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

**Abstract**

We developed a computational method for sample-specific **APA** **i**dentification and **q**uantification (APAIQ) from the conventional RNA-seq data.

**Introduction**

In eukaryotes, the transcription termination is mediated by cleavage of the nascent RNA and followed with the synthesis of poly-adenosines (poly-A) to the 3’end, which is known as polyadenylation. This process is determined by a series of RNA binding proteins (RBP) that recognizes cis elements around polyadenylation site (PAS). PAS motif, hexamers located with 15~40nt upstream of the cleavage site, is one of the most important core elements for PAS definition. They include AAUAAA that present in more than half of the PAS, and its variants (AUUAAA, xxx et al), which has been found in nearly 40% of the PAS in human and mouse genome. UGUA at the downstream of PAS is another core element for PAS definition, while other auxiliary elements located within 100bp flanking PAS also contribute to the formation polyadenylation.

Taking advantage of the sequence surrounding PAS, numerous computational methods have been developed for PAS prediction. The early studies focused on features obtained from statistical analysis of sequences surrounding the PAS. For instance, Hu et al identified 2 core elements and 4 auxiliary elements at 100bp upstream region, and 4 core elements and 5 auxiliary elements at 100bp downstream region of PAS by analyzing large number of PAS on human and mouse (Hu et al). Using these features, support vector machine (SVM) was applied for PAS prediction (Cheng et al. 2006. Bioinf). Later, a hidden Markov models (HMM) was introduced to extract latent features and fed them to an SVM model (Xie et al). More recently, with the development of deep-learning models, convolutional neural network (CNN) has been implemented to extract sequence features automatically, which greatly improved accuracy of the PAS prediction.

However, most mammalian genes use multiple sites for polyadenylation to generate RNA isoforms with varied 3’ends (ref). This alternative polyadenylation (APA) process not only enables single gene to encode multiple proteins, but also greatly increase the complexity of gene regulations via different untranslated regions (UTRs) at the 3’end of mRNA. It has been shown that the usage of different PAS is cell type and even sample specific in general. For instance, the neuron cells tend to use the distal PAS to generate longest isoform, while proliferated cell express short isoforms by using proximal PAS (Ref). Moreover, the dysregulation of APA is associated with human disease, including cancer. APA mediated 3’UTR shortening could activate oncogenes expression through escaping from microRNA regulation, whereas it could also inhibit tumor suppressor in trans by disrupting the competition of microRNA binding (ref).

Thus, even though difference cells may share the identical DNA sequence, the PAS usage could be distinct. The computational methods solely using DNA sequence is very unlikely to predict and quantify sample-specific usages of PAS. In contrast, many experimental technical have been developed for genome-wide identification and quantification of sample-specific APA events by enriching the 3’end of RNA, followed with high-throughput sequencing (3’end-seq). However, these methods are rather laborious and more time and material-consuming comparing to convention RNA sequencing (RNA-seq). More importantly, beyond APA, it’s almost unable to use 3’end-seq data to study other posttranscriptional modifications during RNA metabolic process, such as RNA splicing and editing.

To solve this problem, computational methods have been developed for identification and quantification of APA from RNA-seq data. Almost all of them are designed to detect the changes/transitions in RNA-seq coverage along the genome. Due to the high fluctuation of the coverage, these methods usually get high false positives and low recall on PAS identification. Moreover, these methods are widely used for PAS identification within 3’UTR, whereas it’s very challenge to be applied in other regions due to RNA-seq coverage dropped dramatically in exon-intron boundary. This limits their application for intronic polyadenylation (IPA) identification, while recently studies revealed that IPA are widespread in leukemia to inactivate tumor suppressor by generating truncated protein isoforms (Ref). In addition to identification, methods for APA quantification usually rely on the annotation of PAS, but the current PAS annotation are mainly derived from common cell types or biological samples. There might be tremendous of unannotated PAS in rare cell types or very heterogenous biological samples, such as primary tumor samples from patients.

Here, we developed APAIQ, a computational method which is capable of sample-specific **APA** **i**dentification and **q**uantification from RNA-seq data. By integrating both DNA sequence and RNA-seq coverage into a deep-learning model, our method could accurately and comprehensively identify PAS genome wide in each sample. Our method outperforms all the existed method with higher recall and lower false positives. Meanwhile, our method is able to identify APS globally beyond the 3’UTR. In addition, APAIQ is also able to quantify APA in terms of faithful predict both the expression level and relative usage of each PAS in genes with multiple PAS. Applying APAIQ on large-scale RNA-seq data from hepatocellular carcinoma (HCC) patients identified hundreds of tumor specific APA events, including IPA events, which have not been reported before. Overall, APAIQ is capable of accurate genome-wide APA analysis using the existing RNA-seq data, which will enable to build more comprehensive APA landscape in different cell types, tissues and pathological statues, including cancer.

