frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 30 November 2021
doi: 10.3389/fcell.2021.781285

Identification of Pan-Cancer Biomarkers Based on the Gene Expression Profiles of Cancer Cell Lines

ShiJian Ding[1†], Hao Li[2†], Yu-Hang Zhang[3†], XianChao Zhou[4], KaiYan Feng[5], ZhanDong Li[2], Lei Chen[6], Tao Huang[7,8]* and Yu-Dong Cai[1]**

[1]School of Life Sciences, Shanghai University, Shanghai, China, [2]College of Food Engineering, Jilin Engineering Normal University, Changchun, China, [3]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, United States, [4]Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai, China, [5]Department of Computer Science, Guangdong AIB Polytechnic College, Guangzhou, China, [6]College of Information Engineering, Shanghai Maritime University, Shanghai, China, [7]CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China, [8]CAS Key Laboratory of Tissue Microenvironment and Tumor, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

OPEN ACCESS

Edited by:
Lu Xie,
Shanghai Center For Bioinformation Technology, China

Reviewed by:
Bo Zhou,
Shanghai University of Medicine and Health Sciences, China
Qi Dai,
Zhejiang Sci-Tech University, China

***Correspondence:**
Lei Chen
chen_lei1@163.com
Tao Huang
tohuangtao@126.com
Yu-Dong Cai
cai_yud@126.com

[†]These authors contributed equally to this work

Specialty section:
This article was submitted to Molecular and Cellular Pathology, a section of the journal Frontiers in Cell and Developmental Biology

Received: 22 September 2021
Accepted: 16 November 2021
Published: 30 November 2021

Citation:
Ding S, Li H, Zhang Y-H, Zhou X, Feng K, Li Z, Chen L, Huang T and Cai Y-D (2021) Identification of Pan-Cancer Biomarkers Based on the Gene Expression Profiles of Cancer Cell Lines. Front. Cell Dev. Biol. 9:781285. doi: 10.3389/fcell.2021.781285

There are many types of cancers. Although they share some hallmarks, such as proliferation and metastasis, they are still very different from many perspectives. They grow on different organ or tissues. Does each cancer have a unique gene expression pattern that makes it different from other cancer types? After the Cancer Genome Atlas (TCGA) project, there are more and more pan-cancer studies. Researchers want to get robust gene expression signature from pan-cancer patients. But there is large variance in cancer patients due to heterogeneity. To get robust results, the sample size will be too large to recruit. In this study, we tried another approach to get robust pan-cancer biomarkers by using the cell line data to reduce the variance. We applied several advanced computational methods to analyze the Cancer Cell Line Encyclopedia (CCLE) gene expression profiles which included 988 cell lines from 20 cancer types. Two feature selection methods, including Boruta, and max-relevance and min-redundancy methods, were applied to the cell line gene expression data one by one, generating a feature list. Such list was fed into incremental feature selection method, incorporating one classification algorithm, to extract biomarkers, construct optimal classifiers and decision rules. The optimal classifiers provided good performance, which can be useful tools to identify cell lines from different cancer types, whereas the biomarkers (e.g. NCKAP1, TNFRSF12A, LAMB2, FKBP9, PFN2, TOM1L1) and rules identified in this work may provide a meaningful and precise reference for differentiating multiple types of cancer and contribute to the personalized treatment of tumors.

Keywords: pan-cancer study, feature selection, classification algorithm, decision rule, biomarker

INTRODUCTION

"Cancer" is the term used to describe a series of diseases that is characterized by the spontaneous expansion and spread of somatic cell clones. It is becoming a serious public health problem worldwide. In 2020 alone, over 19.29 million new cases of cancer were diagnosed, and more than 9.58 million people died from cancer (World Health Organization, 2019). The hallmarks of cancer have been extensively described as six biological capabilities, namely, enhanced proliferative signaling, growth suppressor escape, cell death resistance, replicative immortality, angiogenesis induction, and invasion and metastasis activation (Hanahan and Weinberg, 2011). In other words, the pro-oncogenic function is the abnormal expression of various genes based on these six biological capabilities. Therefore, cancer genomic data, particularly gene expression signatures, can provide insight into the occurrence and development of cancers and, importantly, can be used to develop targeted therapies for cancers (Garman et al., 2007).

Although many cancers share the hallmarks of cancer, they are still very different. They grow on different organs and tissue. Pan-cancer studies provide an opportunity to understand the commonalities, heterogeneity, and emergent themes of multiple tumors (Andor et al., 2016). Increased numbers of tumor sample datasets provide scientists with a clear picture of tumors, rare driver events in heterogeneous tumor samples, and new molecular carcinogenic mechanisms that may be readily detected (Weinstein et al., 2013). For example, a study on the genomic predictors of the drug sensitivity of 947 human cancer cell lines based on a cancer cell line encyclopedia revealed known and novel response candidate biomarkers, which may contribute to cancer biology and therapeutic development (Barretina et al., 2012). Another study on long noncoding RNA (lncRNA) in 5185 TCGA tumors demonstrated that although tumor-specific dysregulated lncRNAs are commonly observed in a variety of tumors, genes and pathways could be synergistically regulated in different cancers by the same group of lncRNAs; this information may provide useful ideas for the development of broad-spectrum antineoplastic drugs (Chiu et al., 2018).

The sample size of TCGA based pan-cancer studies is already very large as a multi-omics data source. But it is still not enough to get robust pan-cancer biomarkers if we consider the large variances among cancer patients across cancer types and even within the same cancer. Tumor heterogeneity can be broadly categorized into intratumor heterogeneity and inter tumor heterogeneity (Burrell and Swanton, 2014). Inter tumor heterogeneity refers to the heterogeneity between patients with the same histological tumor type and has been considered to be caused by patient-specific factors, including germline mutations, individualized somatic mutations, and environmental factors. Intratumor heterogeneity can be divided into spatial heterogeneity (different regions of the tumor have different genetic aberrations) and temporal heterogeneity (during disease progression) (Dagogo-Jack and Shaw, 2018). Studies across multiple cancers have suggested that intratumor heterogeneity promotes tumor growth, metastasis, and drug resistance in human cancers (Hyo-eun et al., 2015; Russo et al., 2016). Therefore, treatment strategies with increased effectiveness and durability still need to be developed on the basis of a comprehensive understanding of tumor dynamics.

To get robust pan-biomarkers, there are two approaches: increase the sample size or reduce the variance. TCGA and the following works tried the first approach of increasing sample size. In this study, we would like to try the second approach of reducing the variance by analyzing the cancer cell line data from Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al., 2019). The important genes were extracted by using the Boruta method (Kursa and Rudnicki, 2010). These genes were further analyzed with the max-relevance and min-redundancy (mRMR) method to evaluate their importance and sort them in a feature list. This list was fed into the incremental feature selection (IFS) method (Liu and Setiono, 1998) that combined support vector machine (SVM) (Cortes and Vapnik, 1995) or decision tree (DT) (Safavian and Landgrebe, 1991) to identify important genes and decision rules and build powerful classifiers. Further analysis was performed through a literature review of the top-ranked genes and portion decision rules to confirm the validity and reliability of the results. This study gives new insight into pan-cancer studies and may provide novel targets of tumor-specific therapies.

MATERIALS AND METHODS

Datasets

Xiao et al. (Xiao et al., 2019) downloaded the raw RNA-Seq data from Cancer Cell Line Encyclopedia (CCLE) (Ghandi et al., 2019) and quantified the gene expression levels as Transcripts Per kilobase Million (TPM) using RSEM (Li and Dewey, 2011). We used the processed gene expression data by Xiao et al. (Xiao et al., 2019). The cancer types with sample sizes of less than 10 was removed. Finally, there were 988 cancer cell lines from 20 cancer types. The sample size of each tumor is listed in **Table 1**. For each sample, 57,820 gene features were included. We investigated the expression patterns of genes in different tumor types and whether these tumor types could be distinguished on the basis of expression profiles.

Boruta Feature Selection

The CCLE data involved a large number of genes (features). Obviously, not all genes are associated with the investigated tumor types. Therefore, filtering the important genes is necessary. Here, we applied the Boruta (Kursa and Rudnicki, 2010) method to select a set of relevant features with multiple tumor labels.

The Boruta method is a wrapping algorithm that is based on random forest (RF) and involves the following steps: 1) the new shuffled data are generated by copying the original dataset and shuffling original features; 2) a RF classifier that can output the importance score of each feature is trained by using the new feature matrix as the input; and 3) the features in the original features that are sincerely relevant to the labels are retained, and the shuffled data are removed. Boruta finally selects the relevant features after several iterations of the above three steps.

TABLE 1 | Distribution of samples and decision rules in different cancer cell lines.

Cancer cell line types	Number of cell lines	Number of decision rules	Number of criteria	Number of involved genes
Autonomic ganglia	16	4	42	17
Bone	20	4	46	19
Breast	51	23	325	63
Central nervous system	65	18	219	57
Endometrium	28	16	191	52
Fibroblast	37	3	28	15
Haematopoietic and lymphoid tissue	173	8	96	34
Kidney	32	7	88	38
Large intestine	56	9	123	47
Liver	25	9	115	42
Lung	188	51	740	80
Oesophagus	27	8	145	47
Ovary	47	25	306	67
Pancreas	41	14	208	56
Skin	49	7	83	34
Soft tissue	28	9	126	41
Stomach	37	26	361	72
Thyroid	12	7	79	38
Upper aerodigestive tract	31	11	162	47
Urinary tract	25	16	216	59

The Boruta program that we used in this research was downloaded from https://github.com/scikit-learn-contrib/boruta_py and was set to default parameters for execution.

mRMR

After feature filtering by using the Boruta method, the mRMR (Peng et al., 2005) feature selection method was used to evaluate the importance of the remaining features. This approach has been widely used to analyze complicated systems.

The mRMR method evaluates the importance of target features by using max-relevance and min-redundancy. Features with great relevance to the category labels and low redundancy with other features are considered to be influential. It uses mutual information (MI) to measure relevance and redundancy. The score of the MI between two variables X and Y is calculated as

$$I(X,Y) = \int\int p(x,y) log \frac{p(x,y)}{p(x)p(y)} dxdy, \quad (1)$$

where $p(x,y)$ represents the joint probability distribution function of X and Y, and $p(x)$ and $p(y)$ represent the marginal probability distribution function of X and Y, respectively. The mRMR constructs order feature lists on the basis of the importance of features. Specifically, the program loops several times, and each loop selects a feature that has the greatest correlation with the target variable and the least correlation with the selected features. Finally, a list of sorting features is obtained in accordance with the selected orders.

