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

Yongkang Long, Bin Zhang, Shuye Tian, et al.

Genome Res. published online April 28, 2023
Access the most recent version at doi:10.1101/gr.277177.122

Method

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

Yongkang Long,[1,2,11] Bin Zhang,[1,2,11] Shuye Tian,[3,11] 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,[9] Yvonne Tay,[4,10] Wei Chen,[3] and 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, 23955 Saudi Arabia;* [2]*Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, 23955 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, 518055 China;* [4]*Cancer Science Institute of Singapore, National University of Singapore, 117599 Singapore;* [5]*Shenzhen Haoshi Biotechnology Company, Limited, Bao An District, Shenzhen, 518000 China;* [6]*Computational Systems Biology Lab, Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, the University of Georgia, Athens, Georgia 30605, USA;* [7]*Syneron Technology, Guangzhou, 510535 China;* [8]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190 China;* [9]*Systems Biology Lab for Metabolic Reprogramming, School of Medicine, Southern University of Science and Technology, Shenzhen, 518055 China;* [10]*Department of Biochemistry, Yong Loo Lin School of Medicine, National University of Singapore, 117597 Singapore*

Alternative polyadenylation (APA) enables a gene to generate multiple transcripts with different 3′ ends, which is dynamic across different cell types or conditions. Many computational methods have been developed to characterize sample-specific APA using the corresponding RNA-seq data, but suffered from high error rate on both polyadenylation site (PAS) identification and quantification of PAS usage (PAU), and bias toward 3′ untranslated regions. Here we developed a tool for APA identification and quantification (APAIQ) from RNA-seq data, which can accurately identify PAS and quantify PAU in a transcriptome-wide manner. Using 3′ end-seq data as the benchmark, we showed that APAIQ outperforms current methods on PAS identification and PAU quantification, including DaPars2, Aptardi, mountainClimber, SANPolyA, and QAPA. Finally, applying APAIQ on 421 RNA-seq samples from liver cancer patients, we identified >540 tumor-associated APA events and experimentally validated two intronic polyadenylation candidates, demonstrating its capacity to unveil cancer-related APA with a large-scale RNA-seq data set.

[Supplemental material is available for this article.]

In eukaryotes, the transcription termination is mediated by cleavage of the nascent RNA and followed by the synthesis of non-genomic-templated polyadenosines (poly(A)) to the 3′ end of the RNA, which is known as polyadenylation. This process is controlled by a set of RNA-binding proteins (RBPs) that recognizes *cis* elements surrounding the polyadenylation site (PAS). The PAS motif, a hexamer located 15–40 nt upstream of the cleavage site, is one of the most important core elements for PAS definition (Proudfoot and Brownlee 1976). PAS motifs include AAUAAA that is present in more than half of the PASs, and its variants (AUUAAA et al.), which have been found in nearly 80% of the remaining PASs in the human and mouse genome (Tian et al. 2005). Other elements located within 100 base-pair (bp) flanking PASs also contribute to the formation of polyadenylation (Hu et al. 2005).

Most mammalian genes use multiple sites for polyadenylation, a phenomenon termed alternative polyadenylation (APA), to generate RNA isoforms with different 3′ ends. For instance, >70% of human genes and 60% of mouse genes use multiple PASs (Derti et al. 2012; Xiao et al. 2016). This APA mechanism not only enables a single gene to encode multiple protein isoforms but also greatly increases the complexity of gene expression regulations via different 3′ untranslated regions (3′ UTRs) at the terminal exon. The choice of using a different PAS for each gene is distinct across different cell types. For instance, cells in the brain tissue tend to use the distal PAS to generate long isoforms, whereas proliferating cells prefer short isoforms by using proximal PASs (Sandberg et al. 2008; Miura et al. 2013). Moreover, even for the same type of cells, APA alterations have also been observed under different conditions or upon stimulation (Chang et al. 2015; Zheng et al. 2018).

Dysregulation of APA could be associated with human diseases, including cancer. It has been shown that APA-mediated 3′ UTR shortening could activate oncogene expression via escaping from microRNA regulation, whereas RNA transcripts with shortened 3′ UTRs could also inhibit other transcripts from tumor suppressor by disrupting the competition for shared microRNA binding (Mayr and Bartel 2009; Park et al. 2018). Therefore, it is of the utmost interest to identify the expressed/used PASs and quantify their usages for each gene in different cells and samples. Numerous computational

[11]**These authors contributed equally to this work.**
Corresponding authors: xin.gao@kaust.edu.sa, chenw@sustech.edu.cn, zb.picb@gmail.com

Article published online before print. Article, supplemental material, and publication date are at https://www.genome.org/cgi/doi/10.1101/gr.277177.122. Freely available online through the *Genome Research* Open Access option.

methods have been developed to predict PASs solely based on DNA sequence (Cheng et al. 2006; Xie et al. 2013; Xia et al. 2019), but they are very unlikely to accurately identify sample-specific APA simply because different cells share almost the identical DNA sequence. On the other hand, experimental technologies have been developed for the identification and quantification of sample-specific APA events in a transcriptome-wide manner by enriching the 3′ end of RNA transcripts, followed by high-throughput sequencing (3′ end-seq), such as 3P-seq, Aseq, Poly(A)-seq, and 3′ READS (Jan et al. 2011; Derti et al. 2012; Martin et al. 2012; Hoque et al. 2013). However, these methods are rather laborious, and more time and/or material consuming compared to conventional RNA sequencing (RNA-seq). More importantly, 3′ end-seq data are almost impossible to be used for studying other post-transcriptional RNA processes beyond APA, such as splicing and RNA editing.

Accordingly, many tools have also been developed for the identification and quantification of APA based on RNA-seq data. Almost all of them are designed to detect the drops in RNA-seq read coverage along the gene body. However, due to high fluctuation of the sequencing coverage potentially caused by heterogenous reads mappability, as well as biased amplification efficiency of different RNA fragments during sequencing library preparation, these methods tend to suffer from high false positive rates and low recall on PAS identification. A recent study benchmarking multiple computational tools for APA analysis, including TAPAS (Arefeen et al. 2018), QAPA (Ha et al. 2018), DaPars2 (Li et al. 2021), GETUTR (Kim et al. 2015), and APATrap (Ye et al. 2018), using 3′ end-seq and Pacific Biosciences (PacBio) Iso-Seq data, found that essentially none of them could achieve 50% recall, whereas the false positive rate (FPR) ranges from 30% to 50% (Shah et al. 2021). In addition, these computational methods mainly focus on APA within 3′ UTR. This might be due to the fact that drop of RNA-seq coverage across exon–intron boundaries impacts their detection of splicing-coupled APA, such as intronic polyadenylation (IPA). Nevertheless, it has been reported that IPA is widespread in leukemia to inactivate tumor suppressor by generating truncated protein isoforms (Lee et al. 2018).

Recently, IPAFinder that specifically identifies IPA and compares its usage between two conditions from RNA-seq has been developed (Zhao et al. 2021b). However, a method for comprehensive PAS identification at transcriptome remains largely an unmet need. In addition to PAS identification, current methods for APA quantification also have substantial error rates and they usually rely on the annotation of UTR and PAS in the database, such as QAPA and APAlyzer (Ha et al. 2018; Wang and Tian 2020). To date, the most comprehensive PAS annotation data sets, including PolyA_DB and PolyASite, are mainly derived from common cell lines or tissues (Wang et al. 2018; Herrmann et al. 2020). There might be tremendous unannotated PASs used in less explored biological samples, such as primary tumor samples from individual patients.

Here, to address the above-mentioned limitations, we developed APAIQ, a computational tool that is capable of sample-specific APA identification and quantification (APAIQ) from RNA-seq data in a transcriptome-wide manner.

Results

Identification and quantification of APA from RNA-seq data with APAIQ

APAIQ integrates coverage information from RNA-seq data with genomic sequence through a convolutional neuron network (CNN). It contains two modules, including one for PAS identification and another one that is a regression model for the quantification of expression based on the coverage around the identified or provided PAS position. For PAS identification, a hybrid deep-learning model taking RNA-seq read coverage and DNA sequence was first implemented to predict PAS score at each genomic locus, followed by a customized postprocessing strategy to identify accurate PAS position. To quantify the expression of transcripts using each PAS, a regression model was further trained by taking RNA-seq coverage at its flanking regions (−500 bp to 500 bp) as independent variables/covariates and the expression quantified by 3′ end-seq as dependent variable/response (Methods; Fig. 1A).

We performed RNA 3′ end-seq (QuantSeq 3′ mRNA-seq, Lexogen) on four cell lines, including K562, HepG2, THLE2, and SNU398, to comprehensively characterize the expressed/used PAS in these four samples, respectively. To avoid internal priming potentially caused by the 3′ end-seq, we only used the annotated PAS from GENCODE and PolyA_DB (v3) that is derived from 3′ READS, which claimed to have the ability to overcome the internal primming problem (Hoque et al. 2013). An average of 20,335 PASs in each cell line with a sufficient expression level (reads per million [RPM] > 0.1 and PAU > 0.05) was identified (Methods), in which 10,382 PASs were commonly used among the four cell lines and each cell line has 3500–4000 sample-specific PASs (Fig. 1B). Among the 35,395 PASs expressed in at least one cell line, 35,064 are located within 14,959 protein-coding genes (without any extension), in which 58.9% of the genes use multiple PASs (Fig. 1C). We noted that the RNA-seq read coverage indeed tended to drop downstream from these PASs compared to the upstream regions (Supplemental Fig. S1A). In addition, 87% of the used PASs have canonical PAS motif (AAUAAA) and its variants are located within 100 bp upstream of the cleavage site, whereas the frequency is only ∼20% in the same number of randomly selected genomic loci are at least 50 bp away from any annotated PAS (Supplemental Fig. S1B). These suggested that both RNA-seq read coverage and DNA sequence contain useful information for distinguishing true used PASs from the background.

APAIQ predicts PAS accurately in a genome-wide manner

We first built a binary classification model by integrating RNA-seq coverage and DNA sequence. Using the expressed/used PASs in each cell line as the positive data set, and the same number of randomly selected genomic loci (see definition above) as the negative data set (Methods), we trained a model for binary classification. To evaluate the performance, we applied a cross-validation approach, in which we trained the model using 4/5 of the data sets and made predictions on the rest. As expected, the integrated model achieved 96% TPR/recall with false positive discovery rate (FDR) <7%, which is better than the model that used only RNA-seq coverage (coverage-only) and slightly better than the model that used only DNA sequence (sequence-only) (Supplemental Table S1). It is noteworthy that the sequence-only model achieves a performance similar to that of the integrated model, possibly due to the fact that the negative data set consists of the random regions rather than the unused/unexpressed PASs in each cell line.

