Research Article

Alternative Polyadenylation Modification Patterns Reveal Essential Posttranscription Regulatory Mechanisms of Tumorigenesis in Multiple Tumor Types

Min Li,[1] XiaoYong Pan,[2] Tao Zeng,[3] Yu-Hang Zhang,[4] Kaiyan Feng,[5] Lei Chen ⬤,[6] Tao Huang ⬤,[4] and Yu-Dong Cai ⬤[1]

[1]*School of Life Sciences, Shanghai University, Shanghai 200444, China*
[2]*Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, 200240 Shanghai, China*
[3]*Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China*
[4]*Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China*
[5]*Department of Computer Science, Guangdong AIB Polytechnic, Guangzhou 510507, China*
[6]*College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China*

Correspondence should be addressed to Tao Huang; tohuangtao@126.com and Yu-Dong Cai; cai_yud@126.com

Among various risk factors for the initiation and progression of cancer, alternative polyadenylation (APA) is a remarkable endogenous contributor that directly triggers the malignant phenotype of cancer cells. APA affects biological processes at a transcriptional level in various ways. As such, APA can be involved in tumorigenesis through gene expression, protein subcellular localization, or transcription splicing pattern. The APA sites and status of different cancer types may have diverse modification patterns and regulatory mechanisms on transcripts. Potential APA sites were screened by applying several machine learning algorithms on a TCGA-APA dataset. First, a powerful feature selection method, minimum redundancy maximum relevancy, was applied on the dataset, resulting in a feature list. Then, the feature list was fed into the incremental feature selection, which incorporated the support vector machine as the classification algorithm, to extract key APA features and build a classifier. The classifier can classify cancer patients into cancer types with perfect performance. The key APA-modified genes had a potential prognosis ability because of their significant power in the survival analysis of TCGA pan-cancer data.

1. Introduction

Cancer is one of the most threatening human diseases and ranks second to infectious diseases and cardiovascular diseases. According to statistical data provided by the World Health Organization (WHO) in 2015, cancer accounts for more than 8.8 million deaths worldwide with more than 14 million new cases and a high growth incidence [1]. Among various risk factors of cancer initiation and progression, pathogenic genetic variants and modifications, such as alternative polyadenylation (APA), are remarkable endogenous contributors, directly triggering the malignant phenotype of cancer cells [2].

APA is a specific RNA modification process contributing to gene expression regulation by generating RNA with different $3'$ terminals from a single gene with multiple polyadenylation sites [3]. APA affects biological processes at a transcriptional level in various ways. First, tissue-specific APA can rapidly respond to extracellular cues, regulating the expression level of certain genes as cellular "stress" responses [4]. As evidence confirmed in pancreatic cancer, the APA of ZEB1 rapidly responds to genotoxic stress and promotes gene expression, thereby improving the adaptability of tumor cells in a flexible tumor microenvironment [5]. Second, APA may regulate different metabolisms in living cells by affecting the

subcellular localization of certain protein products. APA contributes to the regulation of 3′UTR-dependent protein localization by modifying the 3′UTR, thereby affecting the widespread trafficking mechanisms for different membrane proteins, including CD47, CD44, and ITGA1 [6, 7]. Third, considering that 3′UTR is involved in multiple splicing events, APA influences posttranscriptional splicing processes and further induces the abnormal production of improper protein isoforms [8–10]. In 2014, a report on the alternative intronic polyadenylation of IL6 *trans*-signaling inhibitor confirmed that different polyadenylation patterns of the same gene (sgp130-E10) may produce different protein isoforms with different biological functions [11]. With numerous regulatory contributions to downstream biological processes, APA is also regulated by various upstream biological mechanisms involving RNA-processing factors and RNA-binding proteins, which constitute a complicated and functional interaction network for posttranscriptional regulation [12].

APA is functionally related to tumorigenesis as a key functional component of pathogen posttranscriptional regulation [13–15]. APA can be involved in tumorigenesis at three levels based on original physical functions: gene expression, protein subcellular localization, and transcription splicing patterns [4]. For example, in gene expression regulation, APA promotes the tumorigenesis of non-small-cell lung cancer by regulating the expression levels of various genes, including PABPN1, CPEB1, and E2F1, and several proliferation markers, such as MKI67, TOP2A, and MCM2 [14].

More than 30% of mRNAs have specific APA sites independent of cell types [4]. Considering that the expression profiles of different cancer types vary, we can infer that the APA sites and status of different cancer types may have diverse modification patterns and regulatory mechanisms on transcripts. Therefore, in this study, we adopted several machine learning algorithms to screen the potential APA sites at the whole genomic level in multiple tumor types and tried to find out the key APA-modified genes that might distinguish different tumor types. The TCGA-APA dataset was first analysed by the feature selection method, minimum redundancy maximum relevancy (mRMR) [16]. A feature list was obtained. Then, the incremental feature selection (IFS) [17], incorporating a support vector machine (SVM) [18], was applied on such a list to extract essential APA features. Most of the key genes corresponding to essential APA features showed a significant power in the survival analysis of TCGA pan-cancer data. Furthermore, the SVM classifier with the extracted essential APA features gave a perfect performance. This study possibly identified tumor-specific APA targets, revealed the irreplaceable role of APA modification patterns for tumorigenesis in multiple tumor types, and proposed APA sites and status as potential tumor biomarkers for the first time.

2. Materials and Methods

2.1. Datasets.
The TCGA-APA dataset was downloaded from Synapse under the accession number of syn7888354 [19]. In the original dataset, 9396 APAs were obtained in 5765

TABLE 1: Summary of the used dataset.

