Frontiers in Bioscience Landmark

Front. Biosci. (Landmark Ed) **2023**; 28(11): 283
https://doi.org/10.31083/j.fbl2811283

Original Research

Differential Expression Analysis Based on Ensemble Strategy on miRNA Profiles of Kidney Clear Cell Carcinoma

Enyang Zhao[1,2], Ziqi Xi[3], Qiong Wu[1,*]

[1]School of Life Science and Technology, Harbin Institute of Technology, 150006 Harbin, Heilongjiang, China
[2]The Second Affiliated Hospital of Harbin Medical University, Harbin Medical University, 150076 Harbin, Heilongjiang, China
[3]College of Aulin, Northeast Forestry University, 150006 Harbin, Heilongjiang, China
*Correspondence: kigo@hit.edu.cn (Qiong Wu)
Academic Editor: Taeg Kyu Kwon
Submitted: 9 February 2023 Revised: 11 March 2023 Accepted: 21 March 2023 Published: 8 November 2023

Abstract

Background: Kidney clear cell carcinoma (KIRC) is the most common type of kidney cancer, accounting for approximately 60–85% of all the kidney cancers. However, there are few options available for early treatment. Therefore, it is extremely important to identify biomarkers and study therapeutic targets for KIRC. **Methods**: Since there are few studies on KIRC, we used a data-driven approach to identify differential genes. Here, we used miRNA gene expression profile data from the TCGA database species of KIRC and proposed a machine learning-based approach to quantify the importance score of each gene. Then, an ensemble method was utilized to find the optimal subset of genes used to predict KIRC by clustering. The most genetic subset was then used to classify and predict KIRC. **Results**: Differential genes were screened by several traditional differential analysis methods, and the selected gene subset showed a better performance. Independent testing sets from the GEO database were used to verify the effectiveness of the optimal subset of genes. Besides, cross-validation was made to verify the effectiveness of the approach. **Conclusions**: Finally, important genes, such as *miR-140* and *miR-210*, were found to be involved in the biochemical processes of KIRC, which also proved the effectiveness of our approach.

Keywords: kidney cancer; miRNA; differential gene; machine learning; ensemble

1. Introduction

Kidney clear cell carcinoma (KIRC) is one of the most common and deadliest urological cancers, accounting for approximately 3% of human malignancies [1] and more than 116,000 deaths per year in patients with KIRC [2].

Despite significant improvements in the diagnosis and treatment of KIRC over the past two decades [3,4], its prognostic chemotherapeutic approach and treatment of metastasis remain limited due to the uncertain cause of its development. Therefore, there is a strong need to identify effective biomarkers to further investigate the pathogenesis of KIRC to facilitate the treatment of this cancer [5,6].

With the development of sequencing technology, microarrays based on high-throughput platforms have been widely used for obtaining intracellular gene expressions, and the corresponding statistical analysis is considered the most promising tool for medical oncology research. It has also become increasingly easy to obtain intracellular gene expression by high-throughput sequencing [7].

After obtaining the gene expressions, there are two main methods for performing differential expression analysis from gene expression profile matrices: parametric and non-parametric methods. Parametric methods capture all the information about the parameters in the data. In this way, valuable genes in data can be predicted by analyzing the model used and its parameters. When parametric methods are applied to differential expression analysis, each expression value for each gene is mapped to a specific distribution, such as Poisson [8,9] or negative binomial [10,11]. As to a non-parametric method, it is not allowed to impose the rigid model to be fitted. Besides, it can obtain more details about data distribution. Since a non-parametric model takes into account the inability to fit the data distribution from a limited set of parameters, the amount of information about the data increases as the amount of data increases.

The commonly used methods for differential expression analysis on expression profile matrices of cancer cells are EdgeR [12–14] and baySeq [15], both of which use negative binomial models. Limma [16], which is considered to be a linear model, is also commonly used. Non-parametric methods such as NOIseq [17] and SAMseq [18] are also in existence. Besides, there are transcript-based assays that can be used for differential gene identification, such as EBSeq [19] and Cuffdiff2 [20].