Although the model achieves good performance in distinguishing true PASs from random genomic loci, PAS identification at the whole genome remains challenging because even a very low false positive rate would result in a huge number of false positive predicted loci. For instance, among 3 billion loci in the human genome, only around 20,000 of them were defined as expressed/used

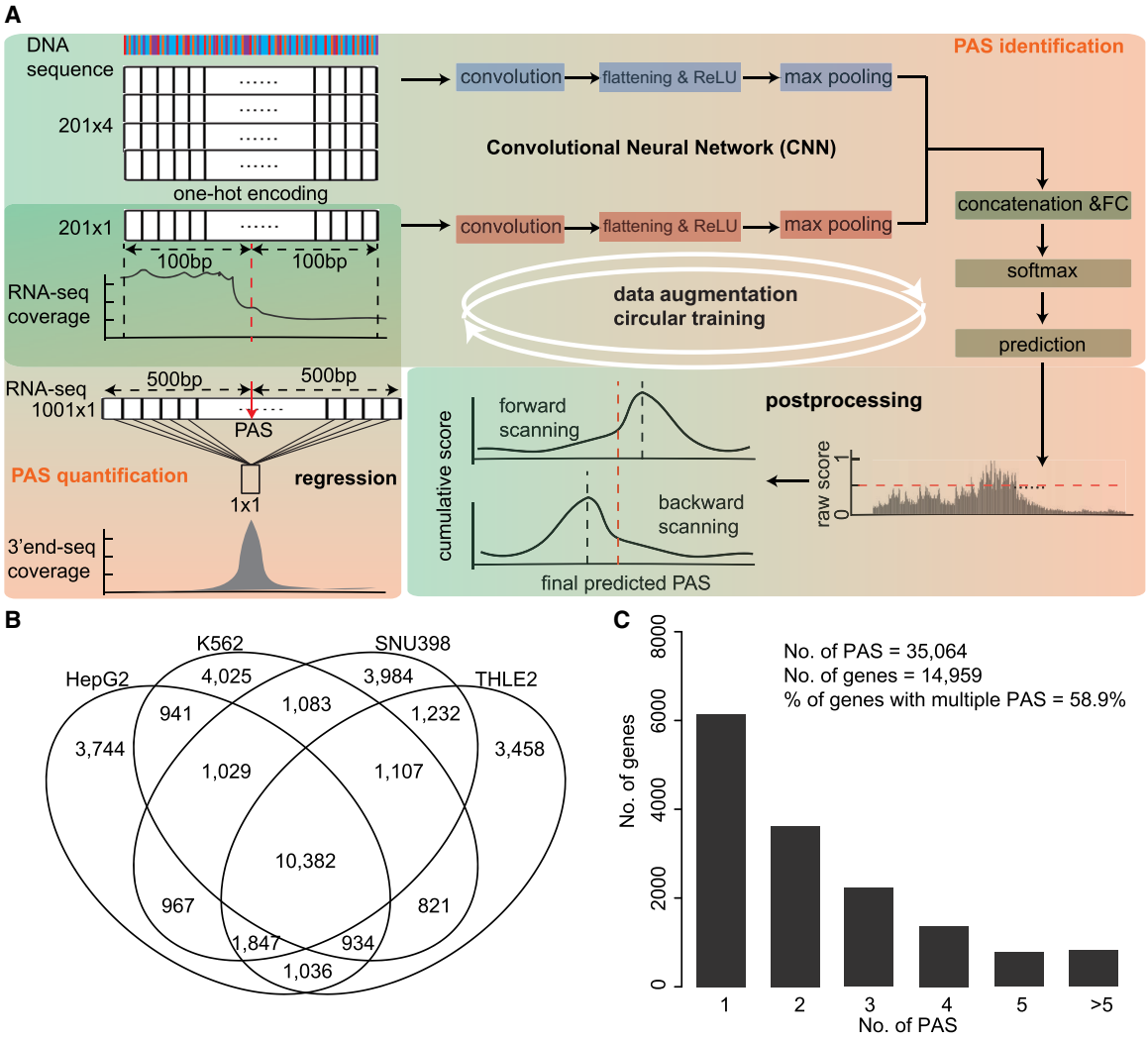

Figure 1. A hybrid deep-learning model for PAS identification. (*A*) A schematic illustration of the framework of APAIQ. (*B*) Venn diagram shows the number of PASs with sufficient expression level in each cell line and their overlaps. (*C*) Numbers of genes with different numbers of the identified PAS.

PASs with sufficient expression level in each single cell line. To further reduce the false positive rate, we developed an enhanced model for genome scanning by introducing three strategies. We first limited the loci for the prediction by requiring the average of RNA-seq coverage at its upstream 100 bp to be higher than RPM of 0.1. In this way, we reduced the total number of scanning loci from 3 billion to several million. Second, we introduced a data augmentation strategy to increase the complexity of the training data set by randomly shifting several bp from true PAS (−12 bp to 12 bp) to augment the positive data set. Finally, we implanted a circular-training-based method, in which we repeatedly replaced the negative data set by the false positive predictions from the previous round, which further increased the power of the model to distinguish positive from negative PAS (Fig. 1A; Methods).

With the enhanced model using these three strategies, we made predictions at each locus in each sample and got a raw prediction score ranging from 0 to 1 that represented the probability of a locus to be a truly used/expressed PAS in this sample. As we considered the locus at the vicinity of each true PAS as positive during data augmentation, the prediction score at these loci could also be relatively high. Accordingly, we designed a postprocessing strategy by scanning the prediction scores to pinpoint the final positions of the true PAS. In this way, we were able to accurately pinpoint the final PAS position (Fig. 1A; Methods). At the end, our model was able to identify ~70% of true PAS (overlapped with predicted PAS within 25 bp) and >80% of the PAS identified by APAIQ was located within 25 bp away from the annotated PAS (3′ end-seq RPM > 0), suggesting that our method achieves both high recall and precision (72.4% with PAS from the ground truth in average) for PAS identification (Fig. 2A). As expected, we also found that the PAS with higher expression levels has higher probabilities to be identified. For instance, >80% of PASs with relatively high expression level (RPM > 5, PAU > 0.05) were successfully identified (Fig. 2B). Even though 20%–30% of the PAS we identified are not from the ground truth, almost half of them (10%–15% of the total identified PASs) are still annotated (within 25 bp), but the expression levels did not pass our threshold (Fig. 2C). Overall, ~85% of the predictions are PAS detected by 3′ end-seq.

As shown in Figure 2D, as an example, there are more than 19 annotated PASs from gene *ELP5*. However, only one was expressed/used in K562 and THLE2, and two of them were expressed/used in HepG2 and SNU398 according to the 3′ end-seq

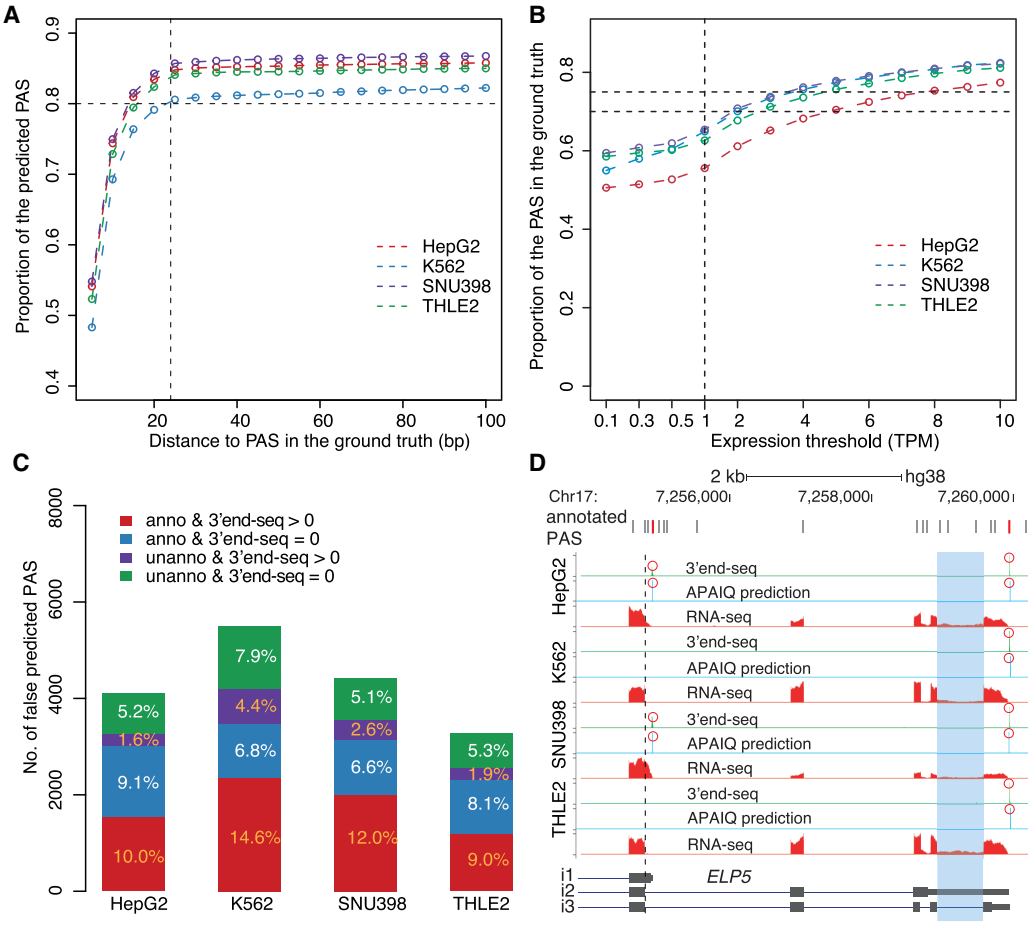

Figure 2. APAIQ predicts PAS comprehensively and accurately. (*A*) Proportion of the predicted PAS within different distances from the PAS in the ground truth. (*B*) The proportion of PAS above different expression thresholds that have been identified by APAIQ. (*C*) Barplot showing the number of the identified PASs that are not overlapped with ground truth divided into four categories. (*D*) Genome Browser illustrating the predicted PAS and the PAS identified by 3′ end-seq in four cell lines.

data. Our model successfully identified them in each cell line, including an intronic PAS that is specifically expressed/used in HepG2 and SNU398 (*ELP5*-i1). The IPA has also been identified by IPAFinder in a previous study (Zhao et al. 2021b), indicating the high reliability of this PAS. To be noted, solely using 3′ end-seq data enabled the identification of these two sample-specific APA events, whereas it was impossible to distinguish isoform *ELP5*-i3 from *ELP5*-i2 because they used the same terminal PAS, whereas *ELP5*-i2 has a retained intron (Fig. 2D). In comparison, by applying APAIQ on RNA-seq data, we could not only identify these two PASs but also retain the ability to determine the splicing patterns of different isoforms, which further highlighted the power of APAIQ in the study of APA coupled with other post-transcriptional RNA processes, such as splicing.