Index	Cancer type	Sample size
1	Bladder urothelial carcinoma (BLCA)	249
2	Breast invasive carcinoma (BRCA)	837
3	Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)	191
4	Glioblastoma multiforme (GBM)	152
5	Head and neck squamous cell carcinoma (HNSC)	422
6	Kidney chromophobe (KICH)	66
7	Kidney renal clear cell carcinoma (KIRC)	446
8	Kidney renal papillary cell carcinoma (KIRP)	195
9	Brain lower-grade glioma (LGG)	486
10	Liver hepatocellular carcinoma (LIHC)	183
11	Lung adenocarcinoma (LUAD)	486
12	Lung squamous cell carcinoma (LUSC)	220
13	Ovarian serous cystadenocarcinoma (OV)	407
14	Prostate adenocarcinoma (PRAD)	370
15	Skin cutaneous melanoma (SKCM)	225
16	Stomach adenocarcinoma (STAD)	282
17	Thyroid carcinoma (THCA)	492

patients with cancer, but several values were missing. APAs with missing values in more than 50% of samples and patients with cancer with missing values in more than 50% of APAs were removed. A total of 7544 APAs and 5709 patients were finally obtained from 17 cancer types. The remaining missing values were imputed using K-NN methods ($K = 3$) by R/Bioconductor package impute. The categories of 17 cancer sites and their corresponding sample sizes are listed in Table 1.

2.2. Feature Selection.
First, mRMR [16] was conducted to rank input features, that is, APA sites, to choose a refined feature set that had better discriminatory power than the original whole set. mRMR is a widely utilized filter-based feature selection method proposed by Peng et al. [16] on the basis of two criteria: (1) relevancy between feature and category must be large and (2) redundancy between features themselves must be small [20–22]. Given a dataset with m features, the mRMR follows the above criteria to select features one by one and added them into a feature list, which is empty initially. In detail, for each of the remaining features, its relevance to targets (class labels) was evaluated by mutual information and its redundancies were assessed to already-selected features. The feature with maximum relevance and minimum redundancy is selected and added to the current feature list. The obtained feature list was called the mRMR feature list. The mRMR program we used was downloaded from http://home.penglab.com/proj/mRMR/index.htm. Default parameters were used to perform such a program.

Second, IFS [17] and SVM [18] were integrated to select discriminatory features and their combination. A series of feature subsets was generated on the basis of the ranked features from mRMR. Then, the classification performance of

SVMs on the samples consisting of the generated feature subsets was evaluated. In the end, the feature subset with the best performance called optimum APA features, such as APA-modified genes, was selected.

SVM is a supervised learning model that can be used to analyse data, recognize feature patterns, and perform classification and regression analysis [18, 23–30]. The SVM constructs a hyperplane with a maximum margin between two groups of samples in a high-dimensional or infinite-dimensional space. SVM is also used to fit nonlinear data by mapping nonlinear data in a low-dimensional space to a high-dimensional space by a kernel trick. SVMs can also be extended for a multiclass problem by learning multiple binary SVM classifiers, and each classifier is used to classify one class from other classes. To quickly implement SVM, the tool "SMO" in Weka [31] was adopted in this study. The training procedures of this type of SVM are optimized by the sequential minimal optimization algorithm [32]. Default parameters were used. The kernel was a polynomial function, and the regularization parameter C was set to 1.

2.3. Performance Measurement. Performance measurement is an effective experimental estimation to assess the generalization performance of machine learning and can be used as an evaluation measurement to estimate the generalization performance of a learned model. In comparing different models, performance measurements should be objective and reflect the accuracy of models. Matthew's correlation coefficient (MCC) [33–38] for measuring multiclass classification performance is applied and formulated as follows:

$$
\begin{aligned}
\text{MCC} &= \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X)\,\text{cov}(Y, Y)}} \\
&= \frac{\sum_{i=1}^{n}\sum_{j=1}^{C}\left(x_{ij} - \bar{x}_j\right)\left(y_{ij} - \bar{y}_j\right)}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{C}\left(x_{ij} - \bar{x}_j\right)^2 \sum_{i=1}^{n}\sum_{j=1}^{C}\left(y_{ij} - \bar{y}_j\right)^2}},
\end{aligned}
\tag{1}
$$

where $\bar{x}_j$ and $\bar{y}_j$ are the means of x_j and y_j, respectively; Y is the truth label; and X is the predicted label. When MCC is 1, the classifier is extremely optimal. When MCC is 0, the learned classifier is not different from a random one. If MCC is −1, the classifier is the worst.

3. Results

In this study, we adopted several machine learning algorithms to analyse the TCGA-APA data. The purpose was to extract essential ATA features that can correctly distinguish different cancer types. The entire procedures are illustrated in Figure 1.

3.1. Results of mRMR. The mRMR was first applied to the TCGA-APA data. All APA features were deeply analysed and sorted in the mRMR feature list. The obtained feature list is provided in Table S1.

3.2. Results of IFS with SVM. Based on the mRMR feature list, the IFS method constructed feature subsets with step ten; that

is, the first ten features comprised the first feature subset; then, the second feature subset further added the next ten features, and so on. On each feature subset, an SVM classifier was built with samples represented by features in the subset. 10-fold crossvalidation was conducted to evaluate the performance of each SVM classifier. The accuracy of each cancer type, overall accuracy, and MCC were counted, which are available in Table S2. To give an overview of the performance of the SVM classifier on different numbers of top features, an IFS curve is plotted in Figure 2(a), in which MCC was set as the y-axis and the number of features as the x-axis. It can be observed that the SVM classifier with lots of APA features always gave a good performance. When the top 60 features were used, the SVM classifier can provide perfect performance with MCC = 1; that is, all cancer patients were classified into the correct cancer type. To investigate whether such perfect performance can be obtained with fewer features, we constructed all possible feature subsets containing 1-60 features. Likewise, an SVM classifier was built on each of these feature subsets. Also, 10-fold crossvalidation was adopted to assess each SVM classifier. Obtained measurements are also provided in Table S2. An IFS curve was also plotted, which is shown in Figure 2(b). It can be observed that when the top 45 features were adopted, the SVM classifier also provided the perfect performance with MCC = 1. Thus, the top 45 APA features were deemed as the optimum features. Furthermore, a perfect SVM classifier was built on these features, which can be a useful tool to discriminate different tumors.

