Received December 2, 2019, accepted December 17, 2019, date of publication December 23, 2019, date of current version January 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2019.2961402

Screening Dys-Methylation Genes and Rules for Cancer Diagnosis by Using the Pan-Cancer Study

YU-HANG ZHANG[1,2], **TAO ZENG**[3], **XIAOYONG PAN**[4,5], **WEI GUO**[6], **ZIJUN GAN**[2], **YUNHUA ZHANG**[7], **TAO HUANG**[2], **AND YU-DONG CAI**[1]

[1] School of Life Sciences, Shanghai University, Shanghai 200444, China
[2] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[3] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China
[4] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China
[5] Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai Jiao Tong University, Shanghai 200240, China
[6] Institute of Health Sciences, Shanghai Jiao Tong University School of Medicine and Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[7] Anhui Province Key Laboratory of Farmland Ecological Conservation and Pollution Prevention, School of Resources and Environment, Anhui Agricultural University, Hefei 230036, China

Corresponding authors: Tao Huang (tohuangtao@126.com) and Yu-Dong Cai (cai_yud@126.com)

This work was supported in part by the Shanghai Municipal Science and Technology Major Project under Grant 2017SHZDZX01, in part by the National Key Research and Development Program of China under Grant 2018YFC0910403, in part by the National Natural Science Foundation of China under Grant 31701151 and Grant 31872418, in part by the Shanghai Sailing Program under Grant 16YF1413800, in part by the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) under Grant 2016245, and in part by the fund of the Key Laboratory of Stem Cell Biology of Chinese Academy of Sciences under Grant 201703.

ABSTRACT Although cancer has long been a major public health problem worldwide, conquering cancer still remains unachievable due to its complexity and diversity. With the development of high-throughput sequencing technologies, the combination of conventional clinical symptoms and new genetic events is becoming effective and has been approved for precise prediction and innovative diagnostic strategies. Epigenetic modification (e.g., DNA methylation) is an important mechanism of transcriptional control in normal or disease functions. Depending on the unique methylation profiles, an important possibility raised is that the characteristic of dys-methylation could provide a new molecular marker system for the identification of the major forms of tumors. In this study, we attempted to distinguish different tumor types. On the basis of DNA methylation data from PanCanAtlas in The Cancer Genome Atlas, we applied mRMR and MCFS methods together to identify the decision rules for distinguishing 33 different tumor types and ranked the features that characterized methylation level. This study highlights the considerable application potential of methylation features in cancer diagnosis and provides insight into novel therapeutic targets.

INDEX TERMS Methylation, rule, cancer diagnosis, pan-cancer.

I. INTRODUCTION

For centuries, cancer has been a major public health problem worldwide. Numerous scientists have devoted themselves to cancer research and attained notable achievements [1]. Nevertheless, the desired goal to conquer cancer remains unachievable because of its complexity and diversity. Cancer can be characterized into different forms in accordance with tumor location, cellular origin, biological process, and genomic alteration profiles, all of which can enhance oncogenesis or affect therapeutic response. Each type of cancer has a unique spectrum of genetic aberrations, including single nucleotide variations, copy number variations, varied gene expression profiles, and different epigenetic alterations [2]. With the development of sequencing technologies, precise and innovative prediction and diagnostic strategies based on the combination of conventional clinical symptoms and new genetic events have become effective and been approved [3]. These strategies are expected to further contribute to cancer treatment in the next decades.

Among genetic events, epigenetic modification (e.g., DNA methylation) is an important mechanism of transcriptional control in mammals. It plays a crucial role in maintaining cellular function on the basis of the proper regulation of gene expression and stable gene silencing in normal cells. However, promoter hypermethylational silencing has been confirmed to give rise to cancer through the transcriptional inhibition of critical growth regulators, such as tumor suppressor genes [4], [5]. Numerous pieces of evidence also point that DNA methylation has a direct role in carcinogenesis.

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

For example, the promoter regions of retinoblastoma and von Hippel–Lindau genes are hypermethylated in unilateral retinoblastoma and renal cancer, respectively [6], [7]. Similarly, the p16INK4a promoter is methylated in lung cancer and mammary epothelial cells [8], and the frequency of aberrant methylation increases with disease progression [9]. This phenomenon indicates that aberrant hypermethylation events act as a primary inactivating event contributing to tumorigenesis.

Certain genes, such as cell cycle inhibitor P16INK4A, exhibit promoter hypermethylation in almost all types of cancers, whereas some others genes would display high frequencies of methylation only in very specific tumor types. For example, GSTP1 is hypermethylated in liver, breast, and kidney cancers but shows little or even no methylation in other tumor types [10]. Similarly, the hypermethylation of BRCA1 is found only in breast and ovarian carcinomas [11], whereas the methylation of mismatch repair gene hMLH1 is restricted to three tumor types: colorectal, endometrial, and gastric tumors [12], [13]. An important possibility raised on the basis of the unique profiles of methylation is that the characteristic of gene methylation alterations might provide a molecular marker system for the identification of the major forms of tumors, which can be considered as potentially powerful strategies for cancer diagnosis.

The Cancer Genome Atlas (TCGA) is a well-known landmark cancer genomics program that aims to catalogue and discover molecular aberrations through large-scale genome sequencing and multidimensional analyses at the DNA, RNA, protein, and epigenetic levels [14]. This project provides access to cancer genomic datasets so that researchers could apply them to improve scientific inquiry, diagnostic methods, and eventual cancer treatments. By now, over 30,000 cancer cases have been recruited, which can be classified into 33 different tumor types. The Pan-Cancer analysis project was launched by TCGA to identify molecular aberrations among distinct cancer types that otherwise would have been missed and to broaden analytical breadth by defining commonalities and differences across cancer types and organs of origin [15].

In this study, we aimed to construct an impactful approach for distinguishing different tumor types. This approach may provide insight into novel therapeutic targets. On the basis of DNA methylation data from PanCanAtlas, we applied the minimum redundancy maximum relevance (mRMR) [16] and Monte Carlo feature selection (MCFS) [17] methods together to identify the decision rules that distinguish 33 different tumor types and ranked the features that were characteristic of methylation level. This study highlights the potential application of methylation features in cancer diagnosis.

II. MATERIALS AND METHODS

A. DATASET

The 450K methylation profiles of TCGA pan-cancers were downloaded from PanCanAtlas (https://gdc.cancer.gov/about-data/publications/pancanatlas). A total of 9,664 samples belonging to 33 cancer types were retrieved. The sample size of each cancer type is listed in **Table 1**. The beta values of 450K methylation probes were used to represent each cancer sample. We explored the methylation landscape of different cancer types and investigated the cancer-specific epigenetic patterns.

TABLE 1. The sample size of each cancer type.

Index	Cancer type	Cancer Name	Sample Size
1	ACC	Adrenocortical carcinoma	79
2	BLCA	Bladder Urothelial Carcinoma	437
3	BRCA	Breast invasive carcinoma	874
4	CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	311
5	CHOL	Cholangiocarcinoma	45
6	COAD	Colon adenocarcinoma	330
7	DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	48
8	ESCA	Esophageal carcinoma	199
9	GBM	Glioblastoma multiforme	154
10	HNSC	Head and Neck squamous cell carcinoma	575
11	KICH	Kidney Chromophobe	65
12	KIRC	Kidney renal clear cell carcinoma	472
13	KIRP	Kidney renal papillary cell carcinoma	315
14	LAML	Acute Myeloid Leukemia	194
15	LGG	Lower Grade GLioma	532
16	LIHC	Liver hepatocellular carcinoma	427
17	LUAD	Lung adenocarcinoma	491
18	LUSC	Lung squamous cell carcinoma	405
19	MESO	Mesothelioma	87
20	OV	Ovarian serous cystadenocarcinoma	10
21	PAAD	Pancreatic adenocarcinoma	194
22	PCPG	Pheochromocytoma and Paraganglioma	186
23	PRAD	Prostate adenocarcinoma	546
24	READ	Rectum adenocarcinoma	101
25	SARC	Sarcoma	265
26	SKCM	Skin Cutaneous Melanoma	475
27	STAD	Stomach adenocarcinoma	395
28	TGCT	Testicular Germ Cell Tumor	155
29	THCA	Thyroid carcinoma	567
30	THYM	Thymoma	126
31	UCEC	Uterine Corpus Endometrial Carcinoma	467
32	UCS	Uterine Carcinosarcoma	57
33	UVM	Uveal Melanoma	80
Total	---	---	9664

B. FEATURE RANKING

Different feature ranking methods can yield different important features in accordance with their principles. Thus, these methods should be evaluated and analyzed together. In this study, we used max relevance (MR) score to filter the original features and remain relevant features with large MR scores. Then we applied two widely used feature selection methods respectively, mRMR [16] and MCFS [17], to rank these relevant features. Finally, we performed incremental feature selection (IFS) [18] on the basis of the two lists of top-ranked features by applying an integrated supervised classifier and obtained the optimal features with the best classification performance in distinguishing samples from 33 cancers.

MR score. If one feature is closely relevant to the output class, then this feature should be important. In this study, we calculated the MR score for each feature, which was defined as the mutual information between class labels and the feature.

mRMR. The mRMR method [16], [19]–[22] measures the contribution of each feature in classification according to its trade-off between the maximum relevance and the minimum redundancy, e.g. one feature has a high rank in selected feature list when it has large relevance with class label and small redundancy with other already selected features. mRMR program was retrieved from http://home.penglab.com/proj/mRMR/index.htm. The output mRMR feature list was used in the following procedures.