The mRMR program used in this research was obtained from http://penglab.janelia.org/proj/mRMR/and executed with default parameters.

Incremental Feature Selection

Even though the mRMR method produces a ranked list of features on the basis of the importance of features, we still cannot determine the influential features. The IFS (Liu and Setiono, 1998) method can determine the optimal number of key features in combination with one classification algorithm. First, IFS generates a series of feature subsets from the above feature list in accordance with the step size. For example, if the step size is 10, the first feature subset will be the top 10 features, and the second subset will be the top 20 features. Next, one classifier is built based on the training set, where the samples are represented by features from each feature subset. The classifiers are evaluated by 10-fold cross-validation (Kohavi, 1995) to obtain evaluation metrics. Finally, the optimal feature subset and the best classifier are determined, and features in this subset are called optimum features.

Decision Tree

DT (Safavian and Landgrebe, 1991) is a model that presents decision rules and classification results in a tree-like structure and is widely used in the biological and biomedical fields. DT is a supervised learning approach that builds a model based on the IF–THEN format. It achieves superior model performance through low computational complexity. The common decision trees are Iterative Dichotomizer 3, C4.5, and Classification and Regression Tree. They use different partition strategies when building a prediction model. In this study, we used the Scikit-learn (Pedregosa et al., 2011) module in Python to construct a DT classifier.

Support Vector Machine

SVM (Cortes and Vapnik, 1995; Zhou J.-P. et al., 2020; Wang et al., 2021) is a supervised learning algorithm in statistical

learning methods that is commonly used in classification and regression problems. SVM maps the data from a low-dimensional space to a high-dimensional space by using a kernel function. Then, a hyperplane with the maximum interval existing in the high-dimensional space makes two classes of samples linearly separable.

In this study, 20 tumor types needed to be classified. This task was a multiclass classification problem. Therefore, we applied the one-versus-rest strategy to train a multiclass SVM, which was split into numerous binary SVMs. For each binary SVM, samples of one class were regarded as positive examples, and samples of all other classes were used as negative examples. We directly used the tool "SMO" in Weka software (Gewehr et al., 2007) in this study. The sequential minimum optimization algorithm (Platt, 1998; Keerthi et al., 2001) was utilized to optimize the training procedure. The kernel function was set as a polynomial function.

Synthetic Minority Oversampling Technique

As can be seen from **Table 1**, the sample sizes for all tumor types were quite different. For example, types "lung" and "hematopoietic and lymphoid tissue" contained 188 and 173 samples, respectively, whereas types "autonomic ganglion" and "bone" had only 16 and 20 samples, respectively. These results indicated that the whole dataset of this study was unbalanced. Accordingly, we adopted the synthetic minority oversampling technique (SMOTE) (Chawla et al., 2002) to balance the dataset when building classifiers. This method uses the k-nearest neighbor algorithm to expand the sample sizes of each minority class. SMOTE first selects random data from one minority class and then finds the k-nearest neighbors in this class. Next, the new sample data are synthesized between the random data and the randomly generated k-nearest neighbor. After SMOTE processing, the sample size of each minority class is equal to that of the majority class. In other words, the sample sizes of the 20 tumor types in this study were equal. In this study, we oversampled data by using the tool "SMOTE" in Weka software (Gewehr et al., 2007).

Performance Measurement

In this study, several multiclass classifiers were used to distinguish samples from 20 tumor types. We adopted 10-fold cross-validation (Kohavi, 1995; Chen et al., 2017; Zhao et al., 2018; Zhou JP. et al., 2020; Jia et al., 2020; Liang et al., 2020; Zhang et al., 2021c; Yang and Chen, 2021; Zhu et al., 2021) to evaluate the performance of each multiclass classifier. We correlation coefficient (MCC) (Matthews, 1975; Gorodkin, 2004; Liu H. et al., 2021; Zhang et al., 2021a; Zhang et al., 2021b; Pan et al., 2021) to measure and evaluate the prediction quality of the results of 10-fold cross-validation. Let X be a matrix representing predicted labels yielded by one classifier and Y be another matrix indicating the actual labels of samples. The calculation formula of MCC is as follows:

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

where $cov(X,Y)$ denotes the correlation coefficient of X and Y, and $\bar{x}_j$ and $\bar{y}_j$ are the average values in the j th column of X and Y, respectively. In addition, C denotes the number of tumor types, and n denotes the total number of samples.

In addition to MCC, we also calculated accuracy on each cancer type and overall accuracy. The accuracy on the ith cancer type was computed by

$$Accuracy_i = \frac{n_i}{N_i} \qquad i = 1, 2, \cdots, 20, \quad (3)$$

where N_i represents the number of samples in the ith cancer type and n_i denotes correctly predicted samples in the ith cancer type. The overall accuracy was calculated by

$$\text{Overall accuracy} = \frac{\sum_{i=1}^{20} n_i}{\sum_{i=1}^{20} N_i} \qquad i = 1, 2, \cdots, 20, \quad (4)$$

MCC was set as the key measurement and others were also provided for reference in this study.

RESULTS

In this study, several computational methods were used to analyze the CCLE dataset of 20 tumor types. The analysis process is shown in **Figure 1**.

Results of the Boruta and max-Relevance and Min-Redundancy Methods

All features were first analyzed by using the Boruta method. A total of 54,634 features were removed, and 3,186 features were retained. These retained features are provided in **Supplementary Table S1**. These 3,186 features were further analyzed by using the mRMR method, and a feature ranking list was generated on the basis of their importance. This list can also be found in **Supplementary Table S1**.

Results of the Incremental Feature Selection Method

The feature list produced by the mRMR method was fed into the IFS method. A series of feature subsets were generated by setting the step size to 10. The DT and SVM were used to build classifiers on each feature subset. Then, all classifiers were evaluated by 10-fold cross-validation to obtain evaluation metrics, such as accuracy on each cancer type, overall accuracy and MCC. The above measurements that were acquired by the two classification algorithms for each subset of features are shown in **Supplementary Table S2**. We plotted the IFS curve to visualize the results. For SVM, the MCC was set as the Y-axis,

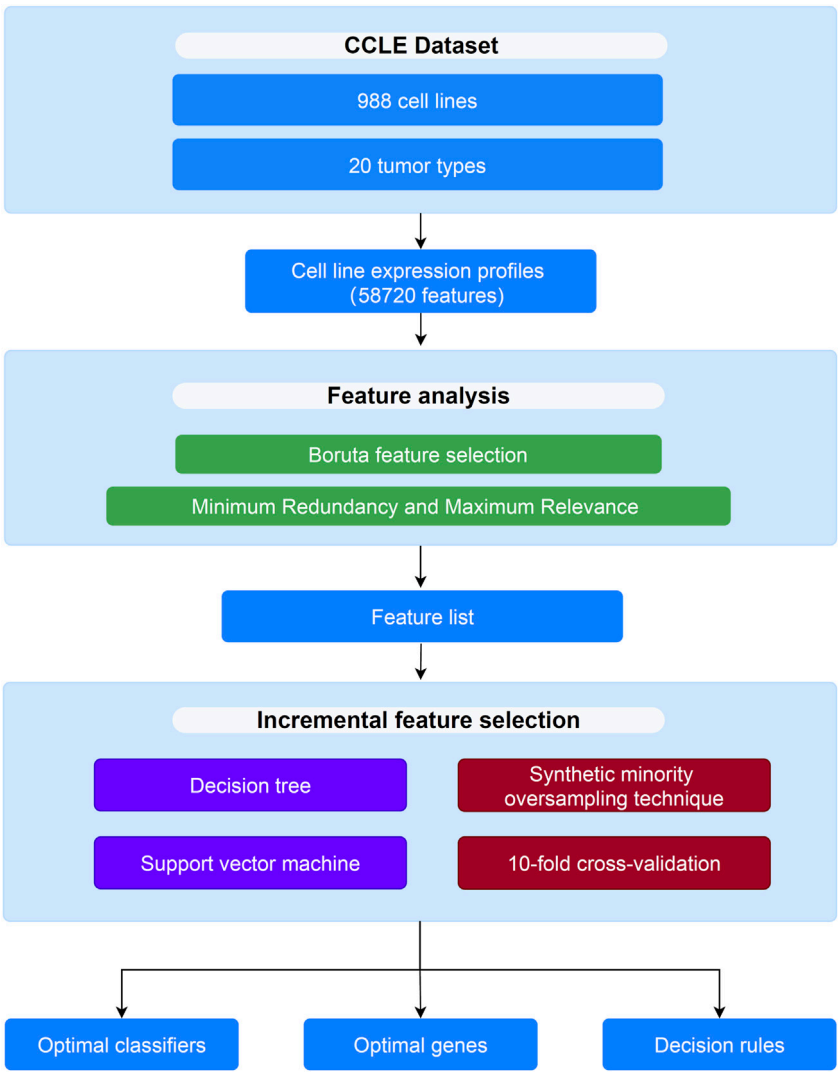

FIGURE 1 | Flow chart to show the entire analysis procedures. The CCLE dataset which includes 988 cell lines and 20 tumor types is analyzed by Boruta and mRMR methods, resulting in a feature list. The list is then fed into the incremental feature selection method to extract optimal genes, build the optimal classifiers and construct decision rules.

and the number of features was set as the X-axis. As shown in **Figure 2**, when the number of features reached 3,130, the highest value of MCC was 0.976. The corresponding overall accuracy was 0.978 (**Table 2**). Accordingly, the best SVM classifier can be built based on these top 3,130 features. Although this classifier provided the highest performance, its efficiency was not very high because an excessively high number of features were used. The IFS results of SVM were carefully checked (**Supplementary Table S2** and **Figure 2**). When the top 400 features were adopted, the MCC reached 0.951, which was only slightly lower than the highest MCC. The overall accuracy was 0.954 (**Table 2**). It was also a little lower than that of the best SVM classifier. It can be concluded that these two SVM classifiers provided almost equal performance. To further confirm this fact, we also investigated the accuracies on 20 cancer types yielded by these two classifiers. A radar graph was plotted, as shown in **Figure 3**. Clearly, the areas inside the curves of two classifiers were almost same, suggesting the equal performance of these two classifiers. However, the number of features was considerably lower. The SVM classifier with these features had drastically higher efficiency than the best SVM classifier. Thus, this classifier could be the proposed classifier for assigning samples to the correct cancer type.

