Supplemental Materials

Accurate transcriptome-wide identification and quantification of alternative polyadenylation from RNA-seq data with APAIQ

Yongkang Long[1,2,¶], Bin Zhang[1,2,¶,*], Shuye Tian[3,¶], Jia Jia Chan[4], Juexiao Zhou[1,2], Zhongxiao Li[1,2], Yisheng Li[3,5], Zheng An[6], Xingyu Liao[1,2], Yu Wang[7], Shiwei Sun[8], Ying Xu[6], Yvonne Tay[4,9], Wei Chen[3, *], Xin Gao[1,2, *]

[1]Computer Science Program, Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

[2]KAUST Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

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

[4]Cancer Science Institute of Singapore, National University of Singapore, Singapore

[5]Shenzhen Haoshi Biotechnology Co., Ltd, Bao An District, Shenzhen, China

[6]Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, the University of Georgia, USA

[7]Syneron Tech, Guangzhou, China

[8]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

[9]Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

¶These authors contributed equally to this work

*Corresponding author

Supplemental information

1. Supplemental Methods

2. Supplemental Tables

3. Supplemental Figures

1. Supplemental Methods

Method S1: Evaluation metrics for PAS identification

we defined any predicted PAS within a distance threshold (default is 25bp) away from the ground truth as true positive (TP), while any annotated PAS that is 1) not expressed (3' end-seq RPM = 0), 2) covered with RNA-seq reads (average RPM at 100bp upstream > 0.05), and 3) not being detected by the method was defined as true negative (TN). The false positives (FP) are those predicted PAS located more than 25bp away from the ground truth. The false negatives (FN) are those PAS from ground truth that are not detected by the methods.

Recall = $\frac{TP}{TP+FN}$

FPR = $\frac{FP}{FP+TN}$

Precision = $\frac{TP}{TP+FP}$

Method S2: Evaluation of APAIQ on an independent dataset

To evaluate the transfer capacity of APAIQ using the pre-trained model, we applied APAIQ to an independent public dataset, which has been used to benchmark computational methods for APA analysis(Shah et al. 2021). The dataset consists of 3' end-seq from 52 HapMap Yoruba human lymphoblastoid cell lines (LCL) (Mittleman et al. 2020), and 87 Yoruba LCL RNA-seq samples. We applied APAIQ to each RNA-seq sample with the pre-trained model for PAS identification and further evaluated the performance of the prediction using the PAS from 3' end-seq data. In total, there are 41,784 identified PAS (Elife_PAS) from the original 3' end-seq study (Mittleman et al. 2020), while the number reduced to 16,341 (GB_PAS) with more stringent filtering in the following benchmark study (Shah et al. 2021). We incorporated these two sources as the ground truth and evaluate the precision and recall of APAIQ predictions with different distance thresholds to define the overlap between predictions and the ground truth.

Method S3: Hyperparameters of the hybrid deep-learning model

We randomly sampled a set of hyperparameters and tested their performance on the validation sets, which take much less computational cost than grid search. The filter size is set to 6 because the length of ployA signal is 6. The number of filters was searched within {16,32,64} and the number of hidden nodes in FC layers is {16,32,64}. We sampled the number of groups from {2,4,8} and dropout rate from {0.25,0.5,0.75}. The L2 regularization rate and learning rate are both sampled from {1e-5,5e-4,1e-4,5e-4,1e-3,5e-3,1e-2,5e-2,1e-1,5e-1}. The range of momentum rate is {0.95,0.98}. Finally, the hypermeters were determined as follows, number of groups = 4, number of filters = 32, number of hidden nodes in FC layer = 64, dropout rate=0.25, all L2 regularization rates=5e-4, momentum rate=0.95, learning rate=1e-3.

2. Supplemental Tables

Table S1: General statistics of APAIQ models on binary classification

Dataset	Model	TP	FP	FN	TN	TPR	FDR	ACC
THLE2	Sequence	19263	1589	995	18669	0.951	0.076	0.936
	Coverage	14762	2238	5496	18020	0.729	0.132	0.809
	Integrated	19627	1769	631	18489	0.969	0.083	0.941
SNU398	Sequence	19829	1461	1162	19530	0.945	0.068	0.938
	Coverage	16094	2291	4897	18700	0.767	0.125	0.829
	Integrated	20195	1278	796	19713	0.962	0.060	0.951

HepG2	Sequence	19141	1564	1191	18768	0.941	0.076	0.932
	Coverage	15288	3100	5044	17232	0.752	0.168	0.800
	Integrated	19481	1304	851	19028	0.958	0.063	0.947
K562	Sequence	18785	1377	1160	18568	0.942	0.068	0.936
	Coverage	15099	3717	4846	16228	0.757	0.198	0.785
	Integrated	19173	1516	772	18429	0.961	0.073	0.943

Table S2: The 20 predicted PAS selected for the experimental validation. The three failed candidates in 3’ RACE experiments were marked in red.

