Am J Cancer Res 2022;12(12):5631-5645
www.ajcr.us /ISSN:2156-6976/ajcr0146465

Original Article

DeepCIA: a novel deep-learning model for cancer type identification using class activation map via transcription factor expression

Seongdo Jeong[1*], Dongjun Lee[2*], Hae Ryoun Park[3,4], Junho Kang[5], Yeuni Yu[5], Jae Joon Hwang[6], Yun Hak Kim[1,7,8]

[1]Research Institute for Convergence of Biomedical Science and Technology, Yangsan Hospital, Pusan National University, Yangsan 50612, Republic of Korea; [2]Department of Convergence Medicine, School of Medicine, Pusan National University, Yangsan 50612, Republic of Korea; [3]Department of Oral Pathology, School of Dentistry, Pusan National University, 49 Busandaehak-ro, Yangsan 50612, Republic of Korea; [4]Periodontal Disease Signaling Network Research Center, School of Dentistry, Pusan National University, Yangsan 50612, Republic of Korea; [5]Medical Research Institute, Pusan National University, Pusan, Republic of Korea; [6]Department of Oral and Maxillofacial Radiology, School of Dentistry, Pusan National University, Dental Research Institute, Yangsan 50610, Republic of Korea; Departments of [7]Anatomy, [8]Biomedical Informatics, School of Medicine, Pusan National University, 49 Busandaehak-ro, Yangsan 50612, Republic of Korea. [*]Equal contributors.

Received September 13, 2022; Accepted November 20, 2022; Epub December 15, 2022; Published December 30, 2022

Abstract: Deep learning methods are powerful analytical tools for large-scale data analysis. Here, we introduce DeepCIA as a novel diagnostic deep-learning model for cancer type identification using a class activation map via transcription factor expression. Although many deep learning researches attempts have recently been made in relation to cancer diagnosis, there are difficulties in using cancer data due to a large-scale problem. Therefore, From The Cancer Genome Atlas (TCGA) and International Cancer Genome Consortium (ICGC) public databases, we selected transcription factor expression profiles of eight cancer types. TCGA included 3496 samples and divided the train and validation sets in an 8:2 ratio. ICGC included 552 samples and was used as a test set for external validation. To compare the performance of 1D-CNN models, we also used SVM and KNN from machine learning. In external validation, 1D-CNN showed a high average accuracy of 98% and was superior to support vector machine (SVM) and k-nearest neighbor (KNN) with a difference in the accuracy of 10-12%. Also, 1D-CNN performed very well in several performance metrics (98.2% Recall, 98.1% Precision, 98.2% F score, 99.8% Specificity, 99.8% AUC, and 99.0% Balanced Accuracy). In each data set evaluation, 1-network, 5-network, and 2-network with high accuracy were selected and visualized through the Class Activation Map. We identified the Cys2Hys2 zinc finger group with the highest distribution across all cancer types. Collectively, DeepCIA can be used as a decision support system for cancer and a classifier for diagnosing unknown primary cancer, while emphasizing its usefulness in cancer diagnosis.

Keywords: Deep learning, cancer classification, class activation map, transcription factor, gene expression

Introduction

Recent advances in deep learning systems based on multi-layer neural networks have enabled large-scale data analysis. Deep learning is a concept derived from artificial intelligence neural networks, and complex structural patterns and corresponding features can be extracted through learning after inputting large-scale data [1]. Convolutional Neural Network (CNN) is a subset of deep learning that uses a mathematical convolution method to extract features from input data by performing intensive computations of nonlinear relationships over numerous hidden layers. CNN has recently proven to be a powerful analytical tool for classification, not only in image data but also in non-image data. For example, there are three-dimensional deep learning systems that predict lung cancer risk using images obtained through computed tomography [2], and deep learning models that predict cancer types based on

gene expression data [3]. Cancer type classification based on gene expression values use 2-D convolutional neural networks and show an accuracy of approximately 95% [4]. In these methods, the size of data transmitted inside the network is continuously reduced by convolution, resulting in low-resolution data, and making it impossible to specify a single gene that affects cancer classification.

Class activation map (CAM) is a visual tool used for image classification. The CAM appears by calculating the weighted sum of the feature maps according to the final convolutional layer, weighting the fully connected layers according to how much each activation contributes to the final score for that class. These CAMs provide qualitative insight into neural networks by visually understanding information trained through network learning and are mainly used to classify medical images and find lesions. Representative studies include chest X-ray abnormalities [5] and localization of diabetic retinopathy lesions.

Cancer is one of the most life-threatening genetic diseases, and early and accurate cancer classification contributes significantly to progress in the medical field. In other words, accurate and rapid diagnosis of cancer types determines the survival and lifespan of cancer patients [6]. Cancer classification has clear limitations in that it is difficult to classify cancer types based only on morphological characteristics, and there is a bias in tumour identification by experts [7, 8]. For this reason, it is estimated that about 10-20% of all cancer cases are misdiagnosed, and it is estimated that more than 40,000 cancer patients die each year due to misdiagnosis or delay in diagnosis [9]. Therefore, several studies highlight the importance of developing cancer diagnosis supporting systems for improving the survival rate and quality of life of cancer patients and for the diagnosis of cancer of unknown primary sites (CUPs) [10]. Recent advances in sequencing technology have created public databases called The Cancer Genome Atlas (TCGA) [11] and the International Cancer Genome Consortium (ICGC) [12]. Therefore, many researchers can easily acquire large-scale gene expression profiles. However, it is very difficult to classify various types of cancer because different platforms of sequencing devices have different quantification methods [13]. While some researchers have tried to distinguish cancer types using gene expression profiles, there have been no notable achievements [14]. Omics data are composed of a high-dimensional structure containing more than 20,000 genes, resulting in a dimensional curse that degrades the performance of cancer classification [15, 16]. Thus, proper gene selection may be an important factor in cancer type classification.

Transcription factors (TFs) are proteins that regulate the transcription of fragment DNA into messenger RNA by binding to a specific DNA region. This plays a central role in the most important cell processes such as intracellular metabolism, cell cycle regulation, and cell differentiation [17]. TFs are also known to be involved in human cancer [18]. Dysregulated transcription factors mediate aberrant gene expression, and transcription factor activity is altered in numerous cancer types via various mechanisms [19]. Further, high expression of some transcription factors is correlated with poor prognosis and chemoresistance [20]. Based on DNA-binding motifs, TFs can be categorized into classical zinc fingers (ZFs) [21], homeodomains [22], and basic helix-loop-helix [23]. Interestingly, Cys2His2 (C2H2)-type zinc fingers (ZFs) are the largest group of all zinc finger motif classes, and zinc fingers can also provide protein-protein and RNA-protein interactions [24]. Proteins containing the C2H2 ZF are trans regulators of gene expression and play an important role in cellular processes such as development, differentiation, and suppression of malignant cell transformation [25]. Furthermore, targeting transcription factors, in combination with other chemotherapeutics, could emerge as a better strategy to treat cancer.

In this paper, we created a new cancer classification model by combining the 1-dimensional convolutional neural network (1D-CNN) with class activation map (CAM), an image visualization technology, based on the expression profiles of 1462 transcription factors included in eight cancer types of TCGA and ICGC. Finally, DNA binding domain groups and transcription factors important for cancer diagnosis and classification were obtained. Our diagnostic model is expected to provide new biological

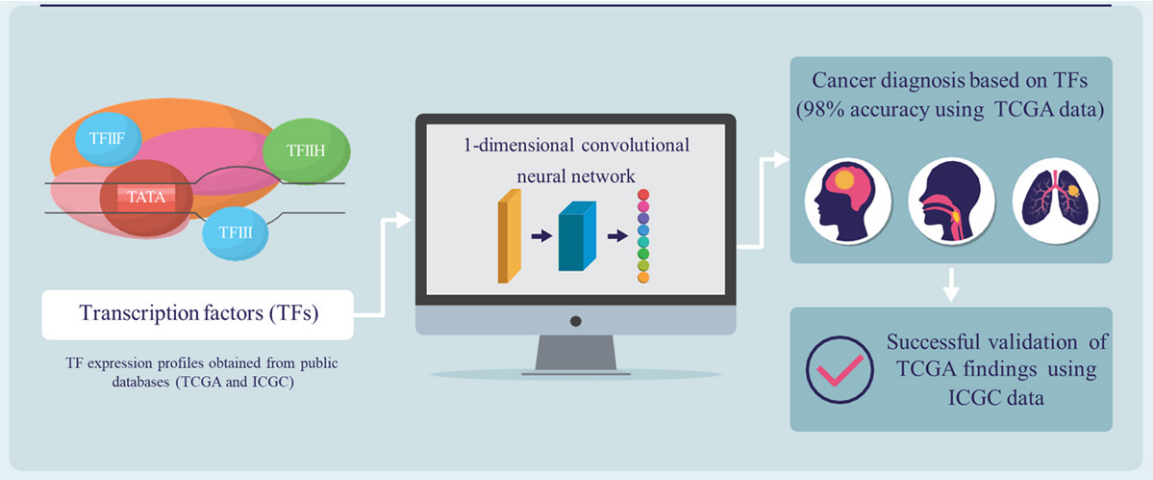

Figure 1. Flowchart summarizing the study design.

insights to cancer biologists. It also suggests the possibility of providing a new perspective on clinical decision support systems and adjuvant therapy for cancer of unknown primary sites.

Materials and methods

Flowchart summarizing the study design

Flowchart summarizing the study design is shown in **Figure 1**: All data were acquired from TCGA and ICGC, and TFs gene selection and pre-processing were performed for input to the network. 1D-CNN learning results obtained visualization using CAM for cancer classification accuracy and cancer type. 1D-CNN was performed using an NVIDIA GeForce RTX 2080 Ti GPU with SF-2000F 14HP power and it takes about 250-300 minutes to train one network. The programming languages used for all work were MATLAB 2020a (MathWorks, Natick, Massachusetts) and R statistical software version 4.0.5.

Dataset

TCGA was downloaded using the GDCquery function of the TCGA biolinks package in R software [26]. The GDCquery function has 12 parameters; of these parameters, project, data, category, data type, and workflow type are used in this study. To develop and validate the classification system, we selected eight cancers using RNA sequencing data from both TCGA and ICGC databases; these include Breast invasive carcinoma (TCGA-BRCA), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (TCGA-DLBC), Head and Neck squamous cell carcinoma (TCGA-HNSC), Clear cell renal cell carcinoma (TCGA-KIRC), Ovarian serous cystadenocarcinoma (TCGA-OV), Pancreatic adenocarcinoma (TCGA-PAAD), Prostate adenocarcinoma (TCGA-PRAD), and Sarcoma (TCGA-SARC). Arguments corresponding to each parameter were set as: data.category = 'Transcriptome Profiling'; data.type = 'Gene Expression Quantification'; workflow.type = 'HTSeq-Counts'. ICGC provides sequence-based gene expression data from the ICGC Data portal (https://dcc.icgc.org/). Therefore, TCGA includes eight cancer types: BRCA, DLBC, HNSC, KIRC, OV, PAAD, PRAD, and SARC. The ICGC includes eight cancer types: breast cancer-Very young women (BRCA-KR), Malignant Lymphoma-DE (MALY-DE), Oral cancer-IN (ORCA-IN), Renal cell cancer-EUFR (RECA-EU), Ovarian cancer-AU (OV-AU), Pancreatic cancer-AU (PACA-AU), Prostate cancer-Adenocarcinoma (PRAD-FR), and Soft tissue cancer-Ewing sarcoma-FR (BOCA-FR). The detailed patients' information of included cohorts is described in Supplementary Table 1. Due to the retrospective nature of this study using only publicly available data, ethics approval was not required.