Differentially expressed gene in bioinformatics research is still under development [21,22]. A typical approach is to use the DESeq2 R package to screen for cytogenetic risk-associated differential expression genes [23]. However, most biomarkers for studying cancers focus only on the screening of individual differential genes and do not take into account the interaction of genes. Yuan [24] proposed a systems biology approach named weighted gene expression network analysis (WGCNA) to characterize the correlation patterns between genes across microarray or

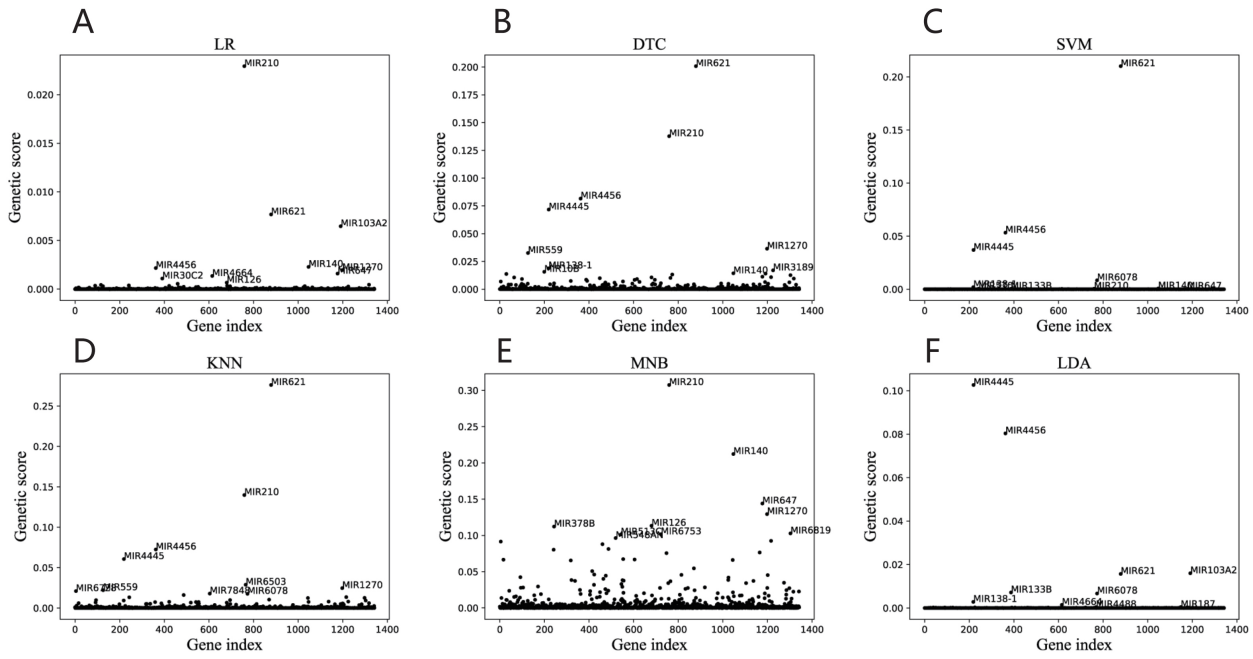

Fig. 1. The computed important scores of miRNAs for each classifier. (A) Gene importance calculated by LR classifier. (B) Gene importance calculated by DTC classifier. (C) Gene importance calculated by SVM classifier. (D) Gene importance calculated by KNN classifier. (E) Gene importance calculated by MNB classifier. (F) Gene importance calculated by LDA classifier. Most genes have gene importance nearly zero. Genes with the highest importance scores are tagged with gene names. Abbreviations: LR, logistic regression; DTC, decision tree classifier; SVM, support vector machine; KNN, support vector machine; MNB, multinomial naïve Bayes; LDA, Fisher's linear discriminant.

RNA sequence data. WGCNA is regarded as a cluster or module for finding highly related genes and identifying phenotypic modules [25]. However, this method also identifies differential genes individually and analyzes intercellular interactions afterward.

miRNAs are a group of endogenous small non-coding RNAs that regulate gene expressions by binding to untranslated regions of target mRNAs [26] and are involved in gene expressions in various biological phenomena, such as homeostasis, development, proliferation, and apoptosis [27,28]. In this paper, the interaction of miRNAs is hypothesized and thought to jointly regulate kidney clear cell carcinoma. A machine learning-based classification method is proposed to quantify the importance of each miRNA; thus, an optimal miRNA combination (including *miR-210*, *miR-621*, *miR-140*, etc.) is discovered from the TCGA database as a statistical signature for diagnosis of KIRC.