Gene	**PAS**	**Type**	**#TCGA samples**	**Band size**	**Expression in HepG2**
SULT2A1	chr19:47870601:-	Terminal exon	374	252	102.4
ADH4	chr4:99123794:-	Terminal exon	313	323	10.6
FGA	chr4:154583927:-	Terminal exon	408	255	15.5
SERIPNA7	chrX:106033200:-	Terminal exon	368	235	18.6
SERIPNA7	chrX:106033424:-	Terminal exon	369	236	13.3
APOB	chr2:21009912:-	Upstream exon	130	331	9.6
PRG4	chr1:186313865:+	Terminal exon	208	192	47.2
SLC9B1	chr1:147184222:-	Intronic	384	-	1.96
IGFBP1	chr7:45893299:+	Terminal exon	170	-	2.75
ASAH2B	chr22:32859644:-	Intronic	238	-	1.65
HSPA5	chr9:125236238:-	Terminal exon	419	384	2579.8

AGT	chr1:230702732:-	Upstream exon	413	281	1821.6
SNRPD3	chr22:24572153:+	Intronic	421	266	356.8
HNRNPA2B1	chr7:26191840:-	Upstream exon	393	225	289.9
POLR2L	chr11:840203:-	Upstream exon	421	280	108.5
ALDH2	chr12:111809981:+	Terminal exon	421	248	69.8
MPC2	chr1:167918068:-	Terminal exon	412	256	64.7
FGB	chr4:154571079:+	Terminal exon	415	250	33.5
PPM1B	chr2:44234066:+	Intronic	418	229	31.6
TMEM184C	chr4:147637705:+	Intronic	414	258	12.2

Table S3: Primer sequences

Primer Name	**sequence**
FLAD1-F1	ACTCAGGACACCAACACCTT
FLAD1-F2	CCACTTTGGACGTAGGCTTG
ERCC-F1	AGGAGCTGGCTAAGATGTGT
ERCC-F2	TATGAGCAGAAACCAGCGGA
FLAD1-qPCR-F	AGGCTGTATACAAACTCGCTGAA
FLAD1-IPA-R	GCTATCTTCTGCCCTCTCAGG
FLAD1-FL-R	CAGTCTTTGCCCCCGTTGA
ERCC-qPCR-F	ATCGCCGCATCAAGAGAAG
ERCC-IPA-R	TATTTGGGGCTCTCTCCTTCC
ERCC-FL-R	GCTGGGGTCATCAGGGTACT
SULT2A1-F1	GGTGCACGCCTATAGTCC
SULT2A1-F2	TTTCTGTTGGTGCTGATATTGCGACTAGGGTTCAGAGAACCAG