In addition to the above SVM algorithm, we used the DT, which is a white-box classification algorithm. In this process, the step size of IFS with the DT was also set to 10, and only the top 400 features in the mRMR list were considered. The IFS results are also available in **Supplementary Table S2**, and the IFS curve is presented in **Figure 4**. The best DT classifier yielded an MCC value of 0.754, which was based on the top 390 features. The

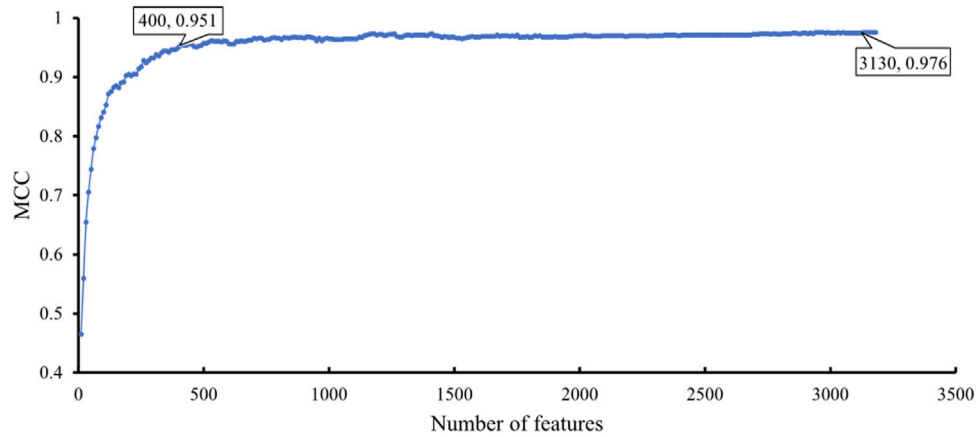

FIGURE 2 | IFS curve with SVM classification algorithm on the different number of features. The SVM provides the highest MCC of 0.976 when the top 3,130 features are adopted. When top 400 features are adopted, SVM provides good performance with MCC of 0.951.

TABLE 2 | Performance of some key classifiers.

Classification algorithm	Number of features	Overall accuracy	MCC
Support vector machine	3,130	0.978	0.976
	400	0.954	0.951
Decision tree	390	0.771	0.754
	100	0.757	0.739

overall accuracy of this classifier was 0.771, as listed in **Table 2**. Likewise, we also wanted to obtain an accepted classifier that used few features and provided high performance. As can be seen from **Figure 4**, the MCC reached 0.739 when the top 100 features were used. The overall accuracy was 0.757 (**Table 2**). They were only a little lower than those of the best DT classifier. Furthermore, the accuracies on 20 cancer types of these two classifiers were also investigated, as illustrated in **Figure 3**. Evidently, these two DT classifiers were almost at the same level. Therefore, these top 100 features were considered to build the DT classifier.

Classification Rules

As mentioned above, the DT classifier with the top 100 features exhibited high performance. Thus, we constructed a DT with these features and all samples. Consequently, we obtained 275 rules, which are presented in **Supplementary Table S3**. The number of decision rules and criteria used for 20 tumor types are shown in **Table 1**. Each cancer type was assigned some decision rules. The cancer type "Lung" was assigned most decision rules, whereas 'Fibroblast' received least rules. The further analysis of these rules can be found in *Analysis of Decision Rules*.

GO and KEGG Enrichment Analysis

As mentioned in *Results of the Incremental Feature Selection Method*, the SVM classifiers with top 400 features gave a little lower performance than the best SVM classifier. However, it had much higher efficiency because much less features were used in this classifier. Thus, these 400 features may be highly related to distinguish different cancer types. Thus, we conducted GO and KEGG enrichment analysis on these features (genes). The results can be found in **Supplementary Table S4**. Some top GO terms and KEGG pathways are illustrated in **Figures 5**, **6**. In *Analysis of Essential Genes*, the discussion on the enrichment analysis results would be given.

DISCUSSION

In this study, we used the Boruta and mRMR methods to analyze features and applied the IFS method combined with SVM and DT to construct classifiers and decision rules. Some essential features (genes) (see **Supplementary Table S1**) were extracted. Furthermore, we obtained several decision rules. In this section, we provide an extensive analysis of these essential genes and decision rules.

Analysis of Essential Genes

Firstly, we performed GO/KEGG enrichment analysis to find whether our 400 selected features were significantly enriched in specific terms. Results were described in *GO and KEGG Enrichment Analysis*, top GO terms and KEGG pathways are illustrated in **Figures 5**, **6**.

We found that the significantly enriched GO terms mainly involve actin organization, cell matrix components and cell polarization. Actin assembly is very important for cell migration, and abnormal regulation of cell migration can drive cancer invasion and metastasis (Yamaguchi and Condeelis, 2007). Different cancers and different differentiation states of cancers often show different patterns of cell migration and the migration of these cancer cells is regulated by various signals. Although it is difficult to use a single strategy to regulate the motility of all cancer cells, inhibiting actin polymerization can inhibit migration of most types of cancer cells (Yamazaki et al.,

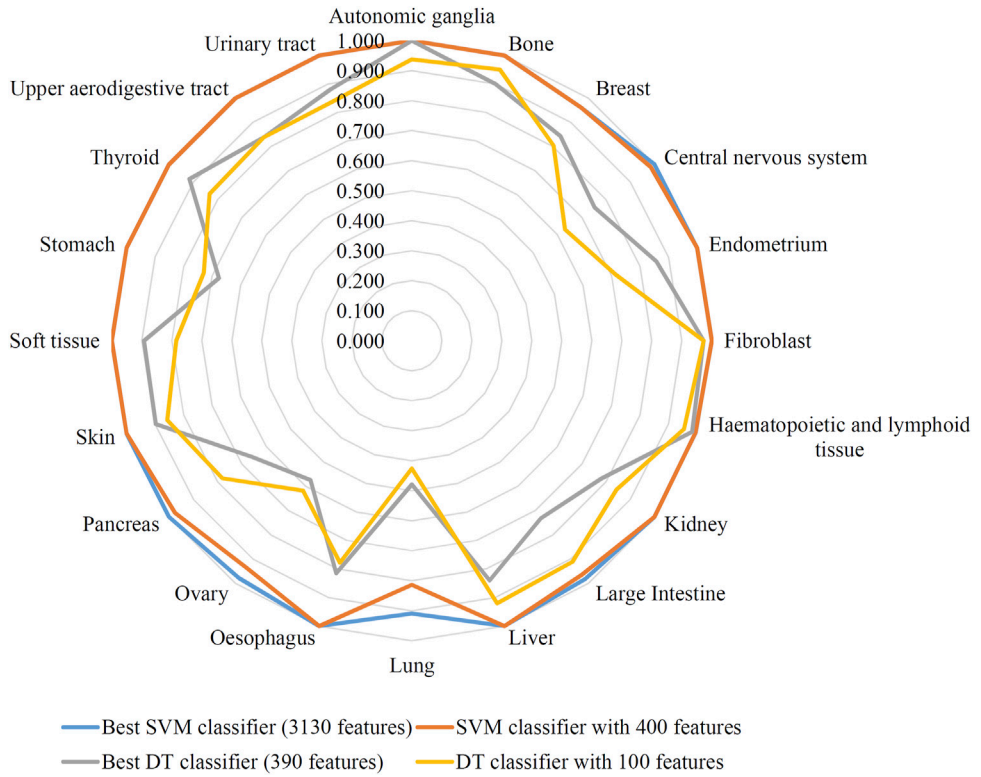

FIGURE 3 | Radar graph to show the performance of two support vector machine (SVM) classifiers and two decision tree (DT) classifiers on 20 cancer types. Two SVM classifiers provide almost equal performance, also for two DT classifiers.

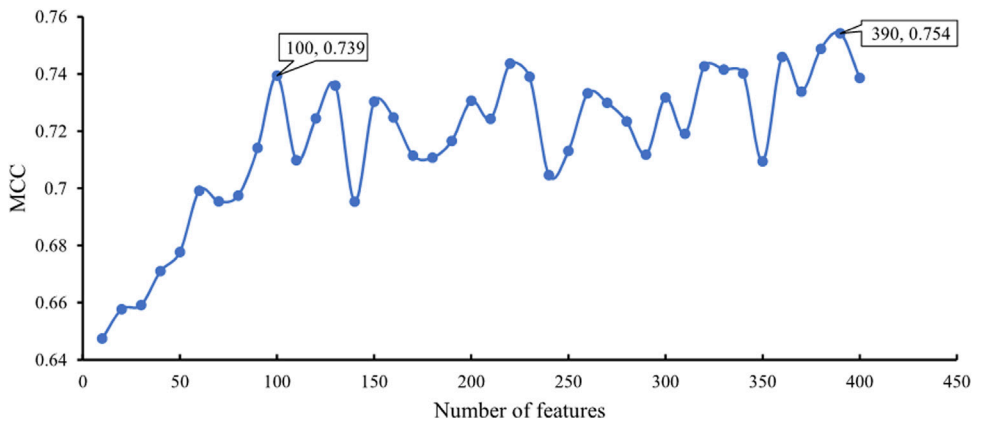

FIGURE 4 | IFS curve with DT classification algorithm on the different number of features. The DT provides the highest MCC of 0.754 when the top 390 features are adopted. DT yields high performance with MCC of 0.739 when only 100 features are used.

2005). The loss of cell polarity has been shown to be related to tumor progression (Wodarz and Näthke, 2007). Generally, aggressive tumors lack polarity, and study have shown that different cancers have different abnormal expression or localization of polar proteins, which may also serve as the basis for our classifier (Ellenbroek et al., 2012). The KEGG results also showed similar results, which are mainly related to migration and actin cytoskeleton. This reflects both the importance of cell migration ability to tumors and the difference in invasion of different tumors.

Secondly, among the 400 selected features (genes), the top-ranked genes were usually highly decisive for distinguishing different cell lines. Therefore, some of them were selected for analysis, which are listed in **Table 3**.