Data pre-processing

For eight cancer types of TCGA, the parts corresponding to the 'Primary Tumor' category were extracted. Genes overlapping in eight can-

Table 1. Number of samples and transcription factors for each cancer

TCGA cohorts	No. of samples	No. of transcription factors
BRCA	1102	1462
DLBC	48	1462
HNSC	500	1462
KIRC	538	1462
OV	374	1462
PAAD	177	1462
PRAD	498	1462
SARC	259	1462
ICGA cohorts	No. of samples	No. of transcription factors
BRCA-KR	50	1462
MALY-DE	105	1462
ORCA-IN	40	1462
RECA-EU	91	1462
OV-AU	93	1462
PACA-AU	91	1462
PRAD-FR	25	1462
BOCA-FR	57	1462

BRCA: Breast Invasive Carcinoma; DLBC: Lymphoid Neoplasm Diffuse Large B-cell Lymphoma; HNSC: Head and Neck Squamous Cell Carcinoma; KIRC: Clear Cell Renal Cell Carcinoma; OV: Ovarian Serous Cystadenocarcinoma; PAAD: Pancreatic Adenocarcinoma; PRAD: Prostate Adenocarcinoma; SARC: Sarcoma; BRCA-KR: Breast Cancer-Very Young Women; MALY-DE: Malignant Lymphoma-DE; ORCA-IN: Oral Cancer-IN; RECA-EU: Renal Cell Cancer-EUFR; OV-AU: Ovarian Cancer-AU; PACA-AU: Pancreatic Cancer-AU; PRAD-FR: Prostate Cancer-Adenocarcinoma; BOCA-FR: Soft Tissue Cancer-Ewing Sarcoma-FR.

cers were processed by setting the median value through the R function 'aggregate'. In ICGC, MALY-DE, RECA-KR, OV-AU, PACA-AU, PRAD-FR, and BOCA-FR are indicated as ensemble gene IDs; hence, they were changed to symbol gene IDs using the R function 'bitr'. Like TCGA, ICGC was also treated by setting overlapping genes as the median value using the R function 'aggregate'. A list of gene and DBD information for transcription factors was provided by Lambert et al [27]. To match the number of input values required for network learning, 1462 transcription factors that are equally included in TCGA and ICGC cancer types were selected (**Table 1**), and a matrix was composed of TFs gene names in rows and samples in columns. Then, to reduce the range of TFs gene expression values, the generated raw read count matrix was applied to each value using $y = \log2(x + 1)$. To perform five-fold cross-validation of network learning, the part corresponding to 20% of each cancer type was passively split and cross-linked for each cancer type. To input data into 1D-CNN, it was necessary to convert the data from the existing one-dimensional matrix data into a two-dimensional grayscale image. So, the image was reconstructed from a 1462 × 1 array to a 38 × 39 image by adding zeros to the last line, then normalized to have pixels in the range [0, 1].

TFs network structure and network train

The 3496 images obtained from TCGA data were divided into a train set and a validation set with a ratio of 8:2. All 552 images obtained from the ICGC data were used as the test set for external validation. The 1D-CNN structure consists of 14 layers as shown in **Figure 2**. A 1D-grayscale image (38 × 39) is input and passed through a convolution layer consisting of a filter size of 1 × 1. The weights initializer built into the convolution layer uses Glorot to initialize the weights. Batch-normalization initializes the parameters by setting the channel offset to zero and the channel scaling to ones. The Relu layer is activated by setting all input values less than zero to zero. The dropout layer sets the dropout probability to 50%. The sizes of the three fully connected layers were 1482, 741, and 8, respectively. The weighted classification layer calculated class weight by dividing the total number of data by the number of samples and applied it to the loss function to minimize the data imbalance [28]. The network was trained for 100-102 epochs using the Adam optimizer with an initial learning rate of 3e-5 and a mini-batch size of 8. Five-fold cross-validation was used to train the deep neural network and to test the performance. To compare cancer classification performance, the 1D-CNN model and its performance were compared using machine learning methods k-nearest neighbor (KNN) and support vector machine (SVM).

Statistical analysis

The analysis was conducted using MATLAB 2020a (MathWorks, Natick, Massachusetts). The MATLAB package 'confusionmat' was used to visualize a confusion matrix chart between groups of actual and predicted cancer types.

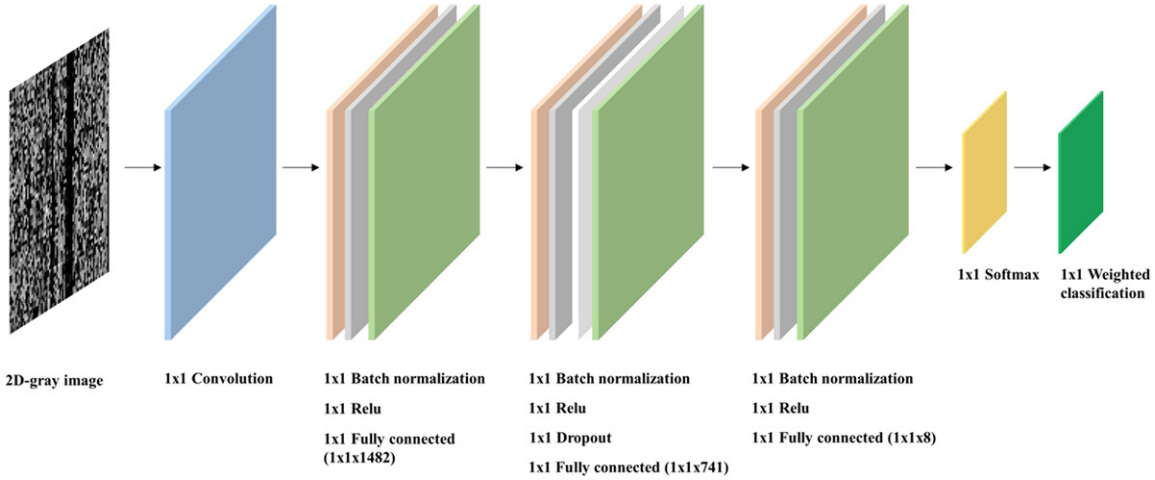

Figure 2. Architecture of 1D-CNN. The network structure consists of 14 layers. The network structure consists of a 1 × 1 convolution layer, 3 batch-normalization layers, 3 relu layers, 3 fully connected layers, 1 dropout, 1 softmax, and 1 weighted classification layer.

The 'multiclass_metrics_common' package was used to evaluate metrics for performance comparison. Using the 'perfcurve' package, we generated an average ROC curve for eight cancer types in each network and obtained area under the curve (AUC) values.

Metrics for performance comparison

Accuracy (1), Precision (2), Recall (Sensitivity) (3), F1-score (4), Specificity (5), Balanced Accuracy (6) and AUC values from the ROC curve were used to evaluate the performance of the models.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall\ (Sensitivity) = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ score = \frac{2 \times (Recall \times Precision)}{Recall + Precision} \quad (4)$$

$$Specificity = \frac{TN}{FP + TN} \quad (5)$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \quad (6)$$

TP, True Positive; FP, False Positive; FN, False Negative; TN, True Negative.

2-D class activation map (CAM)

To obtain a two-dimensional CAM (class activation map) image after network training, we calculated the activation score of the image in a specific 'Relu' layer following the 1 × 1 convolutional layer using the trained neural network and data storage image. Therefore, CAM images for each cancer type were acquired and min-max normalization was performed so that the pixels were in the range [0, 1]. Next, the average CAM image was obtained by calculating the arithmetic mean of the CAM matrix by cancer type, excluding the samples corresponding to misclassification in each data set, and treating values below the threshold of 0.4 as zero values.

Results

Patients' characteristics

The differences in the distribution of age, stage, race, and gender, which are clinical information for cancer patients in TCGA and ICGC, are confirmed in Supplementary Table 1.

Classification performance of cancer type classification using 1462 TFs

The performance evaluation of the SVM, KNN, and 1D-CNN are listed in **Table 2** and Supplementary Tables 2, 3. For the three models, Accuracy, Precision, F-score, Specificity, Recall (sensitivity), AUC, and Balanced Accuracy

Table 2. Performance evaluation of TCGA and ICGC classification trained only by 1462 transcription factors

TCGA train set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.9925	0.9796	0.9862	0.9937	0.9990	0.9998	0.9964
2-network	0.9918	0.9747	0.9837	0.9941	0.9989	0.9999	0.9965
3-network	0.9921	0.9791	0.9862	0.9941	0.9989	0.9997	0.9965
4-network	0.9921	0.9796	0.9866	0.9944	0.9989	0.9999	0.9967
5-network	0.9900	0.9672	0.9783	0.9921	0.9986	0.9998	0.9954
TCGA validation set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.9886	0.9571	0.9717	0.9924	0.9985	0.9998	0.9955
2-network	0.9915	0.9785	0.9844	0.9913	0.9988	0.9993	0.9951
3-network	0.9871	0.9671	0.9763	0.9881	0.9982	0.9988	0.9932
4-network	0.9914	0.9788	0.9863	0.9946	0.9988	0.9997	0.9967
5-network	0.9942	0.9815	0.9874	0.9941	0.9992	1	0.9967
ICGC test set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.9819	0.9782	0.9796	0.9816	0.9974	0.9976	0.9895
2-network	0.9891	0.9874	0.9871	0.9871	0.9985	0.9983	0.9928
3-network	0.9873	0.9862	0.9867	0.9875	0.9982	0.9984	0.9929
4-network	0.9783	0.9753	0.9760	0.9772	0.9969	0.9976	0.9871
5-network	0.9819	0.9782	0.9795	0.9816	0.9974	0.9984	0.9895

were all higher than 98% to 99% in the train and validation set. However, in the test set, the SVM model showed an average accuracy of 88%, the KNN model 85.4%, and the 1D-CNN 98.4% average accuracy (**Figure 3**). When comparing the two machine learning models and 1D-CNN, the 1D-CNN model had the better performance than other models in all aspects. In addition, ROC curves were drawn for individual networks (1D-CNN) in each data set and high AUC values of 99% were obtained (**Figure 4**).

Performance evaluation of 1D-CNN using the confusion matrix

We created a confusion matrix of 1-Network, 5-Network, and 2-Network with the highest accuracy in each dataset and visualized it as a heatmap (**Figure 5**). **Figure 5A** shows that nine samples of BRCA, four samples of HNSC, four samples of KIRC, two samples of OV, and two samples of PAAD were misclassified, and the remaining three cancer types were correctly classified. Next, **Figure 5B** shows that two samples of BRCA, one sample of KIRC, and one sample of PAAD were misclassified, and the remaining five cancer type samples were correctly classified. As shown in **Figure 5C**, one sample of BRCA-KR, two samples of ORCA-IN, and three samples of PACA-AU were misclassified, and the remaining five cancer types were correctly classified.

Visualization of cancer type classification using 1462 TFs

Figure 6 shows the 1-network, 2-network, and 5-network with the highest accuracy among the trained networks from each data set and display visual information for eight cancer types using CAM. In the activation map, blue (low) to red (high) indicates the degree of crystalline influence of major genes by cancer type.