ADH4-F1	GAAGATGCCAGGAGCAATTC
ADH4-F2	TTTCTGTTGGTGCTGATATTGCAGATTGGGTAATGAATACATGGAG
FGA-F1	GATGCAGACCAGTGGGAAG
FGA-F2	TTTCTGTTGGTGCTGATATTGCTTCCCTCAGGGCTGTTCG
SERPINA7-F1	GCTGCCCATAAGGCTGT
SERPINA7-F2	TTTCTGTTGGTGCTGATATTGCCTGCAGCTGTCCCTGAAG
SERPINA7-2-F2	TTTCTGTTGGTGCTGATATTGCGATGTGAGCTTGGACTTGCA
APOB-F1	AGAGACAAGTTTCACATGCC
APOB-F2	TTTCTGTTGGTGCTGATATTGCGATGAGCACTATCATATCCGTGT
PRG4-F1	CACAGGTTAGGAGACGTCG
PRG4-F2	TTTCTGTTGGTGCTGATATTGCCAATACTATAACATTGATGTGCCTAG
HSPA5-F1	AGCAAACTCTATGGAAGTGC
HSPA-F2	TTTCTGTTGGTGCTGATATTGCGAGTTGTAGACACTGATCTGC
AGT-F1	GCCCATTCCTGTTTGCTGT
AGT-F2	TTTCTGTTGGTGCTGATATTGCTGAGTCGACTTTGAGCTGG
SNRPD3-F1	GGTATACATCCGTGGCAGC
SNRPD3-F2	TTTCTGTTGGTGCTGATATTGCTGGCCGCAAGAGGAAGAG
hnRNPA2B1-F1	GGATGAGAGCCCAGAGG
hnRNPA2B1-F2	TTTCTGTTGGTGCTGATATTGCGAGTGTAGAAGCATTCCTTC
POLR2L-F1	TCACTTGTGGCAAGATCGTC
POLR2L-F2	TTTCTGTTGGTGCTGATATTGCATGCGCTGGATGCCCT
ALDH2-F1	AGTCAAAGTGCCTCAGAAG
ALDH2-F2	TTTCTGTTGGTGCTGATATTGCGTGGGTTGGCTGAGGGTA
MPC2-F1	TTCTTTGTGGGGGCAGC
MPC2-F2	TTTCTGTTGGTGCTGATATTGCGAGTTCCTGATCACCTGAAC

FGB-F1	GAAGATCAGGCCCTTCTTCC
FGB-F2	TTTCTGTTGGTGCTGATATTGCCTCTTGCTCACCAAGAAGTAAC
PPM1B-F1	GGTTTACATGTTGTGGATGC
PPM1B-F2	TTTCTGTTGGTGCTGATATTGCGTGCTTTACATGATGATGTG
TMEM184C-F1	GCTACAACATGGATGGGC
TMEM184C-F2	TTTCTGTTGGTGCTGATATTGCGTAGGTCAAAGGATTCAAGGTAG
IGFBP1-F1	ACATCCATGGATGGAGAGGC
IGFBP1-F2	TTTCTGTTGGTGCTGATATTGCCCTGGGTCTCCAGAGATCAG
SLC9B1-F1	CAATCAAATGCAGGTGAGGT
SLC9B1-F2	TTTCTGTTGGTGCTGATATTGCTTGCAGCAGCCCATGAG
ASAH2B-F1	CTCTTTACAGGTTCCTATGG
ASAH2B-F2	TTTCTGTTGGTGCTGATATTGCGAAGCTGAATTCTTGGTCTG
ligation-F	TTTCTGTTGGTGCTGATATTGC
ligation-R	ACTTGCCTGTCGCTCTATCTTC
3RACE-anchor-R	GACCACGCGTATCGATGTCGAC