2. Materials and Methods

2.1 Data

The miRNA expression profiles of KIRC used in this study are obtained from the TCGA database. There are 611 samples, in which cancer and adjacent normal samples are 539 and 72, respectively. The miRNAs with zero variance in expression are excluded and a set with 1343 miRNAs is obtained. A normalization is made on each miRNA using FPKM.

2.2 Finding Differential miRNAs Using Feature Selection

Feature selection is a common method to eliminate redundant and irrelevant variables with the aim of reducing feature dimensionality and discovering a valid subset of genes as a diagnostic biomarker to make data fitting simpler and prediction more accurate [29,30].

The cause of KIRC is generally due to mutation of some genes. However, the probability of mutation is small, which means that only a small number of mutations in cancer and adjacent normal cells lead to differential expressions and even cellular carcinogenesis. Therefore, we speculate that the number of the most differential miRNAs is countable.

In addition, there are often multiple miRNAs that act in combination in a biochemical reaction. That is, there are single miRNAs that do not express differently between cancer and normal tissues, but two or more of these miRNAs that together show differential expressions. In that case, the miRNAs that have a difference in combination cannot be found by conventional methods. To treat this situation, full enumeration is considered. To discover the optimal miRNAs for discrimination between KIRC and normal samples, each miRNA tuple in each dimensional combination is tried. For miRNA expression profiles, the number of miRNAs is usually much larger than the number of samples, so the time complexity of the full permutation is O(n!), which is too computationally intensive.

Thus, a random miRNA selection way is considered. Firstly, m miRNAs are randomly selected every time, and a corresponding classification result such as accuracy is calculated on the expression profiles corresponding to cancer and normal tissues. If one of the n miRNAs is viewed to be important for discrimination between KIRC and normal samples, interference with the miRNA will significantly impact the classification result.

We propose a method to quantify this impact. First of all, 70% of the data set is randomly divided as a training set. Classifiers such as logistic regression (LR), decision tree classifier (DTC), support vector machine (SVM), k-nearest-neighbor (KNN), multinomial naïve Bayes (MNB), and Fisher's linear discriminant (LDA) are trained by using the Python package Scikit-learn [31]. Then, predictions are made on the 30% of the left data. That is, the classification accuracy $Score_1$ is obtained. After that, a one-time permutation on expressions of miRNA i is made. The classifiers are then used to make predictions on the processed data with classification accuracy to be $Score_2$. The importance of the i^{th} miRNA for the accurate classification of KIRC and adjacent normal samples is calculated as follows,

$$impact_i = Score1 - Score2 \tag{1}$$

where $impact_i$ is the importance score of the i^{th} miRNA.

Due to the uneven distribution of positive and negative samples, the classification accuracy is modified and expressed as follows,

$$Accuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{2} \tag{2}$$

where FN, TP, FP and TN represent false negative, true positive, false positive and true negative, respectively. Positive and negative samples refer to KIRC and normal tissues.

The above method is repeated for n rounds in order to count as many miRNA combinations as possible. Then the average importance scores corresponding to all the listed miRNAs are obtained.

Considering the small sample size with unknown distribution, six classifiers (i.e., DTC, MNB, KNN, LDA, LR, and SVM) are utilized for training and validation.

2.3 Clustering

Once the importance score of each miRNA is obtained, how to select the important genes becomes a new problem. If the importance score threshold is set manually, subjective factors will be inevitably introduced. Here we choose to use an unsupervised clustering approach to select miRNAs according to the importance scores.

A clustering algorithm based on probability distribution was used in this study, which is implemented based on the mixture module of the Python Scikit-learn package [31]. It is assumed that the sample points can be divided into k clusters, also known as k components, and each cluster obeys a different Gaussian distribution. The probability of each sample belonging to each distribution is calculated, and the sample is classified into the cluster with the highest probability. Then, based on the expectation maximization (EM) algorithm, the parameters of the Gaussian distribution are updated after several iterations until the model converges to a locally optimal solution.