APAIQ outperforms current methods on PAS identification

Next, we compared APAIQ with several publicly available methods, including Aptardi, DaPars2, mountainClimber, and SANPolyA (Cass and Xiao 2019; Yu and Dai 2020; Li et al. 2021; Lusk et al. 2021), by using our 3′ end-seq data as the benchmark. All these methods were published recently, and they represented different strategies for the identification of PASs, including purely using RNA-seq coverage, DNA sequence, as well as using both coverage and sequence. Among them, DaPars2 and mountainClimber were based on the drops in RNA-seq read coverage, whereas SANPolyA used only DNA sequence and Aptardi used a deep-learning framework with both RNA-seq coverage and DNA sequence features. As SANPolyA was designed for binary classification only, for fair comparisons, we applied our postprocessing strategy to enable it also for genome-wide scanning. First, we performed a simple comparison by using the top 10,000 predictions from each method. We found that APAIQ captured the highest number of true PASs and the distance between the predictions and the ground truth is mainly within 25 bp. Aptardi ranks as the second if the threshold of distance is set to 100 bp (Fig. 3A). As these two methods shared a similar strategy, this result suggests that using deep-learning and integrating RNA-seq coverage and DNA sequence could greatly improve the performance compared with other traditional methods.

Next, we used 25 bp to the ground truth as the distance threshold to calculate recall, FPR, and precisions for more comprehensive evaluation (Supplemental Method S1). As shown in Figure 3B, none of the used published methods achieved 60% recall with 10% FPR, which is consistent with the recent study that benchmarked computational methods for APA analysis based on 3′ end-seq and Iso-Seq data (Shah et al. 2021). As Aptardi scanned

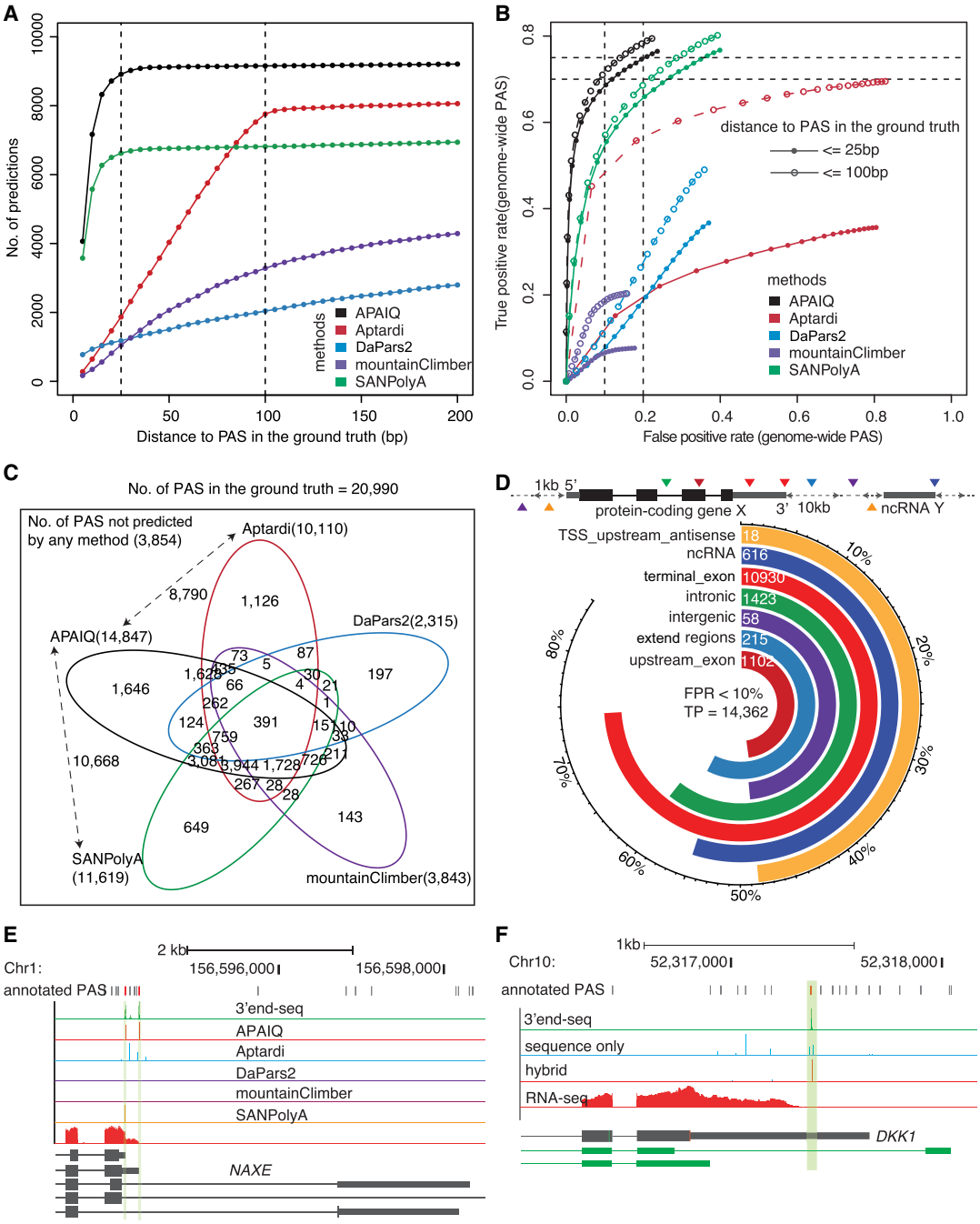

Figure 3. APAIQ outperforms other methods on PAS identification. (*A*) Number of the top 10,000 predicted PASs within different distances from the PAS in the ground truth. (*B*) ROC curves showing the performance of the APAIQ and four other published methods, including Aptardi, DaPars2, mountainClimber, and SANPolyA. (*C*) Venn diagram showing the number of true positives predicted by APAIQ and the four published methods. The number of overlaps between APAIQ and Aptardi, APAIQ, and SANPolyA were indicated by arrows. (*D*) Different categories of the PAS identified by APAIQ. (*E*) An example showing that the IPA event was successfully identified by APAIQ, but not the other three methods. (*F*) Genome Browser showing predictions of PAS from the genes *DKK1* 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).

PAS at a 300 bp window with a 100 bp stepwise, we further relaxed the distance threshold to 100 bp to the ground truth for the evaluation. In this way, Aptardi exhibited a much better performance, whereas its recall was still much lower (~20% lower) than APAIQ with the same FPR, and the same hold (~10% lower) for SANPolyA method coupled with our postprocessing strategy. As both Aptardi and DaPars2 were designed for APA analysis within the last exon, we further evaluated them by only using the PAS within the terminal exon. Again, APAIQ has the highest recall with the same FPR (Supplemental Fig. S1C).

We further analyzed the true positive (TP) predictions of the five methods and found that 18.4% of the PASs in the ground truth were not found by any method. As expected, among those TP specifically predicted by each method, APAIQ has the highest number (1646), followed by Aptardi (1126). Both the TP predicted by Aptardi and SANPolyA are highly overlapped with APAIQ, indicating high consistencies between the deep-learning-based methods (Fig. 3C). Moreover, we classified the PASs into seven categories based on their genomic localization, including terminal exon (including 3′ UTR), upstream exonic regions, intronic regions, transcription start site (TSS) upstream antisense, extended regions, ncRNA, and intergenic regions. Similar to the previous study (Xiao et al. 2016), PAS located within the terminal exon was the most abundant category (73.7%), followed by intronic PAS (17.0%). We found that APAIQ was able to identify PAS from all these categories, whereas PAS from the terminal exon and intronic region has the highest identification rates (Fig. 3D), indicating a higher generalization ability compared to other methods that only focus on 3′ UTR. We also checked the precision of the predictions in each category region and found that the precision for PAS from terminal exons is above 0.8 and precisions for PAS from other genic regions, including intron, upstream exon, and ncRNA, are around 0.6 (Supplemental Fig. S1D). Moreover, the frequency of AAUAAA and other poly(A) signal motifs upstream of the predicted PAS are quite consistent across different types of regions, with ∼50% of them containing AAUAAA, and 80% containing at least one poly(A) signal motif (Supplemental Fig. S1E). PAS from TSS upstream antisense, extended regions, and intergenic regions have relatively low precisions and motif frequency, which might be due to unstable and low expression of RNA products using these PASs. However, this would not impact the general APA analysis, which mainly focuses on PAS from genic regions. Finally, as shown in Figure 3E, only APAIQ was able to successfully identify two PASs from intronic regions of gene *NAXE*.

Synergistic effect of RNA-seq read coverage and DNA sequence on PAS identification

To dissect the contribution of RNA-seq read coverage and DNA sequence to the model, we also compared the integrated model with the coverage-only and the sequence-only models. We found that using two features not only improved precision but also reduced false positive predictions (Supplemental Fig. S1F). As shown in Figure 3F, the sequence-only model identified several annotated PASs from gene *DKK1* that are not expressed/used at all, whereas when adding RNA-seq coverage features, the integrated model successfully discarded those unused PASs. Another example was shown in Supplemental Figure S1G, in which using only RNA-seq coverage falsely identified two PASs from gene *ACOT2* due to the drop of coverage, whereas they were successfully excluded by the integrated model. In addition, a proximal PAS from gene *PHKB* truly used in SNU398 was identified by neither the sequence-only nor the coverage-only model, but was successfully detected by the integrated model (Supplemental Fig. S1H), further demonstrating a synergistic effect by combining these two features together for PAS identification.

High transferability of APAIQ across different cell lines, species, and data sets

To examine whether APAIQ could be applied to samples differing from that used in the training data set, we tested the transferability of our model across different cell lines. Alternately, we trained the model in each of the four cell lines, made predictions on the other three cell lines, and then compared their performances to that in this cell line. We found that the predictions across different cell lines could still achieve a similar recall and precision as that over data from the same cell line (Supplemental Fig. S2A,B). To further test whether APAIQ could be even applied across different species, we tested the model on a data set from the mouse fibroblast (Xiao et al. 2016). The performance is similar to those across different human cell lines (Supplemental Fig. S2C), indicating the generalization power of our method and suggesting potential applications on different RNA-seq samples using the pretrained model.