Cumulative importance of TFs domain binding group using intensity values of CAM

For each cancer type, 146 TFs and 146 DNA binding domains (DBDs) were selected as the top 10% of intensity values in the CAM. For each cancer type in all data sets, only the top 10 overlapping genes are listed (**Table 3**) and the correlation between top 10 TFs and each cancer is described in Supplementary Table 4. The top 10 overlapping transcription factors

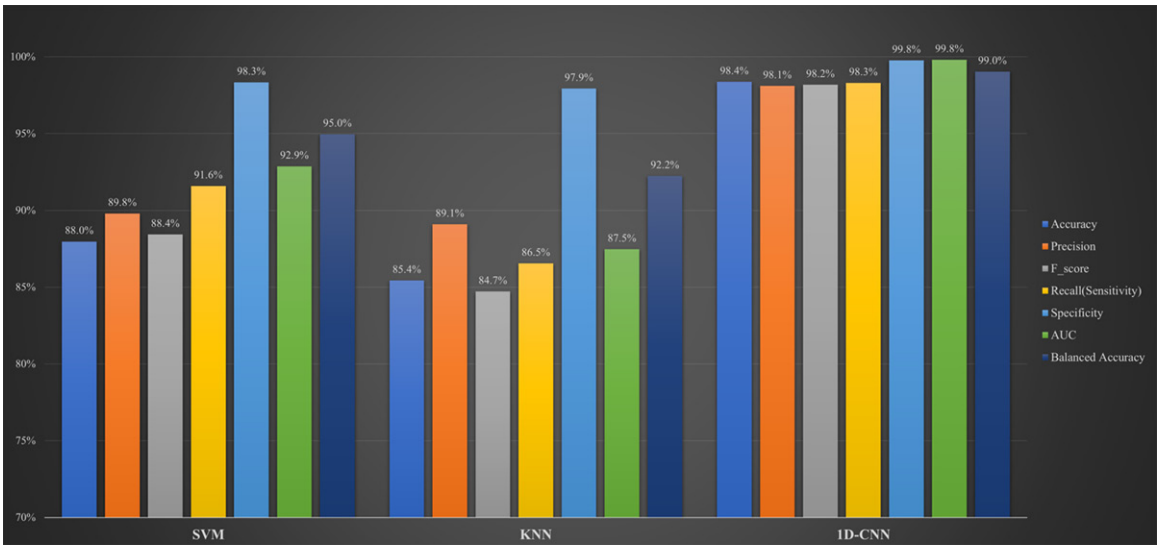

Figure 3. Performance of each classifier in external validation. The graph is the average of 5 individual networks of each model, SVM, KNN, and 1D-CNN. SVM has the lowest 88% accuracy compared to KNN and 1D-CNN. 1D-CNN has the highest accuracy of 98.4%. In addition, 1D-CNN shows the highest number in several performance evaluation metrics.

are the most important genes for the DeepCIA to classify the cancer types (**Table 3** and Supplementary Table 4). Core transcription factors have reported the functional role in each cancer and these genes regulate downstream targets depending on cell-type specificity manners in different cancer [29]. The TFs and DBD for each data set are given in Supplementary Tables 5, 6, 7. By obtaining the cumulative frequency distribution of DBDs for each cancer type, it was possible to obtain the Cys2His2 (C2H2)-type zinc finger (C2H2 ZF) DBD with the highest frequency (**Table 4**).

Discussion

Recently, deep learning, an artificial intelligence technology, has been in the spotlight in the field of big data. In particular, research on a cancer classification technique that combines quantitative mRNA expression data, including genetic information, is ongoing. In the field of deep learning, most studies have used 2D-convolutional neural networks. For data input, the gene expression value matrix was converted into a 2D image and passed through a convolutional neural network. Thus, the size of the data transmitted continues to decrease by convolution and pooling function, resulting in low-resolution and making it difficult to locate accurately analyse the contribution of a single gene to the predicted cancer type. So, we were able to extract the CAM from the front and solve the problem of the CAM resolution to extract the exact gene that affects cancer classification. Additionally, as the gene arrangement structure was compressed from 3D to 2D, there was a problem whereby the genes extracted by the deep learning network were different each time depending on the gene arrangement in 2D image. Therefore, we processed gene data into 1D sequence using 1D-CNN.

We used two well-known machine learning techniques, SVM and KNN, to compare 1D-CNN and network performance. SVM showed 88% accuracy in the origin of cancer using gene expression data and histopathology [30]. And the k-nearest neighbor algorithm is a well-known classification model, and there are studies on metastatic primary site identification [31]. Since the data used in each model were divided into independent datasets for five-cross validation, an accurate comparison was possible. As shown in **Figure 3**, 1D-CNN was able to confirm the performance of 98% or more in all performance indicators on average in the external validation data set. Comparing SVM and KNN in terms of accuracy, we can see a difference of 10-12%. In addition, the 1D-CNN model performs better with a 7-10% difference in various performance indicators.

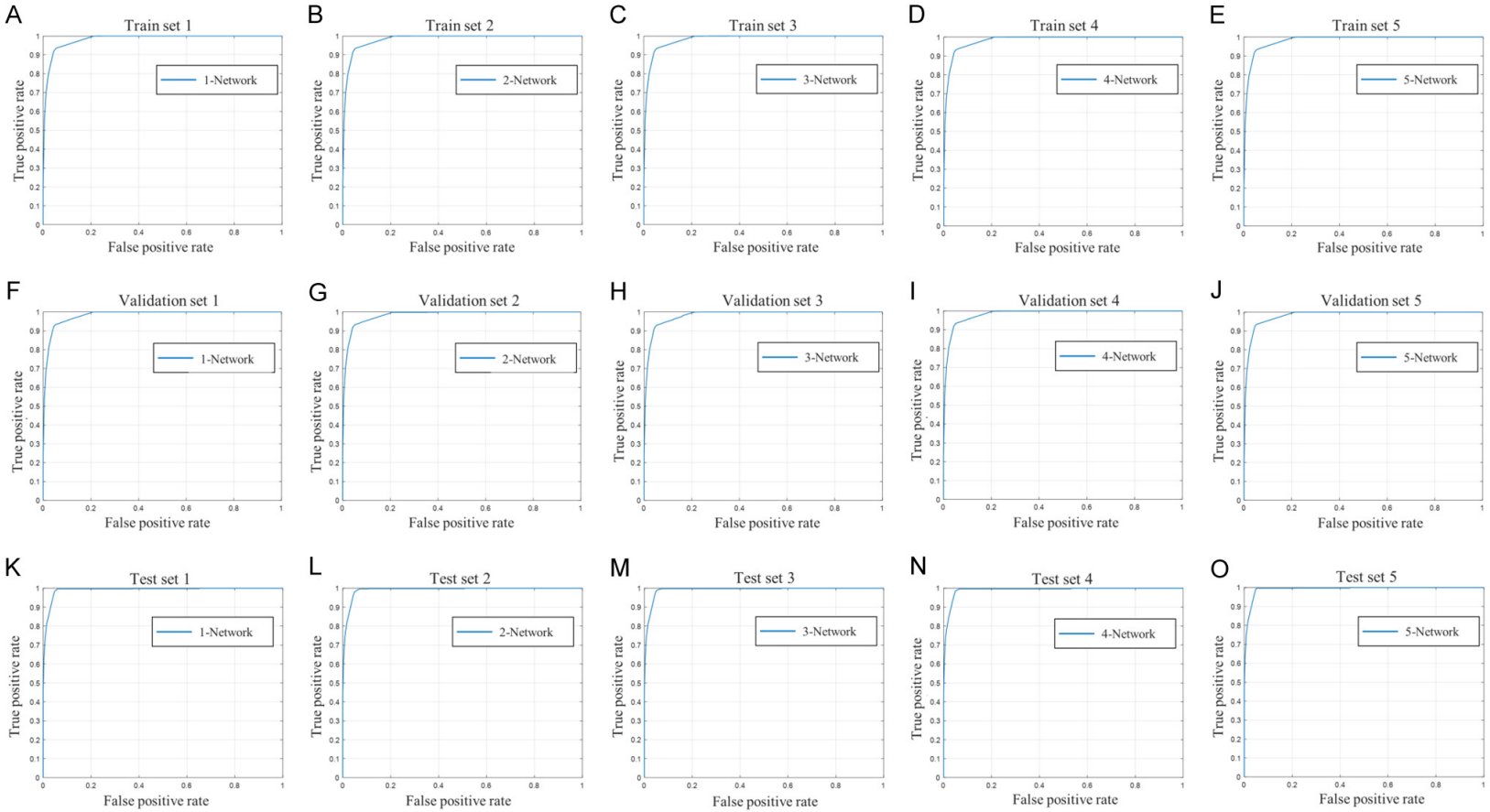

Figure 4. Model performance evaluation through ROC curve of the train, validation, and test set. A-E. ROC curve of 5 networks for the train set. F-J. ROC curve of 5 networks for the validation set. K-O. ROC curve of 5 networks for the test set.

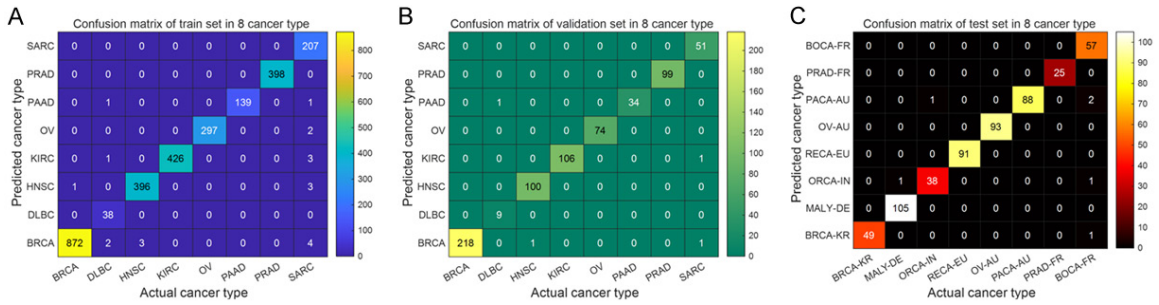

Figure 5. Performance evaluation of the model using the confusion matrix for each dataset. A. Confusion matrix of the train set in eight cancer types. B. Confusion matrix of the validation set in eight cancer types. C. Confusion matrix of the test set in eight cancer types.

Thus, we grafted CAM onto the learned network to visualize each cancer type, and we were able to identify the C2H2 ZF group with the highest distribution in all cancer types.

C2H2 ZF is the largest group of all zinc finger motif classes, and zinc fingers can also provide protein-protein and RNA-protein interactions [24]. Among C2H2 ZNFs, there are a large number of transcription factors with the C-x-C-x-H-x-H motif, which mediates direct interaction with DNA [25]. In addition, GC-rich or GT-rich sequences serve as C2H2-type ZF cis-regulatory elements, and C2H2-type ZNFs also contain other functional domains, such as BTB, POZ, KRAB, and SCAN [32]. We found six C2H2 ZFs as the top 10% of intensity values in the CAM from eight cancer types, namely, KLF6, ZNF395, ZNF703, ZNF704, BCL11B, and ZEB2. KLF6 has been implicated in the regulation of several cellular processes, including development, proliferation, inflammation, apoptosis, differentiation, and cell cycle regulation [33-36]. ZNF395 is implicated in various cancers, such as renal cell carcinomas, osteosarcomas, and Ewing sarcomas [37, 38]. ZNF703 and ZNF704 promote tumour progression in breast and ovarian cancers [39-41]. BCL11B is related to malignant T-cell transformation that occurs in haematological malignancies [42]. ZEB2 promotes the proliferation of primary and metastatic melanoma [43].

Basic leucine zipper (bZIP) TFs are second highest distributed from eight cancer types and have a conserved bZIP domain, which is composed of two structural features located on a contiguous α-helix [44]. The bZIP domain is 60 to 80 amino acids in length with two functional regions, a highly conserved basic region, and a more diversified leucine zipper region [45]. We found five overlapped bZIP ZFs as the top 10% of intensity values in the CAM from eight cancer types, such as NFE2L, XBP1, ATF4, JUN, and FOS. XBP1 and ATF4 are involved in the ER stress response and regulate cell survival and death [46]. JUN-FOS and JUN-ATF dimers are implicated in oncogenesis, and these activities of JUN-FOS and JUN-ATF complexes can be regulated at multiple levels [47].