Finally, the cluster with the highest importance scores is used as the important feature subset. That is

$$p(\mathrm{X}) = \sum_{\mathrm{i}=1}^{\mathrm{k}} \alpha_{\mathrm{i}} * p\left(X \mid \mu_i, \Sigma_i\right) \tag{3}$$

where k is the number of components. α_i is the probability that the sample point belongs to the ith Gaussian distribution, and $p\left(X\mu_i,\ \Sigma_i\right)$ is the probability density function of the ith Gaussian distribution, where μ_i is the mean vector and Σ_i is the covariance matrix.

2.4 Ensemble

After clustering on the importance scores, important feature subsets are obtained from different classifiers. Because the decision boundaries of these classifiers are different, the obtained important feature subsets are different.

To obtain the optimal feature subset, an ensemble strategy is proposed based on the idea of ensemble learning to combine all important feature subsets. The simplest way is to select the important feature subset by the best classifier. But it may lead to model overfitting. The union among these important feature subsets is calculated to get the optimal feature subset.

The combination of the results from individual classifiers can reduce the possibility of feature missed selection caused by a large hypothesis space, reduce the learning algorithm into local optimum, and thus improve the generalization effect of the model. It can also expand the hypothesis space and learn better approximations.

Therefore, this strategy of training the classifiers individually to find the important feature subsets and then combination can find the optimal features that are closer to the real important features and perform better on other data sets.

2.5 Evaluation

Evaluations are performed for the selected genes. One is on the testing set, and the other is on the new data sets built from the TCGA data set with only the selected features. In addition, we made cross-validation to test the validity of the obtained significant genes.

First, the important feature subsets derived from six different classifiers are evaluated. Each time, an important feature subset and all genes are used to train the classifier separately. We use the following evaluation metrics for accuracy evaluation based on the confusion matrix, i.e., TP