MCFS. MCFS is a decision tree-based feature selection method [17], [23]–[25]. It constructs multiple decision trees, and each tree is grown on the bootstrap sample set and feature subset. First, p bootstrap sets are randomly generated by sampling with replacement, and t feature subsets are generated with a much smaller number of features randomly sampled from the original features. In total, $p \times t$ decision trees are grown on the basis of the combination of these bootstrap sets and feature subsets. On the basis of these decision trees, a relative importance (RI) score is calculated for each feature, and highly important features will be selected as (node) features in the decision trees with high frequency. The RI score of each feature is calculated in accordance with the number of splits involving this feature for all $p \times t$ trees and weighted on the basis of the classification accuracy of individual decision trees. In this study, one MCFS implementation is adopted and downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html. After the RI score of each feature was computed, all features were ranked in the decreasing order of their RI scores.

IFS. Given the above feature ranking by mRMR and MCFS, we further selected the optimum features with the best classification performance instead of using an arbitrary or experience-based score cutoff. Thus, we performed IFS with an integrated supervised classifier (i.e., random forest (RF)) to select the optimum features. A series of feature subsets are constructed on the basis of the top 2000 ranked features. Feature subset 1 contains the top one feature, then feature subset 2 contains the top two features, and so on. For each feature subset, a classifier is trained and evaluated on the samples consisting of the features from this feature subset using 10-fold cross-validation. In the end, an optimum feature subset with the best performance is selected.

C. RF

RF is a widely used supervised ensemble classifier consisting of multiple decision trees [20], [26]–[29]. In RF, each decision tree is grown on one bootstrap set and a feature subset randomly selected from the original features. During the growth of individual trees, RF uses out-of-bag error to estimate the generalization ability on the out-of-bag samples, which are the remaining samples not in the bootstrap set but in the original training set. After training multiple decision trees, given a new instance, RF predicts its class label in accordance with the maximum predicted probability of these prior-trained multiple decision trees.

D. RIPPER

Compared with RF, Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [30] can provide a more understandable prediction on the basis of classification rules, thus facilitating the discovery of additional details about the contribution of gene methylation to different cancers.

RIPPER is a greedy rule-learning method that is based on sequential covering. It learns some IF–THEN rules. One rule is that if the methylation values of several genes are greater or smaller than some values in one sample, then a certain prediction about the cancer types of this sample can be made. For binary classification, RIPPER first learns a rule from the training set, and all samples covered by this rule are removed. Then, the next rule is learned on the remaining samples. This learning process is repeated until all samples in the training set are classified on the basis of the produced rules. RIPPER takes a one-vs-all strategy for multiclass classification. It first orders the classes in accordance with the number of samples in the training set. Then, it learns a rule for the least common class, where the positive samples are from this class and the negative samples are from other classes. Next, the samples covered by this rule are removed, and a second rule is learned from the remaining samples for the second least common class, and so on. The last class is a default rule, which predicts a new instance to this class if the above learned rules cannot be satisfied. In this study, we used JRip implemented in MCFS package for RIPPER analysis.

E. PERFORMANCE MEASUREMENT

In this study, we applied RF and RIPPER as the multiclassification classifiers to classify samples from the 33 cancers in TCGA. Performance was evaluated by measurements of individual accuracy, overall accuracy and Matthews correlation coefficient (MCC) [20], [31]–[34] using 10-fold cross-validation [35], [36].

The individual accuracy is defined for each class (cancer type), which is the proportion of correctly predicted samples in one class among all samples in such class. Clearly, these accuracies cannot fully assess the performance of classifiers. The overall accuracy is further employed, which is defined as the proportion of correctly predicted samples among all samples. Its calculation formula is

$$\text{Overall accuracy} = \frac{n}{N}, \tag{1}$$

where n is the number of correctly predicted samples and N stands for the total number of samples. As listed in Table 1, sizes of 33 cancer types are of great differences. The biggest cancer type contains about 87 times samples as many as those in the smallest cancer type. In this case, overall accuracy is not a good measurement. The MCC in multi-class was employed in this study because it is deemed as a balanced measurement even if the class sizes greatly vary. It is defined as

$$MCC = \frac{cov(X, Y)}{\sqrt{cov(X, X)cov(Y, Y)}}, \tag{2}$$

where X and Y stand for two binary matrices. X represents the predicted class of each sample and Y indicates the true class of each sample.

III. RESULTS

In this study, we analyzed the methylation profiles of samples in 33 cancer types. Entire procedures are illustrated in **Fig. 1**.

A. RESULTS OF FEATURE RANKING BY MRMR AND MCFS

We first calculated MR scores for all input features (e.g., methylation sites). We only retained features with MR scores greater than 0.3, i.e., 53,435 important futures of methylation as given in **Tables S1** and **S2**. We further ranked these features by using mRMR and MCFS. Obtained feature lists are given in **Tables S1** and **S2**, respectively.

B. RESULTS OF THE CLASSIFICATION PERFORMANCE OF RF

Given that training on the above ranked feature lists is time consuming, we first selected the top 2,000 features and generated 2,000 ordered feature subsets on the basis of the feature list yielded by mRMR method. Then, we ran IFS with an integrated RF for classifying samples from 33 cancers. Given one feature subset, RF was trained and evaluated on the samples consisting of features from this feature subset with 10-fold cross-validation. The performance measurements corresponding to different numbers of features are given in **Table S3.** For easy observation, a curve was plotted in **Fig. 2A** with MCC as the Y-axis and number of used features as X-axis. It can be observed that when the top 1768 features were used, RF yielded the best MCC value of 0.955 (**Table 2**). The corresponding overall accuracy was 0.957 (**Table 2**). For 33 individual accuracies, seven cancer types received perfect classification and the accuracies on 24 cancer types were larger than 0.950. The detailed performance of such RF classifier on each cancer type is shown in **Fig. 3A**. Furthermore, by carefully checking the performance measurements listed in **Table S3**, the MCC and overall accuracy can reach 0.940 and 0.943 (**Table 2**), respectively, when top 126 features (called optimum set 1) were used, which were only a little lower than those of RF with top 1768 features. Considering the efficiency of the RF classifier, RF with top 126 features was a more proper choice. Its performance on 33 cancer types is illustrated in **Fig. 3A**, which was almost same as those obtained by the RF with top 1768 features.

In addition, we evaluated the top features on the basis of the feature list produced by MCFS in the same way. The performance measurements corresponding to different number of features are given in **Table S4**. A curve was plotted in **Fig. 2B** to show the performance of RF on different feature subsets. When the top 1805 features were used, RF yielded the best MCC value of 0.972 (**Table 2**) in 10-fold cross-validation. The overall accuracy of such RF classifier was 0.973 (**Table 2**). **Fig. 3B** shows the individual accuracies on 33 cancer types. This RF classifier gave perfection classification on ten cancer types and 29 cancer types received the accuracies higher than 0.950. Like to the mRMR results, we also found that RF yielded a high MCC value of 0.950 and overall accuracy of 0.952 (**Table 2**) when using the top 211 features (called optimum set 2). The performance of this classifier on 33 cancer types is illustrated in **Fig. 3B**. They were almost same as those of RF with top 1805 features.

TABLE 2. Performance and optimum number of features of IFS with RF and RIPPER for mRMR and MCFS.

Classifier	Feature selection method	Number of optimum features	MCC	Overall accuracy
RF	mRMR	1768	0.955	0.957
RF	mRMR	126	0.940	0.943
RF	MCFS	1805	0.972	0.973
RF	MCFS	211	0.950	0.952
RIPPER	mRMR	221	0.828	0.835
RIPPER	mRMR	48	0.782	0.791
RIPPER	MCFS	249	0.831	0.838
RIPPER	MCFS	78	0.762	0.772

From the above results, we can see that RF yielded a similar performance when using the top features ranked by mRMR or MCFS, but the number of overlapping features between optimum set 1 and 2 is small in this work. Thus, in the discussion below, we separately analyzed the methylation features in optimum sets 1 and 2.

C. RESULTS OF THE CLASSIFICATION PERFORMANCE OF RIPPER

RIPPER is much more time-consuming than RF. Thus, we ran the IFS with RIPPER on the top 250 features ranked by mRMR and MCFS (i.e., covering the features in optimum sets 1 and 2). First, we evaluated RIPPER on the top features ranked by mRMR. The performance measurements corresponding to different number of features are given in **Table S5**. For easy observation, a curve was plotted in **Fig. 4A**, where Y-axis represents MCC and X-axis stands for the number of used features. The best MCC value of 0.828 (**Table 2**) when using the top 221 features. The overall accuracy yielded by such classifier was 0.835 (**Table 2**). 33 individual accuracies are shown in **Fig. 5A**, where nine accuracies exceeded 0.900. It can be seen that this RIPPER classifier was much inferior to above-mentioned RF classifiers. However, it can provide more insights. Furthermore, RIPPER can provide an MCC value of 0.782 (**Table 2**) when using the top 48 features. The overall accuracy of this classifier was 0.791 (**Table 2**). They were all about 4% lower than those of RIPPER with top 221 features. The individual accuracies are also shown in **Fig. 5A**. On most cancer types, this classifier gave lower performance than RIPPER with top 221 features. However, it can execute with less time because it used much less features. Thus, we believed that this classifier was a more proper choice.