**Results**

**A hybrid deep-learning model for PAS identification**

To identify expressed PAS in each sample with RNA-seq data, the majority of previous computational methods were designed to capture the transitions in RNA-seq converge. On the other hand, computational methods solely using DNA sequence could only predict PAS in general, regardless of the expression specificity across different samples. Here, we develop a hybrid deep-learning model for samples-specific PAS identification using both converge from RNA-seq data and the corresponding DNA sequence. We performed RNA 3’end sequencing on four cell lines, including K562, HepG2, THLE2 and Hep3B, to comprehensively characterize the expressed PAS in sample-specific manner. Using the expressed PAS in each cell line (Methods) as positive dataset, and random loci apart from the PAS as negative dataset, we trained a model for binary classification in each cell line based on CNN (Fig. 1a). We found that the RNA-seq coverage indeed tended to be dropped at the downstream of PAS comparing to upstream (Fig .1b). In addition, xx% of positive dataset have canonical PAS motif (AAUAAA) and its variants, while only xx% of negative dataset have the motifs (Fig. 1c). These further suggested that integrating DNA sequence with RNA-seq converge might have a synergistic effect in sample-specific PAS identification.

**APAIQ predicts PAS accurately in transcriptome wide**

To evaluate the feasibility of integrating two features for PAS identification, we applied a cross validation approaches, in which we trained the model using 1/5 of the datasets and made predictions on the rest. Just as expected, the integrated model achieved xx TPR/recall with FDR less than xx, while models only using DNA sequence or RNA-seq coverage have much worse performance (TPR = xx, FDR = xx, Table 1).

Even though the binary classification model could achieve good performance in the determination of true or false PAS, applying it on genome-wide PAS identification in practical remains challenge since the positive and negative datasets are highly unbalanced. For instance, among 3 billion loci in human genome, only xx of them are truly expressed PAS (positive), while the rest 9x% are negative. To this aim, we extended the model on genome scanning. In brief, we did prediction at each locus and any locus within xx bp away from true PAS were considered as true positive. (Please write a summary about how you did the scanning). Furthermore, even the model could achieve 99% of precision in the binary prediction, the final genome scanning could still get millions of false positives, resulting in extremely high FDRs. To solve this problem, we first limited the loci for the prediction by requiring it covered by RNA-seq reads. In this way, we reduced the total scanning loci from 3 billion to xx million. Additionally, we introduced a circular training strategy during scanning, in which we repeatedly replace the false positive predictions, which further reduced the false positives.

Eventually, our model was able to achieve ~85% recall with FPR < 10%. Interestingly, we found that around 75% of PAS in SNU398 and THLE2 could be detected, while only 60% of PAS in HepG2 and K562 were detected with FDR < 10% (Supp Figs. 1a and 1b). This could be due to RNA-seq data depth? polyA-seq depth?. Even though to obtain ~80% recall, FDRs could be as high as 30% in our prediction, many of them might not be real false positives as they are expressed and annotated, whereas only the expression levels did not pass our threshold (Fig. 2a). To further inspect this, we further grouped annotated PAS with a series of expression levels. Indeed, our model exhibit much stronger capacity of the identification of PAS with higher expression, while it’s not so powerful for lowly expressed PAS (Fig. 2b). We further checked the position of our identified PAS and surprisingly they are very closed to the true PAS, in which >80% of the identified PAS are within 25bp away from the annotation (Fig. 2c). As shown in Fig. 2d as an example, our model was successfully identified PAS a from gene X, which was expressed in xx, xx, but not in xx, while PAS b was expressed in xx xx, but not in xx. Taken all these results together, they suggested that our method is capable for accurate PAS identification using RNA-seq data in genome wide scale.

**APAIQ outperforms other methods on PAS identification**

Since our model showed good performance on PAS identification, we next benchmarked it with several previous published methods, including DaPars2, mountain climber, xx, and APxx. Among them, xx and xx were based on the transition pattern in RNA-seq coverage, while xx and xx relied on DNA sequence, while APXX combined RNA-seq and DNA sequence, but only capable of binary classification. As shown in Fig. 3a, with the same FDR, our method identified the largest number of truly expressed PAS. Moreover, the distance between the identified PAS and true PAS, is smallest in our method comparing to the other 4/5? (Fig. 3b).