3. Supplemental Figures

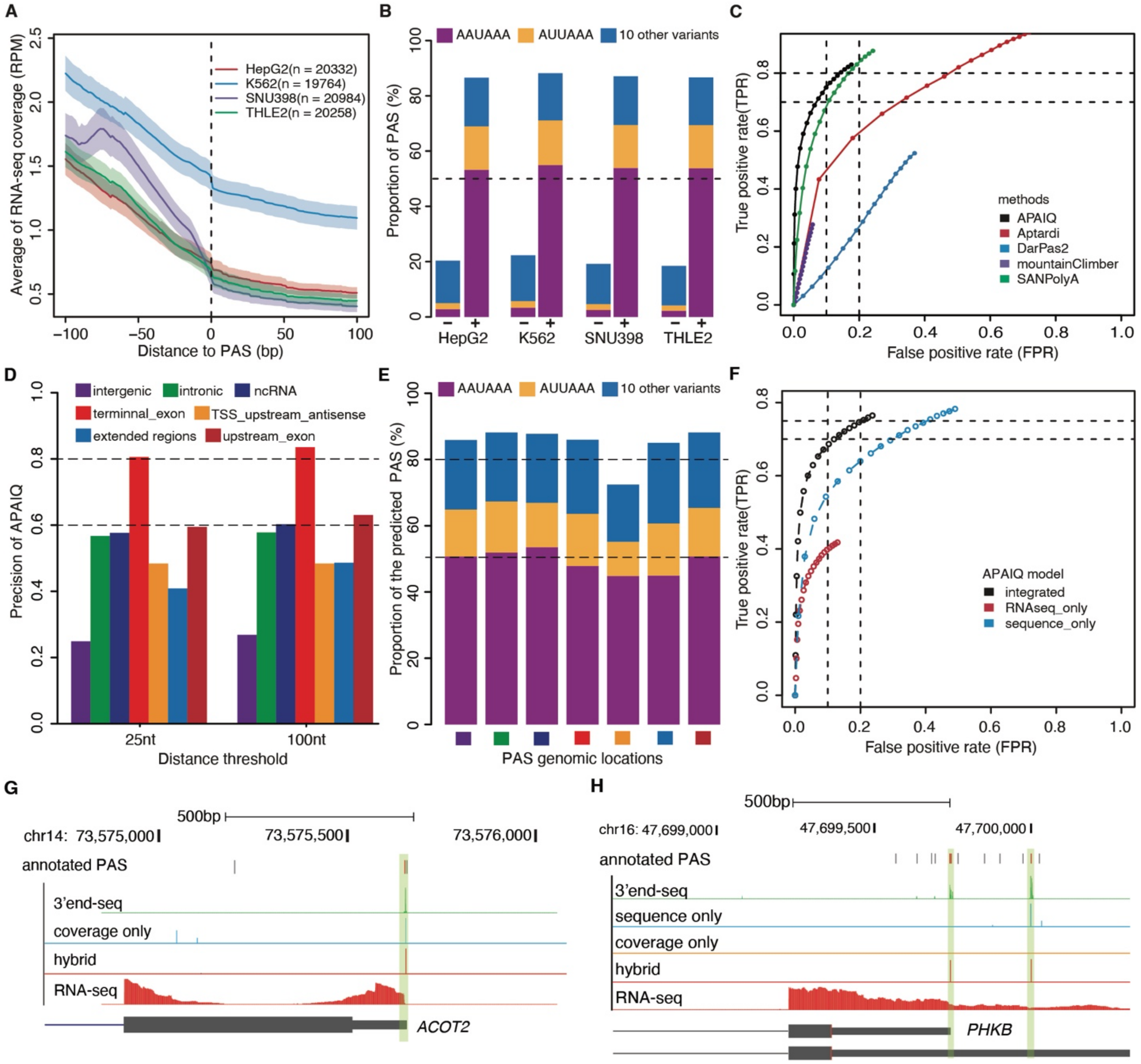

Figure S1: Comprehensive evaluation of APAIQ for PAS identification. A. Distribution of RNA-seq coverage around the expressed/used PAS in four cell lines, including HepG2, K562, SNU398, and THLE2. B. Percentage of the used/expressed PAS (+) harboring PAS motifs in each cell line and their frequency in random background (-). C**.** ROC curves showing the performance of the APAIQ and other 4 published methods, including Aptardi, DaPars2, mountainClimber, and SANPolyA on PAS identification within terminal exon. D. Precision of APAIQ for the

identification of PAS from different genomic location categories. E. Percentage of the predicted PAS harboring PAS motifs in each genomic location category. The colors below X-axis indicate the category. F. ROC curves showing the performance of APAIQ on PAS identification using different training models. RNAseq_only and sequence_only are the models only trained by using RNA-seq coverage and DNA sequence, respectively. The integrated represent the model using boh DNA sequence and RNA-seq coverage as inputs. G and H, Genome browser showing predictions of PAS from two genes based on the hybrid model integrating DNA sequence and RNA-seq coverage, the model only using DNA sequence (sequence-only) and the model only using RNA-seq coverage (coverage-only).