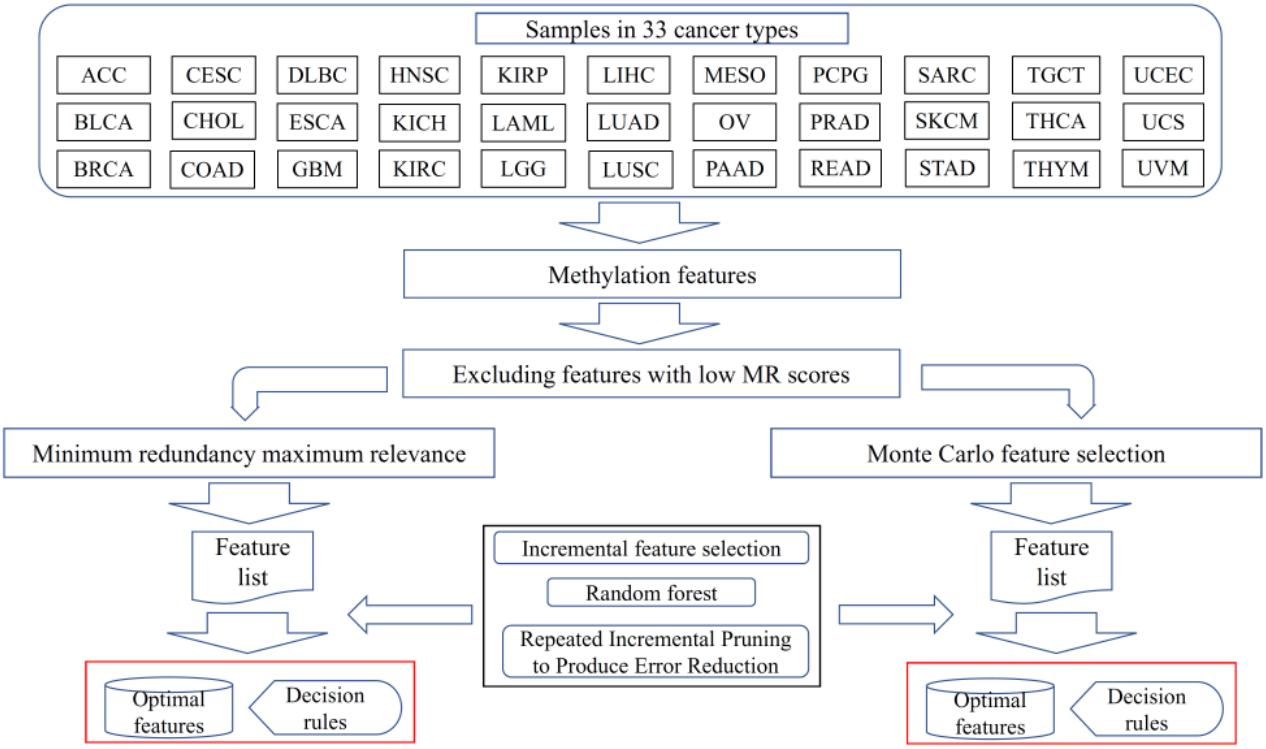


FIGURE 1. Entire procedures to analyze the methylation profiles on samples in 33 cancer types. The irrelevant methylation features that had low MR scores to cancer types were discarded. Remaining features were analyzed by minimum redundancy maximum relevance (mRMR) and Monte Carlo feature selection (MCFS), respectively, result in a feature list, respectively. Each feature list was fed into the incremental feature selection (IFS) method with random forest and Repeated Incremental Pruning to Produce Error Reduction to access optimum features and decision rules.

Similarly, we also evaluated RIPPER on the features ranked by MCFS. The performance measurements corresponding to different number of features are given in **Table S6**. As shown in **Fig. 4B and Table 2**, RIPPER yielded the best MCC value of 0.831 when the top 249 features were used. The overall accuracy was 0.838. Detailed individual accuracies are illustrated in **Fig. 5B**. Eleven accuracies exceeded 0.900. Similarly, when only the top 78 features were used, RIPPER yielded an MCC value of 0.762 and an overall accuracy of 0.772 (**Table 2**). This classifier was a good choice for making prediction because it was faster than the RIPPER with 249 features. Its performance on 33 cancer types is shown in **Fig. 5B**, which was only a little lower than that of the RIPPER with 249 features.

Compared with RF, RIPPER yielded reduced classification performance but produced numerous readable classification rules. As discussed above, for the feature list yielded by the mRMR method, RIPPER with top 48 features was an idea classifier; while for the list yielded by the MCFS method, RIPPER with top 78 features was a good choice. Therefore, the RIPPER was applied on these two sets of features to produce classification rules based on all 9664 cancer samples. 275 classification rules were produced on top 48 features yielded by mRMR method, which are listed in **Table S7**. 302 classification rules learned by RIPPER on the basis of the top 78 features from MCFS, which are provided in **Table S8**.

IV. DISCUSSION

In this study, we constructed two optimal classifiers with high classification accuracy of >0.940 to distinguish various tumor types. These classifiers were built on 126 and 211 selected features identified through the mRMR and MCFS method, respectively. In addition, top features analyzed by the mRMR and MCFS methods were then applied to generate 275 and 302 decision rules via RIPPER (**Tables S7-S8**). These rules could classify the cancer samples into corresponding categories with a high overall accuracy of >0.770. We focused on several top features and decision rules because they have a particular importance in classification, indicating that these features (e.g., the methylation level of these sites) likely play crucial roles in cancer processes. To validate the reliability of our findings, we examined existing experimental evidence through a wide literature review.

A. ANALYSIS OF THE TOP FEATURES IDENTIFIED BY THE MRMR METHOD

IFFO1 (probeID: cg08875705; cg22203219), a member of the intermediate filament family, is correlated with CA-125 levels in patient serum and can be an indicator of disease burden and relapse after tumor resection. The promoter methylation of IFFO1 is a candidate tumor marker given its frequent methylation detected in ovarian tumors [37]. Similarly, the expression and methylation level of IFFO1 are

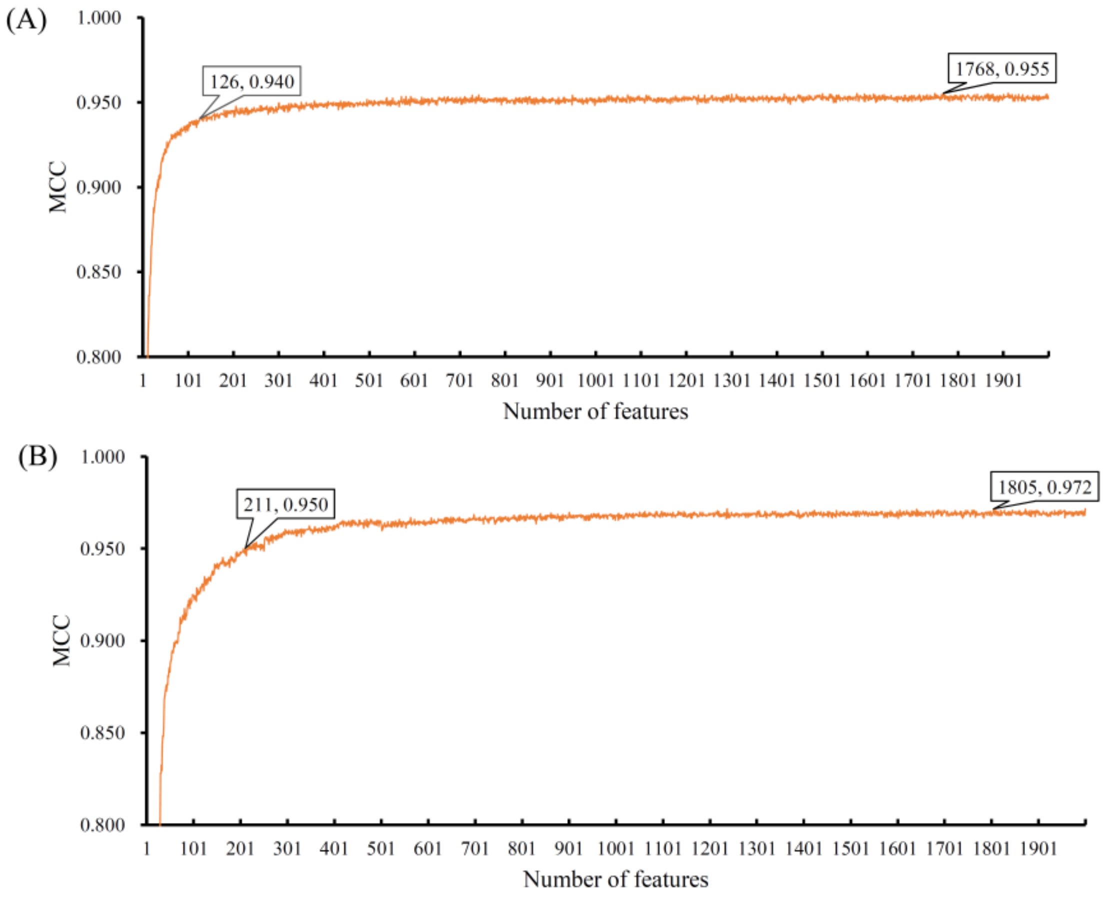


FIGURE 2. Performance of random forest (RF) changes with the number of top features ranked by minimum redundancy maximum relevance (mRMR) and Monte Carlo feature selection (MCFS). A) Performance of RF changes with the number of top features ranked by mRMR, the highest MCC was 0.955 when top 1768 features were adopted; B) Performance of RF changes with the number of top features ranked by MCFS, the best MCC was 0.972 when top 1805 features were used.

related to lung adenocarcinoma, indicating their strong potential as prognostic indicators for lung cancer [38]. These results suggest that IFFO1 has a significant effect in cancer process and exerts its effect through gene methylation at the DNA level.

FNDC3B (probeID: cg04319611), also named as fibronectin type III domain containing 3B, is a protein-coding gene that regulates cell motility. Recent publications have demonstrated that FNDC3B plays an important role in hepatocellular carcinoma by promoting cell migration and tumor metastasis [39]. Meanwhile, FNDC3B is a target of miR-143, the upregulated expression of which promotes the metastasis of prostate cancer cells by modulating FNDC3B expression [40]. Nuclear factor κB can adjust miR-143 to repress FNDC3B, and the downregulation of FNDC3B enhances invasion and migration capability [41].

The protein encoded by **INPP5A** (probeID: cg26549601) is a membrane-associated type I inositol 1,4,5-trisphosphate 5-phosphatase. The overexpression of INPP5A inhibits cell proliferation and invasion capacity and promotes cell apoptosis [42]. A decrease in INPP5A expression levels seems to be a prior event in cutaneous squamous cell carcinoma, indicating that INPP5A may play an important role in tumor development and progression [43].