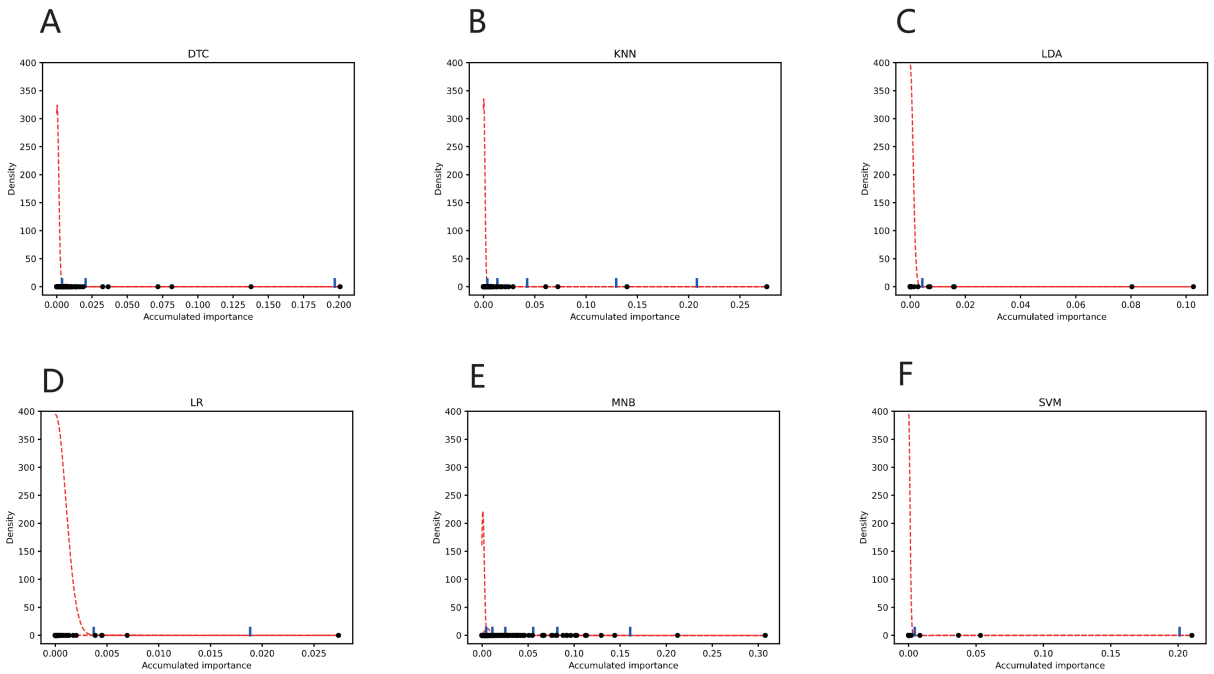

Fig. 2. Clustering results of miRNA important scores using Gaussian mixture clustering. (A) Clustering on gene importance calculated by LR classifier. (B) Clustering on gene importance calculated by DTC classifier. (C) Clustering on gene importance calculated by SVM classifier. (D) Clustering on gene importance calculated by KNN classifier. (E) Clustering on gene importance calculated by MNB classifier. (F) Gene importance calculated by LDA classifier. Gene importance calculated by different classifiers has different Gauss boundaries for clustering.

rate, FP rate, Precision, Recall, F1 measure, and the accuracy after balancing positive and negative samples, as are expressed as follows,

$$\begin{aligned} TP_{rate} &= \frac{TP}{TP+FN}, \\ FP_{rate} &= \frac{FP}{FP+TN}, \\ Precision &= \frac{TP}{TP+FP}, \\ Recall &= \frac{TP}{TP+FN} \\ F1measure &= \frac{2*precision*Recall}{Precision+Recall}. \end{aligned} \tag{4}$$

Second, the ensemble feature subset and the differential genes are evaluated. Three traditional differential analysis methods are used to search for the differential genes. The first two are performed by the R package limma and edgeR, and the last one is performed by a simple two-sample *t*-test. The *p*-value and logFoldChange value of each gene are calculated after the differential analysis to screen out differential genes. The threshold for *p*-value is set to 0.05, and the threshold for logFoldchange is calculated as follows,

$$\log FC_{cutoff} = \overline{|\log FC|} + 2 * \sigma(\overline{|\log FC|}). \tag{5}$$

Then new data sets are built on the selected genes. The same evaluation metrics for accuracy evaluation based on the confusion matrix are used.

Apart from that, one 5-fold cross-validation is made to verify the stability and reliability of the results. The TCGA dataset is divided into five mutually exclusive subsets, and the gene importance is calculated separately after multiple training and testing. Five sets of selected genes are obtained.

2.6 Validation

To validate the effectiveness of the obtained optimal feature subset, independent testing sets of KIRC are selected from the GEO database.

Combinations of any pair of miRNAs are enumerated to see the effect of different combinations on the sample grouping, and then all the important miRNAs are also considered as a combination.

Due to the high dimension of the data, it's difficult to observe the distribution. To see the effect of the optimal feature subset on the sample grouping, the principal component analysis method (PCA) is used. PCA retains most of the effective information while reducing the number of features. Here, only two principal components are finally retained for data visualization.