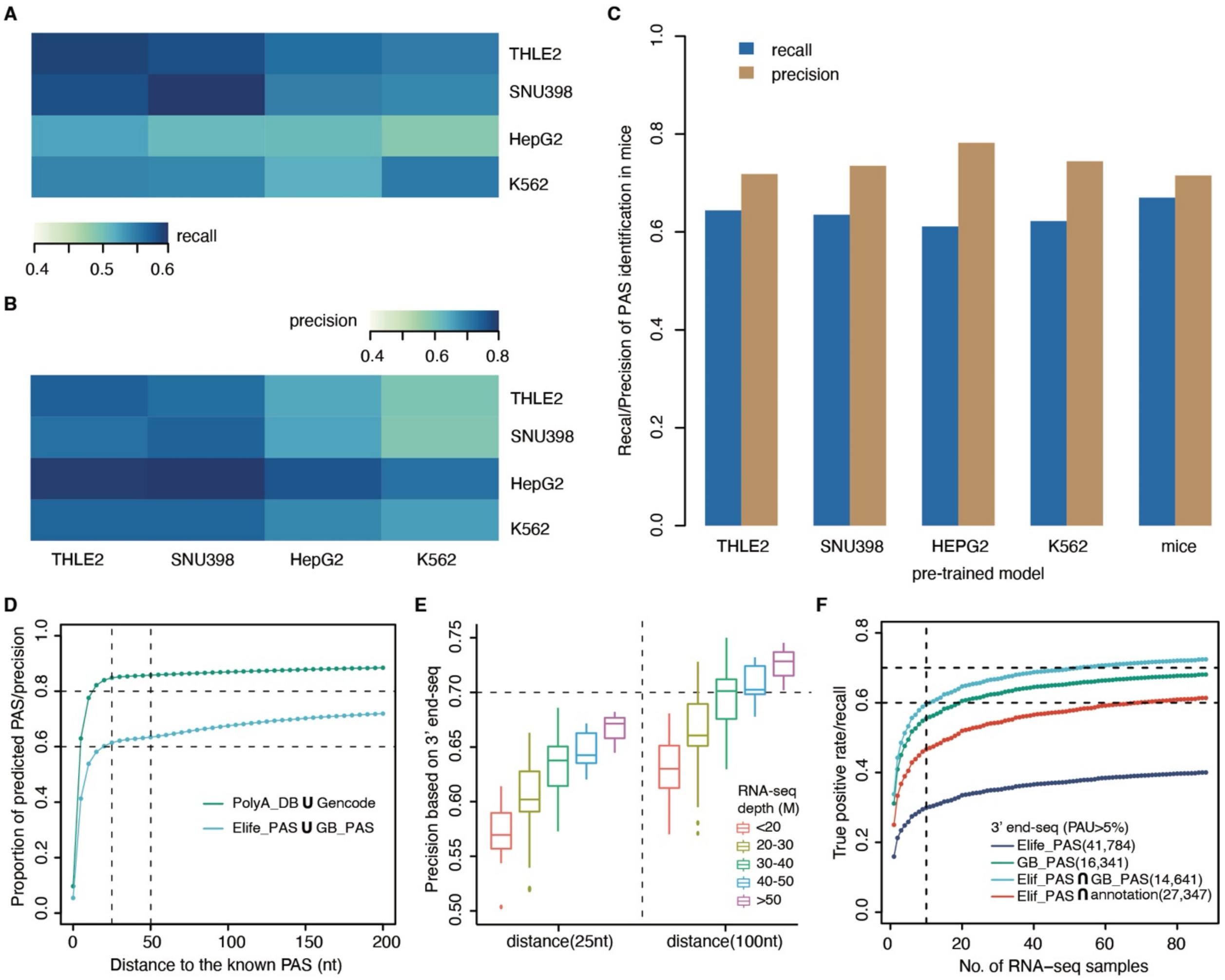

Figure S2: Transfer ability of APAIQ for PAS identification using the pre-trained model. A, B. Heatmap illustrating the recall (A) and precision (B) of PAS identification using the model trained in each cell line and making predictions in the four cell lines. C. recall and precision of PAS predictions on dataset from mouse fibroblast using the pre-trained model from each cell line. D. Precision of the identified PAS by applying to human lymphoblastoid cell line (LCL) based on PAS annotation and PAS derived from 3' end-seq data as the ground truth with different distance thresholds. E. Precision of APAIQ predictions for samples with different sequencing depth with two distance thresholds. F. Recall of APAIQ predictions with different distance thresholds based

on PAS from two source (Elife_PAS and GB_PAS), the intersections between these two sources, and the intersections between Elife_PAS and annotation.