The proto-oncogene **SRC** (probeID: cg24055525), which may play a role in the regulation of embryonic development and cell growth, is upregulated and highly activated in various types of human cancer, e.g., SRC, and affects the development of cell adhesion, invasion, proliferation, and survival [44]. The combined alteration of EGFR and SRC expression in fibroblasts improves the levels of tumorigenicity [45]. The high activation levels of phosphorylated SRC substrates in tumor cells indicate a strong link to metastasis and tumorigenicity induction [46].

Protein produced by **DCPS** (probeID: cg15958289) protects cells with the capability of removing short mRNA fragments containing a cap structure; this process leads to potentially toxic-associated accumulation in the cell. The downregulation of DCPS expression could be a potential cause of the aberrant miRNA profiles observed in cancer [47].

MAPK8IP3 (probeID: cg02330721), which is also known as JIP3, is a scaffold protein implicated in the

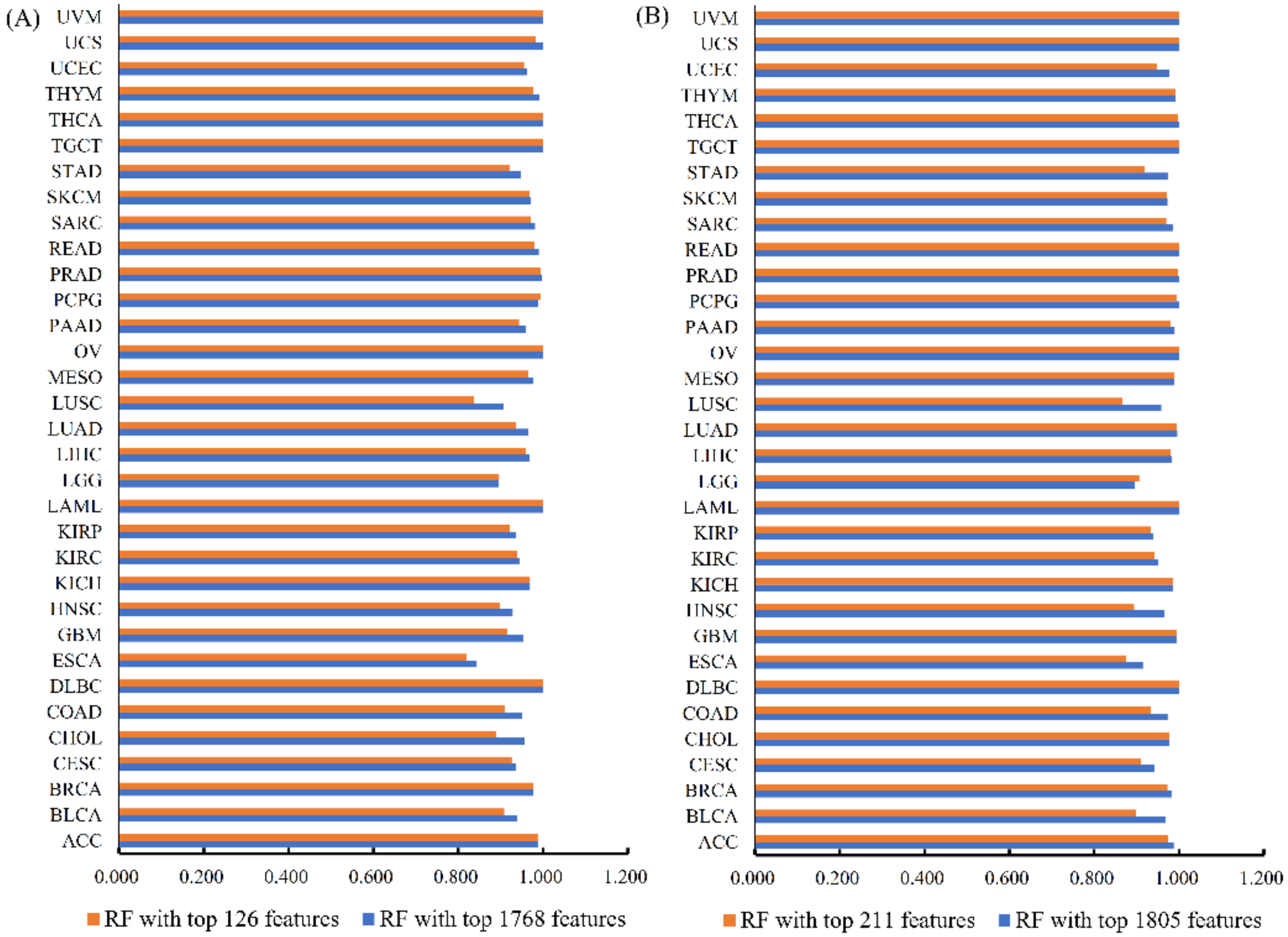


FIGURE 3. Individual accuracies on 33 cancer types yielded by random forest (RF) with some top features yielded by minimum redundancy maximum relevance (mRMR) or Monte Carlo feature selection (MCFS). A) Performance of RF with some top features yielded by mRMR, RFs with top 126 and 1768 features gave similar performance; B) Performance of RF with some top features yielded by MCFS, RFs with top 211 and 1805 features gave similar performance.

JNK pathway [48]. This gene is overexpressed during the initiation and invasion period of pancreatic cancer, thus contributing to the further excessive proliferation of pancreatic cells [49]. The specific mutations of JIP3 influence the adhesion and invasion function of various cancer cells, suggesting the crucial role of JIP3 in cancer [50].

The RNA-coding gene **MIR141** (probeID: cg19794481) or microRNA-141 displays specific expression profiles and can be applied as a biomarker in various cancers. The increased expression of miR-141 is significantly associated with the survival and prognosis of patients with serous ovarian carcinoma [51]. The serum levels of miR-141 can be used to distinguish patients with prostate cancer from healthy individuals and represents a stable blood-based marker for prostate cancer diagnosis [52]. In addition, circulating plasma miR-141 is significantly correlated with stage IV colon cancer and could be applied as an independent prognostic factor for colon cancer [53].

B. ANALYSIS OF THE TOP FEATURES IDENTIFIED BY THE MCFS METHOD

SLC22A23(probeID: cg26673629) belongs to a large family of transmembrane proteins that act as symporters and antiporters to transport organic ions across cell membranes [54]. A prediction model for the recurrence of triple negative breast cancer with high accuracy, sensitivity, and specificity was developed on the basis of several genes, including SLC22A23 [55], and may provide evidence for the potential role of SLC22A23 in carcinogenesis. Meanwhile, the SLC22A23 gene was selected as a differentially expressed gene in a study aiming to identify crucial genes involved in the pathogenesis of laryngeal squamous cell carcinoma [56]. A similar conclusion was also confirmed in another research reporting that the expression of SLC22A23 is associated with laryngeal cancer [57], suggesting that SLC22A23 may be involved in the development of laryngocarcinoma.

Zinc finger protein 500, abbreviated as **ZNF500** (probeID: cg10278046; cg11422964), is a protein-coding gene that may

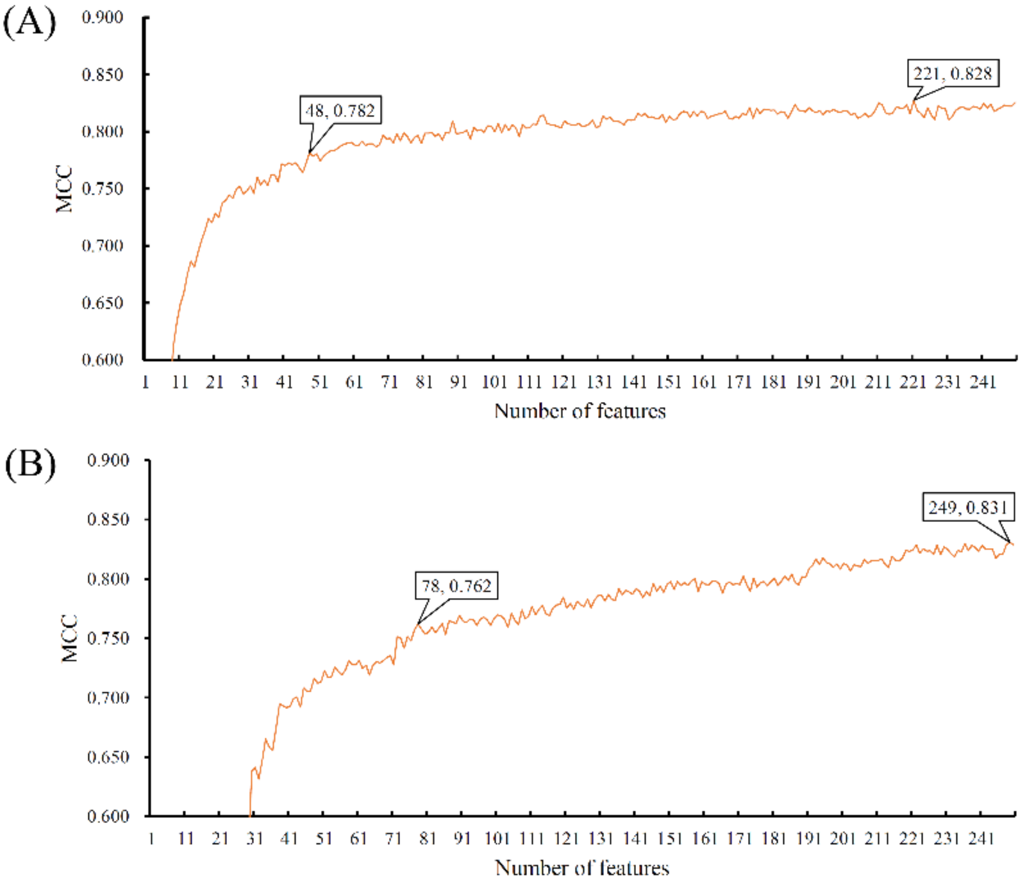


FIGURE 4. Performance of Repeated Incremental Pruning to Produce Error Reduction (RIPPER) changes with the number of top features ranked by minimum redundancy maximum relevance (mRMR) and Monte Carlo feature selection (MCFS). A) Performance of RIPPER changes with the number of top features ranked by mRMR, the highest MCC was 0.828 when top 221 features were involved; B) Performance of RIPPER changes with the number of top features ranked by MCFS, the best MCC was 0.831 when top 249 features were used.

participate in transcriptional regulation. This gene has been identified as a candidate gene for breast cancer risk given that the expression of ZNF500 exhibits consistent allelic expression imbalance with CCV genotype [58]. In fact, a growing body of evidence has revealed the potential roles of the zinc finger protein family in cancer progression. ZNF306 and ZNF309 promote cancer cell growth, migration, and angiogenesis in colorectal cancer [59], [60] and can enhance cell proliferation in multiple myeloma [61]. ZNF388 and ZNF489 play crucial roles in lung cancer by modulating the target gene p53; this modulatory role leads to the migration and invasion of cancer cells [62].