Although the characteristics of the patients included in the cohorts are different even for the same cancer category, the DeepCIA successfully classified the eight cancer types. In addition, it can diagnose the cancer types despite the small number of patients in the training. The most difficult problem in sequencing data research is overcoming the heterogeneity of sequencing platforms. In the current study, we overcome this problem by using the raw read count of TFs genes from different sequencing platforms. In addition, DeepCIA classified cancer types accurately despite differences in the distribution of patient clinical information such as age, gender, stage, and race. Thus, DeepCIA is expected to provide new guidance for diagnostic systems, while also applying it to other diseases. In addition, 2-5% of some cancer diagnoses are classified as cancers of unknown primary origin (CUPs). Extensive diagnostic methods such as imaging, endoscopy, and biopsy are required to designate the primary site of these cancers. However, despite these various diagnostic modalities, determining the origin of primary tumour in CUPs remains a challenging task. Therefore, we expect that the DeepCIA not only provides a new differential diagnosis of the origin of pri-

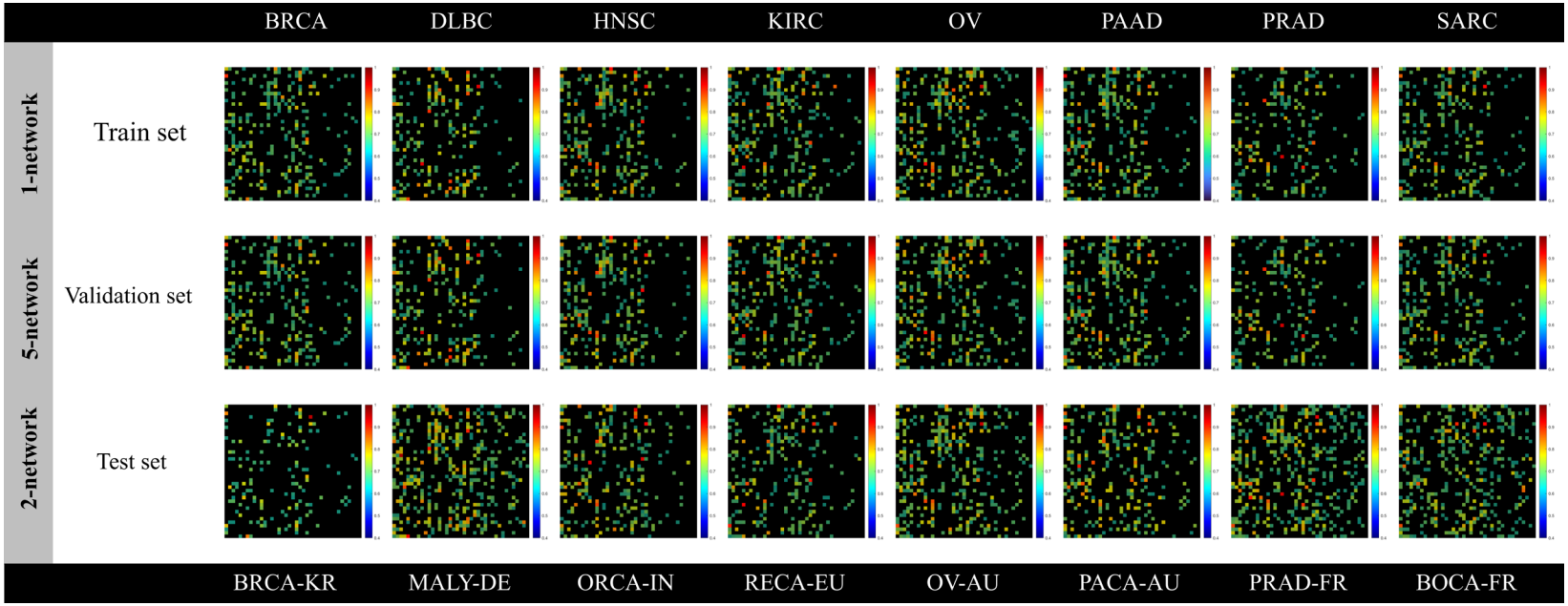

Figure 6. Mean class activation map (CAM) images of each cancer type. The high value (red) in the CAM represents that the gene has contributed greatly to the classification of cancer types. Train and validation set were from TCGA, and test set was from ICGC.

Table 3. Genes and DNA binding domains (DBDs) correspond to the top 10% of the intensity values of the CAM

BRCA-BRCA-KR		DLBC-MALY-DE		HNSC-ORCA-IN		KIRC-RECA-EU	
Gene	DBD	Gene	DBD	Gene	DBD	Gene	DBD
AEBP1	Unknown	HMGA1	AT hook	YBX1	CSD	EPAS1	bHLH
TRPS1	GATA	YBX1	CSD	NFE2L1	bZIP	NFE2L1	bZIP
BHLHE40	bHLH	IRF8	IRF	TP63	p53	KLF6	C2H2 ZF
GATA3	GATA	POU2AF1	Unknown	STAT1	STAT	TSC22D1	Unknown
NFE2L1	bZIP	STAT1	STAT	HMGA1	AT hook	YBX1	CSD
STAT3	STAT	ETS1	Ets	IRF6	IRF	JUN	bZIP
STAT1	STAT	STAT6	STAT	JUNB	bZIP	ZNF395	C2H2 ZF
PBX1	Homeodomain	MYBL2	Myb/SANT	STAT3	STAT	BHLHE40	bHLH
YBX1	CSD	POU2F2	Homeodomain; POU	HIF1A	bHLH	AEBP1	Unknown
ATF4	bZIP	ATF4	bZIP	FOSL2	bZIP	ETS1	Ets
OV-OV-AU		PAAD-PACA-AU		PRAD-PRAD-FR		SARC-BOCA-FR	
Gene	DBD	Gene	DBD	Gene	DBD	Gene	DBD
YBX1	CSD	AEBP1	Unknown	NKX3-1	Homeodomain	YBX1	CSD
HMGA1	AT hook	ELF3	Ets; AT hook	SPDEF	Ets	NFE2L1	bZIP
JUND	bZIP	BHLHE40	bHLH	HOXB13	Homeodomain	JUN	bZIP
PAX8	Paired box	FOS	bZIP	FOXA1	Forkhead	AEBP1	Unknown
JUNB	bZIP	NFE2L1	bZIP	TSC22D1	Unknown	ATF4	bZIP
ELF3	Ets; AT hook	YBX1	CSD	FOS	bZIP	NFIC	SMAD
TSC22D1	Unknown	ATF4	bZIP	ATF4	bZIP	NFIX	SMAD
NFIX	SMAD	TSC22D1	Unknown	NFIX	SMAD	STAT3	STAT
JUN	bZIP	JUNB	bZIP	JUN	bZIP	PRRX1	Homeodomain
STAT1	STAT	STAT3	STAT	NFE2L1	bZIP	EPAS1	bHLH

The table lists only 10 overlapping genes for each cancer type in all data sets.

Table 4. Cumulative distribution of top 10% CAM genes included by DBD groups

DBD	BRCA - BRCA-KR	DLBC - MALY-DE	HNSC - ORCA-IN	KIRC - RECA-EU	OV - OV-AU	PAAD - PACA-AU	PRAD - PRAD-FR	SARC - BOCA-KR	Gene counts
C2H2 ZF	22	14	7	12	11	9	8	10	93
Unknown	14	13	8	11	12	9	11	10	88
bHLH	13	9	9	11	9	8	8	11	78
bZIP	13	6	5	6	6	5	6	6	53
STAT	5	6	4	5	4	4	5	5	38
AT hook	4	4	3	1	3	3	2	2	22
CxxC	2	5	2	2	3	1	1	3	19
Ets	3	5	3	1	0	4	1	1	18
Forkhead	5	2	2	1	1	2	3	1	17
Homeodomain	5	1	1	2	2	1	2	2	16
Rel	1	5	2	2	2	1	2	1	16
HMG/Sox	4	2	0	2	3	0	2	2	15
MBD	2	3	1	1	2	0	3	1	13
MBD; AT hook	1	1	1	2	1	1	2	2	11
Nuclear receptor	2	3	1	3	1	1	0	0	11
GATA	3	1	1	1	1	1	0	1	9
IRF	0	3	2	1	2	0	0	1	9
MADS box	1	2	0	1	1	0	2	2	9
CSD	1	1	1	1	1	1	1	1	8

E2F	1	2	2	0	1	0	1	1	8
HSF	1	1	1	1	1	0	1	1	7
CSL	0	0	1	1	1	1	1	1	6
CUT; Homeodomain	0	1	1	1	1	1	1	0	6
Grainyhead	2	1	0	0	1	0	1	1	6
SAND	0	2	1	1	1	1	0	0	6
CENPB	1	0	1	1	1	0	0	1	5
MBD; CxxC ZF	0	1	0	0	0	0	1	1	3
RFX	0	1	0	0	1	0	0	1	3
Runt	1	1	0	0	0	0	0	1	3
p53	0	1	0	0	0	0	0	1	2
SMAD	2	0	0	0	0	0	0	0	2
AP-2	1	0	0	0	0	0	0	0	1
ARID/BRIGHT	1	0	0	0	0	0	0	0	1
Brinker	1	0	0	0	0	0	0	0	1
C2H2 ZF; AT hook	1	0	0	0	0	0	0	0	1
Ets; AT hook	1	0	0	0	0	0	0	0	1
Homeodomain; POU	0	1	0	0	0	0	0	0	1
Myb/SANT	0	1	0	0	0	0	0	0	1
Paired box	0	1	0	0	0	0	0	0	1
T-box	1	0	0	0	0	0	0	0	1
C2H2 ZF	22	14	7	12	11	9	8	10	93
Unknown	14	13	8	11	12	9	11	10	88
bHLH	13	9	9	11	9	8	8	11	78
bZIP	13	6	5	6	6	5	6	6	53
STAT	5	6	4	5	4	4	5	5	38
AT hook	4	4	3	1	3	3	2	2	22
CxxC	2	5	2	2	3	1	1	3	19
Ets	3	5	3	1	0	4	1	1	18
Forkhead	5	2	2	1	1	2	3	1	17
Homeodomain	5	1	1	2	2	1	2	2	16
Rel	1	5	2	2	2	1	2	1	16

mary tumours, but also the corresponding TFs of the C2H2 ZF and bZIP groups found in our study will play a major role in determining the origin of primary tumour in CUPs.

A limitation of this study was that only eight cancer types were used. TCGA provides 33 cancer types, but ICGC only has eight mRNA expression profiles consistent with TCGA. Therefore, it cannot progress to more types of cancer. Although the DeepCIA was a model trained with TFs from a small number of patients, it was able to distinguish cancer types even if the sequencing platform was different.

Conclusions

In this study, we tried to determine whether eight cancers are classified through quantified TFs expression data. The DeepCIA showed better performance than other models based on SVM or KNN in the external validation. Using the CAM, which has rarely been used for genomic deep learning, we identified the C2H2 ZF and bZIP groups are the most important transcription factors in cancer types classification. We believe that the DeepCIA will be used in both basic cancer research and clinical field.

Acknowledgements

This work was supported by the National Research Foundation of Korea (2021R1A2C4001466, 2022R1A5A2027161 and 2018R1A5A2023879), Research institute for Convergence of biomedical science and technology, Pusan National University Yangsan Hospital (30-2022-006) and the Korea Health Tech-

nology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (HI22C1377). The data reported in this manuscript are tabulated in the main paper and supplementary materials.

Disclosure of conflict of interest

Patents pending by the authors and their institutions.

Address correspondence to: Jae Joon Hwang, Department of Oral and Maxillofacial Radiology, School of Dentistry, Pusan National University, Dental Research Institute, Yangsan 50610, Republic of Korea. Tel: +82-10-5368-6960; Fax: +82-02-360-5029; E-mail: softdent@pusan.ac.kr; Yun Hak Kim, Department of Anatomy and Department of Biomedical Informatics, School of Medicine, Pusan National University, 49 Busandaehak-ro, Yangsan 50612, Republic of Korea. Tel: +82-51-510-8091; Fax: +82-51-510-8049; E-mail: yunhak10510@pusan.ac.kr

References

[1] LeCun Y, Bengio Y and Hinton G. Deep learning. Nature 2015; 521: 436-444.

[2] Ardila D, Kiraly AP, Bharadwaj S, Choi B, Reicher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, Naidich DP and Shetty S. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019; 25: 954-961.

[3] Gao F, Wang W, Tan M, Zhu L, Zhang Y, Fessler E, Vermeulen L and Wang X. DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis 2019; 8: 44.

[4] Mostavi M, Chiu YC, Huang Y and Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics 2020; 13 Suppl 5: 44.

[5] Hwang EJ, Nam JG, Lim WH, Park SJ, Jeong YS, Kang JH, Hong EK, Kim TM, Goo JM, Park S, Kim KH and Park CM. Deep learning for chest radiograph diagnosis in the emergency department. Radiology 2019; 293: 573-580.