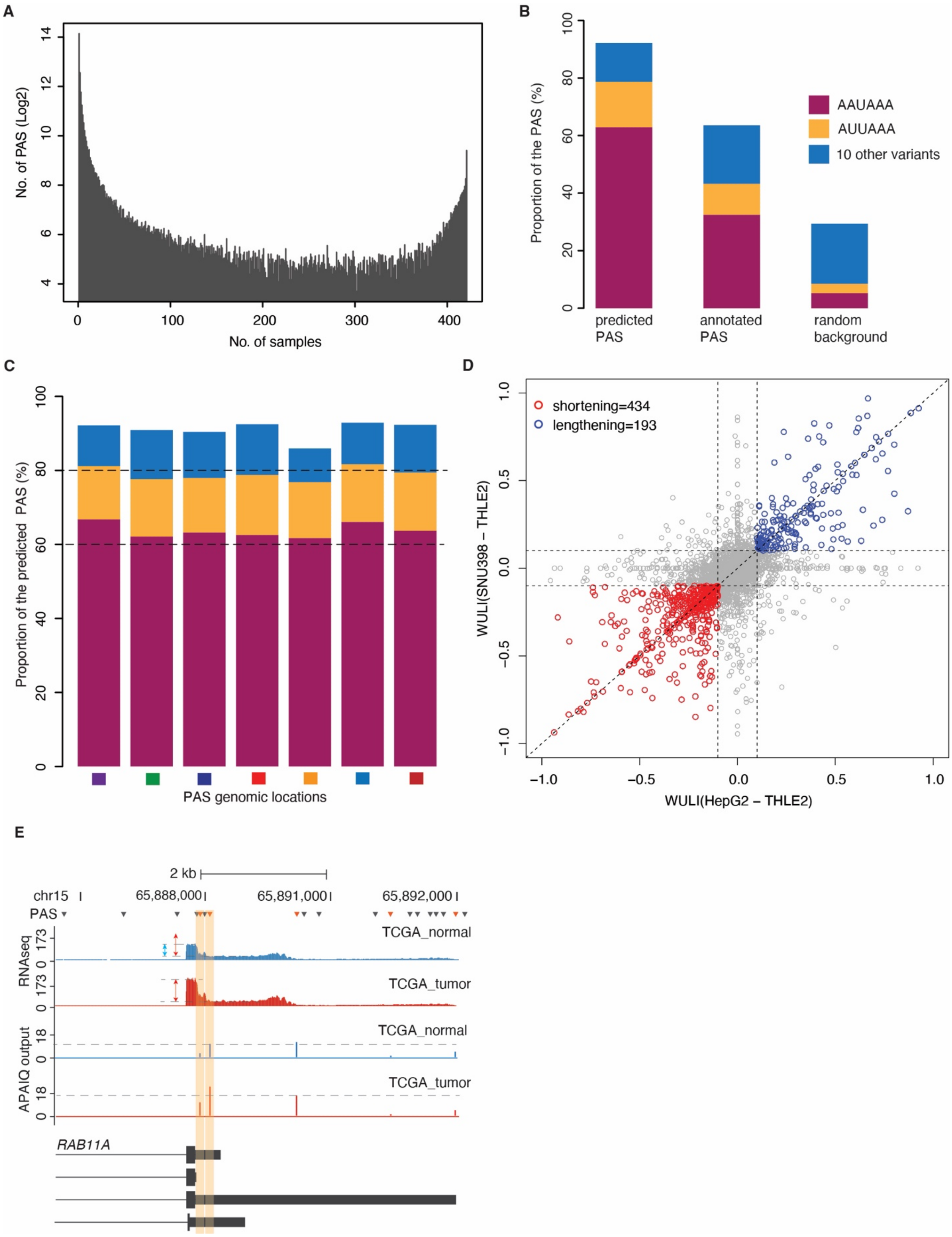
A
No. of PAS (Log2)
4
6
8
10
12
14
0
100
200
300
400
No. of samples
B
Proportion of the PAS (%)
0
20
40
60
80
100
predicted PAS
annotated PAS
random background
AAUAAA
AUUAAA
10 other variants
C
Proportion of the predicted PAS (%)
0
20
40
60
80
100
PAS genomic locations
D
shortening=434
lengthening=193
WULI(SNU398 − THLE2)
WULI(HepG2 − THLE2)
−1.0
−0.5
0.0
0.5
1.0
E
2 kb
chr15
65,888,000
65,891,000
65,892,000
PAS
RNAseq
173
0
TCGA_normal
TCGA_tumor
APAIQ output
18
0
TCGA_normal
TCGA_tumor
RAB11A

Figure S3: Applying APAIQ to RNA-seq data from the liver hepatocellular carcinoma in the cancer genome atlas project (TCGA-LIHC). A. Number of PAS that were identified in different number of samples in TCGA-LIHC dataset. B. Proportion of the PAS with motif AAUAAA, AUUAAA and the other 10 variants. C. Proportion of the PAS with motif AAUAAA, AUUAAA and the other 10 variants in each genomic category. The color of the bar is the same as Supplemental Fig. S3B. The colors below X-axis represent different genomic location categories that has been shown in Supplemental Fig. S1D. D. 3'UTR lengthening and shortening events by comparing WULI of each gene between two liver cancer cell line and one normal liver cell line. X axis indicating the WULI difference between HepG2 and THLE2 and Y axis indicating the WULI difference between SNU398 and THLE2. E. Genome browser illustrating the identified 3'UTR shortening event in tumor compared to normal from gene RAB11A. The two proximal PAS contributing to 3'UTR shortening were highlighted in orange.