3.3. Results of Survival Analysis on Top Ten Features. According to the results in Table S2, the SVM classifier with the top 10 features could reach MCC of 0.9217. This result indicated that the top ten APA features had significant APA patterns with a strong power on discriminating different tumors. These ten features are listed in Table 2. The selected APA-modified genes can discriminate different cancer types so that they would have prognostic power in a pan-cancer manner. Here, relying on the TCGA pan-cancer gene expression data and phenotype data (clinical information) [39], we firstly divided the samples into two parts according to the expression levels (expression quartiles). Using both the high-expression group and low-expression group datasets, we examined each of the top 10 genes for the survival analysis efficacy. The red survival curve shows the group of samples with a higher gene expression level, and the blue survival curve shows the group of samples with a lower expression level. In summary, the TCGA pan-cancer datasets were used to examine each of the top 10 genes based on survival analysis efficiency and are shown in Figure 3.

4. Discussion

4.1. Optimal APA-Associated Genes in Multiple Tumor Types. In this study, we extracted several important APA features as mentioned in Results of IFS with SVM. In addition, according to "Results of Survival Analysis on Top Ten Features," the top ten features can really indicate different cancer types.
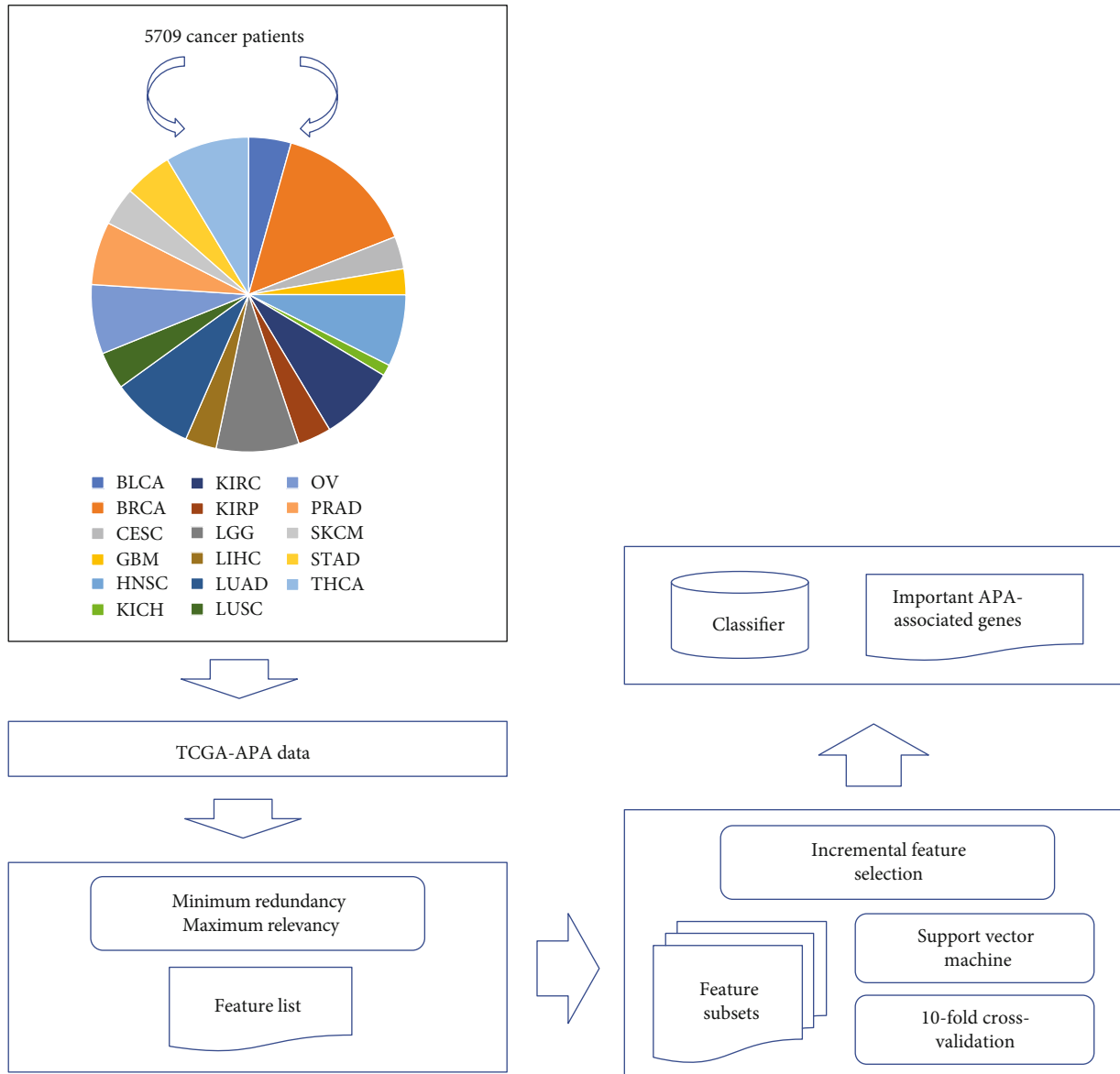
FIGURE 1: Entire procedures for the analysis of TCGA-APA data in different cancer types. The data is first analysed by the minimum redundancy maximum relevancy (mRMR) method, yielding a feature list. Then, incremental feature selection (IFS), incorporating a support vector machine, is applied on the list to extract important APA-associated genes and build an efficient classifier.

Here, we analysed the genes related to these APA features. All these identified genes were reported and validated to have different APA patterns in most of our 17 candidate tumor types. These results validated the efficacy and accuracy of our prediction. The detailed analysis of the APA pattern of the 10 optimal genes in different candidate tumor types is presented as follows. All such 10 optimal genes have been reported to be directly related to APA during tumorigenesis. The major regulatory effects of APA on such genes have been shown at three levels (Figure 4): directly affecting the expression levels and regulating related microRNAs and APA itself as a typical biomarker.