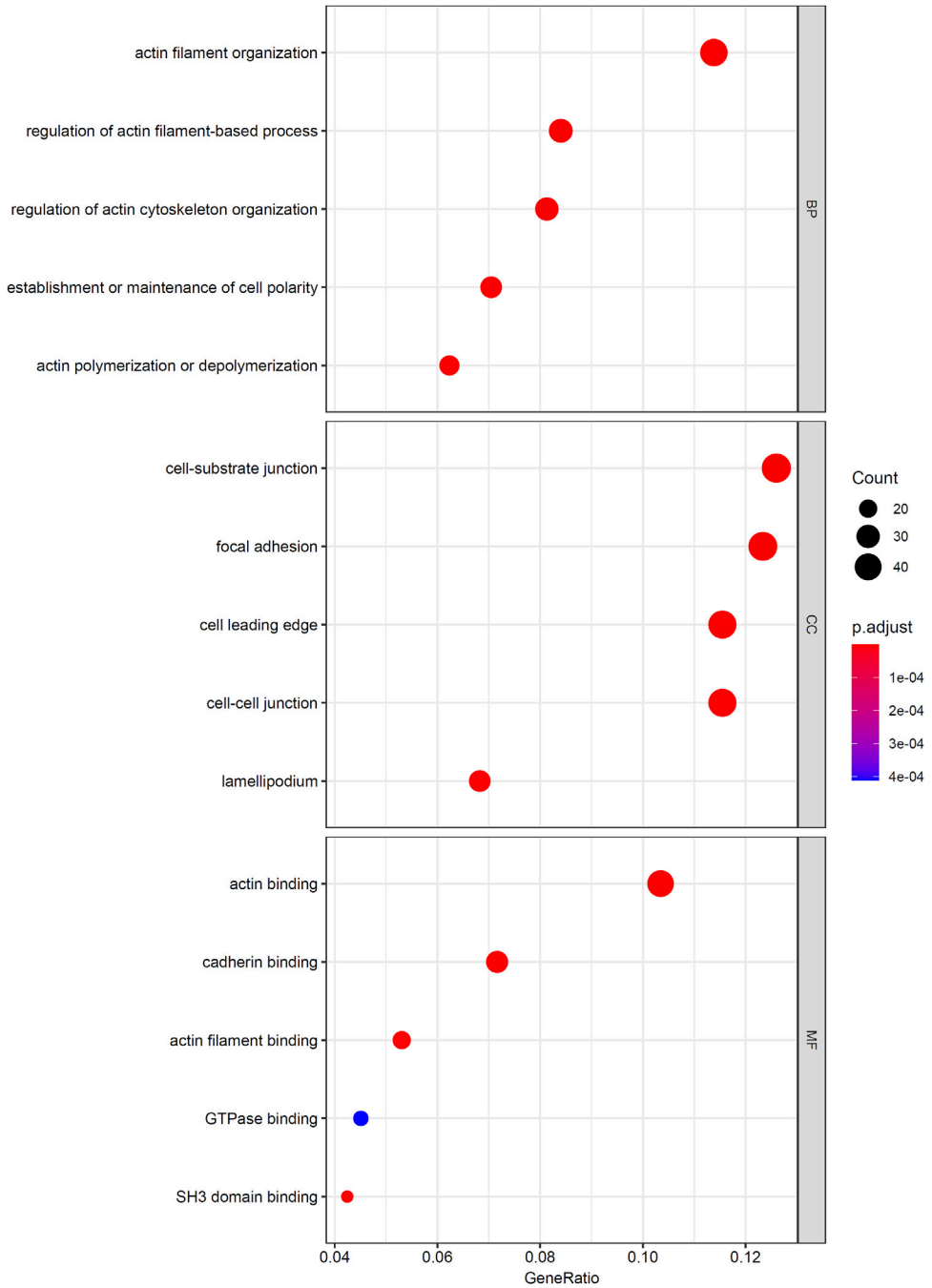

FIGURE 5 | Top enriched GO terms to the top 400 genes in the mRMR feature list.

The highest-ranking feature is NCKAP1 (ENSG00000061676). It encodes the NCK-associated protein 1 as a part of the WAVE (WASF) complex that regulates lamellipodia formation. Past studies have revealed that NCKAP1 is associated with multiple types of human cancer. A previous study showed that the WASF3 gene is a promoter of cell invasion in breast cancer and the Nckap1 can keep WASF3 in an inactive conformation through binding to the WASF homology domain at the N-terminus. The activation of WASF3 depends on the combination with RAC1 which can be prevented by the absence of NCKAP1. Thus, the downregulation of NCKAP1 inhibits the activity of WASF3 and may suppresses metastasis in breast cancer cells. In addition, univariate survival analysis have found that high expression level of NCKAP1 is correlated with short overall survival (Teng et al., 2016). The function of NCKAP1 in liver cancer has recently been clarified. Specifically,

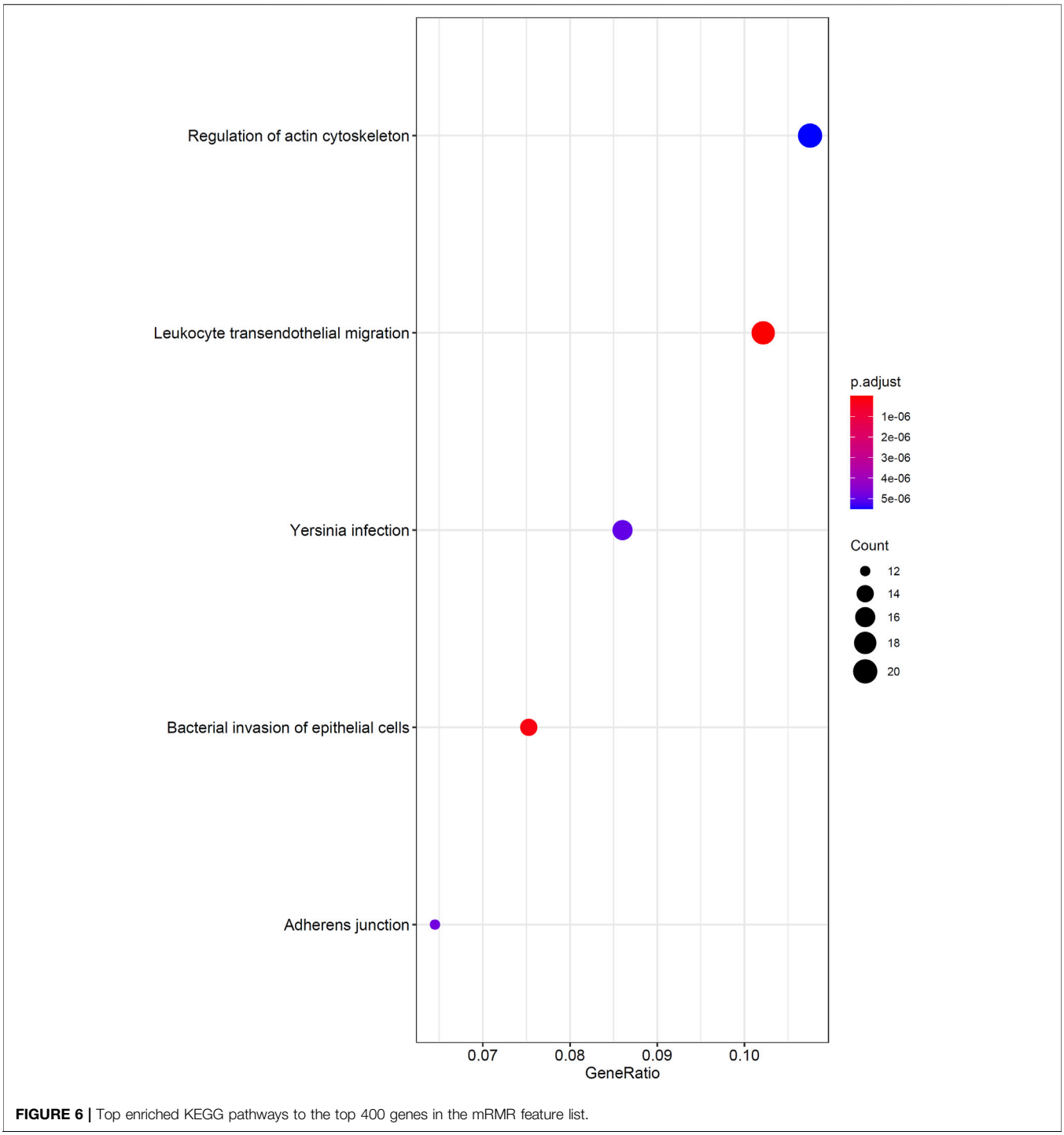

FIGURE 6 | Top enriched KEGG pathways to the top 400 genes in the mRMR feature list.

NCKAP1 can control tumor growth and improve prognosis by enhancing Rb1/p53 activation in hepatocellular carcinoma (HCC) (Zhong et al., 2019). Similarly, a recent study discovered that NCKAP1 is highly expressed in primary non-small-cell lung cancer (NSCLC) and is significantly associated with histologic tumor grade, metastasis, and poor survival rate. It is also related to the HSP90-mediated invasion and metastasis of NSCLC by stimulating MMP9 activation and the epithelial–mesenchymal transition (EMT) (Xiong et al., 2019). In conclusion, NCKAP1 is aberrantly expressed in a variety of cancer types and could be a biomarker and potential therapeutic target.

TNFRSF12A (ENSG00000006327) encodes the receptor of TNFSF12/TWEAK, which is also known as fibroblast growth factor-inducible molecule 14. It can promote endothelial cell proliferation and angiogenesis. Studies have demonstrated that

TABLE 3 | Information of essential genes.

Ensembl ID	Gene symbol	Description
ENSG00000061676	NCKAP1	NCK Associated Protein 1
ENSG00000006327	TNFRSF12A	TNF Receptor Superfamily Member 12A
ENSG00000172037	LAMB2	Laminin Subunit Beta 2
ENSG00000122642	FKBP9	FKBP Prolyl Isomerase 9
ENSG00000070087	PFN2	Profilin 2
ENSG00000141198	TOM1L1	Target Of Myb1 Like 1 Membrane Trafficking Protein

TNFRSF12A is highly expressed in breast cancer, and a high TNFRSF12A level associated with matrix metalloproteinase (MMP)-9 overexpression is related to cancer progression; thus, TNFRSF12A-targeting therapy could improve survival rates in cancer (Yang et al., 2018). Furthermore, through modulating the expression of MMP-9, the overexpression of TNFRSF12A can promote prostate cancer progression and result in poor treatment outcomes (Huang et al., 2011). TNFRSF12A is also highly expressed in human HCC, and *in vivo* experiments have revealed that TNFRS12A knockdown can inhibit cancer cell proliferation and migration (Wang et al., 2017). TNFRSF12A has also been demonstrated to be highly expressed in NSCLC and contribute to NSCLC cell migration and invasion *in vitro* (Whitsett et al., 2012). Other studies have also confirmed that TNFRSF12A is overexpressed in melanomas, gliomas, and esophageal and pancreatic cancers (Han et al., 2005; Tran et al., 2006; Watts et al., 2007; Zhou et al., 2013). Interestingly, in certain tumor types, TNFRSF12A exhibits a low expression level. A study on TCGA data suggested that the downregulation of TNFRSF12A in thyroid cancer could be a potential molecular biomarker for the prediction of poor prognosis (Wu et al., 2020). Therefore, TNFRSF12A has different expression patterns in different cancers and could be a remarkable feature for distinguishing different cancer cell lines. In addition, it could also be a critical therapeutic target, and preclinical studies have shown that the use of inhibitors in cancer with high TNFRSF12A expression has certain effects (Wajant, 2013).