[6] Bradley CJ, Given CW and Roberts C. Disparities in cancer diagnosis and survival. Cancer 2001; 91: 178-188.

[7] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286: 531-537.

[8] Tan AC and Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Appl Bioinformatics 2003; 2 Suppl: S75-83.

[9] Newman-Toker DE, Wang Z, Zhu Y, Nassery N, Saber Tehrani AS, Schaffer AC, Yu-Moe CW, Clemens GD, Fanai M and Siegal D. Rate of diagnostic errors and serious misdiagnosis-related harms for major vascular events, infections, and cancers: toward a national incidence estimate using the "Big Three". Diagnosis 2021; 8: 67-84.

[10] Moreira MW, Rodrigues JJ, Korotaev V, Al-Muhtadi J and Kumar N. A comprehensive review on smart decision support systems for health care. IEEE Systems Journal 2019; 13: 3536-3545.

[11] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C and Stuart JM. The cancer genome atlas pan-cancer analysis project. Nat Genet 2013; 45: 1113-1120.

[12] International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, Eerola I, Gerhard DS, Guttmacher A, Guyer M, Hemsley FM, Jennings JL, Kerr D, Klatt P, Kolar P, Kusada J, Lane DP, Laplace F, Youyong L, Nettekoven G, Ozenberger B, Peterson J, Rao TS, Remacle J, Schafer AJ, Shibata T, Stratton MR, Vockley JG, Watanabe K, Yang H, Yuen MM, Knoppers BM, Bobrow M, Cambon-Thomsen A, Dressler LG, Dyke SO, Joly Y, Kato K, Kennedy KL, Nicolás P, Parker MJ, Rial-Sebbag E, Romeo-Casabona CM, Shaw KM, Wallace S, Wiesner GL, Zeps N, Lichter P, Biankin AV, Chabannon C, Chin L, Clément B, de Alava E, Degos F, Ferguson ML, Geary P, Hayes DN, Hudson TJ, Johns AL, Kasprzyk A, Nakagawa H, Penny R, Piris MA, Sarin R, Scarpa A, Shibata T, van de Vijver M, Futreal PA, Aburatani H, Bayés M, Botwell DD, Campbell PJ, Estivill X, Gerhard DS, Grimmond SM, Gut I, Hirst M, López-Otín C, Majumder P, Marra M, McPherson JD, Nakagawa H, Ning Z, Puente XS, Ruan Y, Shibata T, Stratton MR, Stunnenberg HG, Swerdlow H, Velculescu VE, Wilson RK, Xue HH, Yang L, Spellman PT, Bader GD, Boutros PC, Campbell PJ, Flicek P, Getz G, Guigó R, Guo G, Haussler D, Heath S, Hubbard TJ, Jiang T, Jones SM, Li Q, López-Bigas N, Luo R, Muthuswamy L, Ouellette BF, Pearson JV, Puente XS, Quesada V, Raphael BJ, Sander C, Shibata T, Speed TP, Stein LD, Stuart JM, Teague JW, Totoki Y, Tsunoda T, Valencia A, Wheeler DA, Wu H, Zhao S, Zhou G, Stein LD, Guigó R, Hubbard TJ, Joly Y,

Jones SM, Kasprzyk A, Lathrop M, López-Bigas N, Ouellette BF, Spellman PT, Teague JW, Thomas G, Valencia A, Yoshida T, Kennedy KL, Axton M, Dyke SO, Futreal PA, Gerhard DS, Gunter C, Guyer M, Hudson TJ, McPherson JD, Miller LJ, Ozenberger B, Shaw KM, Kasprzyk A, Stein LD, Zhang J, Haider SA, Wang J, Yung CK, Cros A, Liang Y, Gnaneshan S, Guberman J, Hsu J, Bobrow M, Chalmers DR, Hasel KW, Joly Y, Kaan TS, Kennedy KL, Knoppers BM, Lowrance WW, Masui T, Nicolás P, Rial-Sebbag E, Rodriguez LL, Vergely C, Yoshida T, Grimmond SM, Biankin AV, Bowtell DD, Cloonan N, deFazio A, Eshleman JR, Etemadmoghadam D, Gardiner BB, Kench JG, Scarpa A, Sutherland RL, Tempero MA, Waddell NJ, Wilson PJ, McPherson JD, Gallinger S, Tsao MS, Shaw PA, Petersen GM, Mukhopadhyay D, Chin L, DePinho RA, Thayer S, Muthuswamy L, Shazand K, Beck T, Sam M, Timms L, Ballin V, Lu Y, Ji J, Zhang X, Chen F, Hu X, Zhou G, Yang Q, Tian G, Zhang L, Xing X, Li X, Zhu Z, Yu Y, Yu J, Yang H, Lathrop M, Tost J, Brennan P, Holcatova I, Zaridze D, Brazma A, Egevard L, Prokhortchouk E, Banks RE, Uhlén M, Cambon-Thomsen A, Viksna J, Ponten F, Skryabin K, Stratton MR, Futreal PA, Birney E, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Martin S, Reis-Filho JS, Richardson AL, Sotiriou C, Stunnenberg HG, Thoms G, van de Vijver M, van't Veer L, Calvo F, Birnbaum D, Blanche H, Boucher P, Boyault S, Chabannon C, Gut I, Masson-Jacquemier JD, Lathrop M, Pauporté I, Pivot X, Vincent-Salomon A, Tabone E, Theillet C, Thomas G, Tost J, Treilleux I, Calvo F, Bioulac-Sage P, Clément B, Decaens T, Degos F, Franco D, Gut I, Gut M, Heath S, Lathrop M, Samuel D, Thomas G, Zucman-Rossi J, Lichter P, Eils R, Brors B, Korbel JO, Korshunov A, Landgraf P, Lehrach H, Pfister S, Radlwimmer B, Reifenberger G, Taylor MD, von Kalle C, Majumder PP, Sarin R, Rao TS, Bhan MK, Scarpa A, Pederzoli P, Lawlor RA, Delledonne M, Bardelli A, Biankin AV, Grimmond SM, Gress T, Klimstra D, Zamboni G, Shibata T, Nakamura Y, Nakagawa H, Kusada J, Tsunoda T, Miyano S, Aburatani H, Kato K, Fujimoto A, Yoshida T, Campo E, López-Otín C, Estivill X, Guigó R, de Sanjosé S, Piris MA, Montserrat E, González-Díaz M, Puente XS, Jares P, Valencia A, Himmelbauer H, Quesada V, Bea S, Stratton MR, Futreal PA, Campbell PJ, Vincent-Salomon A, Richardson AL, Reis-Filho JS, van de Vijver M, Thomas G, Masson-Jacquemier JD, Aparicio S, Borg A, Børresen-Dale AL, Caldas C, Foekens JA, Stunnenberg HG, van't Veer L, Easton DF, Spellman PT, Martin S, Barker AD, Chin L, Collins FS, Compton CC, Ferguson ML, Gerhard DS, Getz G, Gunter C, Guttmacher A, Guyer M, Hayes DN, Lander ES, Ozenberger B, Penny R, Peterson J, Sander C, Shaw KM, Speed TP, Spellman PT, Vockley JG, Wheeler DA, Wilson RK, Hudson TJ, Chin L, Knoppers BM, Lander ES, Lichter P, Stein LD, Stratton MR, Anderson W, Barker AD, Bell C, Bobrow M, Burke W, Collins FS, Compton CC, DePinho RA, Easton DF, Futreal PA, Gerhard DS, Green AR, Guyer M, Hamilton SR, Hubbard TJ, Kallioniemi OP, Kennedy KL, Ley TJ, Liu ET, Lu Y, Majumder P, Marra M, Ozenberger B, Peterson J, Schafer AJ, Spellman PT, Stunnenberg HG, Wainwright BJ, Wilson RK and Yang H. International network of cancer genome projects. Nature 2010; 464: 993-998.

[13] Chu Y and Corey DR. RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther 2012; 22: 271-274.

[14] Li Y, Kang K, Krahn JM, Croutwater N, Lee K, Umbach DM and Li L. A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data. BMC Genomics 2017; 18: 508.

[15] Lin WJ and Chen JJ. Class-imbalanced classifiers for high-dimensional data. Brief Bioinform 2013; 14: 13-26.

[16] Yang S and Naiman DQ. Multiclass cancer classification based on gene expression comparison. Stat Appl Genet Mol Biol 2014; 13: 477-496.

[17] Latchman DS. Transcription factors: an overview. Int J Biochem Cell Biol 1997; 29: 1305-1312.

[18] Darnell JE Jr. Transcription factors as targets for cancer therapy. Nat Rev Cancer 2002; 2: 740-749.

[19] Bushweller JH. Targeting transcription factors in cancer-from undruggable to reality. Nat Rev Cancer 2019; 19: 611-624.

[20] Vishnoi K, Viswakarma N, Rana A and Rana B. Transcription factors in cancer development and therapy. Cancers (Basel) 2020; 12: 2296.

[21] Wolfe SA, Nekludova L and Pabo CO. DNA recognition by Cys2His2 zinc finger proteins. Annu Rev Biophys Biomol Struct 2000; 29: 183-212.

[22] Scott MP, Tamkun JW and Hartzell GW 3rd. The structure and function of the homeodomain. Biochim Biophys Acta 1989; 989: 25-48.

[23] Jones S. An overview of the basic helix-loop-helix proteins. Genome Biol 2004; 5: 226.

[24] Razin SV, Borunova VV, Maksimenko OG and Kantidze OL. Cys2His2 zinc finger protein family: classification, functions, and major members. Biochemistry (Mosc) 2012; 77: 217-226.

[25] Jen J and Wang YC. Zinc finger proteins in cancer progression. J Biomed Sci 2016; 23: 53.