To further confirm the transfer capacity of APAIQ using the pretrained model, we applied APAIQ to an independent public data set (Supplemental Method S2), which has been used to benchmark computational methods for APA analysis (Shah et al. 2021). It turned out that >80% of the predicted PASs are within 25 nt to the PAS from annotation (union between GECODE and PolyA_DB), whereas >60% of the predicted PASs are located within 25 nt to PAS in the ground truth and ∼70% are within 100 nt (Supplemental Fig. S2D). This is slightly lower than the precision achieved from the four cell lines (precision of around 0.72 using 25 nt as the threshold). We noticed that in general the sequencing depth of these LCL RNA-seq samples are lower than that of the four cell lines (∼100 M). Thus, we inspected the precision in samples with different sequencing depth. We found that the precision is positively correlated with sequencing depth, for which the samples with sequencing depth higher than 50 M have precisions above 0.7 using 100 nt as the threshold (Supplemental Fig. S2E).

Because the PASs in the ground truth are derived from the integration of more than 50 3′-seq samples, we further checked the recall by cumulating the predicted PASs from multiple RNA-seq samples. With the increasing number of RNA-seq samples, the predictions could reach 70% recall for the PASs shared by Elife_PAS and GB_PAS (Supplemental Fig. S2F). The relative low recall for Elife_PAS is likely due to its high false positive rate as only 65.4% (27,347 out of 41,784) of the Elife_PAS are overlapped with annotation. Overall, using LCL data as an independent resource, we showed that APAIQ could achieve similar performance as that from the four cell lines with our pretrained model, indicating the robustness and good transfer capacity of APAIQ for PAS identification based on canonical RNA-seq data.

Accurate quantification of PAS usage by APAIQ

After the identification of PASs, we further aimed to quantify the expression of each PAS based on RNA-seq data. To do so, we introduced a regression model, in which we considered the expression of the PAS (Y) as a dependent variable/response and the coverages at each locus within 500 bp flanking the PAS (X) as predictors/covariates. The expression Y was estimated by using 3′ end-seq data and the coverage X was derived from RNA-seq data (Fig. 1A). Our prediction achieved an average Pearson correlation coefficient of 0.75 in the four cell lines. Furthermore, we trained the model from each cell line and made predictions in the other three cell lines separately. It turned out that the correlation coefficients derived between different cell lines were comparable to those from the same cell line. This result is expected as RNA-seq coverage around each PAS should be determined by its expression profile in each sample and the features in coverage variations learned by the model should be consistent across different samples. Overall, this result also suggests a good transferability of the regression model (Fig. 4A).

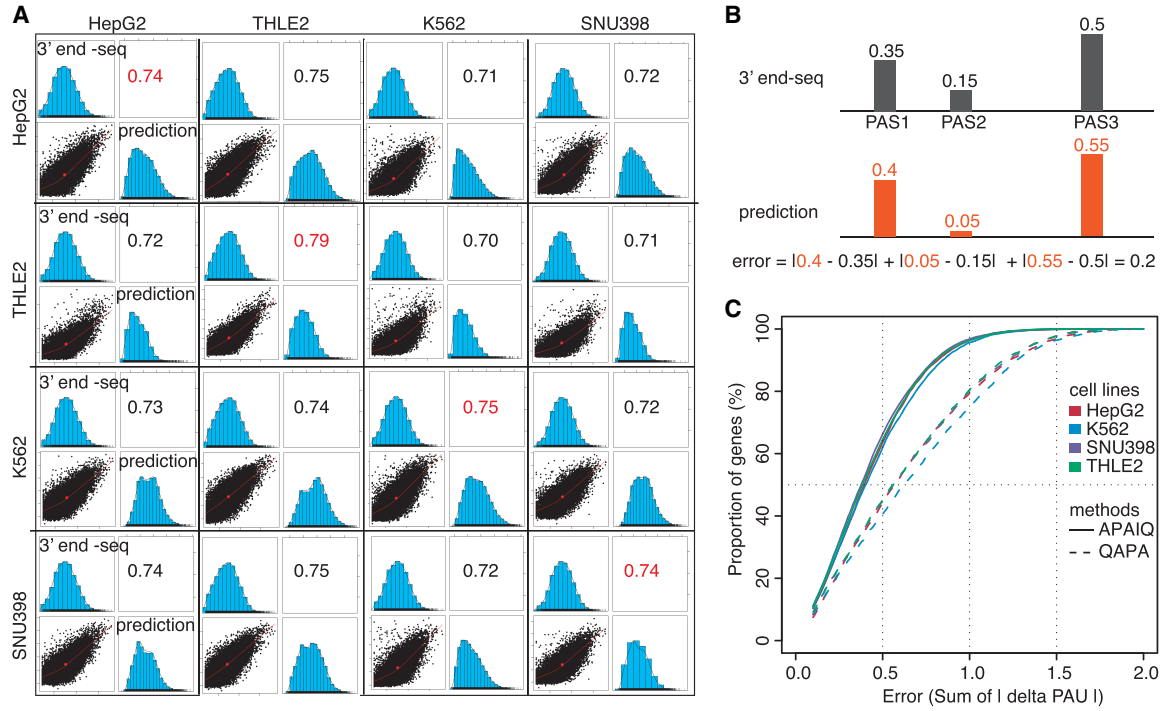

Figure 4. Accurate quantification of APA by APAIQ. (*A*) Correlation of the expression of the PAS predicted by APAIQ and that quantified by 3′ end-seq data. The row indicates cell lines in which the model was trained, whereas the column indicates cell lines in which the prediction was performed. (*B*) Schematic illustration of the calculation of the error of the PAS usage (PAU) between the prediction from APAIQ and that quantified by 3′ end-seq data. (*C*) Cumulative plot of the error of PAU predicted by APAIQ and QAPA in the four cell lines.

Among an average of 11,010 genes expressed in each cell line, ~50% of them used multiple PASs. For these genes, we calculated the PAU by using its expression divided by the total expression of all the PASs from the corresponding gene. In a recent study, a metric (error) was introduced by summing up the PAU difference across all the PASs from each gene, to measure the concordance of PAU between prediction and quantification by 3′ end-seq data (Fig. 4B). With this method, it showed that QAPA achieved the lowest error in APA quantification (Shah et al. 2021). Therefore, we further calculated the error of our predictions and compared to the results from QAPA. It turned out that APAIQ could predict PAU faithfully, with a much lower error rate than QAPA (Fig. 4C).

Application of APAIQ to a TCGA RNA-seq data set identifies tumor-associated APA events

Finally, we applied APAIQ on 421 RNA-seq samples (371 tumor and 50 adjacent normal) from a liver hepatocellular carcinoma cohort in The Cancer Genome Atlas (TCGA-LIHC). An average of 11,432 PASs were identified in each sample and significantly more (P-value < 3.63×10^{-15}, Wilcoxon test) PASs were identified in tumor samples than that in the adjacent normal samples (Fig. 5A), suggesting higher transcriptional activities, abnormalities, and/or heterogeneities of cells in tumor samples. Among these identified PASs in each sample, 91% are overlapped with annotation (within 25 bp away from the annotation), indicating the high quality of the identified PASs.

By checking the sample-wise heterogeneity of the identified PASs, we observed a typic bimodal distribution, in which most PASs were either only detected in very few samples or detected in almost all the samples (Supplemental Fig. S3A). Among them, 47,949 PASs were detected in more than one sample, including 36,319 annotated and 11,630 unannotated ones. Similar to the results from the cell lines (Fig. 3B), we were able to identify PASs from all categories of genomic regions, in which the PASs from terminal exon and intronic regions are the most abundant (51.8% and 18.4% among 47,949 identified PASs). Compared to the identified PASs that are overlapped with annotation, those unannotated ones are most frequently located within intronic regions (30.2%), with only 16.2% from terminal exon, suggesting that PASs within terminal exon might be better annotated than those from other regions, and IPA annotation needs improvement (Fig. 5B). Higher frequency of unannotated PASs identified in intronic regions could also be due to the fact that the precision for the predicted PASs in intronic region is lower than that in terminal exon. To further evaluate the quality of these identified PASs, we checked the frequency of canonical PAS motif (AAUAAA) and its variants and found that >90% of the identified PASs have the motifs, which is even higher than that from annotation (Supplemental Fig. S3B). In addition, the motif frequency for predictions across multiple genomic locations is quite similar (Supplemental Fig. S3C), which is consistent with the results in the cell line.

To quantify APA events and compare them between tumor and normal, we further applied APAIQ to calculate the expression level of each PAS. In total, 36,904 PASs from 15,326 protein-coding genes were sufficiently expressed (predicted RPM > 1) in more than one sample. Among these genes, 59.3% used multiple PASs (Fig. 5C), which is slightly higher than 46% that has been reported in a recent study based on 3′ end-seq data from lung cancer patients (Zingone et al. 2021). For PASs located within terminal exon, we defined weighted 3′ UTR length index (WULI), a metric