LAMB2 (ENSG00000172037) encodes a subunit of laminins, which are one of the major glycoproteins present in the basement membrane of the extracellular matrix and are related to tumor angiogenesis, invasion, and metastasis. A previous study revealed that the downregulation of LAMB2 caused by HE4 gene interference results in the invasion and metastasis of ovarian cancer cells (Zhuang et al., 2014). Studies on pancreatic cancer have demonstrated that the lack of basement membrane continuity, which is determined by limited laminin expression, is associated with poor postoperative outcomes. In other words, in pancreatic cancer, the downregulation of LAMB2 is correlated with poor prognosis (Van Der Zee et al., 2012).

FKBP9 (ENSG00000122642), which encodes FKBP prolyl isomerase 9, is known to be associated with chaperonin-mediated protein folding and protein metabolism. A recent study has found that FKBP9 could be an independent prognostic marker for predicting the poor prognosis of patients with prostate cancer; that high FKBP9 levels and short biochemical-recurrence-free survival are significantly correlated (p = 0.041); and that FKBP9 may be a cancer promoter that enhances prostate cancer progression (Jiang et al., 2020). Another study found that FKBP9 is a critical factor for promoting the malignant behaviors of glioblastoma cells; high FKBP9 level is related to poor prognosis and could confer malignant cells with the capability to resist endoplasmic reticulum stress inducers (Xu et al., 2020). Other studies have also confirmed that FKBP9 is connected with other cancers, such as colorectal and breast cancers (Bianchini et al., 2006; Chang et al., 2020). Thus, FKBP9 may be an effective feature of many cancer cell lines.

PFN2 (ENSG00000070087) encodes an actin monomer-binding protein. It participates in regulating actin aggregation in response to extracellular signals and cell motility. Recently, PFN2 has emerged as a key regulator of cancer development and progression. PFN2 has been reported to be highly expressed in triple-negative breast cancer (TNBC); it could promote the proliferation, migration, and invasion of TNBC cells and may be partially responsible for the worsened survival associated with high PFN2 levels (Ling et al., 2021). In esophageal squamous cell carcinoma, a high PFN2 level is related to short overall survival. Moreover, PFN2 expression is positively associated with tumor invasion depth and lymph node metastasis (Cui et al., 2016). Another study demonstrated that PFN2 is highly expressed in head and neck squamous cell cancer (HNSCC) tissues and cell lines and that the activation of the PI3K/Akt/β-catenin signaling pathway by PFN2 results in the proliferation and metastasis promotion of HNSCC, whereas PFN2 knockdown produces the opposite effects (Zhou et al., 2019). However, another study suggested a different result: the degree of tumor metastasis is negatively associated with PFN2 expression level likely because of the enhancement in EMT induced by low PFN2 levels considering that enhanced EMT may increase migratory capabilities (Zhang H. et al., 2018). Other studies have also found that PFN2 has different expression patterns and effects in NSCLC, small cell lung cancer, and gastric cancer (Hippo et al., 2002; Yan et al., 2017; Cao et al., 2020). In conclusion, PFN2 plays an important role in a variety of cancers and could be an important biomarker for different cancer cells, as well as an attractive therapeutic target.

TOM1L1 (ENSG00000141198) encodes the target of myb1-like 1 membrane trafficking protein. ERBB2-induced breast cancer cell invasion has been documented to be caused by the TOM1L1-derived membrane delivery of MT1-MMP, and ERBB2 and TOM1L1 are frequently coamplified in the breast (Chevalier

et al., 2016). Other studies have also found that TOM1L1 is related to colorectal cancer and is highly expressed in bladder cancer (Emaduddin et al., 2008; Zhang Y. et al., 2018).

As analyzed above, the selected genes from our results showed strong expression differences in multiple cancer cells. These genes could be good therapeutic targets. By the same token, distinct gene expression patterns could also be remarkably decisive features for different cancer cell lines.

Analysis of Decision Rules

Previously, we constructed 275 decision rules on the basis of the top 100 selected features and all cell lines. Each rule contained several criteria. The numbers of rules and criteria for each cancer type are listed in **Table 1**. In addition, the number of genes involving rules for each cancer type is also listed in this table. In the following, we provide our interpretation and experimental evidence for some rules based on published literature. These evidences indicate the effects of the high/low expression of key genes on tumors which also found to have similar expression patterns in the decision rules of the corresponding tumor cell line (relatively high/low expression level compared to other cell lines).

The 23 rules for identifying breast cancer cell lines included 325 criteria, which involved 62 genes. These genes have considerable experimental support, and here we show some evidence. LDHB (ENSG00000111716) encodes the B subunit of lactate dehydrogenase enzyme, which participates in glycolysis. A study found that LDHB is specifically upregulated in basal-like TNBC, and the loss of LDBH arrests tumor growth *in vivo* (Cui et al., 2015). One study discovered that PAX8 (ENSG00000125618) is the best discriminatory marker between ovarian and breast carcinomas (Nonaka et al., 2008). The same study also reported that PAX8 is negatively expressed in serous carcinoma but is positively expressed in breast carcinomas. This expression pattern is in agreement with our decision rules. RAB34 (ENSG00000109113) regulates the spatial distribution of lysosomes, secretion, and micropinocytosis and is expressed at high levels in breast cancer cell lines. A recent study has found that RAB34 is overexpressed in breast cancer and that the high expression of RAB34 is closely linked to breast cancer cell adhesion, migration, and invasion (Sun et al., 2018).

Among the 275 rules, 51 could identify lung cancer cell lines with 740 criteria involving 80 genes. Here, we provide clear experimental evidence for some genes that are well established in the literature. SOX10 (ENSG00000100146) is a transcription factor that encodes genes involved in the regulation of embryonic development and cell-fate decisions. Studies have demonstrated that SOX10 is usually overexpressed in multiple cancers. It can activate stem/progenitor cells through the Wnt/β-catenin signaling pathway and induces mesenchymal transformation expression (Zhou et al., 2014; Miettinen et al., 2015). However, it appeared in all the lung cancer decision rules with a low expression. A recent experimental study on 1085 NSCLC tumor tissue samples has given direct support for our results. A microarray analysis study revealed that SOX10 is negatively expressed in NSCLC, with only 5 (<1%) cases showing positive results (Kriegsmann et al., 2018). ARHGAP30 (ENSG00000186517) encodes Rho GTPase-activating protein 30, which plays an important role in cell adhesion and cytoskeleton organization regulation. It is downregulated in lung cancer cell lines. Moreover, a low ARHGAP30 level is associated with the activation of Wnt/β-catenin signaling pathways and further leads to lung cancer cell proliferation, migration, and invasion (Mao and Tong, 2018). CTDSPL/RBSP3 (ENSG00000144677) was also downregulated in our rules. It has been reported to be a tumor-suppressor gene in multiple cancers (Kashuba et al., 2009) and to be downregulated in lung cancer (Senchenko et al., 2010). TSPAN4 (ENSG00000214063) was highly expressed in our rules. The transcriptional product of TSPAN4 is a circular RNA that is upregulated in lung adenocarcinoma; circ-TSPAN4 can promote metastasis by increasing the expression of ZEB1 (Ying et al., 2019). In our rules, S100A13 (ENSG00000189171) was required to be relatively highly expressed. As has been seen in another study, S100A13 is overexpressed in NSCLC, especially in the advanced stage. High S100A13 level is strongly associated with tumor angiogenesis and poor prognosis (Miao et al., 2018).

Liver cancer cell lines had nine decision rules containing 115 criteria. These criteria involved 42 genes. The expression patterns of many genes in these rules have been confirmed in several other studies. A1BG-AS1 (ENSG00000268895) is a RNA gene, and its transcriptional product is a lncRNA. A study found that A1BG-AS1 inhibits HCC cell proliferation, migration, and invasion *in vitro*. Clinical association analysis revealed that A1BG-AS1 is downregulated in HCC, and low A1BG-AS1 level is also associated with advanced tumor stage, microvascular invasion, and high tumor grade (Bai et al., 2019). CMTM4 (ENSG00000183723) also showed a low expression level in our rules. As found in other studies, CMTM4 plays a tumor-suppressor role in HCC, wherein it inhibits tumor activities by regulating cell growth and cell cycle (Bei et al., 2017). Thus, consistent with our results, CMTM4 showed negative expression in HCC. AKAP1 (ENSG00000121057) encodes A-kinase anchoring protein 1 and plays an important role in the regulation of mitochondrial function and oxidative metabolism. A previous study identified that AKAP1 is overexpressed in HHC; this expression pattern also provides supporting evidence for our decision rules. AKAP1 may contribute to tumor progression and result in poor overall and disease-free survival rates in patients with HCC (Yu et al., 2018).

The identification of ovarian cancer cell lines had 25 rules with 306 criteria. These criteria involved 67 genes. The validity of our results was supported by other studies that have confirmed some of these genes. As mentioned in the rules for breast cancer cell lines, PAX8 is highly expressed in ovarian cancer and could be a remarkable feature for discriminating between breast cancer and ovarian cancer (Nonaka et al., 2008). In addition, another study found that the knockdown of PAX8 significantly reduces cancer cell proliferation, migration, and invasion (Di Palma et al., 2014). GNAI2 (ENSG00000114353) encodes heterotrimeric G protein, which plays a direct role in regulating the cAMP response element-binding protein. In agreement with our findings, the results of the direct sequencing and qPCR analysis of 589 human ovarian cancer revealed that 85.9% (506) of patients

have decreased GNAI2 messaging (Raymond et al., 2014). SRPX (ENSG00000101955) is reported to be a tumor-suppressor gene and is downregulated in multiple cancer cells and tissues (Tambe et al., 2016). This result is consistent with our decision rules for endometrial, pancreatic, and urinary tract cancers. However, one difference is worth noting: we found that SRPX was overexpressed in most rules for ovarian cancer. The overexpression of SRPX has been affirmed by a recent study based on clinical specimens, wherein the upregulation of SRPX is associated with tumor invasion and migration activity in ovarian cancer (Liu et al., 2019).