[26] Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G and Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016; 44: e71.
[27] Lambert SA, Jolma A, Campitelli LF, Das PK, Yin Y, Albu M, Chen X, Taipale J, Hughes TR and Weirauch MT. The human transcription factors. Cell 2018; 172: 650-665.
[28] Johnson JM and Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data 2019; 6: 27.
[29] Chen Y, Xu L, Lin RY, Müschen M and Koeffler HP. Core transcriptional regulatory circuitries in cancer. Oncogene 2020; 39: 6633-6646.
[30] Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y and Xu W. Applications of support vector machine (SVM) learning in cancer genomics. Cancer Genomics Proteomics 2018; 15: 41-51.
[31] Hu LY, Huang MW, Ke SW and Tsai CF. The distance function effect on k-nearest neighbor classification for medical datasets. Springerplus 2016; 5: 1304.
[32] Cassandri M, Smirnov A, Novelli F, Pitolli C, Agostini M, Malewicz M, Melino G and Raschellà G. Zinc-finger proteins in health and disease. Cell Death Discov 2017; 3: 17071.
[33] Ito G, Uchiyama M, Kondo M, Mori S, Usami N, Maeda O, Kawabe T, Hasegawa Y, Shimokata K and Sekido Y. Krüppel-like factor 6 is frequently down-regulated and induces apoptosis in non-small cell lung cancer cells. Cancer Res 2004; 64: 3838-3843.
[34] Bieker JJ. Krüppel-like factors: three fingers in many pies. J Biol Chem 2001; 276: 34355-34358.
[35] Matsumoto N, Kubo A, Liu H, Akita K, Laub F, Ramirez F, Keller G and Friedman SL. Developmental regulation of yolk sac hematopoiesis by Kruppel-like factor 6. Blood 2006; 107: 1357-1365.
[36] Tetreault MP, Yang Y and Katz JP. Krüppel-like factors in cancer. Nat Rev Cancer 2013; 13: 701-713.
[37] Tsukahara T, Nabeta Y, Kawaguchi S, Ikeda H, Sato Y, Shimozawa K, Ida K, Asanuma H, Hirohashi Y, Torigoe T, Hiraga H, Nagoya S, Wada T, Yamashita T and Sato N. Identification of human autologous cytotoxic T-lymphocyte-defined osteosarcoma gene that encodes a transcriptional regulator, papillomavirus binding factor. Cancer Res 2004; 64: 5442-5448.
[38] Yabe H, Tsukahara T, Kawaguchi S, Wada T, Sato N, Morioka H and Yabe H. Overexpression of papillomavirus binding factor in Ewing's sarcoma family of tumors conferring poor prognosis. Oncol Rep 2008; 19: 129-134.
[39] Holland DG, Burleigh A, Git A, Goldgraben MA, Perez-Mancera PA, Chin SF, Hurtado A, Bruna A, Ali HR, Greenwood W, Dunning MJ, Samarajiwa S, Menon S, Rueda OM, Lynch AG, McKinney S, Ellis IO, Eaves CJ, Carroll JS, Curtis C, Aparicio S and Caldas C. ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. EMBO Mol Med 2011; 3: 167-180.
[40] Wang S, Wang C, Hu Y, Li X, Jin S, Liu O, Gou R, Zhuang Y, Guo Q, Nie X, Zhu L, Liu J and Lin B. ZNF703 promotes tumor progression in ovarian cancer by interacting with HE4 and epigenetically regulating PEA15. J Exp Clin Cancer Res 2020; 39: 264.
[41] Yang C, Wu J, Liu X, Wang Y, Liu B, Chen X, Wu X, Yan D, Han L, Liu S, Shan L and Shang Y. Circadian rhythm is disrupted by ZNF704 in breast carcinogenesis. Cancer Res 2020; 80: 4114-4128.
[42] Huang X, Du X and Li Y. The role of BCL11B in hematological malignancy. Exp Hematol Oncol 2012; 1: 22.
[43] Vandamme N, Denecker G, Bruneel K, Blancke G, Akay Ö, Taminau J, De Coninck J, De Smedt E, Skrypek N, Van Loocke W, Wouters J, Nittner D, Köhler C, Darling DS, Cheng PF, Raaijmakers MIG, Levesque MP, Mallya UG, Rafferty M, Balint B, Gallagher WM, Brochez L, Huylebroeck D, Haigh JJ, Andries V, Rambow F, Van Vlierberghe P, Goossens S, van den Oord JJ, Marine JC and Berx G. The EMT transcription factor ZEB2 promotes proliferation of primary and metastatic melanoma while suppressing an invasive, mesenchymal-like phenotype. Cancer Res 2020; 80: 2983-2995.
[44] Gai WX, Ma X, Qiao YM, Shi BH, Ul Haq S, Li QH, Wei AM, Liu KK and Gong ZH. Characterization of the bZIP transcription factor family in pepper (capsicum annuum l.): CabZIP25 positively modulates the salt tolerance. Front Plant Sci 2020; 11: 139.
[45] Wang Y, Zhang Y, Zhou R, Dossa K, Yu J, Li D, Liu A, Mmadi MA, Zhang X and You J. Identification and characterization of the bZIP transcription factor family and its expression in response to abiotic stresses in sesame. PLoS One 2018; 13: e0200850.
[46] Hetz C. The unfolded protein response: controlling cell fate decisions under ER stress and beyond. Nat Rev Mol Cell Biol 2012; 13: 89-102.
[47] van Dam H and Castellazzi M. Distinct roles of Jun: Fos and Jun: ATF dimers in oncogenesis. Oncogene 2001; 20: 2453-2464.

Supplementary Table 1. Cohort information of TCGA and ICGC patients by eight cancer types

TCGA-BRCA			ICGC-BRCA-KR		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	1101	58 (18.9)	Age (years)	50	31.8 (2.6)
	N	Percentage (%)		N	Percentage (%)
Sex	1101		Sex	50	
Female	1089	98.9	Female	50	100.0
Male	12	1.1	Male	-	-
Race	1007		Race		
White	761	75.5	White	-	-
Asian	61	6	Asian	-	-
Black	184	18.2	Black	-	-
American Indian	1	0.1	American Indian	-	-
AJCC_pathologic_stage	1090		AJCC_pathologic_stage		
I-II	809	74.2	I-II	-	-
III-IV	269	24.6	III-IV	-	-
X	12	1.1	X	-	-
pT stage	1101		pT stage	47	
T0	-	-	T0	-	-
T1	280	25.4	T1	19	40.4
T2	639	58.0	T2	24	51.1
T3	139	12.6	T3	3	6.4
T4	40	3.6	T4	1	2.1
TX	3	0.2	TX		
pN stage	1101		pN stage		
N0	521	47.3	N0	-	-
N1	364	33	N1	-	-
N2	120	10.8	N2	-	-
N3	76	6.9	N3	-	-
NX	20	1.8	NX		
pM stage	1101		pM stage		
M0	914	83.0	M0	-	-
M1	22	1.9	M1	-	-
MX	165	14.9	MX	-	-
primary_diagnosis	749		primary_diagnosis	48	
IDC	522	69.6	IDC	44	91.7
ILC	136	18.1	DCIS	3	6.3
Mixed	91	12.1	Adenocarcinoma	1	2.1

TCGA-DLBC			ICGC-MALY-DE		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	48	55 (15.4)	Age (years)	105	49.1 (23.7)
	N	Percentage (%)		N	Percentage (%)
Sex	48		Sex	105	
Female	26	54.2	Female	42	40.0
Male	22	45.8	Male	63	60.0
Race	48		Race		
White	29	60.4	White	-	-
Asian	18	37.5	Asian	-	-
Black	1	2.1	Black	-	-
ANN_arbor_clinical_stage	42		ANN_arbor_clinical_stage	94	
I-II	25	59.5	I-II	27	28.7
III-IV	17	40.5	III-IV	67	71.3
X	-	-	X	-	-

pT stage			pT stage		
T0	-	-	T0	-	-
T1	-	-	T1	-	-
T2	-	-	T2	-	-
T3	-	-	T3	-	-
T4	-	-	T4	-	-
TX	-	-	TX	-	-
pN stage	-	-	pN stage	-	-
N0	-	-	N0	-	-
N1	-	-	N1	-	-
N2	-	-	N2	-	-
N3	-	-	N3	-	-
NX	-	-	NX	-	-
pM stage	48		pM stage		
M0	-	-	M0	-	-
M1	-	-	M1	-	-
MX	48	100.0	MX	-	-
primary_diagnosis	48	-	primary_diagnosis		
Malignant lymphoma	48	100.0	-	-	-

TCGA-HNSC			ICGC-ORCA-IN		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	499	57.15 (18.6)	Age (years)	40	47.9 (10.8)
	N	Percentage (%)		N	Percentage (%)
Sex	500		Sex	40	
Female	133	26.6	Female	6	15
Male	367	73.4	Male	34	85
Race	485		Race		
White	426	87.8	White	-	-
Asian	10	2.1	Asian	-	-
Black	47	9.7	Black	-	-
American Indian	2	0.4	American Indian	-	-
AJCC_pathologic_stage	432		AJCC_pathologic_stage		
I-II	95	22.0	I-II	-	-
III-IV	337	78.0	III-IV	-	-
X	-	-	X	-	-
pT stage	478		pT stage	40	
T0	1	0.2	T0	-	-
T1	45	9.4	T1	-	-
T2	132	27.6	T2	1	2.5
T3	96	20.1	T3	3	7.5
T4	171	35.8	T4	36	90.0
TX	33	6.9	TX		
pN stage	476		pN stage	40	
N0	171	35.9	N0	10	25.0
N1	65	13.7	N1	19	47.5
N2	164	34.5	N2	11	27.5
N3	7	1.5	N3	-	-
NX	69	14.5	NX	-	-
pM stage	248		pM stage	40	
M0	186	75.0	M0	40	100.0
M1	1	0.4	M1	-	-
MX	61	24.6	MX	-	-

primary_diagnosis	500		primary_diagnosis		
Basaloid squamous cell carcinoma	10	2.0	-	-	-
Squamous cell carcinoma	490	98.0	-	-	-

TCGA-KIRC			ICGC-RECA-AU		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	538	58.9 (17.4)	Age (years)	91	60.4 (10.0)
	N	Percentage (%)		N	Percentage (%)
Sex	538		Sex	91	
Female	186	34.6	Female	39	42.9
Male	352	65.4	Male	52	57.1
Race	531		Race		
White	467	87.9	White	-	-
Asian	8	1.5	Asian	-	-
Black	56	10.5	Black	-	-
AJCC_pathologic_stage	535		AJCC_pathologic_stage		
I-II	330	61.7	I-II	-	-
III-IV	205	38.3	III-IV	-	-
X	-	-	X	-	-
pT stage	538		pT stage	91	
T0	-	-	T0	-	-
T1	277	51.5	T1	54	59.3
T2	71	13.2	T2	13	14.3
T3	179	33.3	T3	22	24.2
T4	11	2.0	T4	2	2.2
TX	-	-	TX	-	-
pN stage	538		pN stage	91	
N0	241	44.8	N0	79	86.8
N1	16	3.0	N1	2	2.2
N2	-	-	N2	-	-
N3	-	-	N3	-	-
NX	281	52.2	NX	10	11.0
pM stage	536		pM stage	91	
M0	428	79.9	M0	81	89.0
M1	78	14.6	M1	9	9.9
MX	30	5.6	MX	1	1.1
primary_diagnosis	535		primary_diagnosis		
Clear cell adenocarcinoma	524	97.9	-	-	-
Renal cell carcinoma	14	2.6	-	-	-

TCGA-OV			ICGC-OV-AU		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	374	59.9 (15.2)	Age (years)	93	59.6 (8.6)
		Percentage (%)			Percentage (%)
Sex	374		Sex	93	
Female	374	100.0	Female	93	100.0
Male	-	-	Male		
Race	360		Race		
White	324	90.0	White	-	-
Asian	11	3.1	Asian	-	-
Black	25	6.9	Black	-	-
American Indian	2	0.6	American Indian		
Native Hawaiian	1	0.3	Native Hawaiian		

FIGO_stage	371		FIGO_stage	93	
I-II	22	5.9	I-II	-	-
III-IV	349	94.1	III-IV	93	100.0
X	-	-	X	-	-
pT stage			pT stage		
T0	-	-	T0	-	-
T1	-	-	T1	-	-
T2	-	-	T2	-	-
T3	-	-	T3	-	-
T4	-	-	T4	-	-
TX	-	-	TX	-	-
pN stage			pN stage		
N0	-	-	N0	-	-
N1	-	-	N1	-	-
N2	-	-	N2	-	-
N3	-	-	N3	-	-
NX	-	-	NX	-	-
pM stage			pM stage		
M0	-	-	M0	-	-
M1	-	-	M1	-	-
MX	-	-	MX	-	-
primary_diagnosis	374		primary_diagnosis		
Serous cystadenocarcinoma	371	99.2	-	-	-
Papillary serous cystadenocarcinoma	3	0.8	-	-	-

TCGA-PAAD			ICGC-PACA-AU		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	177	62.5 (14.1)	Age (years)	91	65.6 (10.6)
		Percentage (%)			Percentage (%)
Sex	177		Sex	90	
Female	80	45.2	Female	43	47.8
Male	97	54.8	Male	47	52.2
Race	173		Race		
White	156	90.2	White	-	-
Asian	11	6.4	Asian	-	-
Black	6	3.5	Black	-	-
AJCC_pathologic_stage	174		AJCC_pathologic_stage		
I-II	167	96.0	I-II	-	-
III-IV	7	4.0	III-IV	-	-
X	-	-	X	-	-
pT stage	176		pT stage	86	
T0	-	-	T0	-	-
T1	7	4.0	T1	1	1.2
T2	24	13.6	T2	8	9.3
T3	141	80.1	T3	74	86.0
T4	3	1.7	T4	1	1.2
TX	1	0.6	TX	2	2.3
pN stage	176		pN stage	86	86.0
N0	49	27.8	N0	26	30.2
N1	123	69.9	N1	58	67.4
N2	-	-	N2	-	-
N3	-	-	N3	-	-
NX	4	2.3	NX	2	2.3

pM stage	177		pM stage	89	
M0	79	44.6	M0	6	6.7
M1	4	2.3	M1	7	7.9
MX	94	53.1	MX	76	85.4
primary_diagnosis	177		primary_diagnosis	89	
Infiltrating duct carcinoma	142	80.2	Pancreatic Ductal Adenocarcinoma	73	82.0
Adenocarcinoma	20	11.3	Acinar Cell Carcinoma	2	2.2
Neuroendocrine carcinoma	8	4.5	Adenosquamous carcinoma	4	4.5
Mucinous adenocarcinoma	5	2.8	Intraductal Papillary Mucinous Neoplasm with invasion	8	9.0
-	-	-	Mucinous Non-cystic carcinoma	1	1.1
-	-	-	Undifferentiated (anaplastic) carcinoma	1	1.1