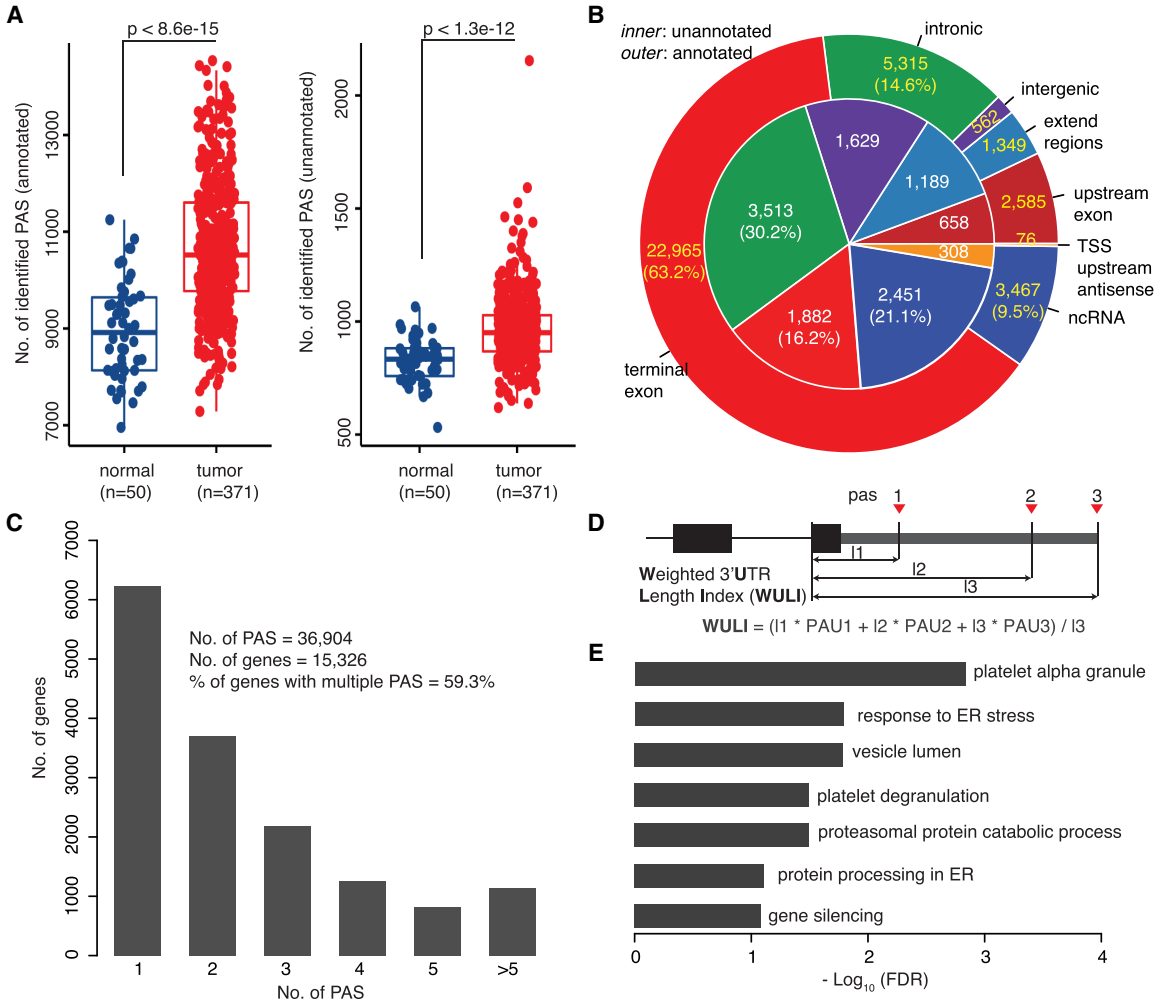

Figure 5. Applying APAIQ on a large-scale RNA-seq data set identifies tumor-associated APA events. (*A*) Boxplot showing the number of annotated and novel PASs identified by APAIQ in each RNA-seq sample from TCGA-LIHC. *P*-values were derived from Wilcoxon test. (*B*) Proportion of the identified PASs from a different genomic category. The *outer* circle indicates the identified PASs that are overlapped with the annotation and the *inner* pie presents the identified PASs that are not overlapped with the annotation. (*C*) Numbers of genes with different numbers of the identified PASs. (*D*) A schematic illustration of the calculation of weighted 3′ UTR length index (WULI). PAU stands for PAS usage and the l1, l2, and l3 represent the length from start to the terminal exon to each PAS. (*E*) Barplot showing the BH-adjusted *P*-value (FDR) of Gene Ontology terms and KEGG pathways enriched for genes with 3′ UTR shortening in tumor compared to normal.

to evaluate 3′ UTR length for each gene in each sample (Fig. 5D). Compared to a previous method using only the two most highly expressed PASs (distal and proximal) to measure 3′ UTR lengthening and shortening (Hoque et al. 2013), WULI calculated an average of lengths by taking into account both the UTR length and usage of each isoform. To check the feasibility of the method, we first applied WULI to identify 3′ UTR shortening and lengthening events by comparing two liver cancer cell lines (HepG2 and SNU398) to normal liver cell line (THLE2) separately. A consistent 3′ UTR shortening trend was observed in these two comparisons (HepG2 vs. THLE2 and SNU398 vs. THLE2) (Supplemental Fig. S3D).

Next, we analyzed the data from 50 matched tumor and normal samples, in which 245 shortening and 155 lengthening events were identified. As shown in Supplemental Figure S3E as an example, there are 16 annotated PASs located within the terminal exon of gene *RAB11A* and four of them have been used. The identified 3′ UTR shortening was mainly contributed by the increased usage of two proximal PASs in tumor samples compared to that in normal samples. We found that those genes with 3′ UTR shortening were enriched for functions related to platelet alpha granule and response to endoplasmic reticulum (ER) stress (Fig. 5E), whereas no significant functional enrichment was identified for genes with 3′ UTR lengthening. ER stress is frequently observed in cancer, including hepatocellular carcinoma (HCC), due to its high demand of protein synthesis. Our finding is consistent with previous observations that genes related to the response of ER stress are activated in HCC and correlated with poor prognosis (Shuda et al. 2003; Pavlović and Heindryckx 2021). On the other hand, platelet alpha granule might be critical for tumor favored microenvironment (Pavlovic et al. 2019).

For each intronic PAS, we calculated its PAU and compared that between tumor and normal samples. As a result, 68 up-regulated and 73 down-regulated IPA events were identified and the genes with the dysregulated IPA are enriched in protein activation cascade GO term (GO:0072376, odds ratio = 9.86, FDR = 0.001) and

complement and coagulation cascades pathway (hsa04610, odds ratio = 8.32, FDR = 0.001).

Validation of the predicted tumor-associated APA events in a liver cancer cell line

To experimentally validate APAIQ predictions, we selected 20 identified PASs that are not overlapped with any annotated transcript 3′ ends and performed 3′ RACE experiments to examine whether these PASs were expressed/used in a liver colon cancer cell line, HepG2 (Supplemental Table S2). Seventeen out of 20 of them were successfully validated and the failed three candidates might be due to their relative low expression level in the cell line (Supplemental Fig. S4A,B). Furthermore, we conducted a ligation-based assay and confirmed that 16 out of the 17 candidates indeed have poly(A) tail, whereas the only failed one has the lowest expression among these 17 candidates (Supplemental Fig. S4C,D).

In addition, among the tumor-associated IPA events, the top significant one enriched in cancer is from the gene flavin adenine dinucleotide synthetase 1 (*FLAD1*), which is overexpressed in multiple cancer types, including gastric, breast, and liver cancer, and correlated with poor prognosis (Jia et al. 2019; Hu et al. 2020; Ye et al. 2020). We found that the identified IPA is barely or almost not used in normal samples, whereas usage of the IPA in tumor samples is comparable to the canonical PAS (Fig. 6A). Compared to the full-length isoform, transcripts with this IPA encode a much shorter protein isoform with distinct amino acids (294 amino acids vs. 587 amino acids). The canonical protein isoform encoded by *FLAD1* is a key enzyme in flavin adenine dinucleotide biosynthesis, whereas for the short protein isoform, it awaits future investigation to explore its functions in cancer.

Another IPA candidate that is highly used in tumor samples but barely used in normal samples is from the gene *ERCC1*, which encodes a protein with DNA repair functions (Fig. 6B). The transcripts using this intronic PAS could generate a protein with a varied C-terminal (ERCC1-201) compared to canonical protein isoforms (ERCC1-202). Using Alpha-fold (Jumper et al. 2021), we predicted structures of these two proteins-isoforms and found that a typic alpha helix structure at the C-terminal of ERCC1-202 would be disrupted in ERCC1-201 (Supplemental Fig. S5), which might impact its function. Indeed, a previous study reported that the C-terminal of ERCC1-202 contains the domain for double-strand DNA binding and interacts with its cofactor ERCC4 (previously known as XPF), and therefore only ERCC1-202 has full capacity for nucleotide excision repair (Friboulet et al. 2013). These results suggested that tumor cells with IPA of *ERCC1* would have dysregulated DNA repair functions, which might contribute to HCC development.

Finally, we validated these two candidates by performing a 3′ rapid amplification of cDNA ends (3′ RACE) experiment in the liver cancer cell line HepG2. As shown in Figure 6C–F, we successfully detected the IPA transcript from both genes. Moreover, using real-time quantitative reverse transcription PCR (qRT-PCR), we showed that the relative expression levels of the transcript isoforms using IPA compared to the canonical ones are consistent with what we estimated by APAIQ with tumor samples from the TCGA-LIHC data set (Fig. 6C,D). These results further confirmed the reliability of APAIQ. Taken together, our results demonstrated that APAIQ is a powerful tool to facilitate APA analysis and its functional understanding using large-scale public RNA-seq data sets, such as that from TCGA, the Genotype-Tissue Expression (GTEx), and the ENCyclopedia Of DNA Elements (ENCODE) projects.

Discussion

There are only around 20,000 genes in the human genome, whereas 10 times or even more numbers of transcripts were transcribed from these gene loci, which are mainly mediated by a series of RNA metabolic processes, including alternative transcriptional start site (TSS), alternative splicing, and APA. With the emerging of high-throughput sequencing technology, gene expression, alternative splicing, and RNA editing have been extensively studied by directly using the RNA-seq data. However, utilization of conventional RNA-seq data for comprehensive APA analysis remains limited due to the shortage and limitations of current methods for APA identification and quantification. The main limitations of the current methods include the following: (1) RNA-seq coverage-based methods are impacted by frequent fluctuation of coverages in transcriptome, resulting in high false positives and bias toward 3′ UTR; (2) DNA sequence-based methods are unable to distinguish PAS expressed versus unexpressed in specific samples; (3) the recently published method, Aptardi, which is the first method to integrate both RNA-seq coverage and DNA sequence, was designed for scanning 3′ UTR with a 300 bp window. Thus, it can only determine whether or not a window contains truly expressed/used PASs and this resolution for PAS identification is relatively low.

A recent systematic evaluation of 11 methods using simulated and public 3′ end-seq data showed that TAPAS has the best performance, followed by DaPars (Chen et al. 2020). Based on a more recent study that benchmarked five computational methods for APA analysis with 3′ end-seq and Iso-Seq, none of them were able to identify >50% of the expressed PAS except QAPA which requires the annotated PAS as input. Among the five, the second generation of DaPars (DaPars2) identified slightly more annotated PAS than TAPAS. For the quantification of PAS usage, this study showed that QAPA has the lowest error rate (Shah et al. 2021).

Here, we generated a data set for benchmarking, which includes the matched 3′ end-seq and RNA-seq data from four human cell lines. To be noted, the in-house 3′ end-seq data have two independent sources. The 3′ end sequencing of HepG2, SNU398, and THLE2 cell lines was performed in one lab and K562 in another lab. The matched RNA-seq data of the four cell lines were from four independent sources (Methods). However, the performance of APAIQ on these four cell lines is quite consistent (Figs. 2A,B, 4A,C), indicating the robustness of our methods. By using 3′ end-seq data as the benchmark, we showed that APAIQ can identify >70% of the expressed PAS with FPR <10%. As we found that PASs with higher expression level achieved higher recall (Fig. 2B), suggesting that the false negatives could be enriched for lowly expressed PASs. The false positives without 3′ end-seq supports could be due to either strong PAS-related *cis*-elements around these loci or the RNA-seq read coverage having clear drop patterns. The former would not become false positive anymore in the differential APA analysis as the input DNA sequences are identical, whereas the latter might be due to some biases introduced by RNA-seq, which still needs further detailed analysis.