At the same time, we noted the exclusive genes for some cancer cell line may be quite important. For example, CD276 (ENSG00000103855) only been shown in the rules of lung cancer cell lines. CD276 (B7-H3) encode a member of the immunoglobulin superfamily and is an important immune checkpoint member of the B7/CD28 families. It is induced by antigen presenting cells and participates in the regulation of T cell-mediated immune response (Picarda et al., 2016). Studies have found that CD276 is associated with Mycoplasma pneumoniae pneumonia. It is up-regulated in patients' plasma and may be involved in the progression of pneumonia by increasing the concentration of TNF-α and the activation of neutrophils (Chen et al., 2013). At the same time, CD276 is also abnormally expressed in a variety of tumors and participates in tumor proliferation, apoptosis, differentiation, invasion and interepithelial transformation. Usually CD276 is up-regulated in tumors and is associated with poor prognosis of patients (Liu S. et al., 2021). In NSCLC, a previous meta-analysis found that the high expression of CD276 was significantly associated with patients' lymph node metastasis and advanced TNM staging (Wu et al., 2016). Other studies have also found that the expression of CD276 is related to the smoking history and pathological types of patients. Usually, the expression of CD276 in patients with lung adenocarcinoma or smoking history is associated with a shorter overall survival (Inamura et al., 2017; Zhang and Hao, 2019). Although CD276 is highly expressed in a variety of tumors, and its molecular mechanism to promote cancer progression is not clear, our results show that it may be more important for lung cancer. At the same time, other study also found that CD276 was up-regulated in tumor cells of lung cancer patients treated with trametinib, which can achieve better therapeutic effect after combined B7-H3 × T cell bispecific antibody treatment, and this also proves that CD276 is a potential therapeutic target for lung cancer.

We provided pieces of evidence for some decisive genes in the decision rules for four classes of cancer cell lines in the preceding discussion. Although these genes also have different expression patterns in other cancer cell lines and a large number of remaining genes have not yet been explained in detail, we can confirm the reliability of our results from the substantial evidence that we presented. Notably, some distinctive and decisive genes in our rules have not previously been investigated by other researchers. These genes may give new insight into tumor growth and progression, as well as novel potential therapeutic targets.

CONCLUSION

This study gave a computational investigation on the cell line gene expression data of cancer cell lines. Several machine learning algorithms were applied on such data. On one hand, we constructed efficient classifiers, which can be latent tools to identify different cancer types. On the other hand, a new set of potential biomarkers (*NCKAP1*, *TNFRSF12A*, *LAMB2*, *FKBP9*, *PFN2*, *TOM1L1*) and expression rules for the identification of different cancers at the transcriptome level were discovered. These biomarkers and rules can be useful materials to uncover mechanism underlying different cancer types, thereby improving our understanding on cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://figshare.com/articles/dataset/scRNA-seq_Datasets/7174922?file=14847260.

AUTHOR CONTRIBUTIONS

LC, TH and Y-DC designed the study. SD, HL and Y-HZ performed the experiments. SD, Y-HZ, XZ, KF and ZL analyzed the results. HL and Y-HZ wrote the manuscript. All authors contributed to the research and reviewed the manuscript.

FUNDING

This work was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (XDB38050200, XDA26040304), National Key R&D Program of China (2018YFC0910403), the Fund of the Key Laboratory of Tissue Microenvironment and Tumor of Chinese Academy of Sciences (202002).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2021.781285/full#supplementary-material

REFERENCES

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., et al. (2016). Pan-cancer Analysis of the Extent and Consequences of Intratumor Heterogeneity. *Nat. Med.* 22, 105–113. doi:10.1038/nm.3984

Bai, J., Yao, B., Wang, L., Sun, L., Chen, T., Liu, R., et al. (2019). lncRNA A1BG-AS1 Suppresses Proliferation and Invasion of Hepatocellular Carcinoma Cells by Targeting miR-216a-5p. *J. Cel. Biochem.* 120, 10310–10322. doi:10.1002/jcb.28315

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., et al. (2012). The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* 483, 603–607. doi:10.1038/nature11003

Bei, C., Zhang, Y., Wei, R., Zhu, X., Wang, Z., Zeng, W., et al. (2017). Clinical Significance of CMTM4 Expression in Hepatocellular Carcinoma. *Onco. Targets Ther.* 10, 5439–5443. doi:10.2147/ott.s149786

Bianchini, M., Levy, E., Zucchini, C., Pinski, V., Macagno, C., De Sanctis, P., et al. (2006). Comparative Study of Gene Expression by cDNA Microarray in Human Colorectal Cancer Tissues and normal Mucosa. *Int. J. Oncol.* 29, 83–94. doi:10.3892/ijo.29.1.83

Burrell, R. A., and Swanton, C. (2014). Tumour Heterogeneity and the Evolution of Polyclonal Drug Resistance. *Mol. Oncol.* 8, 1095–1111. doi:10.1016/j.molonc.2014.06.005

Cao, Q., Liu, Y., Wu, Y., Hu, C., Sun, L., Wang, J., et al. (2020). Profilin 2 Promotes Growth, Metastasis, and Angiogenesis of Small Cell Lung Cancer through Cancer-Derived Exosomes. *Aging* 12, 25981–25999. doi:10.18632/aging.202213

Chang, Y.-S., Chang, C.-M., Lin, C.-Y., Chao, D.-S., Huang, H.-Y., and Chang, J.-G. (2020). Pathway Mutations in Breast Cancer Using Whole-Exome Sequencing. *Oncol. Res.* 28, 107–116. doi:10.3727/096504019x15698362825407

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *jair* 16, 321–357. doi:10.1613/jair.953

Chen, Z.-R., Zhang, G.-B., Wang, Y.-Q., Yan, Y.-D., Zhou, W.-F., Zhu, C.-H., et al. (2013). Soluble B7-H3 Elevations in Hospitalized Children with Mycoplasma Pneumoniae Pneumonia. *Diagn. Microbiol. Infect. Dis.* 77, 362–366. doi:10.1016/j.diagmicrobio.2013.09.006

Chen, L., Wang, S., Zhang, Y.-H., Li, J., Xing, Z.-H., Yang, J., et al. (2017). Identify Key Sequence Features to Improve CRISPR sgRNA Efficacy. *IEEE Access* 5, 26582–26590. doi:10.1109/access.2017.2775703

Chevalier, C., Collin, G., Descamps, S., Touaitahuata, H., Simon, V., Reymond, N., et al. (2016). TOM1L1 Drives Membrane Delivery of MT1-MMP to Promote ERBB2-Induced Breast Cancer Cell Invasion. *Nat. Commun.* 7, 10765. doi:10.1038/ncomms10765

Chiu, H. S., Somvanshi, S., Patel, E., Chen, T. W., Singh, V. P., Zorman, B., et al. (2018). Pan-cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Rep.* 23, 297–312.e12. doi:10.1016/j.celrep.2018.03.064

Cortes, C., and Vapnik, V. (1995). Support-Vector Networks. *Mach Learn.* 20, 273–297. doi:10.1007/bf00994018

Cui, J., Quan, M., Jiang, W., Hu, H., Jiao, F., Li, N., et al. (2015). Suppressed Expression of LDHB Promotes Pancreatic Cancer Progression via Inducing Glycolytic Phenotype. *Med. Oncol.* 32, 143. doi:10.1007/s12032-015-0589-8

Cui, X.-B., Zhang, S.-M., Xu, Y.-X., Dang, H.-W., Liu, C.-X., Wang, L.-H., et al. (2016). PFN2, a Novel Marker of Unfavorable Prognosis, Is a Potential Therapeutic Target Involved in Esophageal Squamous Cell Carcinoma. *J. Transl Med.* 14, 137. doi:10.1186/s12967-016-0884-y

Dagogo-Jack, I., and Shaw, A. T. (2018). Tumour Heterogeneity and Resistance to Cancer Therapies. *Nat. Rev. Clin. Oncol.* 15, 81–94. doi:10.1038/nrclinonc.2017.166

Di Palma, T., Lucci, V., De Cristofaro, T., Filippone, M. G., and Zannini, M. (2014). A Role for PAX8 in the Tumorigenic Phenotype of Ovarian Cancer Cells. *BMC cancer* 14, 292. doi:10.1186/1471-2407-14-292

Ellenbroek, S. I., Iden, S., and Collard, J. G. (2012). Cell Polarity Proteins and Cancer. *Semin. Cancer Biol.* 22, 208–215. doi:10.1016/j.semcancer.2012.02.012

Emaduddin, M., Edelmann, M. J., Kessler, B. M., and Feller, S. M. (2008). Odin (ANKS1A) Is a Src Family Kinase Target in Colorectal Cancer Cells. *Cell Commun Signal* 6, 7. doi:10.1186/1478-811x-6-7

Garman, K. S., Nevins, J. R., and Potti, A. (2007). Genomic Strategies for Personalized Cancer Therapy. *Hum. Mol. Genet.* 16, R226–R232. doi:10.1093/hmg/ddm184

Gewehr, J. E., Szugat, M., and Zimmer, R. (2007). BioWeka Extending the Weka Framework for Bioinformatics. *Bioinformatics* 23, 651–653. doi:10.1093/bioinformatics/btl671

Ghandi, M., Huang, F. W., Jané-Valbuena, J., Kryukov, G. V., Lo, C. C., Mcdonald, E. R., 3rd, et al. (2019). Next-generation Characterization of the Cancer Cell Line Encyclopedia. *Nature* 569, 503–508. doi:10.1038/s41586-019-1186-3

Gorodkin, J. (2004). Comparing Two K-Category Assignments by a K-Category Correlation Coefficient. *Comput. Biol. Chem.* 28, 367–374. doi:10.1016/j.compbiolchem.2004.09.006

Han, H., Nagle, R., and Von Hoff, D. D. (2005). Overexpression of FN14/TWEAK Receptor in Pancreatic Cancer. *Cancer Res.* 65, 554–555.