Five of the top 10 features identified by the MCFS method belong to the **IFFO1** gene (cg17198308, cg00363813, cg08875705, cg23737737, and cg00983904), indicating that five methylation sites with particular classification importance are located in different regions of one gene. Notably, the two top-ranked features identified by mRMR method were also annotated in INFFO1. All these results strongly suggest that the methylation of IFFO1 gene has an important effect in cancer.

Genes in the ETS family are implicated in diverse cancers, such as sarcoma, acute myeloid leukemia, and chronic myelomoncytic leukemia [63]. The protein-coding gene **ELF3** (probeID: cg26328757) has been identified as a member of the ETS gene family. The expression of ELF3 in lung carcinoma and adenovarcinoma is higher than in normal tissues [64]. Another study, which conducted experiments with overexpressed ELF3 in human breast ductal carcinoma in situ, confirmed that ELF3 can differentially activate several malignancy-associated gene promoters [65]. A recent work found that ELF3 is recurrently amplified and upregulated in colorectal cancer tissues. Moreover, ELF3 drives β-catenin transactivation and is significantly associated with the poor survival of colorectal cancer patients [66].

The protein encoded by **PPM1F** (probeID: cg12894883) is a member of the PP2C family of Ser/Thr protein phosphatases. The overexpression of PPM1F rescues the miR149-mediated inhibition of cell migration and invasion in hepatocellular carcinoma, suggesting its facilitation effect and potential therapeutic target for HCC treatment [67].

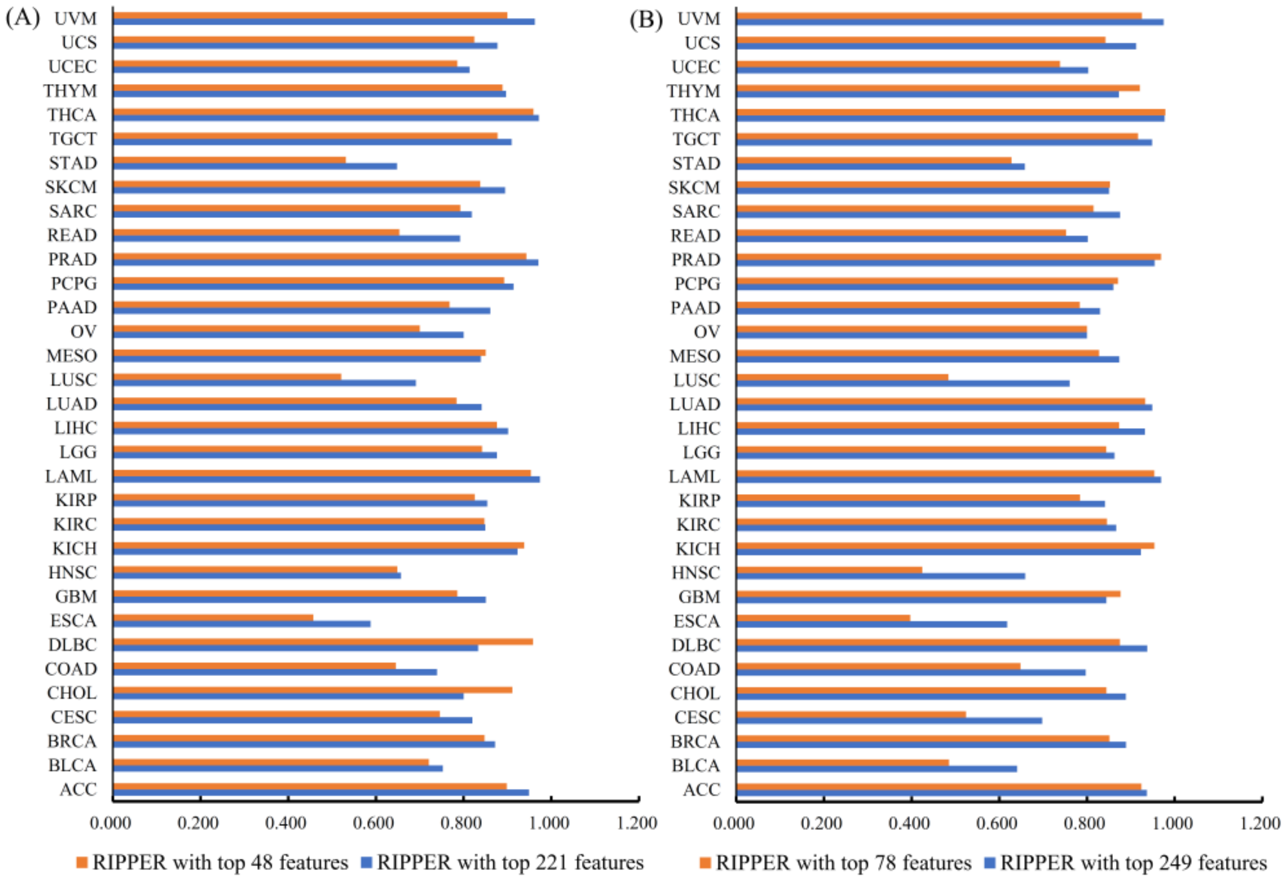


FIGURE 5. Individual accuracies on 33 cancer types yielded by Repeated Incremental Pruning to Produce Error Reduction (RIPPER) with some top features yielded by minimum redundancy maximum relevance (mRMR) or Monte Carlo feature selection (MCFS). A) Performance of RIPPER with some top features yielded by mRMR, RIPPERs with top 48 and 221 features gave similar performance; B) Performance of RIPPER with some top features yielded by MCFS, RIPPERs with top 78 and 249 features gave similar performance.

C. ANALYSIS OF DECISION RULES IDENTIFIED ON THE BASIS OF THE MRMR METHOD

A total of 275 decision rules for distinguishing 33 different types of cancers were identified on the basis of the mRMR method. Among them, the top 10, which contains 81 criteria, could be used to identify pancreatic adenocarcinoma; the next five rules could be used to identify lymphoid neoplasm diffuse large B-cell lymphoma; the following six rules could identify kidney chromophobe; and cholangiocarcinoma could be identified by 10 rules. Using this approach, we could classify diverse cancer samples into their corresponding category on the basis of decision rules.

In the 16 rules identifying sarcoma, two criteria involved one feature (cg10518264) located in the **HLA-DMB** gene region and requiring high methylation. HLA-DMB belongs to HLA class II beta chain paralogues. A nested case–control study found that a mutation in HLA-DMB increases the risk of developing Kaposi's sarcoma [68], suggesting the repression role of HLA-DMB in sarcoma. This finding is consistent with the rules predicted in our study that the aberrant expression of HLA-DMB caused by high methylation is an indicator of sarcoma.

The feature (cg01835695) involved in five rules identifying cervical squamous cell carcinoma was required to be in low methylation levels, showing that the hypomethylation of the annotated gene, **MGAT1**, is an indicator of cervical carcinoma. In vitro experiments have found that MGAT1, a member of the glycosyltransferase family, plays a crucial role in cancer progression [69]. The migration and invasion capability of HeLa cell was inhibited as demonstrated by MGAT1 knockdown experiment, indicating that MGAT1 likely promotes cervical carcinoma [70]. This conclusion confirmed our decision rules that low methylation levels of MGAT1 lead to a high risk of cervical carcinoma.

Among the 81 criteria involved in 10 rules that could be used to identify pancreatic adenocarcinoma, one feature (cg18440692) was required at low methylation levels in two criteria. This feature represents the methylation degree of one site located in **FAM53B**, a protein-coding gene that acts as a regulator of the Wnt signaling pathway. A research aiming to identify lncRNA biomarkers in pancreatic cancer showed that FAM53B antisense RNA1 (FAM53B-AS1) is positively associated with overall survival[71]. It implies that the upregulation of FAM53B may enhance the progression of pancreatic adenocarcinoma given that antisense RNA acts as an mRNA inhibitor and reduces protein expression. It is consistent with the decision rules that require the low methylation levels of FAM53B for identifying pancreatic cancer.

D. ANALYSIS OF DECISION RULES IDENTIFIED ON THE BASIS OF THE MCFS METHOD

We constructed 302 rules from 78 features involved in 1,572 criteria on the basis of the MCFS method. Among the 17 decision rules that could indicate bladder urothelial carcinoma, four rules attracted our attention. Two features (cg25042226 and cg06881093) are both required in low methylation levels in the decision rules. These features are located in the region of **PAX8**, which encodes a member of the paired box family of transcription factors. Immunohistochemical and RT-PCR studies have shown that PAX8 is expressed in the majority of bladder urothelial neoplasis but not in normal adult urothelial epithelium [72]. Moreover, the expression of PAX8 may contribute to urothelial tumorigenesis through p53 given that PAX8 inhibits p53 expression [73], and the absence of p53 promotes urothelial cell proliferation [74]. The rules that required hypomethylation lead to the high expression of PAX8, which may enhance the development of bladder urothelia carcinoma.