Apart from PAS identification, APAIQ also achieves a much lower error rate on APA quantification compared to QAPA (Fig. 4C). For instance, with APAIQ more than half of genes have an error margin of PAU <0.4%, and ~75% of the PASs have an error margin lower than 0.6. In comparison, these error margins were 0.6 (50% genes) and 0.95 (75% genes), respectively, when using QAPA, the best tool thus far for APA quantification (Shah et al.

2021). For a major PAS with PAU > 0.65, the probability to falsely determine it as a minor PAS (PAU < 0.35, error > 0.6) is <25% using APAIQ, whereas this could be as high as 50% using QAPA. Therefore, using the same RNA-seq data, APAIQ could improve the accuracy of APA quantification compared to current methods.

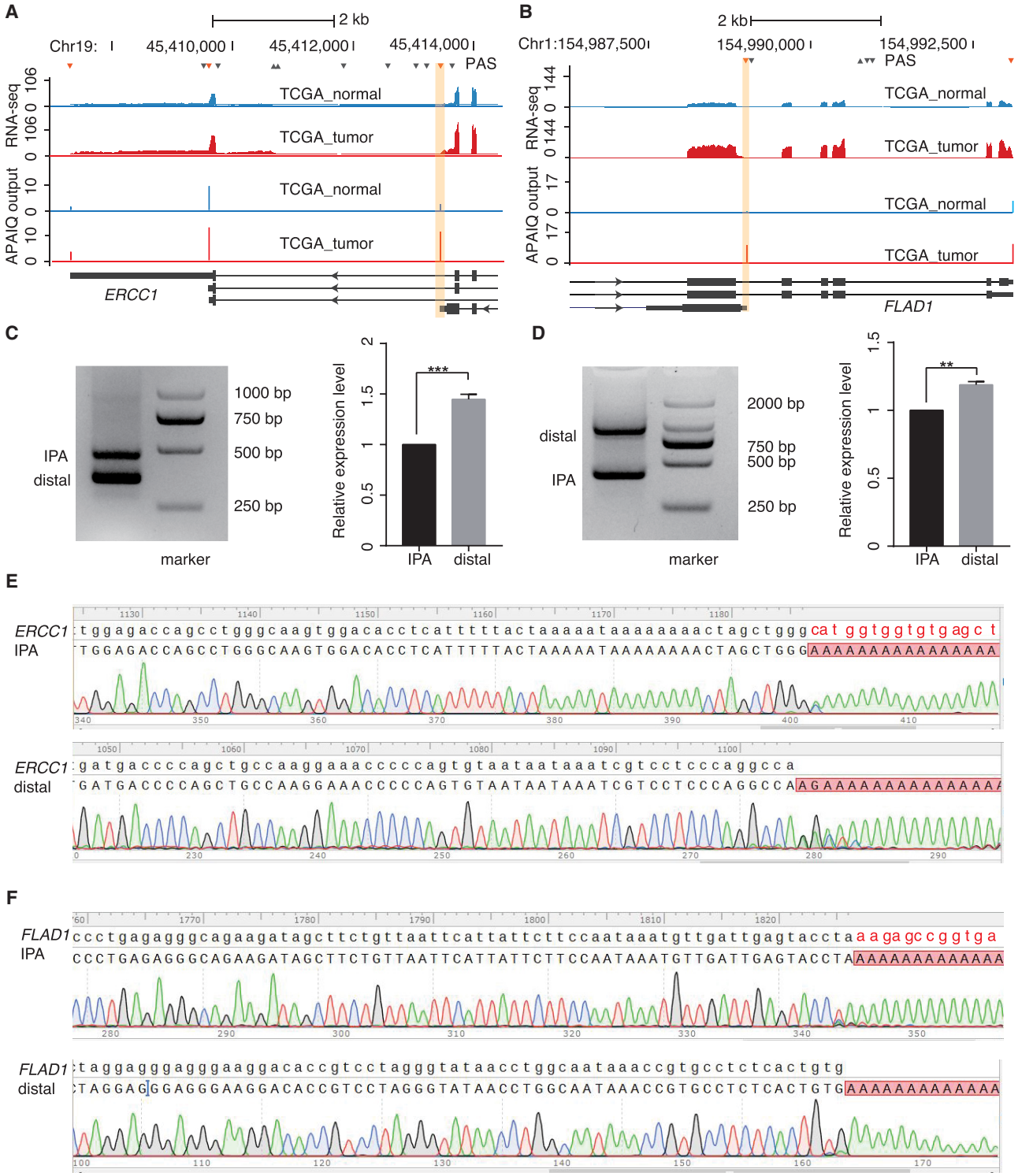

Figure 6. Experimental validation of the predicted APA events. (*A*,*B*) Genome Browser showing an identified tumor-associated IPA event (highlighted in orange) from genes *ERCC1* (*A*) and *FLAD1* (*B*). (*C*,*D*) The gel of 3′ RACE experiment (*left bottom*) and qRT-PCR results of the two isoforms from genes *ERCC1* (*C*) and *FLAD1* (*D*). (*E*,*F*) Sanger sequencing results of the amplified transcripts from *ERCC1* (*E*) and *FLAD1* (*F*) by using 3′ RACE experiment. Genomic sequences downstream from cleavage site identified by Sanger sequencing for two IPA were marked in red.

As we used the common human genome assembly (hg38) to extract DNA sequence for PAS identification, the genetic variations and polymorphisms were neglected. This could be a limitation of our current method, especially when applying it on tumor samples that frequently harbored a large number of DNA mutations. A later version incorporating small nucleotide variation (SNV) and structure variations could identify PASs more accurately, and meanwhile potentially provide direct links to the identification of APA-related variations.

Most of the previous methods, such as DaPars2 (Li et al. 2021), GETUTR (Kim et al. 2015), QAPA (Ha et al. 2018), and Aptardi (Lusk et al. 2021), were designed for APA analysis within the terminal exon, in which APA has been extensively studied. For instance, utilizing RNA-seq data from TCGA, the landscape of APA regulating 3′ UTR lengthening and shortening has been characterized by several previous studies (Xia et al. 2014; Feng et al. 2018). On the other hand, global studies of APA using RNA-seq data from other regions, such as intronic regions, that could alter coding sequence (CDS) and the coding potential, have just recently been addressed (Zhao et al. 2021a). APAIQ could greatly facilitate such a kind of analysis as it can identify and quantify APA in a transcriptome-wide manner, which was demonstrated by the application of APAIQ on RNA-seq data from the TCGA-LIHC cohort. Indeed, we identified more than 100 tumor-associated/specific IPA events, including the validated FLAD1-IPA and ERCC1-IPA, which likely contribute to liver cancer development by altering the coding sequence of functional domains. Future applications of APAIQ on the complete TCGA data set would build a comprehensive atlas of RNA transcript 3′ ends in human cancer, which would facilitate a complete understanding of the functional relevance of APA regulation in tumor development.

Methods

Cell lines and high-throughput RNA sequencing

Total RNA was extracted from the cell lines, including THLE2, HepG2, K562, and SNU398, using TRIzol reagent followed by column purification using the PureLink RNA Mini Kit (Thermo Fisher Scientific). Next, the extracted RNA was prepared for library with QuantSeq 3′mRNA-seq Library Prep Kit REV for Illumina (Lexogen) and sent for sequencing.

3′ RACE and qRT-PCR

Total RNA was extracted using RNA Isolator Total RNA Extraction Reagent (Vazyme R401-01). For qRT-PCR, the first-strand cDNA was synthesized using the HiScript III 1st Strand cDNA Synthesis Kit (Vazyme R312-02) with oligo(dT) as reverse transcription (RT) primer. Quantitative PCR (qPCR) was performed using Hieff qPCR SYBR Green Master Mix (Yeasen).

For the 3′ RACE assay, cDNA was synthesized using anchored oligo(dT) (GACCACGCGTATCGATGTCGACTTTTTTTTTTTTTTT TTVN) as RT primer. Then cDNA was amplified by nest PCR using gene specific forward primers and anchor reverse primer (anchor-R). PCR products were separated by agarose gel and purified by Gel DNA Exaction mini kit (Vazyme). Purified PCR products were sent for Sanger sequencing and visualized by SnapGene Viewer. Primer sequences were listed in Supplemental Table S3.

Ligation assay

For the ligation assay, ribosomal RNA and tRNA were first removed from 2 μg total RNA, using the ZYMO RNA Clean & Concentrator-5 kit (Zymo R1015). The rRNA-depleted RNA (5 μL) was then mixed with 1 μL 50 pmol Universal miRNA cloning linker (NEB S1315S; 5′ rAppCTGTAGGCACCATCAAT–NH2 3′) at 65°C for 5 min, and then ligated using T4 RNA Ligase 2, truncated KQ (NEB M0373S) at 16°C for 10 h. After ligation with 3′ adaptor, cDNA was synthesized by SMARTer PCR cDNA Synthesis Kit (TaKaRa 634926). The 3′ cDNA end of individual gene was amplified by nest PCR using gene specific forward primers and ligation reverse primer (ligation-R). PCR products were separated by agarose gel and sent for Sanger sequencing. Primer sequences are listed in Supplemental Table S3.

Sequencing data processing

Clean reads of RNA-seq data from four cell lines, including both the in-house and published data sets (obtained from the NCBI BioProject database [https://www.ncbi.nlm.nih.gov/bioproject/] PRJNA495931, PRJNA56 2266, and library ENCLB471LNG and ENCLB352YLJ from the ENCODE Project) (The ENCODE Project Consortium 2004), were aligned to human reference genome (hg38) with the transcriptome annotation (https://www.gencodegenes.org) using STAR (Dobin et al. 2013). We used standard parameters from the ENCODE Project (https://www.encodeproject.org) for the alignment and only the unique mapped reads were kept for further analysis. The parameter "- -outWigType bedGraph" from STAR was used to generate files of the RNA-seq coverage in the bedGraph format, in which the coverage at each genomic locus was normalized to reads per million (RPM) using the total uniquely mapped reads.

For 3′ end sequencing data, using the same criteria as in the previous study (Tian et al. 2022), Illumina sequencing adaptor and the leading T at the forward read (read 1) was removed. Next, the clean forward reads were aligned to reference genome (hg38) using STAR (Dobin et al. 2013).

Positive and negative data set preparation

To get training and testing data sets for the deep-learning model, we first built the general annotation using 289,565 PASs from polyADB3 and 184,617 PASs extracted from the transcript end based on GENCODE annotation. Next, in each cell line, we mapped the reads from 3′ end sequencing data to the annotation with "BEDTools" to get the expressed PAS in each sample (sample-specific PASs). In brief, for each PAS, any reads with 3′ end located within 25 bp were counted and then normalized to RPM using total uniquely mapped reads. If two annotated PASs were located within 50 bp, only the one with higher expression was kept. Because the used 3′ end-seq might introduce internal primming, we only include the expressed and annotated PASs in downstream analysis. Next, in each sample, by overlapping the PAS with gene annotation, we calculated the relative usage of each PAS by dividing the expression of the PAS to the sum of all the PASs from the same gene. Any PAS with expression level higher/no less than 0.1 RPM and usage higher/no less than 0.05 was considered as expressed/used.