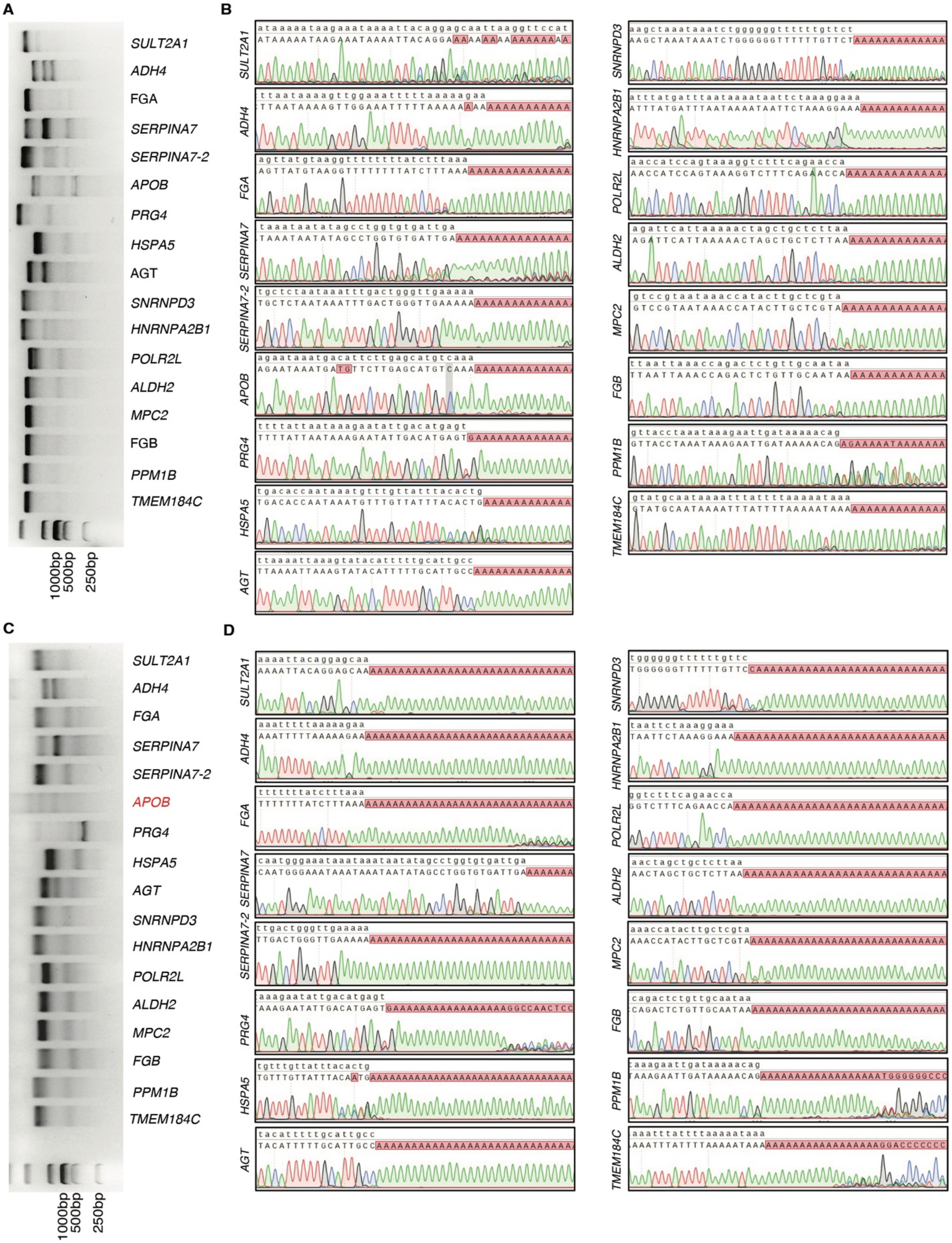

Figure S4: Experimental validation of the predicted PAS. A The gel of 3’ RACE experimental

results for 17 transcripts using the predicted PAS. B Sanger sequencing of the 3’ ends for the amplified transcripts using the predicted PAS. C The gel of ligation-based experimental results for 17 transcripts using the predicted PAS. APOB was marked in red as no clear band were detected in the gel. D Sanger sequencing of the 3’ ends for the transcripts in ligation-based experiment.