TCGA-PRAD			ICGC-PRAD-FR		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	498	60.3 (10.7)	Age (years)	25	63.3 (6.1)
		Percentage (%)			Percentage (%)
Sex	498		Sex	25	
Female	-	-	Female	-	
Male	498	100.0	Male	25	100.0
Race	484		Race		
White	414	85.5	White	-	-
Asian	12	2.5	Asian	-	-
Black	57	11.8	Black	-	-
American Indian	1	0.2	American Indian	-	-
Primary_gleason_grade	498		Primary_gleason_grade		
Pattern 2	1	0.2	Pattern 2	-	-
Pattern 3	199	40.0	Pattern 3	-	-
Pattern 4	249	50.0	Pattern 4	-	-
Pattern 5	49	9.8	Pattern 5	-	-
pT stage	491		pT stage	25	
T0	-	-	T0	-	-
T1	-	-	T1	11	44.0
T2	189	38.5	T2	14	56.0
T3	292	59.5	T3	-	-
T4	10	2.0	T4	-	-
TX	-	-	TX	-	-
pN stage	425		pN stage	25	
N0	347	81.6	N0	25	100.0
N1	78	18.4	N1	-	-
N2	-	-	N2	-	-
N3	-	-	N3	-	-
NX	-	-	NX	-	-
pM stage			pM stage	25	
M0	-	-	M0	25	100.0
M1	-	-	M1	-	-
MX	-	-	MX	-	-
primary_diagnosis	498		primary_diagnosis		
Adenocarcinoma	485	97.4	-	-	-
Infiltrating duct carcinoma	9	1.8	-	-	-
Adenocarcinoma	3	0.6	-	-	-
Mucinous adenocarcinoma	1	0.2	-	-	-

TCGA-SARC			ICGC-BOCA-FR		
Variable	N	Mean (SD*)	Variable	N	Mean (SD*)
Age (years)	259	58.5 (19.0)	Age (years)	57	16.6 (8.4)
		Percentage (%)			Percentage (%)
Sex	259		Sex	57	
Female	141	54.4	Female	26	45.6
Male	118	45.6	Male	31	54.4
Race	250		Race		
White	226	90.4	White	-	-
Asian	6	2.4	Asian	-	-
Black	18	7.2	Black	-	-
AJCC_pathologic_stage			AJCC_pathologic_stage		
I-II	-	-	I-II	-	-
III-IV	-	-	III-IV	-	-
X	-	-	X	-	-
pT stage			pT stage	55	
T0	-	-	T0	-	-
T1	-	-	T1	18	32.7
T2	-	-	T2	-	-
T3	-	-	T3	-	-
T4	-	-	T4	-	-
TX	-	-	TX	37	67.3
pN stage			pN stage		
N0	-	-	N0	-	-
N1	-	-	N1	-	-
N2	-	-	N2	-	-
N3	-	-	N3	-	-
NX	-	-	NX	-	-
pM stage			pM stage		
M0	-	-	M0	-	-
M1	-	-	M1	-	-
MX	-	-	MX	-	-
primary_diagnosis	259		primary_diagnosis		
fibromatosis	2	0.8	-	-	-
Dedifferentiated liposarcoma	57	22.0	-	-	-
Fibromyxosarcoma	25	9.7	-	-	-
Giant cell sarcoma	3	1.2	-	-	-
Leiomyosarcoma	101	39.0	-	-	-
Liposarcom	1	0.4	-	-	-
Malignant fibrous histiocytoma	12	4.6	-	-	-
Malignant peripheral nerve sheath tumor	9	3.5	-	-	-
Myxoid leiomyosarcoma	3	1.2	-	-	-
Pleomorphic liposarcoma	2	0.8	-	-	-
Synovial sarcoma	10	3.9	-	-	-
Undifferentiated sarcoma	34	13.1	-	-	-

*SD: standard deviation.

Supplementary Table 2. Performance evaluation of TCGA and ICGC classification using SVM

TCGA train set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.9982	0.9973	0.9978	0.9984	0.9998	0.9987	0.9991
2-network	0.9986	0.9979	0.9982	0.9985	0.9998	0.9989	0.9992
3-network	0.9982	0.9973	0.9981	0.9990	0.9998	0.9988	0.9994
4-network	0.9986	0.9979	0.9982	0.9985	0.9998	0.999	0.9992
5-network	0.9982	0.9971	0.9976	0.9982	0.9998	0.9984	0.9990
TCGA validation set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.9957	0.9932	0.9947	0.9964	0.9994	0.9961	0.9979
2-network	0.9943	0.9911	0.9930	0.9954	0.9992	0.995	0.9973
3-network	0.9886	0.9841	0.9853	0.9869	0.9983	0.9914	0.9926
4-network	0.9928	0.9914	0.9935	0.9959	0.9989	0.9952	0.9974
5-network	0.9971	0.9964	0.9973	0.9983	0.9996	0.9986	0.9989
ICGC test set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.8859	0.9027	0.8918	0.9196	0.9842	0.9424	0.9519
2-network	0.8551	0.8848	0.8592	0.8997	0.9799	0.9057	0.9398
3-network	0.9040	0.9088	0.9074	0.9318	0.9867	0.9317	0.9593
4-network	0.8859	0.9040	0.8914	0.9203	0.9842	0.9314	0.9522
5-network	0.8678	0.8896	0.8717	0.9080	0.9817	0.9324	0.9449

Supplementary Table 3. Performance evaluation of TCGA and ICGC classification using KNN

TCGA train set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.9943	0.9919	0.9931	0.9943	0.9992	0.9956	0.9967
2-network	0.9925	0.9870	0.9894	0.9918	0.9989	0.9925	0.9953
3-network	0.9946	0.9926	0.9940	0.9954	0.9992	0.9953	0.9973
4-network	0.9939	0.9922	0.9927	0.9932	0.9991	0.9941	0.9962
5-network	0.9929	0.9905	0.9917	0.9930	0.9990	0.9941	0.9960
TCGA validation set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.9943	0.9935	0.9940	0.9946	0.9991	0.9925	0.9969
2-network	0.9957	0.9932	0.9950	0.9970	0.9994	0.9962	0.9982
3-network	0.9857	0.9809	0.9814	0.9821	0.9979	0.9873	0.9900
4-network	0.9900	0.9864	0.9887	0.9911	0.9985	0.9911	0.9948
5-network	0.9957	0.9951	0.9955	0.9958	0.9994	0.9993	0.9976
ICGC test set	Accuracy	Precision	F_score	Recall (sensitivity)	Specificity	AUC	Balanced Accuracy
1-network	0.8659	0.8979	0.8561	0.8729	0.9810	0.8838	0.9270
2-network	0.8478	0.8923	0.8388	0.8588	0.9787	0.8902	0.9188
3-network	0.8533	0.8882	0.8491	0.8666	0.9788	0.8575	0.9227
4-network	0.8514	0.8865	0.8418	0.8622	0.9789	0.8773	0.9206
5-network	0.8533	0.8897	0.8500	0.8667	0.9792	0.8647	0.9230

Supplementary Table 4. The functional role of top 10 TF genes in each cancer

Type	Gene	Description
BRCA-BRCA-KR	TRPS1, STAT3	Upregulated in cancer (diagnostic marker) [1]
	GATA3	Tumor suppressor [2]
	BHLHE40	Predicting disease outcome and metastatic risk [3]
	PBX1	Prognostic marker for ER-positive, luminal A, and luminal B subtypes [4]
	YBX1	Overexpression is associated with unfavorable outcome [5]
	ATF4	Upregulated ATF4 by HER2 promotes tumor cell migration [6]
DLBC-MALY-DE	IRF8	IRF8 upregulation was detected in DLBCL tumor tissues [7]
	ETS1	Regulation of immune cell function [8]
	ATF4	Overexpression of ATF4 results with MYC dysregulation in lymphoma progression [9]
	STAT6	Primary mediastinal B-Cell lymphoma (PMBL) pathogenesis [10]
HNSC-ORCA-IN	TP63	TP63 overexpression leads to head and neck squamous cell carcinoma progression and metastasis [11]
	HMGA1	Overexpression is associated with tumourigenesis [12]
	JUNB	JunB promotes cell invasion, migration and distant metastasis in head and neck squamous cell carcinoma [13]
	HIF1A	HIF-1α overexpression in oral squamous cell carcinoma [14]
KIRC-RECA-EU	EPAS1	Upregulated in renal cell carcinoma [15]
	KLF6	Tumor suppressor [16]
	YBX1	Overexpression is associated with migration, invasion, and adhesion in renal cell carcinoma [17]
	Jun	Inducing malignant transformation in renal cell carcinoma [18]
	ZNF395	Overexpression is associated with renal cell carcinoma proliferation, migration, and invasion [19]
	AEBP1	Novel candidate biomarker for diabetic kidney disease [20]
	ETS1	Ets-1 is involved in angiogenesis in renal cell carcinoma [21]
OV-OV-AU	YBX1	YB-1 activation is a powerful marker of outcomes for ovarian cancer patients [22]
	HMGA1	Overexpression is associated with epithelial ovarian carcinomas [23]
	JUND, JUNB, JUN	The regulation of cell proliferation, apoptosis and angiogenesis in cancer [24]
	PAX8	Expression of PAX8 associated with ovarian carcinomas [25]
	ELF3	Novel biomarker for the prognosis of ovarian cancer [26]
	STAT1	Prognostic biomarker for High-Grade Serous Ovarian Cancer [27]
PAAD-PACA-AU	FOS	c-fos expression is associated with pancreatic cancer (PC) progression and dismal prognosis [28]
	YBX1	Overexpression of YBX1 promotes pancreatic ductal adenocarcinoma growth [29]
	ATF4	ATF4 is overexpressed in PDAC and associated with a poor prognosis [30]
	STAT3	Targeting STAT3 in Cancer Immunotherapy [31]
PRAD-PRAD-FR	SPDEF	Upregulation of SPDEF is associated with poor prognosis in prostate cancer [32]
	HOXB13	HOX13 expression is associated with carcinogenesis of prostate cancer [33]
	FOXA1	FOXA1 promotes tumor progression in prostate cancer [34]
	NFIX	NFI interact with FOXA1 to regulate prostate-specific gene expression [35]
	JUN	long-term c-Jun overexpression also down-regulates Androgen receptor (AR) expression [36]
	NFE2L1	Candidate biomarkers for prostate cancer [37]
SARC-BOCA-FR	JUN	JUN oncogene amplification and overexpression in sarcoma [38]
	ATF4	ATF4 activation causes proteasome inhibitor bortezomib-induced osteosarcoma cell death [39]
	STAT3	STAT3 is activated in a subset of the Ewing sarcoma family of tumours [40]
	PRRX1	PRRX1 promotes malignant properties in human osteosarcoma [41]

References

[1] Ai D, Yao J, Yang F, Huo L, Chen H, Lu W, Soto LMS, Jiang M, Raso MG, Wang S, Bell D, Liu J, Wang H, Tan D, Torres-Cabala C, Gan Q, Wu Y, Albarracin C, Hung MC, Meric-Bernstam F, Wistuba II, Prieto VG, Sahin AA and Ding Q. TRPS1: a highly sensitive and specific marker for breast carcinoma, especially for triple-negative breast cancer. Mod Pathol 2021; 34: 710-719.