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: the Next Generation. *Cell* 144, 646–674. doi:10.1016/j.cell.2011.02.013

Hippo, Y., Taniguchi, H., Tsutsumi, S., Machida, N., Chong, J. M., Fukayama, M., et al. (2002). Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays. *Cancer Res.* 62, 233–240.

Huang, M., Narita, S., Tsuchiya, N., Ma, Z., Numakura, K., Obara, T., et al. (2011). Overexpression of Fn14 Promotes Androgen-independent Prostate Cancer Progression through MMP-9 and Correlates with Poor Treatment Outcome. *Carcinogenesis* 32, 1589–1596. doi:10.1093/carcin/bgr182

Hyo-eun, C. B., Ruddy, D. A., Radhakrishna, V. K., Caushi, J. X., Zhao, R., Hims, M. M., et al. (2015). Studying Clonal Dynamics in Response to Cancer Therapy Using High-Complexity Barcoding. *Nat. Med.* 21, 440–448. doi:10.1038/nm.3841

Inamura, K., Yokouchi, Y., Kobayashi, M., Sakakibara, R., Ninomiya, H., Subat, S., et al. (2017). Tumor B7-H3 (CD276) Expression and Smoking History in Relation to Lung Adenocarcinoma Prognosis. *Lung Cancer* 103, 44–51. doi:10.1016/j.lungcan.2016.11.013

Jia, Y., Zhao, R., and Chen, L. (2020). Similarity-Based Machine Learning Model for Predicting the Metabolic Pathways of Compounds. *IEEE Access* 8, 130687–130696. doi:10.1109/access.2020.3009439

Jiang, F.-N., Dai, L.-J., Yang, S.-B., Wu, Y.-D., Liang, Y.-X., Yin, X.-L., et al. (2020). Increasing of FKBP9 Can Predict Poor Prognosis in Patients with Prostate Cancer. *Pathol. - Res. Pract.* 216, 152732. doi:10.1016/j.prp.2019.152732

Kashuba, V. I., Pavlova, T. V., Grigorieva, E. V., Kutsenko, A., Yenamandra, S. P., Li, J., et al. (2009). High Mutability of the Tumor Suppressor Genes RASSF1 and RBSP3 (CTDSPL) in Cancer. *PloS one* 4, e5231. doi:10.1371/journal.pone.0005231

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput.* 13, 637–649. doi:10.1162/089976601300014493

Kohavi, R. (1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, Quebec, Canada (Lawrence Erlbaum Associates), 1137–1145.

Kriegsmann, M., Kriegsmann, K., Harms, A., Longuespée, R., Zgorzelski, C., Leichsenring, J., et al. (2018). Expression of HMB45, MelanA and SOX10 Is Rare in Non-small Cell Lung Cancer. *Diagn. Pathol.* 13, 68. doi:10.1186/s13000-018-0751-7

Kursa, M. B., and Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *J. Stat. Softw.* 36, 1–13. doi:10.18637/jss.v036.i11

Li, B., and Dewey, C. N. (2011). RSEM: Accurate Transcript Quantification from RNA-Seq Data with or without a Reference Genome. *BMC Bioinf.* 12, 323. doi:10.1186/1471-2105-12-323

Liang, H., Chen, L., Zhao, X., and Zhang, X. (2020). Prediction of Drug Side Effects with a Refined Negative Sample Selection Strategy. *Comput. Math. Methods Med.* 2020, 1573543. doi:10.1155/2020/1573543

Ling, Y., Cao, Q., Liu, Y., Zhao, J., Zhao, Y., Li, K., et al. (2021). Profilin 2 (PFN2) Promotes the Proliferation, Migration, Invasion and Epithelial-To-Mesenchymal Transition of Triple Negative Breast Cancer Cells. *Breast Cancer* 28, 368–378. doi:10.1007/s12282-020-01169-x

Liu, H., and Setiono, R. (1998). Incremental Feature Selection. *Appl. Intell.* 9, 217–230. doi:10.1023/a:1008363719778

Liu, C. L., Pan, H. W., Torng, P. L., Fan, M. H., and Mao, T. L. (2019). SRPX and HMCN1 Regulate Cancer-Associated Fibroblasts to Promote the Invasiveness of Ovarian Carcinoma. *Oncol. Rep.* 42, 2706–2715. doi:10.3892/or.2019.7379

Liu, H., Hu, B., Chen, L., and Lu, L. (2021a). Identifying Protein Subcellular Location with Embedding Features Learned from Networks. *Curr. Proteomics* 18, 646–660. doi:10.2174/1570164617999201124142950

Liu, S., Liang, J., Liu, Z., Zhang, C., Wang, Y., Watson, A. H., et al. (2021b). The Role of CD276 in Cancers. *Front. Oncol.* 11, 847. doi:10.3389/fonc.2021.654684

Mao, X., and Tong, J. (2018). ArhgaP30 Suppressed Lung Cancer Cell Proliferation, Migration, and Invasion through Inhibition of the Wnt/β-Catenin Signaling Pathway. *Onco Targets Ther.* 11, 7447–7457. doi:10.2147/ott.s175255

Matthews, B. W. (1975). Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta (Bba) - Protein Struct.* 405, 442–451. doi:10.1016/0005-2795(75)90109-9

Miao, S., Qiu, T., Zhao, Y., Wang, H., Sun, X., Wang, Y., et al. (2018). Overexpression of S100A13 Protein Is Associated with Tumor Angiogenesis and Poor Survival in Patients with Early-Stage Non-Small Cell Lung Cancer. *Thorac. Cancer* 9, 1136–1144. doi:10.1111/1759-7714.12797

Miettinen, M., Mccue, P. A., Sarlomo-Rikala, M., Biernat, W., Czapiewski, P., Kopczynski, J., et al. (2015). Sox10-A Marker for Not Only Schwannian and Melanocytic Neoplasms but Also Myoepithelial Cell Tumors of Soft Tissue. *Am. J. Surg. Pathol.* 39, 826–835. doi:10.1097/pas.0000000000000398

Nonaka, D., Chiriboga, L., and Soslow, R. A. (2008). Expression of Pax8 as a Useful Marker in Distinguishing Ovarian Carcinomas from Mammary Carcinomas. *Am. J. Surg. Pathol.* 32, 1566–1571. doi:10.1097/pas.0b013e31816d71ad

Pan, X., Li, H., Zeng, T., Li, Z., Chen, L., Huang, T., et al. (2021). Identification of Protein Subcellular Localization with Network and Functional Embeddings. *Front. Genet.* 11, 626500. doi:10.3389/fgene.2020.626500

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-Learn: Machine Learning in Python. *J. Machine Learn. Res.* 12, 2825–2830.

Hanchuan Peng, H., Fuhui Long, F., and Ding, C. (2005). Feature Selection Based on Mutual Information Criteria of max-dependency, max-relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Machine Intell.* 27, 1226–1238. doi:10.1109/tpami.2005.159

Picarda, E., Ohaegbulam, K. C., and Zang, X. (2016). Molecular Pathways: Targeting B7-H3 (CD276) for Human Cancer Immunotherapy. *Clin. Cancer Res.* 22, 3425–3431. doi:10.1158/1078-0432.ccr-15-2428

J. Platt (Editor) (1998). *Fast Training of Support Vector Machines Using Sequential Minimal Optimization* (Cambridge, MA: MIT Press).

Raymond, J. R., Appleton, K. M., Pierce, J. Y., and Peterson, Y. K. (2014). Suppression of GNAI2 Message in Ovarian Cancer. *J. Ovarian Res.* 7, 6. doi:10.1186/1757-2215-7-6

Russo, M., Siravegna, G., Blaszkowsky, L. S., Corti, G., Crisafulli, G., Ahronian, L. G., et al. (2016). Tumor Heterogeneity and Lesion-specific Response to Targeted Therapy in Colorectal Cancer. *Cancer Discov.* 6, 147–153. doi:10.1158/2159-8290.cd-15-1283

Safavian, S. R., and Landgrebe, D. (1991). A Survey of Decision Tree Classifier Methodology. *IEEE Trans. Syst. Man. Cybern.* 21, 660–674. doi:10.1109/21.97458

Senchenko, V. N., Anedchenko, E. A., Kondratieva, T. T., Krasnov, G. S., Dmitriev, A. A., Zabarovska, V. I., et al. (2010). Simultaneous Down-Regulation of Tumor Suppressor Genes RBSP3/CTDSPL, NPRL2/G21 and RASSF1A in Primary Non-small Cell Lung Cancer. *BMC cancer* 10, 75. doi:10.1186/1471-2407-10-75

Sun, L., Xu, X., Chen, Y., Zhou, Y., Tan, R., Qiu, H., et al. (2018). Rab34 Regulates Adhesion, Migration, and Invasion of Breast Cancer Cells. *Oncogene* 37, 3698–3714. doi:10.1038/s41388-018-0202-7

Tambe, Y., Hasebe, M., Kim, C. J., Yamamoto, A., and Inoue, H. (2016). The Drs Tumor Suppressor Regulates Glucose Metabolism via Lactate Dehydrogenase-B. *Mol. Carcinog.* 55, 52–63. doi:10.1002/mc.22258

Teng, Y., Qin, H., Bahassan, A., Bendzunas, N. G., Kennedy, E. J., and Cowell, J. K. (2016). The WASF3-NCKAP1-CYFIP1 Complex Is Essential for Breast Cancer Metastasis. *Cancer Res.* 76, 5133–5142. doi:10.1158/0008-5472.can-16-0562

Tran, N. L., Mcdonough, W. S., Savitch, B. A., Fortin, S. P., Winkles, J. A., Symons, M., et al. (2006). Increased Fibroblast Growth Factor-Inducible 14 Expression Levels Promote Glioma Cell Invasion via Rac1 and Nuclear Factor-κB and Correlate with Poor Patient Outcome. *Cancer Res.* 66, 9535–9542. doi:10.1158/0008-5472.can-06-0418

Van Der Zee, J. A., Van Eijck, C. H., Hop, W. C., Biermann, K., Dicheva, B. M., Seynhaeve, A. L., et al. (2012). Tumour Basement Membrane Laminin Expression Predicts Outcome Following Curative Resection of Pancreatic Head Cancer. *Br. J. Cancer* 107, 1153–1158. doi:10.1038/bjc.2012.373