4.1.1. Genes Directly Regulated by APA at Expression Level in Multiple Tumor Types. The first examined APA-modified

gene is COPS7A, which contains six potential APA sites [40]. The transcripts of COPS7A have quite different APA sites in various tissues; thus, the gene has different APA-modified patterns in various tumor types [40]. Based on mRNA sequencing data from the TCGA database, recent publications have confirmed that COPS7A has a specific expression pattern in multiple tumor types, supporting our inference from an independent aspect [41]. As a detailed case, variant bAug10 with two unique APA sites, or the transcript of COPS7A, is specifically expressed in the colon and the ovary that distinguish tumor types derived from the two tissues from other tumor types [40]. This result validated the efficacy and accuracy of our prediction.

DPM2 is another predicted gene with a unique APA status in multiple tumor types. APA regulates the specific
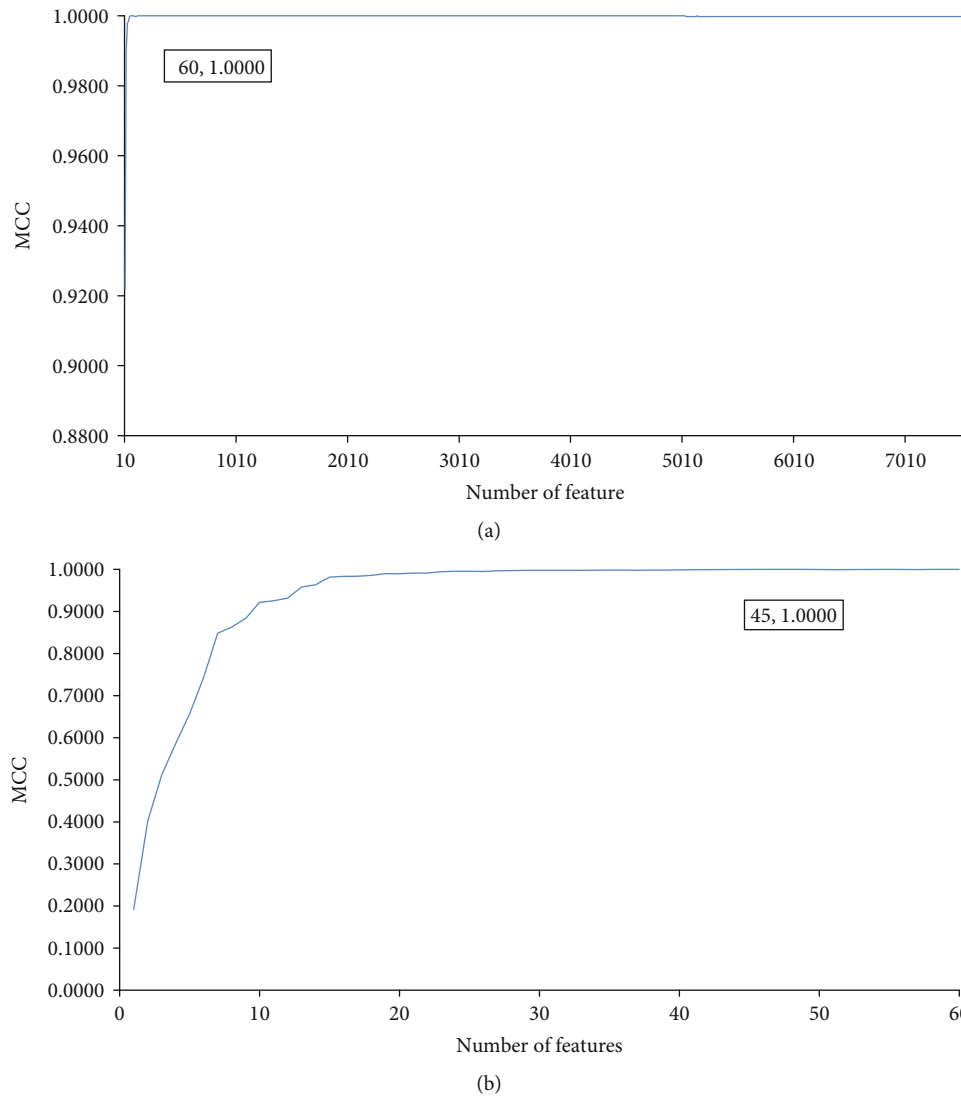
(a)

(b)

FIGURE 2: IFS curves to illustrate the performance of the support vector machine on different numbers of top features. (a) IFS curve with step 10; (b) IFS curve with 1-60 features. The perfect performance is obtained when the top 45 features are used.

TABLE 2: Top 10 features selected by the mRMR method.

Feature index	Gene name	Score
1	NM_001164095|COPS7A|chr12|+	0.91171
2	NM_001165415|LDHA|chr11|+	0.73417
3	NM_001242795|NUP93|chr16|+	0.68580
4	NM_003863|DPM2|chr9|-	0.66010
5	NM_153365|TAPT1|chr4|-	0.65060
6	NM_003377|VEGFB|chr11|+	0.63976
7	NM_001256661|TEAD2|chr19|-	0.63742
8	NM_001114394|PAPD4|chr5|+	0.62491
9	NM_001001701|C4orf3|chr4|-	0.61409
10	NM_002129|HMGB2|chr4|-	0.61438

biological function of the polyadenylation signal sequence and further contributes to the biosynthesis of dolichol phosphate-mannose in multiple mammalian cell subtypes [42]. Considering that dolichol phosphate mannose has different expression patterns in multiple cancer types, such as glioma and head and neck cancer, we can regard the APA status of DPM2 as a potential biomarker for the identification of different tumor types [43].

TEAD2, another predicted gene with an APA-modified pattern, may also have different APA modification patterns in various tumor types. A specific pattern of polyadenylation (AATAAA) on TEAD-2 regulates the expression of our predicted gene TEAD-2 involved in early mouse development, implying that this gene is regulated by APA modification [44]. In terms of the contribution of TEAD-2 to tumorigenesis, the APA modification of TEAD-2 may be functionally related to liver cancer development [45], suggesting that this gene may be a potential biomarker for the identification of a particular tumor type.