A criterion that required the hypermethylation of **CUX1** (cg06010390) to identify pancreatic adenocarcinoma is found to have experimental support. The transcription factor CUX1 is a regulator of cell differentiation and cell cycle progression. The accumulated evidence revealed that reduced CUX1 expression facilitates tumor initiation. The partial knockdown of CUX1 in human cord blood progenitors caused an increase in engraftment upon transplantation into immunodeficient mice [75]. Increased tumor formation was found with CUX1 knockdown in T-cell acute lymphoblastic leukemia cells after subcutaneous injection into immunodeficient mice [76]. These results have established that CUX1 acts as a tumor suppresser gene and that the reduced expression of CUX1 can promote tumor cell proliferation.

In the 17 rules identifying lung squamous cell carcinoma, three criteria require that the important gene **DPP9** (cg01098142) remain at low mehtylation levels. DPP9 encodes a protein that is a member of the ubiquitous atypical serine proteases family, which has been linked to various diseases, including type 2 diabetes, obesity, and cancer [77]. A recent study reported that the expression levels of DPP9 are significantly increased in non-small cell lung cancer (NSCLC) tissues compared with adjacent normal tissues and are highly correlated with a poor overall survival rate in patients with NSCLC [78]. Furthermore, loss-of-function experiments demonstrated that the downregulation of DPP9 suppresses the proliferation, migration, and invasion of NSCLC cells, suggesting the potential tumor promotion role of DPP9 in NSCLC. These results confirmed our decision rules that the upregulation of DPP9 by demethylation could contribute to the progression of lung carcinoma.

We applied the mRMR and MCFS methods together to identify diverse decision rules from the PanCanAtlas data with the aim to distinguish different tumor types on the methylation level. Our decision rules can distinguish 33 different tumor types. The ranked features can be used to characterize the methylation levels of different cancer sites and provide insight into the novel therapeutic targets. This study highlights the wide application potential of methylation features in cancer diagnosis.

V. CONCLUSION

In this study, the methylation profiles on samples in 33 cancer types were deeply analyzed by some machine learning algorithms. Some important methylation features and classification rules were accessed. Genes related to important methylation features and rules were extensively discussed. Hopefully, the new findings reported in this study may give new insights on pan-cancer study based on methylation.

ACKNOWLEDGMENT

(Yu-Hang Zhang, Tao Zeng, and Xiaoyong Pan contributed equally to this work.)

Supporting information: Table S1: The genes with MR scores greater than 0.3 and the list on them yielded by mRMR; Table S2: The genes with MR scores greater than 0.3 and their RI scores calculated by MCFS; Table S3: The classification performance of RF when using different number of top features ranked by mRMR; Table S4: The classification performance of RF when using different number of top features ranked by MCFS; Table S5: The classification performance of RIPPER when using different number of top features ranked by mRMR; Table S6: The classification performance of RIPPER when using different number of top features ranked by MCFS; Table S7: The learned classification rules by RIPPER based on the top 48 features ranked by mRMR; Table S8. The learned classification rules by RIPPER based on the top 78 features ranked by MCFS.

REFERENCES

[1] D. Hanahan and R. A. Weinberg, "The hallmarks of cancer," *Cell*, vol. 100, no. 1, pp. 57–70, 2000.

[2] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, p. 719, 2009.

[3] A. A. Friedman, A. Letai, D. E. Fisher, and K. T. Flaherty, "Precision medicine for cancer with next-generation functional diagnostics," *Nature Rev. Cancer*, vol. 15, no. 12, p. 747, 2015.

[4] S. B. Baylin and J. G. Herman, "DNA hypermethylation in tumorigenesis: Epigenetics joins genetics," *Trends Genet.*, vol. 16, no. 4, pp. 168–174, 2000.

[5] P. A. Jones, "DNA methylation errors and cancer," *Cancer Res.*, vol. 56, no. 11, pp. 2463–2467, 1996.

[6] C. Stirzaker, D. S. Millar, C. L. Paul, P. M. Warnecke, J. Harrison, P. C. Vincent, M. Frommer, and S. J. Clark, "Extensive DNA methylation spanning the Rb promoter in retinoblastoma tumors," *Cancer Res.*, vol. 57, pp. 2229–2237, 1997.

[7] J. G. Herman, F. Latif, Y. Weng, M. I. Lerman, B. Zbar, S. Liu, D. Samid, D. S. Duan, J. R. Gnarra, and W. M. Linehan, "Silencing of the VHL tumor-suppressor gene by DNA methylation in renal carcinoma," *Proc. Nat. Acad. Sci. USA*, vol. 91, no. 21, pp. 9700–9704, 1994.

[8] S. A. Belinsky, K. J. Nikula, W. A. Palmisano, R. Michels, G. Saccomanno, E. Gabrielson, S. B. Baylin, and J. G. Herman, "Aberrant methylation of $p16^{INK4a}$ is an early event in lung cancer and a potential biomarker for early diagnosis," *Proc. Nat. Acad. Sci. USA*, vol. 95, no. 20, pp. 11891–11896, 1998.

[9] D. J. Wong, S. A. Foster, D. A. Galloway, and B. J. Reid, "Progressive region-specific de novo methylation of the p16 CpG island in primary human mammary epithelial cell strains during escape from M_0 growth arrest," *Mol. Cellular Biol.*, vol. 19, no. 8, pp. 5642–5651, 1999.

[10] M. Esteller, P. G. Corn, S. B. Baylin, and J. G. Herman, "A gene hypermethylation profile of human cancer," *Cancer Res.*, vol. 61, no. 8, pp. 3225–3229, 2001.

[11] J. C. Tchou, X. H. Lin, D. Freije, W. B. Isaacs, J. D. Brooks, A. Rashid, A. M. De Marzo, Y. Kanai, S. Hirohashi, and W. G. Nelson, "GSTP1 CpG island DNA hypermethylation in hepatocellular carcinomas," *Int. J. Oncol.*, vol. 16, pp. 663–739, 2000.

[12] M. F. Kane, M. Loda, G. M. Gaida, J. Lipman, R. Mishra, H. Goldman, J. M. Jessup, and R. Kolodner, "Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines," *Cancer Res.*, vol. 57, no. 5, pp. 808–811, 1997.

[13] A. S. Fleisher, M. Esteller, S. Wang, G. Tamura, H. Suzuki, J. Yin, T.-T. Zou, J. M. Abraham, D. Kong, K. N. Smolinski, Y.-Q. Shi, M.-G. Rhyu, S. M. Powell, S. P. James, K. T. Wilson, J. G. Herman, and S. J. Meltzer, "Hypermethylation of the hMLH1 gene promoter in human gastric cancers with microsatellite instability," *Cancer Res.*, vol. 59, no. 5, pp. 1090–1095, 1999.

[14] K. Tomczak, P. Czerwi ska, and M. Wiznerowicz, "The cancer genome atlas (TCGA): An immeasurable source of knowledge," *Contemp. Oncol.*, vol. 19, no. 1A, p. A68, 2015.

[15] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature Genet.*, vol. 45, no. 10, p. 1113, 2013.

[16] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[17] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski, "Monte Carlo feature selection for supervised classification," *Bioinformatics*, vol. 24, pp. 110–117, Jan. 2008.

[18] H. Liu and R. Setiono, "Incremental feature selection," *Appl. Intell.*, vol. 9, no. 3, pp. 217–230, Nov./Dec. 1998.

[19] L. Chen, X. Pan, X. Hu, Y.-H. Zhang, S. Wang, T. Huang, and Y.-D. Cai, "Gene expression differences among different MSI statuses in colorectal cancer," *Int. J. Cancer*, vol. 143, no. 7, pp. 1731–1740, Oct. 2018.

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

[21] T. Wang, L. Chen, and X. Zhao, "Prediction of drug combinations with a network embedding method," *Combinat. Chem. High Throughput Screening*, vol. 21, no. 10, pp. 789–797, 2018.

[22] L. Chen, S. Wang, Y.-H. Zhang, J. Li, Z.-H. Xing, J. Yang, T. Huang, and Y.-D. Cai, "Identify key sequence features to improve CRISPR sgRNA efficacy," *IEEE Access*, vol. 5, pp. 26582–26590, 2017.

[23] X. Pan, L. Chen, K.-Y. Feng, X.-H. Hu, Y.-H. Zhang, X.-Y. Kong, T. Huang, and Y.-D. Cai, "Analysis of expression pattern of snoRNAs in different cancer types with machine learning algorithms," *Int. J. Mol. Sci.*, vol. 20, p. 2185, May 2019.

[24] L. Chen, X. Pan, Y.-H. Zhang, X. Kong, T. Huang, and Y.-D. Cai, "Tissue differences revealed by gene expression profiles of various cell lines," *J. Cellular Biochem.*, vol. 120, pp. 7068–7081, May 2019.

[25] L. Chen, J. Li, Y.-H. Zhang, K. Feng, S. Wang, Y. Zhang, T. Huang, X. Kong, and Y.-D. Cai, "Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method," *J. Cell Biochem.*, vol. 119, pp. 3394–3403, Apr. 2018.

[26] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

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

[28] 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.

[29] R. Zhao, L. Chen, B. Zhou, Z.-H. Guo, S. Wang, and Aorigele, "Recognizing novel tumor suppressor genes using a network machine learning strategy," *IEEE Access*, vol. 7, pp. 155002–155013, 2019.

[30] W. W. Cohen, "Fast effective rule induction," in *Proc. 12th Int. Conf. Mach. Learn.*, 1995, pp. 115–123.

