCLB471LNG, ENCLB352YLJ; k562: from Chen lab). The mouse BL6 fibroblast RNA-Seq are available on ENA accession: PRJEB7211, run accession from ERR597280 to ERR597282. The polyadenylation sites from PolyA\_DB3 and gencode release 38 are available respectively on <https://exon.apps.wistar.org/PolyA_DB/v3/> and <https://www.gencodegenes.org/>.

**Code Availability**

The open source APAIQ is freely available at https://githun.com/ijayden-lung/APAIQ