Wajant, H. (2013). The TWEAK-Fn14 System as a Potential Drug Target. *Br. J. Pharmacol.* 170, 748–764. doi:10.1111/bph.12337

Wang, T., Ma, S., Qi, X., Tang, X., Cui, D., Wang, Z., et al. (2017). Knockdown of the Differentially Expressed Gene TNFRSF12A Inhibits Hepatocellular Carcinoma Cell Proliferation and Migration *In Vitro*. *Mol. Med. Rep.* 15, 1172–1178. doi:10.3892/mmr.2017.6154

Wang, Y., Xu, Y., Yang, Z., Liu, X., and Dai, Q. (2021). Using Recursive Feature Selection with Random Forest to Improve Protein Structural Class Prediction for Low-Similarity Sequences. *Comput. Math. Methods Med.* 2021, 5529389. doi:10.1155/2021/5529389

Watts, G. S., Tran, N. L., Berens, M. E., Bhattacharyya, A. K., Nelson, M. A., Montgomery, E. A., et al. (2007). Identification of Fn14/TWEAK Receptor as a Potential Therapeutic Target in Esophageal Adenocarcinoma. *Int. J. Cancer* 121, 2132–2139. doi:10.1002/ijc.22898

Weinstein, J. N., Collisson, E. A., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., et al. (2013). The Cancer Genome Atlas Pan-Cancer Analysis Project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764

Whitsett, T. G., Cheng, E., Inge, L., Asrani, K., Jameson, N. M., Hostetter, G., et al. (2012). Elevated Expression of Fn14 in Non-small Cell Lung Cancer Correlates with Activated EGFR and Promotes Tumor Cell Migration and Invasion. *Am. J. Pathol.* 181, 111–120. doi:10.1016/j.ajpath.2012.03.026

Wodarz, A., and Näthke, I. (2007). Cell Polarity in Development and Cancer. *Nat. Cel Biol.* 9, 1016–1024. doi:10.1038/ncb433

World Health Organization (2019). International Agency for Research on Cancer. Available at https://gco.iarc.fr/today (Accessed January 4, 2021).

Wu, S., Zhao, X., Wu, S., Du, R., Zhu, Q., Fang, H., et al. (2016). Overexpression of B7-H3 Correlates with Aggressive Clinicopathological Characteristics in Non-small Cell Lung Cancer. *Oncotarget* 7, 81750–81756. doi:10.18632/oncotarget.13177

Wu, Z.-H., Niu, X., Wu, G.-H., and Cheng, Q. (2020). Decreased Expression of TNFRSF12A in Thyroid Gland Cancer Predicts Poor Prognosis. *Medicine* 99, e21882. doi:10.1097/md.0000000000021882

Xiao, Z., Dai, Z., and Locasale, J. W. (2019). Metabolic Landscape of the Tumor Microenvironment at Single Cell Resolution. *Nat. Commun.* 10, 3763. doi:10.1038/s41467-019-11738-0

Xiong, Y., He, L., Shay, C., Lang, L., Loveless, J., Yu, J., et al. (2019). Nck-associated Protein 1 Associates with HSP90 to Drive Metastasis in Human Non-small-cell Lung Cancer. *J. Exp. Clin. Cancer Res.* 38, 122. doi:10.1186/s13046-019-1124-0

Xu, H., Liu, P., Yan, Y., Fang, K., Liang, D., Hou, X., et al. (2020). FKBP9 Promotes the Malignant Behavior of Glioblastoma Cells and Confers Resistance to Endoplasmic Reticulum Stress Inducers. *J. Exp. Clin. Cancer Res.* 39, 44. doi:10.1186/s13046-020-1541-0

Yamaguchi, H., and Condeelis, J. (2007). Regulation of the Actin Cytoskeleton in Cancer Cell Migration and Invasion. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1773, 642–652. doi:10.1016/j.bbamcr.2006.07.001

Yamazaki, D., Kurisu, S., and Takenawa, T. (2005). Regulation of Cancer Cell Motility through Actin Reorganization. *Cancer Sci.* 96, 379–386. doi:10.1111/j.1349-7006.2005.00062.x

Yan, J., Ma, C., and Gao, Y. (2017). MicroRNA-30a-5p Suppresses Epithelial-Mesenchymal Transition by Targeting Profilin-2 in High Invasive Non-small Cell Lung Cancer Cell Lines. *Oncol. Rep.* 37, 3146–3154. doi:10.3892/or.2017.5566

Yang, Y., and Chen, L. (2021). Identification of Drug–Disease Associations by Using Multiple Drug and Disease Networks. *Curr. Bioinf.* 16. doi:10.2174/1574893616666210825115406

Yang, J., Min, K.-W., Kim, D.-H., Son, B. K., Moon, K. M., Wi, Y. C., et al. (2018). High TNFRSF12A Level Associated with MMP-9 Overexpression Is Linked to Poor Prognosis in Breast Cancer: Gene Set Enrichment Analysis and Validation in Large-Scale Cohorts. *PloS one* 13, e0202113. doi:10.1371/journal.pone.0202113

Ying, X., Zhu, J., and Zhang, Y. (2019). Circular RNA Circ-TSPAN4 Promotes Lung Adenocarcinoma Metastasis by Upregulating ZEB1 via Sponging miR-665. *Mol. Genet. Genomic Med.* 7, e991. doi:10.1002/mgg3.991

Yu, J., Zhang, Y., Zhou, D., Wu, J., and Luo, R. (2018). Higher Expression of A-Kinase Anchoring-Protein 1 Predicts Poor Prognosis in Human Hepatocellular Carcinoma. *Oncol. Lett.* 16, 131–136. doi:10.3892/ol.2018.8685

Zhang, C., and Hao, X. (2019). Prognostic Significance of CD276 in Non-small Cell Lung Cancer. *Open Med.* 14, 805–812. doi:10.1515/med-2019-0076

Zhang, H., Yang, W., Yan, J., Zhou, K., Wan, B., Shi, P., et al. (2018a). Loss of Profilin 2 Contributes to Enhanced Epithelial-Mesenchymal Transition and Metastasis of Colorectal Cancer. *Int. J. Oncol.* 53, 1118–1128. doi:10.3892/ijo.2018.4475

Zhang, Y., Fang, L., Zang, Y., and Xu, Z. (2018b). Identification of Core Genes and Key Pathways via Integrated Analysis of Gene Expression and DNA Methylation Profiles in Bladder Cancer. *Med. Sci. Monit.* 24, 3024–3033. doi:10.12659/msm.909514

Zhang, Y.-H., Li, H., Zeng, T., Chen, L., Li, Z., Huang, T., et al. (2021a). Identifying Transcriptomic Signatures and Rules for SARS-CoV-2 Infection. *Front. Cel Dev. Biol.* 8, 627302. doi:10.3389/fcell.2020.627302

Zhang, Y.-H., Li, Z., Zeng, T., Chen, L., Li, H., Huang, T., et al. (2021b). Detecting the Multiomics Signatures of Factor-specific Inflammatory Effects on Airway Smooth Muscles. *Front. Genet.* 11, 599970. doi:10.3389/fgene.2020.599970

Zhang, Y.-H., Zeng, T., Chen, L., Huang, T., and Cai, Y.-D. (2021c). Determining Protein-Protein Functional Associations by Functional Rules Based on Gene Ontology and KEGG Pathway. *Biochim. Biophys. Acta (Bba) - Proteins Proteomics* 1869, 140621. doi:10.1016/j.bbapap.2021.140621

Zhao, X., Chen, L., and Lu, J. (2018). A Similarity-Based Method for Prediction of Drug Side Effects with Heterogeneous Information. *Math. Biosci.* 306, 136–144. doi:10.1016/j.mbs.2018.09.010

Zhong, X.-P., Kan, A., Ling, Y.-H., Lu, L.-H., Mei, J., Wei, W., et al. (2019). NCKAP1 Improves Patient Outcome and Inhibits Cell Growth by Enhancing Rb1/p53 Activation in Hepatocellular Carcinoma. *Cell Death Dis.* 10, 369. doi:10.1038/s41419-019-1603-4

Zhou, H., Ekmekcioglu, S., Marks, J. W., Mohamedali, K. A., Asrani, K., Phillips, K. K., et al. (2013). The TWEAK Receptor Fn14 Is a Therapeutic Target in Melanoma: Immunotoxins Targeting Fn14 Receptor for Malignant Melanoma Treatment. *J. Invest. Dermatol.* 133, 1052–1062. doi:10.1038/jid.2012.402

Zhou, D., Bai, F., Zhang, X., Hu, M., Zhao, G., Zhao, Z., et al. (2014). SOX10 Is a Novel Oncogene in Hepatocellular Carcinoma through Wnt/β-catenin/TCF4 cascade. *Tumor Biol.* 35, 9935–9940. doi:10.1007/s13277-014-1893-1

Zhou, K., Chen, J., Wu, J., Xu, Y., Wu, Q., Yue, J., et al. (2019). Profilin 2 Promotes Proliferation and Metastasis of Head and Neck Cancer Cells by Regulating PI3K/AKT/β-Catenin Signaling Pathway. *Oncol. Res.* 27, 1079–1088. doi:10.3727/096504019x15579146061957

Zhou, J. P., Chen, L., and Guo, Z. H. (2020a). iATC-NRAKEL: An Efficient Multi-Label Classifier for Recognizing Anatomical Therapeutic Chemical Classes of Drugs. *Bioinformatics* 36, 1391–1396. doi:10.1093/bioinformatics/btz757

Zhou, J.-P., Chen, L., Wang, T., and Liu, M. (2020b). iATC-FRAKEL: a Simple Multi-Label Web Server for Recognizing Anatomical Therapeutic Chemical Classes of Drugs with Their Fingerprints Only. *Bioinformatics* 36, 3568–3569. doi:10.1093/bioinformatics/btaa166

Zhu, Y., Hu, B., Chen, L., and Dai, Q. (2021). iMPTCE-Hnetwork: A Multilabel Classifier for Identifying Metabolic Pathway Types of Chemicals and Enzymes with a Heterogeneous Network. *Comput. Math. Methods Med.* 2021, 6683051. doi:10.1155/2021/6683051

Zhuang, H., Tan, M., Liu, J., Hu, Z., Liu, D., Gao, J., et al. (2014). Human Epididymis Protein 4 in Association with Annexin II Promotes Invasion and Metastasis of Ovarian Cancer Cells. *Mol. Cancer* 13, 243. doi:10.1186/1476-4598-13-243

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.