[31] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta-Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975.

[32] J. Gorodkin, "Comparing two K-category assignments by a K-category correlation coefficient," *Comput. Biol. Chem.*, vol. 28, nos. 5–6, pp. 367–374, Dec. 2004.

[33] L. Chen, C. Chu, Y.-H. Zhang, M. Zheng, L. Zhu, X. Kong, and T. Huang, "Identification of drug-drug interactions using chemical interactions," *Current Bioinform.*, vol. 12, no. 6, pp. 526–534, 2017.

[34] H. Cui and L. Chen, "A binary classifier for the prediction of EC numbers of enzymes," *Current Proteomics*, vol. 16, no. 5, pp. 381–389, 2019.

[35] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, 1995, pp. 1137–1145.

[36] 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*, 2019.

[37] M. Campan, M. Moffitt, S. Houshdaran, H. Shen, M. Widschwendter, G. Daxenbichler, T. Long, C. Marth, I. A. Laird-Offringa, M. F. Press, L. Dubeau, K. D. Siegmund, A. H. Wu, S. Groshen, U. Chandavarkar, L. D. Roman, A. Berchuck, C. L. Pearce, and P. W. Laird, "Genome-scale screen for DNA methylation-based detection markers for ovarian cancer," *PLoS ONE*, vol. 6, p. e28141, Dec. 2011.

[38] N. N. Feng, Y. Wang, M. Zheng, X. Yu, H. Lin, R.-N. Ma, O. Shi, X. Zheng, M. Gao, H. Yu, L. Garmire, and B. Qian, "Genome-wide analysis of DNA methylation and their associations with long noncoding RNA/mRNA expression in non-small-cell lung cancer," *Epigenomics*, vol. 9, pp. 137–153, Feb. 2017.

[39] C. H. Lin, Y.-W. Lin, Y.-C. Chen, C.-C. Liao, Y.-S. Jou, M.-T. Hsu, and C.-F. Chen, "FNDC3B promotes cell migration and tumor metastasis in hepatocellular carcinoma," *Oncotarget*, vol. 7, pp. 49498–49508, Aug. 2016.

[40] X. Fan, X. Chen, W. Deng, G. Zhong, Q. Cai, and T. Lin, "Up-regulated microRNA-143 in cancer stem cells differentiation promotes prostate cancer cells metastasis by modulating FNDC3B expression," *BMC Cancer*, vol. 13, p. 61, Feb. 2013.

[41] X. Zhang, S. Liu, T. Hu, S. Liu, Y. He, and S. Sun, "Up-regulated microRNA-143 transcribed by nuclear factor kappa B enhances hepatocarcinoma metastasis by repressing fibronectin expression," *Hepatology*, vol. 50, pp. 490–499, Aug. 2009.

[42] M. Yang, X. Zhai, T. Ge, C. Yang, and G. Lou, "miR-181a-5p promotes proliferation and invasion and inhibits apoptosis of cervical cancer cells via regulating inositol polyphosphate-5-phosphatase a (INPP5A)," *Oncol. Res.*, vol. 26, pp. 703–712, Jun. 2018.

[43] A. Sekulic, S. Y. Kim, G. Hostetter, S. Savage, J. G. Einspahr, A. Prasad, P. Sagerman, C. Curiel-Lewandrowski, R. Krouse, G. T. Bowden, J. Warneke, D. S. Alberts, M. R. Pittelkow, D. DiCaudo, B. J. Nickoloff, J. M. Trent, and M. Bittner, "Loss of inositol polyphosphate 5-phosphatase is an early event in development of cutaneous squamous cell carcinoma," *Cancer Prevention Res.*, vol. 3, pp. 1277–1283, Oct. 2010.

[44] R. B. Irby and T. J. Yeatman, "Role of SRC expression and activation in human cancer," *Oncogene*, vol. 19, pp. 5636–5642, Nov. 2000.

[45] W. Mao, R. Irby, D. Coppola, L. Fu, M. Wloch, J. Turner, H. Yu, R. Garcia, R. Jove, and T. J. Yeatman, "Activation of c-SRC by receptor tyrosine kinases in human colon cancer cells with high metastatic potential," *Oncogene*, vol. 15, pp. 3083–3090, Dec. 1997.

[46] M. T. Brown and J. A. Cooper, "Regulation, substrates and functions of SRC," *Biochim. Biophys. Acta*, vol. 1287, pp. 121–149, Jun. 1996.

[47] O. Meziane, S. Piquet, G. D. Bossé, D. Gagné, E. Paquet, C. Robert, M. A. Tones, and M. J. Simard, "The human decapping scavenger enzyme DcpS modulates microRNA turnover," *Sci. Rep.*, vol. 5, p. 16688, Nov. 2015.

[48] T. Sun, N. Yu, L.-K. Zhai, N. Li, C. Zhang, L. Zhou, Z. Huang, X.-Y. Jiang, Y. Shen, and Z.-Y. Chen, ''c-Jun NH_2-terminal kinase (JNK)-interacting protein-3 (JIP3) regulates neuronal axon elongation in a kinesin- and JNK-dependent manner,'' *J. Biol. Chem.*, vol. 288, pp. 14531–14543, May 2013.

[49] C. L. Standen, N. J. Kennedy, R. A. Flavell, and R. J. Davis, ''Signal transduction cross talk mediated by Jun N-terminal kinase-interacting protein and insulin receptor substrate scaffold protein complexes,'' *Mol. Cell Biol.*, vol. 29, pp. 4831–4840, Sep. 2009.

[50] T. Takino, M. Nakada, H. Miyamori, Y. Watanabe, T. Sato, D. Gantulga, K. Yoshioka, K. M. Yamada, and H. Sato, ''JSAP1/JIP3 cooperates with focal adhesion kinase to regulate c-Jun N-terminal kinase and cell migration,'' *J. Biol. Chem.*, vol. 280, pp. 37772–37781, Nov. 2005.

[51] E. J. Nam, H. Yoon, S. W. Kim, H. Kim, Y. T. Kim, J. H. Kim, J. W. Kim, and S. Kim, ''MicroRNA expression profiles in serous ovarian carcinoma,'' *Clin. Cancer. Res.*, vol. 14, pp. 2690–2695, May 2008.

[52] P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin, and M. Tewari, ''Circulating microRNAs as stable blood-based markers for cancer detection,'' *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 30, pp. 10513–10518, Jul. 2008.

[53] K. X. Chen, L. Zhang, D. E. Cogdell, H. Zheng, A. J. Schetter, M. Nykter, C. C. Harris, K. Chen, S. R. Hamilton, and W. Zhang, ''Circulating plasma MiR-141 is a novel biomarker for metastatic colon cancer and predicts poor prognosis,'' *Cancer Res.*, vol. 6, no. 3, p. e17745, Apr. 2011.

[54] J. A. Jacobsson, T. Haitina, J. Lindblom, and R. Fredriksson, ''Identification of six putative human transporters with structural similarity to the drug transporter SLC22 family,'' *Genomics*, vol. 90, pp. 595–609, Nov. 2007.

[55] L. H. Chen, W.-H. Kuo, M.-H. Tsai, P.-C. Chen, C. K. Hsiao, E. Y. Chuang, L.-Y. Chang, F.-J. Hsieh, L.-C. Lai, and K.-J. Chang, ''Identification of prognostic genes for recurrent risk prediction in triple negative breast cancer patients in Taiwan,'' *PLoS ONE*, vol. 6, no. 11, p. e28222, 2011.

[56] N. Buyru, S. Ekizoglu, E. Karaman, and T. Ulutin, ''Expression of SLC22A23 gene in laryngeal carcinoma,'' *Cancer Res.*, vol. 77, pp. 2146–2146, Jul. 2017.

[57] S. Ekizoglu, D. Seven, T. Ulutin, J. Guliyev, and N. Buyru, ''Investigation of the SLC22A23 gene in laryngeal squamous cell carcinoma,'' *BMC Cancer*, vol. 18, p. 477, Apr. 2018.

[58] M. M. Marjaneh, H. Sivakumaran, K. Hillman, S. Kaufmann, N. Hussein, L. Lima, S. Ham, S. Kar, J. Beesley, L. Fachal, D. Easton, A. M. Dunning, A. Moller, G. Chenevix-Trench, S. Edwards, and J. D. French, ''High-throughput allelic expression imbalance analyses identify candidate breast cancer risk genes,'' *bioRxiv*, 2019.

[59] L. Yang, S. R. Hamilton, A. Sood, T. Kuwai, L. Ellis, A. Sanguino, G. Lopez-Berestein, and D. D. Boyd, ''The previously undescribed ZKSCAN3 (ZNF306) is a novel 'driver' of colorectal cancer progression,'' *Cancer Res.*, vol. 68, pp. 4321–4330, Jun. 2008.

[60] L. Yang, L. Zhang, Q. Y. Wu, and D. D. Boyd, ''Unbiased screening for transcriptional targets of ZKSCAN3 identifies integrin β4 and vascular endothelial growth factor as downstream targets,'' *J. Biol. Chem.*, vol. 283, pp. 35295–35304, Dec. 2008.

[61] L. Yang, H. Wang, S. M. Kornblau, D. A. Graber, N. Zhang, J. A. Matthews, M. Wang, D. M. Weber, S. K. Thomas, J. J. Shah, L. Zhang, G. Lu, M. Zhao, R. Muddasani, S.-Y. Yoo, K. A. Baggerly, and R. Z. Orlowski, ''Evidence of a role for the novel zinc-finger transcription factor ZKSCAN3 in modulating Cyclin D2 expression in multiple myeloma,'' *Oncogene*, vol. 30, pp. 1329–1340, Mar. 2011.