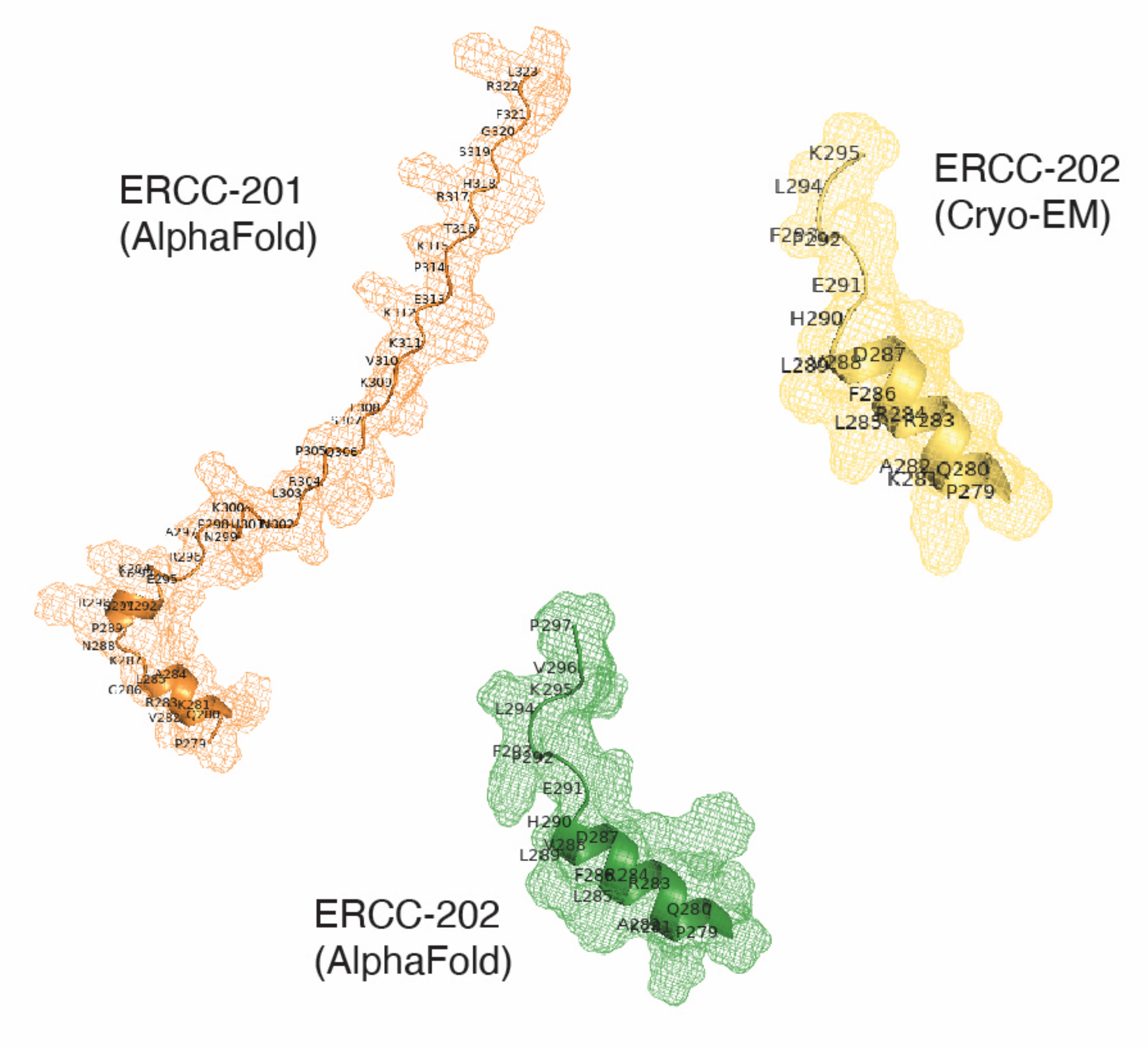

Figure S5: 3D Structure of the C-terminal of two ERCC1 protein isoforms.