For each expressed PAS in each cell line, we extracted the DNA sequence and the RNA-seq coverage from 100 bp upstream to 100 bp downstream regions. To avoid discrepancy between RNA-seq and 3′ end-seq data from the same sample/cell line, we further filtered out the PAS by requiring that the average RNA-seq coverage at 100 bp upstream should be no less than 0.05 RPM. In this way, we finally got ~20,000 used/expressed PASs as the positive data set (ground truth) in each cell line (Fig. 1B). The same number of sites with average RNA-seq coverage no less than 0.05 RPM at 100 bp upstream regions, and meanwhile being at least 50 bp far

away from any true PAS, was randomly selected from the genome as negative data set.

A hybrid deep-learning model using both DNA sequence and RNA-seq coverage

As shown in Figure 1A, we built a hybrid deep-learning model that contains two independent convolutional neuron networks, which take DNA sequence and RNA-seq coverage as input, respectively. The DNA sequence underwent one-hot encoding to 201 × 4 matrix and the normalized RNA-seq coverage (RPM values) was converted to 201 × 1 matrix. Both of them went through a convolutional layer consisting of 32 filters with a kernel size of 6. These were followed by a group normalization layer with group size 4. A rectified linear unit (ReLU) was applied to the normalized results as the activation function. After a max-pooling layer with pooling window setting as 6, features were flattened into one-dimensional array and fed to the fully connected layer. Then, two features were concatenated together and fed into another fully connected layer followed by a softmax activation function to approximate the probability function. The final output is a prediction score between 0 and 1, whereas dropout was introduced after the max-pooling layer as regularization to reduce over-fitting problems (Supplemental Method S3).

Training and evaluation of the deep-learning model on binary classification

In each sample, we applied a cross-validation strategy to evaluate our learning model, in which we trained the model with 80% of the positive data set and negative data set and tested it in the rest of the data set. We measured a series of metrics, including accuracy, recall, FPR, and FDR, for the evaluation.

An enhanced model for genome scanning

Because our model achieved extraordinary performance on binary classification (Supplemental Table S1), we wanted to apply it for genome scanning. We introduced a data augmentation strategy, in which we set the training epoch as 500 and the mini-batch size as 32. In each epoch, the sites randomly shifted from the true PAS from −12 to 12 bp were considered as the positive samples, whereas the sites more than 50 bp away from any true PAS were considered as negatives. To evaluate the model in each epoch, we split the genome into blocks with length of about one million base pair (bp) and we got ∼3000 blocks in each sample. These blocks were further divided into five groups and a fivefold cross-validation was applied for the evaluation.

Eventually, we selected the model with the highest accuracy, and utilized it to scan the genome with a stepwise one bp. Any site with coverage lower than 0.05 RPM at the upstream 100 bp was ignored. In this way, for each site, we used the window from upstream 100 bp to downstream 100 bp as input and obtained a score between 0 and 1.

As the model was trained by randomly shifting from −12 to 12 bp to the true PAS, sites close to the true PAS would also get a relatively high score (>0.5). To further find the precise position of the PAS and reduce the total number of the predicted sites, we introduced a clustering method to convert site-based score (Sc) to a cumulative cluster-based score (Cc). In brief, we first scanned the prediction score at each site in the forward direction based on genomic coordinates. The initial Cc was set as 0 and any continuous sites with prediction score higher than 0.5 were merged into the cluster and Cc would be accumulated with the prediction score of the site, whereas any sites with score lower than 0.5 would give a penalty P (default = 1) to Cc. The cluster was ended when Cc dropped to 0 and the site that obtained maximum Cc within each cluster was reported as the peak summit. We found that these summits showed a systematic bias toward the downstream of the true PAS. To correct this bias, we repeated the scanning in the reverse direction based on genomic coordinates, which reported summits showing bias toward the upstream of the true PAS. We finally used the middle position between the forward scanned summit and the backward scanned summit as the putative PAS.

Benchmarking computational methods for APA analysis with 3′ end-seq data

For APA identification, any predicted PAS within 25 bp away from the ground truth (annotated PAS with expression > 0.1 RPM and usage > 0.05) was considered as true positive, whereas any annotated PAS that is not expressed (RPM = 0) while covered with RNA-seq reads (average RPM at 100 bp upstream > 0.05), and meanwhile not being detected by the method, was defined as true negative (TN). The false positives (FP) are those predicted PASs located >25 bp away from the ground truth. A ROC curve was generated to illustrate the performance of APAIQ and four other methods, including Aptardi, DaPars2, mountainClimber, and SANPolyA. As SANPolyA is designed for binary classification, we did the prediction at each locus and implemented the same forward and backward scanning strategy to pinpoint the final PAS. As Aptardi scanned PAS with a 300 bp window and stepwise is 100 bp, we further relaxed the definition of TP and FP for those PASs within 100 bp and apart from the ground truth, respectively, and repeated the comparative analysis again.

For APA quantification, we used the same metric, termed as error, as described in the previous study (Shah et al. 2021), to measure the concordance between the prediction of the PAS usage and the quantification from the 3′ end-seq. In brief, we calculated the usage of each PAS (PAU) by using its expression divided by the total expression of all the PASs from the corresponding gene. Next, the difference between the predicted PAU and the PAU quantified by 3′ end-seq was calculated (delta PAU). For each gene, error was derived by summing up the absolute value of delta PAU of all the PASs from each gene. Cumulative distribution of error among all the genes with multiple PASs was generated and the results from APAIQ were further compared to those from QAPA.

Analysis of TCGA RNA-seq data with APAIQ

RNA-seq data of 421 samples from the TCGA-LIHC cohort were downloaded from the Genomic Data Commons (GDC) using the GDC Data Transfer Tool (https://gdc.cancer.gov/access-data/gdc-data-transfer-tool) and aligned to the reference human genome (hg38) using STAR (v2.7.9) with two pass model (Dobin et al. 2013). RNA-seq coverage at each genomic locus was derived from the BAM file by using BEDTools (Quinlan and Hall 2010). For each sample, APAIQ was conducted to predict the PAS, and any PASs detected in more than one sample were kept and merged to build a comprehensive reference. For each PAS in the reference, regression model was further applied to quantify the expression level.

Based on the quantified expression level (TPM) of each PAS, we calculated the PAS usage (PAU). For each PAS in each gene, we divided the expression of the PAS by the sum of expression of all the PASs from this gene to get the PAU. To evaluate the effect of PAS usage on 3′ UTR length, instead of using all PASs, we calculated the usage only for the PAS located on the terminal exon (expression of the PAS divided by the sum of the expression of all the PASs located on the same terminal exon). We further defined weighted 3′ UTR length index (WULI) by using usage of each

PAS within the terminal exon and their corresponding terminal exon length. In this way, each gene was assigned a WULI, which is a value from 0 to 1, and the larger value indicates that the isoforms with greater 3′ UTR length were used.

We further compared PAU of each PAS or WULI of each gene in the 50 TCGA tumor samples to that in the 50 matched adjacent normal samples with Wilcoxon test. The Benjamini–Hochberg method was applied to adjust the *P*-values from multiple comparisons. The significant PAU or 3′ UTR changes were obtained by requiring the BH-adjusted *P*-value (FDR) to be smaller than 0.1. As we only used the matched tumor and normal samples from the same patient, the age, gender, ethnicity, and other variables are matched between tumor and normal samples, which are unlikely to confound the results.

Software availability

The open source code of APAIQ is freely available as the Supplemental Code file and at GitHub (https://github.com/ijayden-lung/APAIQ). The compiled version can be found at https://anaconda.org/joshuachou/apaiq.

Data access

The raw 3′ end-seq data and K562 RNA-seq data generated in this study have been submitted to NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number PRJNA794041.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank all past and present members in the Structure and Functional Bioinformatics Group for their assistance and constructive feedback on this project. We also thank KAUST-HPC for providing generous support on computational resources. This work was supported by King Abdullah University of Science and Technology (KAUST) Office Administration (ORA) under Award Nos. FCC/1/1976-44-1, FCC/1/1976-44-01, FCC/1/1976-45-01, URF/1/4098-01-01, URF/1/4352-01-01, URF/1/4379-01-01, URF/1/4663-01-01, REI/1/5202-01-01, and REI/1/4940-01-01; National Key Research and Development Program of China (Grant No. 2021YFF1201000); National Nature Science Foundation of China (Grant Nos. 62002388, 32100431, 31970601); Shenzhen Science and Technology Program (Grant No. KQTD20180411143432337); and Shenzhen–Hong Kong Institute of Brain Science–Shenzhen Fundamental Research Institutions (Grant No. 2021SHIBS0002).

Author contributions: W.C., X.G., and B.Z. designed the study. Y. Long constructed the deep-learning model, wrote the original code, and prepared the data for the analysis. B.Z. performed the data analysis and modified the source code of APAIQ. S.T. prepared high-throughput sequencing library for K562 and performed experimental validation in HepG2. J.J.C. prepared the sequencing library for HepG2, SNU398, and THLE2. J.Z. helped to build the deep-learning model and compiled the source code to bio-conda. Z.L., Z.A., and Y.X. provided RNA-seq data from the TCGA project. Y. Li helped the primary analysis of K562 3′ end-seq data. W.C. and Y.T. provided the cell line samples. X.L. helped to modify the source code. Y.W. and S.S. helped to predict protein structures. Y. Long, B.Z., W.C., and X.G. wrote the manuscript.