The next gene in our top-ranked prediction list is HMGB2, which is a member of the nonhistone chromosomal high-mobility group protein family. A specific study on non-
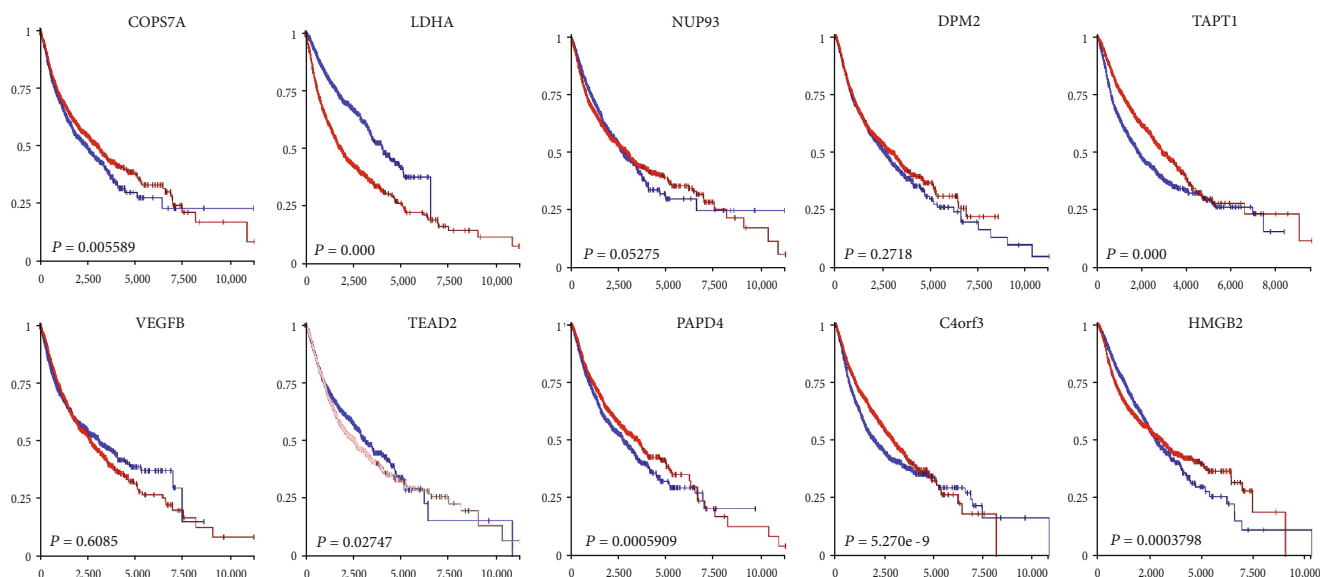
FIGURE 3: Survival analysis of the top 10 features in pan-cancer cohorts. Among the top 10 genes, seven genes had significant survival risks in the pan-cancer manner (red for the high-expression group, and blue for the low-expression group). Genes LDHA and TEAD2 are shown to predict a poor prognosis, and genes COPS7A, TAPT1, PAPD4, C4orf3, and HMGB2 are shown to indicate a good prognosis.
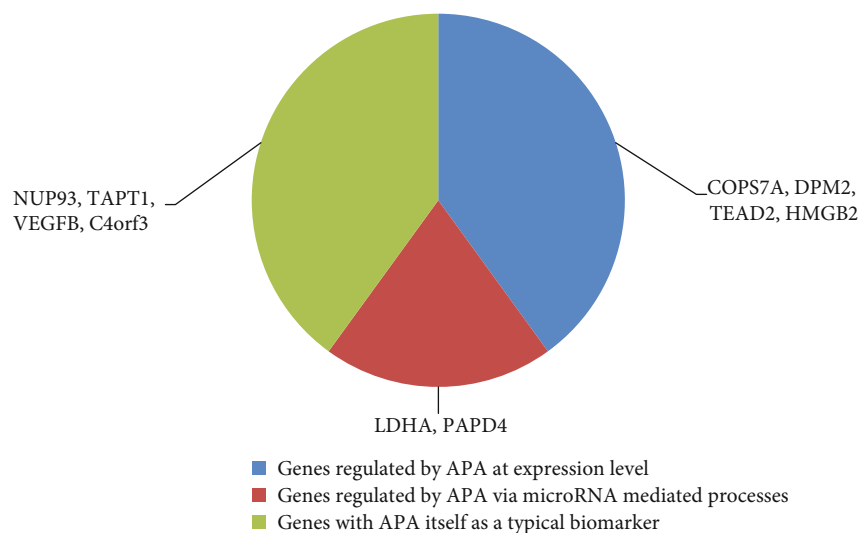

FIGURE 4: Summary of the APA-associated genes' regulatory methods on multiple tumor subtypes. Here, we summarized the three major subgroups of contributions that APA-associated genes contribute to different cancer subtypes together with the number of genes among the top 10 genes that contribute to each subgroup.

small-cell lung cancer transcriptome confirmed early in 2008 that the polyadenylation pattern may directly affect the progression of lung cancer [46]. The APA modification of HMGB2 may also be involved in thyroid cancer cells [46]. Therefore, in candidate tumor types, HMGB2, an APA site-targeting gene, may be differentially expressed or regulated in many tumor types, validating the efficacy and accuracy of our prediction.

4.1.2. Genes Regulated by APA via MicroRNA-Mediated Processes. LDHA is another gene with a differentially APA-modified pattern in candidate tumor types. With 14 potential APA sites, LDHA is differentially APA modified in different

tissues [47]. A recent study on hepatocellular carcinoma cells has confirmed that APA-modified LDHA may directly participate in tumorigenesis by regulating the biological functions of microRNAs, validating the specific contribution of APA modification to LDHA [47]. Therefore, in the candidate tumor types, the APA pattern of LDHA may contribute to the identification of LIHC on the basis of the abovementioned evidence.

The predicted gene PAPD4 is another specific biomarker for the identification of different tumor types. In contrast to other predicted genes, PAPD4 can participate in the polyadenylation of target mRNAs, indicating its specific contribution to APA [48]. In terms of the contribution

of APA to PAPD4, a study in 2014 reported that the APA modification of PAPD4 regulated by HBx may directly contribute to the HBV-related dysregulation of miR-122 [49]. Biological processes, that is, HBV-related dysregulation of miRNA, are functionally associated with hepatocellular carcinoma (LIHC), reflecting the cancer subtyping potential of PAPD4 [50].