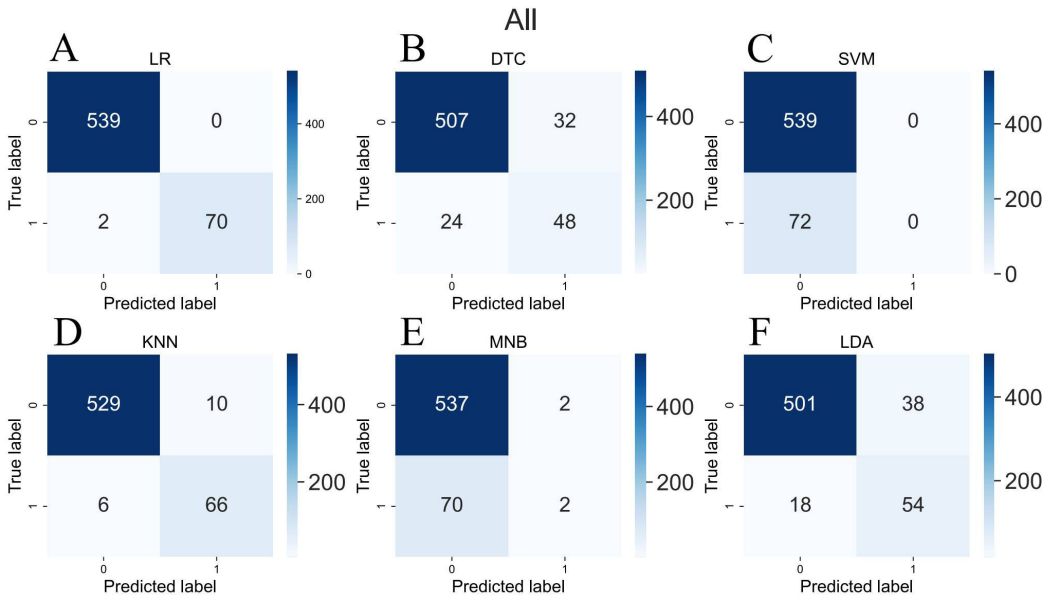

Fig. 3. Heatmap of the confusion matrix using all genes. (A) The confusion matrix of LR. (B) The confusion matrix of DTC. (C) The confusion matrix of SVM. (D) The confusion matrix of KNN. (E) The confusion matrix of MNB. (F) The confusion matrix of LDA.

3. Result

In this paper, we analyzed the miRNA expression profiles of KIRC and assumed that only countable miRNAs significantly influenced the classification results. Six classifiers were used to quantify the importance of each miRNA separately using a premutation-based approach. And a density-descent clustering method was performed to find the important miRNAs in KIRC.

3.1 Feature Selection Results

The data were equally divided into halves, each of which corresponded to the training set and testing set. In the training set, 70% of the samples were randomly selected to train the six classifiers and the remaining out-of-bag samples were used to calculate the importance scores of each miRNA according to Eqn. 1. The procedure was repeated 700,000 times. The average importance score of each miRNA was obtained, as shown in Fig. 1. It can be seen that the miRNAs with high important scores account for only a fraction of all the miRNAs, about 1%, with *miR-621*, *miR-210*, and *miR4456* ranking high in most of the classifiers. Finding common miRNAs using different classification methods indicates the validity of our method. The phenomenon that different miRNAs appear with high important scores indicates that sample distribution affects the classification results according to different classifiers with different classification boundaries.

3.2 Clustering Results

For miRNAs that are not involved in the cancer process and are expressed at similar levels to non-cancerous cells, they are not important in the prediction of KIRC. According to the previous hypothesis, only a small number of variants of miRNAs cause KIRC. Therefore, there will be a large number of miRNAs with low important scores. In contrast, the important miRNAs are viewed as outliers. This can be verified in Fig. 2, which shows the effectiveness of the clustering method. The important genes are selected based on the Gaussian boundaries presented by short blue horizontal lines shown in Fig. 2.

3.3 Classification Results

To verify the validity of the found miRNAs, six classifiers were separately trained on the training data, and the classification accuracies were calculated on the independent test set. The classification results are shown in Table 1. It can be seen that LR and LDA using all the miRNAs for classification are more efficient than those using the selected miRNAs. And the higher accuracy of LR also means that there may be overfitting. However, DTC, SVM, KNN, and MNB achieve better classification results considering the selected miRNAs. Especially, there is no good hyperplane using SVM with all the miRNAs considered, while it shows much better classification results under the selected miRNAs. The corresponding confusion matrices are shown in Figs. 3,4.

To compare the optimal feature subset with the differential genes found by traditional differential analysis methods, four new datasets were constructed and used for evaluation.

Volcano plots of the differential genes are shown in Fig. 5. The differential genes are divided into up-regulated