References

Arefeen A, Liu J, Xiao X, Jiang T. 2018. TAPAS: tool for alternative polyadenylation site analysis. *Bioinformatics* **34:** 2521–2529. doi:10.1093/bioinformatics/bty110

Cass AA, Xiao X. 2019. mountainClimber identifies alternative transcription start and polyadenylation sites in RNA-seq. *Cell Syst* **9:** 393–400.e6. doi:10.1016/j.cels.2019.07.011

Chang J-W, Zhang W, Yeh H-S, De Jong EP, Jun S, Kim K-H, Bae SS, Beckman K, Hwang TH, Kim K-S. 2015. mRNA 3′-UTR shortening is a molecular signature of mTORC1 activation. *Nat Commun* **6:** 7218. doi:10.1038/ncomms8218

Chen M, Ji G, Fu H, Lin Q, Ye C, Ye W, Su Y, Wu X. 2020. A survey on identification and quantification of alternative polyadenylation sites from RNA-seq data. *Brief Bioinformatics* **21:** 1261–1276. doi:10.1093/bib/bbz068

Cheng Y, Miura RM, Tian B. 2006. Prediction of mRNA polyadenylation sites by support vector machine. *Bioinformatics* **22:** 2320–2325. doi:10.1093/bioinformatics/btl394

Derti A, Garrett-Engele P, MacIsaac KD, Stevens RC, Sriram S, Chen R, Rohl CA, Johnson JM, Babak T. 2012. A quantitative atlas of polyadenylation in five mammals. *Genome Res* **22:** 1173–1183. doi:10.1101/gr.132563.111

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29:** 15–21. doi:10.1093/bioinformatics/bts635

The ENCODE Project Consortium. 2004. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306:** 636–640. doi:10.1126/science.1105136

Feng X, Li L, Wagner EJ, Li W. 2018. TC3A: The Cancer 3′ UTR Atlas. *Nucleic Acids Res* **46:** D1027–D1030. doi:10.1093/nar/gkx892

Friboulet L, Olaussen KA, Pignon J-P, Shepherd FA, Tsao M-S, Graziano S, Kratzke R, Douillard J-Y, Seymour L, Pirker R, et al. 2013. ERCC1 isoform expression and DNA repair in non–small-cell lung cancer. *N Engl J Med* **368:** 1101–1110. doi:10.1056/NEJMoa1214271

Ha KC, Blencowe BJ, Morris Q. 2018. QAPA: a new method for the systematic analysis of alternative polyadenylation from RNA-seq data. *Genome Biol* **19:** 45. doi:10.1186/s13059-018-1414-4

Herrmann CJ, Schmidt R, Kanitz A, Artimo P, Gruber AJ, Zavolan M. 2020. PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3′ end sequencing. *Nucleic Acids Res* **48:** D174–D179. doi:10.1093/nar/gkz918

Hoque M, Ji Z, Zheng D, Luo W, Li W, You B, Park JY, Yehia G, Tian B. 2013. Analysis of alternative cleavage and polyadenylation by 3′ region extraction and deep sequencing. *Nat Methods* **10:** 133–139. doi:10.1038/nmeth.2288

Hu J, Lutz CS, Wilusz J, Tian B. 2005. Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation. *RNA* **11:** 1485–1493. doi:10.1261/rna.2107305

Hu P, Pan Y, Wang C, Zhang W, Huang H, Wang J, Zhang N. 2020. FLAD1 is up-regulated in Gastric Cancer and is a potential prediction of prognosis. *Int J Med Sci* **17:** 1763–1772. doi:10.7150/ijms.48162

Jan CH, Friedman RC, Ruby JG, Bartel DP. 2011. Formation, regulation and evolution of *Caenorhabditis elegans* 3′ UTRs. *Nature* **469:** 97–101. doi:10.1038/nature09616

Jia X, Wang C, Huang H, Zhang P, Yao Z, Xu L. 2019. FLAD1 is overexpressed in breast cancer and is a potential predictor of prognosis and treatment. *Int J Clin Exp Med* **12:** 3138–3152.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596:** 583–589. doi:10.1038/s41586-021-03819-2

Kim M, You B-H, Nam J-W. 2015. Global estimation of the 3′ untranslated region landscape using RNA sequencing. *Methods* **83:** 111–117. doi:10.1016/j.ymeth.2015.04.011

Lee S-H, Singh I, Tisdale S, Abdel-Wahab O, Leslie CS, Mayr C. 2018. Widespread intronic polyadenylation inactivates tumour suppressor genes in leukaemia. *Nature* **561:** 127–131. doi:10.1038/s41586-018-0465-8

Li L, Huang K-L, Gao Y, Cui Y, Wang G, Elrod ND, Li Y, Chen YE, Ji P, Peng F. 2021. An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* **53:** 994–1005. doi:10.1038/s41588-021-00864-5

Lusk R, Stene E, Banaei-Kashani F, Tabakoff B, Kechris K, Saba LM. 2021. Aptardi predicts polyadenylation sites in sample-specific transcriptomes using high-throughput RNA sequencing and DNA sequence. *Nat Commun* **12:** 1652. doi:10.1038/s41467-021-21894-x

Martin G, Gruber AR, Keller W, Zavolan M. 2012. Genome-wide analysis of pre-mRNA 3′ end processing reveals a decisive role of human cleavage factor I in the regulation of 3′ UTR length. *Cell Rep* **1:** 753–763. doi:10.1016/j.celrep.2012.05.003

Mayr C, Bartel DP. 2009. Widespread shortening of 3′ UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell* **138:** 673–684. doi:10.1016/j.cell.2009.06.016

Miura P, Shenker S, Andreu-Agullo C, Westholm JO, Lai EC. 2013. Widespread and extensive lengthening of 3′ UTRs in the mammalian brain. *Genome Res* **23:** 812–825. doi:10.1101/gr.146886.112

Park HJ, Ji P, Kim S, Xia Z, Rodriguez B, Li L, Su J, Chen K, Masamha CP, Baillat D, et al. 2018. 3′ UTR shortening represses tumor-suppressor genes in trans by disrupting ceRNA crosstalk. *Nat Genet* **50:** 783–789. doi:10.1038/s41588-018-0118-8

Pavlović N, Heindryckx F. 2021. Exploring the role of endoplasmic reticulum stress in hepatocellular carcinoma through mining of the human protein atlas. *Biology (Basel)* **10:** 640. doi:10.3390/biology10070640

Pavlovic N, Rani B, Gerwins P, Heindryckx F. 2019. Platelets as key factors in hepatocellular carcinoma. *Cancers (Basel)* **11:** 1022. doi:10.3390/cancers11071022

Proudfoot N, Brownlee G. 1976. 3′ non-coding region sequences in eukaryotic messenger RNA. *Nature* **263:** 211–214. doi:10.1038/263211a0

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26:** 841–842. doi:10.1093/bioinformatics/btq033

Sandberg R, Neilson JR, Sarma A, Sharp PA, Burge CB. 2008. Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites. *Science* **320:** 1643–1647. doi:10.1126/science.1155390

Shah A, Mittleman BE, Gilad Y, Li YI. 2021. Benchmarking sequencing methods and tools that facilitate the study of alternative polyadenylation. *Genome Biol* **22:** 291. doi:10.1186/s13059-021-02502-z

Shuda M, Kondoh N, Imazeki N, Tanaka K, Okada T, Mori K, Hada A, Arai M, Wakatsuki T, Matsubara O, et al. 2003. Activation of the ATF6, XBP1 and grp78 genes in human hepatocellular carcinoma: a possible involvement of the ER stress pathway in hepatocarcinogenesis. *J Hepatol* **38:** 605–614. doi:10.1016/S0168-8278(03)00029-1

Tian B, Hu J, Zhang H, Lutz CS. 2005. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Res* **33:** 201–212. doi:10.1093/nar/gki158

Tian S, Zhang B, He Y, Sun Z, Li J, Li Y, Yi H, Zhao Y, Zou X, Li Y. 2022. CRISPR-iPAS: a novel dCAS13-based method for alternative polyadenylation interference. *Nucleic Acids Res* **50:** e26. doi:10.1093/nar/gkac108

Wang R, Tian B. 2020. APAlyzer: a bioinformatics package for analysis of alternative polyadenylation isoforms. *Bioinformatics* **36:** 3907–3909. doi:10.1093/bioinformatics/btaa266

Wang R, Nambiar R, Zheng D, Tian B. 2018. PolyA_DB 3 catalogs cleavage and polyadenylation sites identified by deep sequencing in multiple genomes. *Nucleic Acids Res* **46:** D315–D319. doi:10.1093/nar/gkx1000

Xia Z, Donehower LA, Cooper TA, Neilson JR, Wheeler DA, Wagner EJ, Li W. 2014. Dynamic analyses of alternative polyadenylation from RNA-seq reveal a 3′-UTR landscape across seven tumour types. *Nat Commun* **5:** 5274. doi:10.1038/ncomms6274

Xia Z, Li Y, Zhang B, Li Z, Hu Y, Chen W, Gao X. 2019. DeeReCT-PolyA: a robust and generic deep learning method for PAS identification. *Bioinformatics* **35:** 2371–2379. doi:10.1093/bioinformatics/bty991

Xiao MS, Zhang B, Li YS, Gao Q, Sun W, Chen W. 2016. Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation. *Mol Syst Biol* **12:** 890. doi:10.15252/msb.20167375

Xie B, Jankovic BR, Bajic VB, Song L, Gao X. 2013. Poly(A) motif prediction using spectral latent features from human DNA sequences. *Bioinformatics* **29:** i316–i325. doi:10.1093/bioinformatics/btt218

Ye C, Long Y, Ji G, Li QQ, Wu X. 2018. APAtrap: identification and quantification of alternative polyadenylation sites from RNA-seq data. *Bioinformatics* **34:** 1841–1849. doi:10.1093/bioinformatics/bty029

Ye C, Zhang X, Chen X, Cao Q, Zhang X, Zhou Y, Li W, Hong L, Xie H, Liu X, et al. 2020. Multiple novel hepatocellular carcinoma signature genes are commonly controlled by the master pluripotency factor OCT4. *Cell Oncol* **43:** 279–295. doi:10.1007/s13402-019-00487-3

Yu H, Dai Z. 2020. SANPolyA: a deep learning method for identifying poly(A) signals. *Bioinformatics* **36:** 2393–2400. doi:10.1093/bioinformatics/btz970

Zhao Z, Xu Q, Wei R, Huang L, Wang W, Wei G, Ni T. 2021a. Comprehensive characterization of somatic variants associated with intronic polyadenylation in human cancers. *Nucleic Acids Res* **49:** 10369–10381. doi:10.1093/nar/gkab772

Zhao Z, Xu Q, Wei R, Wang W, Ding D, Yang Y, Yao J, Zhang L, Hu Y-Q, Wei G, et al. 2021b. Cancer-associated dynamics and potential regulators of intronic polyadenylation revealed by IPAFinder using standard RNA-seq data. *Genome Res* **31:** 2095–2106. doi:10.1101/gr.271627.120

Zheng D, Wang R, Ding Q, Wang T, Xie B, Wei L, Zhong Z, Tian B. 2018. Cellular stress alters 3′ UTR landscape through alternative polyadenylation and isoform-specific degradation. *Nat Commun* **9:** 2268. doi:10.1038/s41467-018-04730-7

Zingone A, Sinha S, Ante M, Nguyen C, Daujotyte D, Bowman ED, Sinha N, Mitchell KA, Chen Q, Yan C, et al. 2021. A comprehensive map of alternative polyadenylation in African American and European American lung cancer patients. *Nat Commun* **12:** 5605. doi:10.1038/s41467-021-25763-5

Received August 3, 2022; accepted in revised form February 28, 2023.