4.1.3. Genes with APA Itself as a Typical Biomarker. The next predicted gene with differential APA patterns in different tumor types is NUP93. The transcript information from the NCBI AceView database supports NUP93 with five validated APA-modified sites, confirming the potential of APA-mediated transcription regulation on this gene [40]. In our study, NUP93 was APA modified in gastrointestinal diseases, including pancreatic cancer; thus, the APA status of NUP93 may be a potential indicator for the identification of pancreatic cancer [51]. The APA status of NUP93 in glioma, in addition to pancreatic cancer, is also tumor-specific [52, 53]. Therefore, with a unique APA status in pancreatic cancer and glioma, the predicted NUP93 may be an effective indicator for the identification of different tumor types.

TAPT1 has specific APA patterns in multiple tumor types. As a transmembrane protein, the APA modification of this gene affects the stability of the encoded protein's transmembrane structure [54, 55]. In terms of the contribution of APA to TAPT1 in different tumor types, APA modification affects the $3'$ UTR of TAPT1 in hepatocellular carcinoma cell lines [56], indicating that this specific pattern of APA modification may contribute to the identification of LIHC from other tumor types.

VEGFB is functionally modified by APA under multiple physical and pathological conditions [57]. With five validated APA sites, the abnormal APA-modified transcripts of VEGFB contribute to the pathogenesis of chronic liver disease and even LIHC, and this finding was consistent with our prediction [58]. Similarly, the specific APA modification of VEGFB may be involved in the functional regulation of CPEB1 and CPEB4 [58], further contributing to the tumorigenesis of multiple tumor types, including cervical, ovarian, and glioma cancers [59]. Therefore, the specific pattern of APA modification on our predicted VEGFB might also be an applicable biomarker of different tumor types, validating the efficacy and accuracy of our prediction.

C4orf3 is a predicted gene that may be functionally related to APA-mediated tumorigenesis. It is APA modified and forms a functional fusion transcript named KLHL2-C4orf3 fusion transcript, which is merged by the second, third, and fourth exons of KLHL2 and the first intron of C4orf3 [60]. Therefore, the modified APA sites of C4orf3 may be a potential APA target of the fusion transcript KLHL2-C4orf3. Considering that this fusion transcript is functionally related to lung adenocarcinoma but not to other tumor types, specific APA-modified patterns of C4orf3-induced transcript or fusion transcript may be potential biomarkers for the identification of lung adenocarcinoma [60].

4.2. Optimal APA-Associated Genes in Pan-Cancer Survival Analysis. APA-modified genes can participate in multiple cancer development and progression. As such, APA-modified genes play important roles in pan-cancer tumorigenesis, including potential prognosis power in pan-cancer cohorts. Thus, the survival risk of each APA gene (Figure 4) was evaluated on the basis of the TCGA pan-cancer gene expression data and phenotype data [39]. For each APA gene, the pan-cancer samples were divided into two groups based on expression quartiles, where one group of the samples had a high gene expression (red survival curve) and the other group of the samples had a low gene expression (blue survival curve). Of the 10 APA genes, 7 had significant survival risks in a pan-cancer manner. LDHA and TEAD2 were oncogenes whose high expression levels indicated a poor survival expectation. By contrast, other genes, including COPS7A, TAPT1, PAPD4, C4orf3, and HMGB2, have a tumor suppressor effect, and their high expression levels characterize patients with a good survival potential. Thus, 70% of our 10 APA genes had satisfactory prognostic power in a pan-cancer way, thereby supporting our analysis efficiency.

Overall, this study examined 10 potential gene biomarkers with differential APA-modified patterns in different tumor types. The 10 identified biomarkers were validated by recent publications, reflecting the efficacy and accuracy of the study. The presented computational approach might contribute to the identification of potential APA target sites at a whole genome level and provide a new approach to reveal the significant role of APA-induced RNA modification underlying tumorigenesis.

5. Conclusions

This study analysed the APA sites for multiple tumor types using several computational methods. Several key APA-modified genes were extracted, which can distinguish different tumor types; that is, they can be potential tumor biomarkers.

Data Availability

Previously reported data were used to support this study and are available at Synapse. These prior studies (and datasets) are cited at relevant places within the text as references [19].

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

Supplementary Materials

Supplementary 1. Table S1: ranked gene list from the mRMR method.

Supplementary 2. Table S2: classification performance yielded by the IFS method and SVM classifiers with different numbers of features.

References

[1] R. Bark, C. Mercke, E. Munck-Wikland, N. A. Wisniewski, and L. Hammarstedt-Nordenvall, "Cancer of the gingiva," *European Archives of Oto-Rhino-Laryngology*, vol. 273, no. 6, pp. 1335–1345, 2016.

[2] L. B. Alexandrov, "Signatures of mutational processes in human cancer," *Molecular Cancer Research*, vol. 15, 2017.

[3] B. Tian and J. L. Manley, "Alternative polyadenylation of mRNA precursors," *Nature Reviews Molecular Cell Biology*, vol. 18, no. 1, pp. 18–30, 2017.

[4] Y. Lin, Z. Li, F. Ozsolak et al., "An in-depth map of polyadenylation sites in cancer," *Nucleic Acids Research*, vol. 40, no. 17, pp. 8460–8471, 2012.

[5] I. Passacantilli, V. Panzeri, P. Bielli et al., "Alternative polyadenylation of ZEB1 promotes its translation during genotoxic stress in pancreatic cancer cells," *Cell Death & Disease*, vol. 8, no. 11, article e3168, 2017.

[6] B. Berkovits and C. Mayr, "Alternative $3'$ UTRs act as scaffolds to regulate membrane protein localization and function," *FEBS Journal*, vol. 282, pp. 37–37, 2015.