Moreover, as xx and xx were designed to detect APA events within the 3’UTR, we classified the expressed PAS into xx groups based on their genomic localizations, including terminal exon (including 3’UTR), upstream exonic regions, upstream intronic regions, and intergenic regions. We found that our method outperformed all these published methods in each genomic region (Table S2). As shown in Fig. 3c, our method identified a PAS within intronic regions, while all other method failed to do so.

To dissect the contribution of RNA-seq converge and DNA sequence to the model, we also compared the integrated model with coverage- and DNA-only model. We found that, using two features not only improved precision, but also reduced FDR in the prediction. An example was shown in Fig. 3d, using DNA-only model, identified an annotated PAS that are not expressed, while adding the RNA-seq coverage features, the integrated model was successfully discarded this PAS. Another example in Fig. 3e, RNA-seq coverage identified a false PAS from gene x due to the drastic drop of coverage, while there is none of sequence features supporting this PAS, which was also successfully predicted as false by the integrated model (Fig. 3f). In addition, some PAS were neither identified in DNA-only, nor coverage-only model, but finally identified in the integrated model, further confirmed a synergistic effect by combining these two features for PAS identification.

**Generalization abilities of APAIQ on PAS identification**

To examine whether the deep-learning model could be applied on different samples, we tested the model across different cell lines. Simply, we trained the model in each cell line and made predictions on the samples from other three cell lines. We found that it showed relatively good transferring abilities. For instance, the model trained in xx, could still get xx recall with xx precision in xx cell lines (Fig. 3e/Fig. S). To further tested whether the model could be even applied on another species, we tested the model on another dataset from mouse (Fig. 3f/Fig. S). Still, the model showed good performance, indicating a good generalization ability of our method on different samples.

**Accurately quantification of APA by APAIQ**

After the identification of PAS based on RNA-seq data, we further asked whether is possible 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) is a function of the coverages at each locus within 100bp flanking the PAS (X). The expression level could be estimated by 3’end-sequencing data and the coverage at each locus were derived from RNA-seq data. Our prediction achieved an average Pearson correlation 0.77 in four cell lines (Fig. 4a).

Among xx genes with expressed PAS based on 3’end-seq data, xx genes have more than one PAS. We further examined the ability of our method for APA quantification. For each PAS, we calculated their relative usage using its expression divided by the total expression of this gene. The usage of PAS showed a typic bimodal distribution, in which most of PAS were either predominant PAS or minor PAS within a gene (Fig. 4b). Our method could still predict the relative usage of these PAS faithfully with a Pearson correlation as high as 0.7x, which is much accurate than previous published method QAPA, xx, and xx (Fig. 4c).

**Applying APAIQ on large-scale RNA-seq data identifies tumor-specific APA events**

Finally, to test our method, we applied it on xx RNA-seq samples from The Cancer Genome Atlas Liver Hepatocellular Carcinoma (TCGA-LIHC) (Fig. 5a). In total, xx PAS were identified and xx of them are overlapped with annotation. Among them, xx are within terminal exon/3’UTR and xx of them are located at upstream exon and intronic regions (Fig. 5b). To further evaluate the quality of the identified PAS, we first checked the frequency of canonical PAS motif (AAUAAA) and its variants and found that xx% of the identified PAS has the motif. In addition, the nucleotide composition around the identified PAS share a similar patter as the results from 3’end sequencing data reported in previous study (Fig. 5c).

Next, we also quantified the expression of each PAS and comparing them between samples from tumor and normal. Finally, we identified xx PAS which are significantly upregulated in tumor, including xx were only detected in tumor samples. We also correlated the expression of each PAS with the patient overall survival rate, and xx of them are associated with poor prognosis, while xx of them are associated with good survival. These xx PAS are from nn genes and very interestingly, expression of many of these genes are not associated with prognosis, while the PAS was, suggesting different isoforms with the varied 3’end might have distinct functions in cancer progression. As shown in Fig. 5e, an intronic PAS from gene X were identified in xx% of tumor samples but absent in all normal samples. This IPA events might produce a truncated protein isoform, which might contribute to liver cancer development.

**Discussion**

**Methods and Materials**