[2] Chu IM, Lai WC, Aprelikova O, El Touny LH, Kouros-Mehr H and Green JE. Expression of GATA3 in MDA-MB-231 triple-negative breast cancer cells induces a growth inhibitory response to TGFβ. PLoS One 2013; 8: e61125.

[3] Kiss Z, Mudryj M and Ghosh PM. Non-circadian aspects of BHLHE40 cellular function in cancer. Genes Cancer 2020; 11: 1-19.

[4] Ao X, Ding W, Ge H, Zhang Y, Ding D and Liu Y. PBX1 is a valuable prognostic biomarker for patients with breast cancer. Exp Ther Med 2020; 20: 385-394.

[5] Wang X, Guo XB, Shen XC, Zhou H, Wan DW, Xue XF, Han Y, Yuan B, Zhou J, Zhao H, Zhi QM and Kuang YT. Prognostic role of YB-1 expression in breast cancer: a meta-analysis. Int J Clin Exp Med 2015; 8: 1780-91.
[6] Zeng P, Sun S, Li R, Xiao ZX and Chen H. HER2 upregulates ATF4 to promote cell migration via activation of ZEB1 and downregulation of E-cadherin. Int J Mol Sci 2019; 20: 2223.
[7] Zhong W, Xu X, Zhu Z, Du Q, Du H, Yang L, Ling Y, Xiong H and Li Q. Increased expression of IRF8 in tumor cells inhibits the generation of Th17 cells and predicts unfavorable survival of diffuse large B cell lymphoma patients. Oncotarget 2017; 8: 49757-49772.
[8] Priebe V, Sartori G, Napoli S, Chung EYL, Cascione L, Kwee I, Arribas AJ, Mensah AA, Rinaldi A, Ponzoni M, Zucca E, Rossi D, Efremov D, Lenz G, Thome M and Bertoni F. Role of ETS1 in the transcriptional network of diffuse large b cell lymphoma of the activated B cell-like type. Cancers 2020; 12: 1912.
[9] Tameire F, Verginadis II, Leli NM, Polte C, Conn CS, Ojha R, Salas Salinas C, Chinga F, Monroy AM, Fu W, Wang P, Kossenkov A, Ye J, Amaravadi RK, Ignatova Z, Fuchs SY, Diehl JA, Ruggero D and Koumenis C. ATF4 couples MYC-dependent translational activity to bioenergetic demands during tumour progression. Nat Cell Biol 2019; 21: 889-899.
[10] Ritz O, Guiter C, Dorsch K, Dusanter-Fourt I, Wegener S, Jouault H, Gaulard P, Castellano F, Möller P and Leroy K. STAT6 activity is regulated by SOCS-1 and modulates BCL-XL expression in primary mediastinal B-cell lymphoma. Leukemia 2008; 22: 2106-2110.
[11] Lakshmanachetty S, Balaiya V, High WA and Koster MI. Loss of TP63 promotes the metastasis of head and neck squamous cell carcinoma by activating MAPK and STAT3 signaling. Mol Cancer Res 2019; 17: 1279-1293.
[12] Wang Y, Hu L, Zheng Y and Guo L. HMGA1 in cancer: cancer classification by location. J Cell Mol Med 2019; 23: 2293-2302.
[13] Hyakusoku H, Sano D, Takahashi H, Hatano T, Isono Y, Shimada S, Ito Y, Myers JN and Oridate N. JunB promotes cell invasion, migration and distant metastasis of head and neck squamous cell carcinoma. J Exp Clin Cancer Res 2016; 35: 6.
[14] Zhou J, Huang S, Wang L, Yuan X, Dong Q, Zhang D and Wang X. Clinical and prognostic significance of HIF-1α overexpression in oral squamous cell carcinoma: a meta-analysis. World J Surg Oncol 2017; 15: 104.
[15] Xia G, Kageyama Y, Hayashi T, Kawakami S, Yoshida M and Kihara K. Regulation of vascular endothelial growth factor transcription by endothelial PAS domain protein 1 (EPAS1) and possible involvement of EPAS1 in the angiogenesis of renal cell carcinoma. Cancer 2001; 91: 1429-1436.
[16] Gao Y, Li H, Ma X, Fan Y, Ni D, Zhang Y, Huang Q, Liu K, Li X, Wang L, Gu L, Yao Y, Ai Q, Du Q, Song E and Zhang X. KLF6 suppresses metastasis of clear cell renal cell carcinoma via transcriptional repression of E2F1. Cancer Res 2017; 77: 330-342.
[17] Wang Y, Su J, Wang Y, Fu D, Ideozu JE, Geng H, Cui Q, Wang C, Chen R, Yu Y, Niu Y and Yue D. The interaction of YBX1 with G3BP1 promotes renal cell carcinoma cell metastasis via YBX1/G3BP1-SPP1-NF-κB signaling axis. J Exp Clin Cancer Res 2019; 38: 386.
[18] Koo AS, Chiu R, Soong J, Dekernion JB and Belldegrun A. The expression of C-jun and junB mRNA in renal cell cancer and in vitro regulation by transforming growth factor beta 1 and tumor necrosis factor alpha 1. J Urol 1992; 148: 1314-1318.
[19] Zhao C, Wood CG, Karam JA, et al. The role of ZNF395 in renal cell carcinoma proliferation, migration, and invasion. American Society of Clinical Oncology 2016.
[20] Tao Y, Wei X, Yue Y, Wang J, Li J, Shen L, Lu G, He Y, Zhao S, Zhao F, Weng Z, Shen X and Zhou L. Extracellular vesicle-derived AEBP1 mRNA as a novel candidate biomarker for diabetic kidney disease. J Transl Med 2021; 19: 326.
[21] Mikami S, Oya M, Mizuno R, Murai M, Mukai M and Okada Y. Expression of Ets-1 in human clear cell renal cell carcinomas: implications for angiogenesis. Cancer Sci 2006; 97: 875-882.
[22] Panupinthu N, Yu S, Zhang D, Zhang F, Gagea M, Lu Y, Grandis JR, Dunn SE, Lee HY and Mills GB. Self-reinforcing loop of amphiregulin and Y-box binding protein-1 contributes to poor outcomes in ovarian cancer. Oncogene 2014; 33: 2846-2856.
[23] Masciullo V, Baldassarre G, Pentimalli F, Berlingieri MT, Boccia A, Chiappetta G, Palazzo J, Manfioletti G, Giancotti V, Viglietto G, Scambia G and Fusco A. HMGA1 protein over-expression is a frequent feature of epithelial ovarian carcinomas. Carcinogenesis 2003; 24: 1191-1198.
[24] Eckhoff K, Flurschütz R, Trillsch F, Mahner S, Jänicke F and Milde-Langosch K. The prognostic significance of Jun transcription factors in ovarian cancer. J Cancer Res Clin Oncol 2013; 139: 1673-1680.
[25] Nonaka D, Chiriboga L and Soslow RA. Expression of pax8 as a useful marker in distinguishing ovarian carcinomas from mammary carcinomas. Am J Surg Pathol 2008; 32: 1566-1571.
[26] Xu H, Wang H, Li G, Jin X and Chen B. The immune-related gene ELF3 is a novel biomarker for the prognosis of ovarian cancer. Int J Gen Med 2021; 14: 5537-5548.

[27] Josahkian JA, Saggioro FP, Vidotto T, Ventura HT, Candido Dos Reis FJ, de Sousa CB, Tiezzi DG, de Andrade JM, Koti M and Squire JA. Increased STAT1 expression in high grade serous ovarian cancer is associated with a better outcome. Int J Gynecol Cancer 2018; 28: 459-465.
[28] Guo JC, Li J, Zhao YP, Zhou L, Cui QC, Zhou WX, Zhang TP and You L. Expression of c-fos was associated with clinicopathologic characteristics and prognosis in pancreatic cancer. PLoS One 2015; 10: e0120332.
[29] Liu Z, Li Y, Li X, Zhao J, Wu S, Wu H and Gou S. Overexpression of YBX1 promotes pancreatic ductal adenocarcinoma growth via the GSK3B/Cyclin D1/Cyclin E1 pathway. Mol Ther Oncolytics 2020; 17: 21-30.
[30] Wei L, Lin Q, Lu Y, Li G, Huang L, Fu Z, Chen R and Zhou Q. Cancer-associated fibroblasts-mediated ATF4 expression promotes malignancy and gemcitabine resistance in pancreatic cancer via the TGF-β1/SMAD2/3 pathway and ABCC1 transactivation. Cell Death Dis 2021; 12: 334.
[31] Zou S, Tong Q, Liu B, Huang W, Tian Y and Fu X. Targeting STAT3 in cancer immunotherapy. Mol Cancer 2020; 19: 145.
[32] Meiners J, Schulz K, Möller K, Höflmayer D, Burdelski C, Hube-Magg C, Simon R, Göbel C, Hinsch A, Reiswich V, Weidemann S, Izbicki JR, Sauter G, Jacobsen F, Möller-Koop C, Mandelkow T, Blessin NC, Lutz F, Viehweger F, Lennartz M, Fraune C, Heinzer H, Minner S, Bonk S, Huland H, Graefen M, Schlomm T and Büscheck F. Upregulation of SPDEF is associated with poor prognosis in prostate cancer. Oncol Lett 2019; 18: 5107-5118.
[33] Park CK, Shin SJ, Cho YA, Joo JW and Cho NH. HoxB13 expression in ductal type adenocarcinoma of prostate: clinicopathologic characteristics and its utility as potential diagnostic marker. Sci Rep 2019; 9: 20205.
[34] Gerhardt J, Montani M, Wild P, Beer M, Huber F, Hermanns T, Müntener M and Kristiansen G. FOXA1 promotes tumor progression in prostate cancer and represents a novel hallmark of castration-resistant prostate cancer. Am J Pathol 2012; 180: 848-861.
[35] Grabowska MM, Elliott AD, DeGraff DJ, Anderson PD, Anumanthan G, Yamashita H, Sun Q, Friedman DB, Hachey DL, Yu X, Sheehan JH, Ahn JM, Raj GV, Piston DW, Gronostajski RM and Matusik RJ. NFI transcription factors interact with FOXA1 to regulate prostate-specific gene expression. Mol Endocrinol 2014; 28: 949-964.
[36] Hsu CC and Hu CD. Transcriptional activity of c-Jun is critical for the suppression of AR function. Mol Cell Endocrinol 2013; 372: 12-22.
[37] Nikitina AS, Sharova EI, Danilenko SA, Butusova TB, Vasiliev AO, Govorov AV, Prilepskaya EA, Pushkar DY and Kostryukova ES. Novel RNA biomarkers of prostate cancer revealed by RNA-seq analysis of formalin-fixed samples obtained from Russian patients. Oncotarget 2017; 8: 32990-33001.
[38] Mariani O, Brennetot C, Coindre JM, Gruel N, Ganem C, Delattre O, Stern MH and Aurias A. JUN oncogene amplification and overexpression block adipocytic differentiation in highly aggressive sarcomas. Cancer Cell 2007; 11: 361-374.
[39] Luo J, Xia Y, Yin Y, Luo J, Liu M, Zhang H, Zhang C, Zhao Y, Yang L and Kong L. ATF4 destabilizes RET through nonclassical GRP78 inhibition to enhance chemosensitivity to bortezomib in human osteosarcoma. Theranostics 2019; 9: 6334-6353.
[40] Lai R, Navid F, Rodriguez-Galindo C, Liu T, Fuller CE, Ganti R, Dien J, Dalton J, Billups C and Khoury JD. STAT3 is activated in a subset of the Ewing sarcoma family of tumours. J Pathol 2006; 208: 624-632.
[41] Joko R, Yamada D, Nakamura M, Yoshida A, Takihira S, Takao T, Lu M, Sato K, Ito T, Kunisada T, Nakata E, Ozaki T and Takarada T. PRRX1 promotes malignant properties in human osteosarcoma. Transl Oncol 2021; 14: 100960.