[7] B. D. Berkovits and C. Mayr, "Alternative $3'$ UTRs act as scaffolds to regulate membrane protein localization," *Nature*, vol. 522, no. 7556, pp. 363–367, 2015.

[8] Y. Zhu, X. Wang, E. Forouzmand et al., "Molecular mechanisms for CFIm-mediated regulation of mRNA alternative polyadenylation," *Molecular Cell*, vol. 69, no. 1, pp. 62–74.e4, 2018.

[9] W. Ma, C. Chen, Y. Liu et al., "Coupling of microRNA-directed phased small interfering RNA generation from long noncoding genes with alternative splicing and alternative polyadenylation in small RNA-mediated gene silencing," *The New Phytologist*, vol. 217, no. 4, pp. 1535–1550, 2018.

[10] K. Meyer, T. Köster, C. Nolte et al., "Adaptation of iCLIP to plants determines the binding landscape of the clock-regulated RNA-binding protein AtGRP7," *Genome Biology*, vol. 18, no. 1, p. 204, 2017.

[11] J. Sommer, C. Garbers, J. Wolf et al., "Alternative intronic polyadenylation generates the interleukin-6 trans-signaling inhibitor sgp130-E10," *Journal of Biological Chemistry*, vol. 289, no. 32, pp. 22140–22150, 2014.

[12] D. H. Zheng and B. Tian, "RNA-binding proteins in regulation of alternative cleavage and polyadenylation," in *Systems Biology of RNA Binding Proteins*, G. Yeo, Ed., vol. 825 of Advances in Experimental Medicine and Biology, pp. 97–127, Springer, New York, NY, USA, 2014.

[13] T. Sheng, H. Li, W. Zhang et al., "NUDT21 negatively regulates PSMB2 and CXXC5 by alternative polyadenylation and contributes to hepatocellular carcinoma suppression," *Oncogene*, vol. 37, no. 35, pp. 4887–4900, 2018.

[14] J. Ichinose, K. Watanabe, A. Sano et al., "Alternative polyadenylation is associated with lower expression of PABPN1 and poor prognosis in non-small cell lung cancer," *Cancer Science*, vol. 105, no. 9, pp. 1135–1141, 2014.

[15] A. R. Morris, A. Bos, B. Diosdado et al., "Alternative cleavage and polyadenylation during colorectal cancer development," *Clinical Cancer Research*, vol. 18, no. 19, pp. 5256–5266, 2012.

[16] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[17] H. A. Liu and R. Setiono, "Incremental feature selection," *Applied Intelligence*, vol. 9, no. 3, pp. 217–230, 1998.

[18] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[19] Y. Xiang, Y. Ye, Y. Lou et al., "Comprehensive characterization of alternative polyadenylation in human cancer," *Journal of the National Cancer Institute*, vol. 110, no. 4, pp. 379–389, 2018.

[20] L. Chen, X. Pan, X. H. Hu et al., "Gene expression differences among different MSI statuses in colorectal cancer," *International Journal of Cancer*, vol. 143, no. 7, pp. 1731–1740, 2018.

[21] J. R. Li, L. Lu, Y.-. H. Zhang et al., "Identification of synthetic lethality based on a functional network by using machine learning algorithms," *Journal of Cellular Biochemistry*, vol. 120, no. 1, pp. 405–416, 2019.

[22] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information," *Mathematical Biosciences*, vol. 306, pp. 136–144, 2018.

[23] X. Y. Pan and H. B. Shen, "Robust prediction of B-factor profile from sequence using two-stage SVR based on random forest feature selection," *Protein and Peptide Letters*, vol. 16, no. 12, pp. 1447–1454, 2009.

[24] A. H. Mirza, C. H. B. Berthelsen, S. E. Seemann et al., "Transcriptomic landscape of lncRNAs in inflammatory bowel disease," *Genome Medicine*, vol. 7, no. 1, p. 39, 2015.

[25] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 383–391, 2019.

[26] L. Chen, S. Wang, Y. H. Zhang et al., "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[27] Y.-D. Cai, S. Zhang, Y. H. Zhang et al., "Identification of the gene expression rules that define the subtypes in glioma," *Journal of Clinical Medicine*, vol. 7, no. 10, p. 350, 2018.

[28] J.-P. Zhou, L. Chen, and Z.-H. Guo, "iATC-NRAKEL: an efficient multi-label classifier for recognizing anatomical therapeutic chemical classes of drugs," *Bioinformatics*, vol. 36, no. 5, pp. 1391–1396, 2020.

[29] J.-P. Zhou, L. Chen, T. Wang, and M. Liu, "iATC-FRAKEL: a simple multi-label web server for recognizing anatomical therapeutic chemical classes of drugs with their fingerprints only," *Bioinformatics*, vol. 36, no. 11, pp. 3568-3569, 2020.

[30] J. Che, L. Chen, Z. H. Guo, S. Wang, and Aorigele, "Drug target group prediction with multiple drug networks," *Combinatorial Chemistry & High Throughput Screening*, vol. 23, no. 4, pp. 274–284, 2020.

[31] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*, Morgan Kaufmann Pub, 2005.

[32] J. Platt, "Sequential minimal optimizaton: a fast algorithm for training support vector machines," Technical Report MSR-TR-98-14, 1998.

[33] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Computational Biology and Chemistry*, vol. 28, no. 5-6, pp. 367–374, 2004.

[34] L. Chen, C. Chu, Y. H. Zhang et al., "Identification of drug-drug interactions using chemical interactions," *Current Bioinformatics*, vol. 12, no. 6, pp. 526–534, 2017.

[35] X. Zhao, L. Chen, Z. H. Guo, and T. Liu, "Predicting drug side effects with compact integration of heterogeneous networks," *Current Bioinformatics*, vol. 14, no. 8, pp. 709–720, 2019.

[36] X. Pan, L. Chen, K. Y. Feng et al., "Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms," *International Journal of Molecular Sciences*, vol. 20, no. 9, p. 2185, 2019.