[62] J. Jen, L. L. Lin, F. Y. Lo, H. T. Chen, S. Y. Liao, Y. A. Tang, W.-C. Su, R. Salgia, C.-L. Hsu, H.-C. Huang, H.-F. Juan, and Y.-C. Wang, ''Oncoprotein ZNF322A transcriptionally deregulates alpha-adducin, cyclin D1 and p53 to promote tumor growth and metastasis in lung cancer,'' *Oncogene*, vol. 36, p. 5219, Sep. 2017.

[63] B. Wasylyk, S. L. Hahn, and A. Giovane, ''The Ets family of transcription factors,'' *Eur. J. Biochem.*, vol. 211, pp. 7–18, Jan. 1993.

[64] M. J. Tymms, A. Y. N. Ng, R. S. Thomas, B. C. Schutte, J. Zhou, H. J. Eyre, G. R. Sutherland, A. Seth, M. Rosenberg, T. Papas, C. Debouck, and I. Kola, ''A novel epithelial-expressed ETS gene, ELF3: Human and murine cDNA sequences, murine genomic organization, human mapping to 1q32.2 and expression in tissues and cancer,'' *Oncogene*, vol. 15, pp. 2449–2462, Nov. 1997.

[65] K. L. Eckel, J. J. Tentler, G. J. Cappetta, S. E. Diamond, and A. Gutierrez-Hartmann, ''The epithelial-specific ETS transcription factor ESX/ESE-1/Elf-3 modulates breast cancer-associated gene expression,'' *DNA Cell Biol.*, vol. 22, pp. 79–94, 2003.

[66] J. Wang, Z.-F. Chen, H.-M. Chen, M.-Y. Wang, X. Kong, Y.-C. Wang, T.-T. Sun, J. Hong, W. Zou, J. Xu, and J.-Y. Fang, ''Elf3 drives β-catenin transactivation and associates with poor prognosis in colorectal cancer,'' *Cell Death Disease*, vol. 5, p. e1263, May 2014.

[67] G. Luo, Y.-L. Chao, B. Tang, B.-S. Li, Y.-F. Xiao, R. Xie, S.-M. Wang, Y.-Y. Wu, H. Dong, X.-D. Liu, and S.-M. Yang, ''miR-149 represses metastasis of hepatocellular carcinoma by targeting actin-regulatory proteins PPM1F,'' *Oncotarget*, vol. 6, no. 35, p. 37808, 2015.

[68] B. Aissani, A. K. Boehme, H. W. Wiener, S. Shrestha, L. P. Jacobson, and R. A. Kaslow, ''SNP screening of central MHC-identified HLA-DMB as a candidate susceptibility gene for HIV-related Kaposi's sarcoma,'' *Genes Immunity*, vol. 15, no. 6, p. 424, 2014.

[69] K. S. Lau and J. W. Dennis, ''N-Glycans in cancer progression,'' *Glycobiology*, vol. 18, no. 10, pp. 750–760, 2008.

[70] R. B. Zavareh, M. A. Sukhai, R. Hurren, M. Gronda, X. Wang, C. D. Simpson, N. Maclean, F. Zih, T. Ketela, C. J. Swallow, J. Moffat, D. R. Rose, H. Schachter, and A. D. Schimmer, ''Suppression of cancer progression by MGAT1 shRNA knockdown,'' *PLoS ONE*, vol. 7, no. 9, p. e43721, 2012.

[71] B. Wu, K. Wang, J. Fei, Y. Bao, X. Wang, Z. Song, F. Chen, J. Gao, and Z. Zhong, ''Novel three-incRNA signature predicts survival in patients with pancreatic cancer,'' *Oncol. Rep.*, vol. 40, pp. 3427–3437, Sep. 2018.

[72] L. Pellizzari, C. Puppin, L. Mariuzzi, F. Saro, M. Pandolfi, R. Di Lauro, C. A. Beltrami, and G. Damante, ''PAX8 expression in human bladder cancer,'' *Oncol. Rep.*, vol. 16, pp. 1015–1020, Nov. 2006.

[73] A. P. Read, ''Pax genes—Paired feet in three camps,'' *Nature Genet.*, vol. 9, pp. 333–334, Apr. 1995.

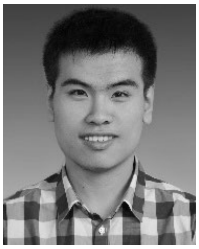
[74] N. J. Shaw, N. T. Georgopoulos, J. Southgate, and L. K. Trejdosiewicz, ''Effects of loss of p53 and p16 function on life span and survival of human urothelial cells,'' *Int. J. Cancer*, vol. 116, no. 4, pp. 634–639, 2005.

[75] M. E. McNerney, C. D. Brown, X. Wang, E. T. Bartom, S. Karmakar, C. Bandlamudi, S. Yu, J. Ko, B. P. Sandall, T. Stricker, J. Anastasi, R. L. Grossman, J. M. Cunningham, M. M. Le Beau, and K. P. White, ''CUX1 is a haploinsufficient tumor suppressor gene on chromosome 7 frequently inactivated in acute myeloid leukemia,'' *Blood*, vol. 121, pp. 975–983, Feb. 2013.

[76] C. C. Wong, I. Martincorena, A. G. Rust, M. Rashid, C. Alifrangis, L. B. Alexandrov, J. C. Tiffen, C. Kober, Chronic Myeloid Disorders Working Group of the International Cancer Genome Consortium, A. R. Green, C. E. Massie, J. Nangalia, S. Lempidaki, H. Dohner, K. Dohner, S. J. Bray, U. McDermott, E. Papaemmanuil, P. J. Campbell, and D. J. Adams, ''Inactivating CUX1 mutations promote tumorigenesis,'' *Nature Genet.*, vol. 46, pp. 33–38, Jan. 2014.

[77] C. Olsen and N. Wagtmann, ''Identification and characterization of human DPP9, a novel homologue of dipeptidyl peptidase IV,'' *Gene*, vol. 299, pp. 185–193, Oct. 2002.

[78] Z. Tang, J. Li, Q. Shen, J. Feng, H. Liu, W. Wang, L. Xu, G. Shi, X. Ye, M. Ge, X. Zhou, and S. Ni, ''Contribution of upregulated dipeptidyl peptidase 9 (DPP9) in promoting tumoregenicity, metastasis and the prediction of poor prognosis in non-small cell lung cancer (NSCLC),'' *Int. J. Cancer*, vol. 140, no. 7, pp. 1620–1632, 2017.



YU-HANG ZHANG was born in Jinzhou, Liaoning, China, in 1992. He received the B.S. degree in medical laboratory from the Shanghai Jiao Tong University School of Medicine, in 2014, and the Ph.D. degree in genetics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, in 2019. He is the author of more than 40 articles. His research interests include machine learning, liquid biopsy, and tumor immunotherapy. He was a recipient of the Merit Student from the University of Chinese Academy of Sciences, in 2017.

TAO ZENG received the B.S., M.S., and Ph.D. degrees from Wuhan University, Wuhan, China, in 2003, 2006, and 2010, respectively. Since 2013, he has been an Associate Professor at the Key Laboratory of Systems Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences. He is currently with the Shanghai Research Center for Brain Science and Brain-Inspired Intelligence, Shanghai, China. His research interests include bioinformatics, network biology, computational biology, machine learning, and graph theory.

XIAOYONG PAN received the Ph.D. degree in major bioinformatics from Copenhagen University, Denmark, in 2017, and the master's degree from Shanghai Jiao Tong University, in 2011. He is currently an Assistant Professor with Shanghai Jiao Tong University. Before that, he held a postdoctoral position at the Erasmus Medical Center, Rotterdam, Netherlands, from 2016 to 2018. His research interests include bioinformatics, deep learning, and electronic health record. He is also a Lead Guest Editor of IEEE Access.

WEI GUO received the B.S. degree in life science and technology from Wuhan University, in 2015. After that, he joined the Laboratory of Molecular Genetics in Shanghai Jiao Tong University School of Medicine & Institute of Health Sciences for his master and Ph.D. education. His research interests include bioinformatics, microbiome, and cancers.

ZIJUN GAN was born in Zhejiang, China, in 1993. She received the B.S. degree in biopharmaceutical from Zhejiang A&F University, in 2016, and the master's degree in genetics from the Shanghai Institute for Biological Sciences, in 2019. Her research interests include bioinformatics and genetics.

YUNHUA ZHANG received the Ph.D. degree in major bioinformatics from Nanjing University, in 2008, and the master's degree from Anhui Agriculture University, in 2003. He is currently a Professor with Anhui Agriculture University. Before that, he was a Visiting Scholar with UT State University, from 2017 to 2018. His research interests include environmental biology and molecular biology.

TAO HUANG received the B.S. degree in bioinformatics from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the Ph.D. degree in bioinformatics from the Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China, in 2012. Since 2014, he has been an Associate Professor and the Director of the Bioinformatics Core Facility, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai. From 2012 to 2014, he was a Postdoctoral Fellow with the Department of Genetics and Genomics Sciences, Icahn School of Medicine at Mount Sinai, New York City, USA. His research interests include bioinformatics, computational biology, systems genetics, and big data research. He has published over 100 articles. His works have been cited for over 3000 times with an H-index of 26 and an i10-index of 64. He has been a reviewer for over 20 journals and an editor/guest editor of seven journals and books.

YU-DONG CAI has been a Professor in bioinformatics with the School of Life Science, Shanghai University, since 2015. His main interests include various areas of systems biology and bioinformatics, such as protein-protein interaction, disease biomarkers prediction, drug-target interaction, and protein functional sites prediction. He has published over 200 peer-reviewed scientific articles, including invited reviews. His works have been cited for more than 7500 times, with H-index of 51. He is the Editorial Board Member of *Biochimica et Biophysica Acta* (BBA) - Proteins and Proteomics and *Biochemistry Research International*. He has been a Guest Editor of *Computational Proteomics*, *Systems Biology & Clinical Implications*, *Biochimica et Biophysica Acta* (BBA) - Proteins and Proteomics.

• • •