[37] X. Zhang, L. Chen, Z. H. Guo, and H. Liang, "Identification of human membrane protein types by incorporating network embedding methods," *IEEE Access*, vol. 7, pp. 140794–140805, 2019.

[38] H. Liang, L. Chen, X. Zhao, and X. Zhang, "Prediction of drug side effects with a refined negative sample selection strategy," *Computational and Mathematical Methods in Medicine*, vol. 2020, Article ID 1573543, 16 pages, 2020.

[39] R. Neapolitan, C. M. Horvath, and X. Jiang, "Pan-cancer analysis of TCGA data reveals notable signaling pathways," *BMC Cancer*, vol. 15, no. 1, p. 516, 2015.

[40] D. Thierry-Mieg and J. Thierry-Mieg, "AceView: a comprehensive cDNA-supported gene and transcripts annotation," *Genome Biology*, vol. 7, article S12, 2006.

[41] C. A. Wicker and T. Izumi, "Analysis of RNA expression of normal and cancer tissues reveals high correlation of COP9 gene expression with respiratory chain complex components," *BMC Genomics*, vol. 17, no. 1, p. 983, 2016.

[42] Y. Maeda, S. Tomita, R. Watanabe, K. Ohishi, and T. Kinoshita, "DPM2 regulates biosynthesis of dolichol phosphate-mannose in mammalian cells: correct subcellular localization and stabilization of DPM1, and binding of dolichol phosphate," *The EMBO Journal*, vol. 17, no. 17, pp. 4920–4929, 1998.

[43] C. Lindskog, "The potential clinical impact of the tissue-based map of the human proteome," *Expert Review of Proteomics*, vol. 12, no. 3, pp. 213–215, 2015.

[44] K. J. Kaneko and M. L. DePamphilis, "Soggy, a spermatocyte-specific gene, lies 3.8 kb upstream of and antipodal to TEAD-2, a transcription factor expressed at the beginning of mouse development," *Nucleic Acids Research*, vol. 28, no. 20, pp. 3982–3990, 2000.

[45] C. Heinzle, Z. Erdem, J. Paur et al., "Is fibroblast growth factor receptor 4 a suitable target of cancer therapy?," *Current Pharmaceutical Design*, vol. 20, no. 17, pp. 2881–2898, 2014.

[46] A. Tanney, G. R. Oliver, V. Farztdinov et al., "Generation of a non-small cell lung cancer transcriptome microarray," *BMC Medical Genomics*, vol. 1, no. 1, p. 20, 2008.

[47] X. Li, P. Lu, B. Li et al., "Sensitization of hepatocellular carcinoma cells to irradiation by miR-34a through targeting lactate dehydrogenase-A," *Molecular Medicine Reports*, vol. 13, no. 4, pp. 3661–3667, 2016.

[48] R. Yamagishi, T. Tsusaka, H. Mitsunaga, T. Maehata, and S. I. Hoshino, "The STAR protein QKI-7 recruits PAPD4 to regulate post-transcriptional polyadenylation of target mRNAs," *Nucleic Acids Research*, vol. 44, no. 6, pp. 2475–2490, 2016.

[49] F. Peng, X. Xiao, Y. Jiang et al., "HBx down-regulated Gld2 plays a critical role in HBV-related dysregulation of miR-122," *PLoS One*, vol. 9, no. 3, article e92998, 2014.

[50] J. Zhou, L. Yu, X. Gao et al., "Plasma microRNA panel to diagnose hepatitis B virus-related hepatocellular carcinoma," *Journal of Clinical Oncology*, vol. 29, no. 36, pp. 4781–4788, 2011.

[51] A. R. Bauer, *Method for the early detection of pancreatic cancer and other gastrointestinal disease conditions*, 2007.

[52] V. Patil, J. Pal, and K. Somasundaram, "Elucidating the cancer-specific genetic alteration spectrum of glioblastoma derived cell lines from whole exome and RNA sequencing," *Oncotarget*, vol. 6, no. 41, pp. 43452–43471, 2015.

[53] M. Delaleau and K. L. Borden, "Multiple export mechanisms for mRNAs," *Cell*, vol. 4, no. 3, pp. 452–473, 2015.

[54] E. E. Creemers, A. Bawazeer, A. P. Ugalde et al., "Genome-wide polyadenylation maps reveal dynamic mRNA 3'-end formation in the failing human heart," *Circulation Research*, vol. 118, no. 3, pp. 433–438, 2016.

[55] L. Wang, R. D. Dowell, and R. Yi, "Genome-wide maps of polyadenylation reveal dynamic mRNA 3'-end formation in mammalian cell lineages," *RNA*, vol. 19, no. 3, pp. 413–425, 2013.

[56] Z. Qiu, K. Zou, L. Zhuang et al., "Hepatocellular carcinoma cell lines retain the genomic and transcriptomic landscapes of primary human cancers," *Scientific Reports*, vol. 6, no. 1, article 27411, 2016.

[57] M. H. Dijkstra, E. Pirinen, J. Huusko et al., "Lack of cardiac and high-fat diet induced metabolic phenotypes in two independent strains of V*egf-b* knockout mice," *Scientific Reports*, vol. 4, article 6238, 2014.

[58] V. Calderone, J. Gallego, G. Fernandez-Miranda et al., "Sequential functions of CPEB1 and CPEB4 regulate pathologic expression of vascular endothelial growth factor and angiogenesis in chronic liver disease," *Gastroenterology*, vol. 150, no. 4, pp. 982–997.e30, 2016, e30.

[59] C. N. Hansen, Z. Ketabi, M. W. Rosenstierne, C. Palle, H. C. Boesen, and B. Norrild, "Expression of CPEB, GAPDH and U6snRNA in cervical and ovarian tissue during cancer development," *APMIS*, vol. 117, no. 1, pp. 53–59, 2009.

[60] Y. Hong, W. J. Kim, C. Y. Bang, J. C. Lee, and Y. M. Oh, "Identification of alternative splicing and fusion transcripts in non-small cell lung cancer by RNA sequencing," *Tuberculosis and Respiratory Diseases*, vol. 79, no. 2, pp. 85–90, 2016.