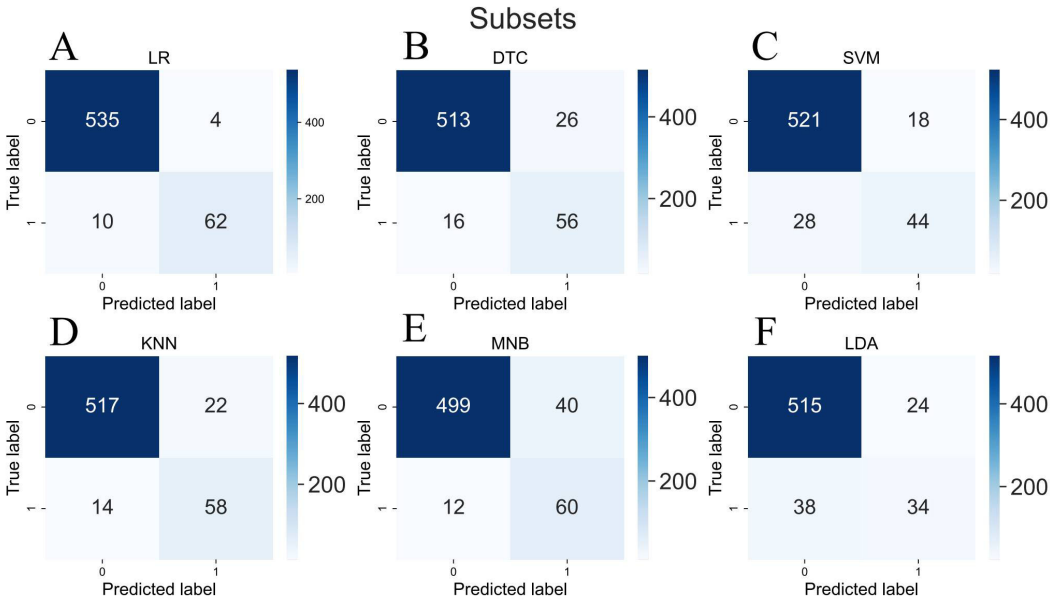

Fig. 4. Heatmap of confusion matrix using the gene subsets. (A) The confusion matrix of LR. (B) The confusion matrix of DTC. (C) The confusion matrix of SVM. (D) The confusion matrix of KNN. (E) The confusion matrix of MNB. (F) The confusion matrix of LDA. Python package seaborn was utilized to draw the heatmap.

Table 1. Classification results on the independent test set.

	Gene sets	TP rate	FP rate	Precision	Recall	F1 measure	Balance accuracy
LR	All	0.972	0	1	0.972	0.986	0.986
	Subsets	0.861	0.007	0.939	0.861	0.899	0.927
DTC	ALL	0.667	0.059	0.6	0.667	0.632	0.804
	Subsets	0.778	0.048	0.683	0.778	0.727	0.865
SVM	ALL	0	0	Nan	0	0.5	0.5
	Subsets	0.661	0.033	0.71	0.611	0.657	0.789
KNN	ALL	0.917	0.019	0.868	0.917	0.892	0.949
	Subsets	0.806	0.04	0.707	0.806	0.753	0.949
MNB	ALL	0.028	0.004	0.5	0.028	0.053	0.881
	Subsets	0.833	0.074	0.6	0.833	0.698	0.88
LDA	ALL	0.75	0.07	0.587	0.75	0.659	0.84
	Subsets	0.472	0.044	0.586	0.472	0.523	0.714

Abbreviations: TP, true positive; FP, false positive; F1, F1-measure.

genes and down-regulated genes. Blue points in Fig. 5 represent down-regulated genes, and red points in Fig. 5 represent down-regulated genes. The top 10 differential genes are labeled with gene names.

By making a comparison between the proposed method with the prevailing feature selection methods such as edgeR, limma, and *t*-test, the quantitative assessment metrics of the classification results on the four data sets are shown in Table 2, and the corresponding ROC curves and AUCs are shown in Fig. 6.

From Table 2 we can see that the significant genes found by this method achieved the same or even better classification results than traditional methods, using only 15 selected genes, which is much smaller than the number of genes found by traditional differential analysis.

3.4 Validation Results

Each pair of important miRNAs was enumerated and GSE independent testing sets GSE109368, GSE151423, and GSE 151419 were used to verify the effectiveness of each gene combination. After enumeration, it is found that some pair of *miR-210*, *miR-140* and *miR-1270* are more effective in dividing the sample group as shown in Fig. 7.

All the important miRNAs were also treated as a combination, the features were compressed into two dimensions by PCA as shown in Fig. 8.

From Figs. 7,8, it can be found that the cancer samples show some degree of separation from the normal samples, which verifies the validity of the important miRNAs.

Table 2. Classification results on the selected genes.

	TP rate	FP rate	Precision	Recall	F1 measure	Balance accuracy	Number of features
ensemble	0.895	0.006	0.944	0.895	0.919	0.985	15
t-test	0.579	0	1	0.579	0.733	0.943	25
limma	0.895	0.012	0.895	0.895	0.895	0.965	55
edgeR	0.737	0	1	0.737	0.848	0.988	73

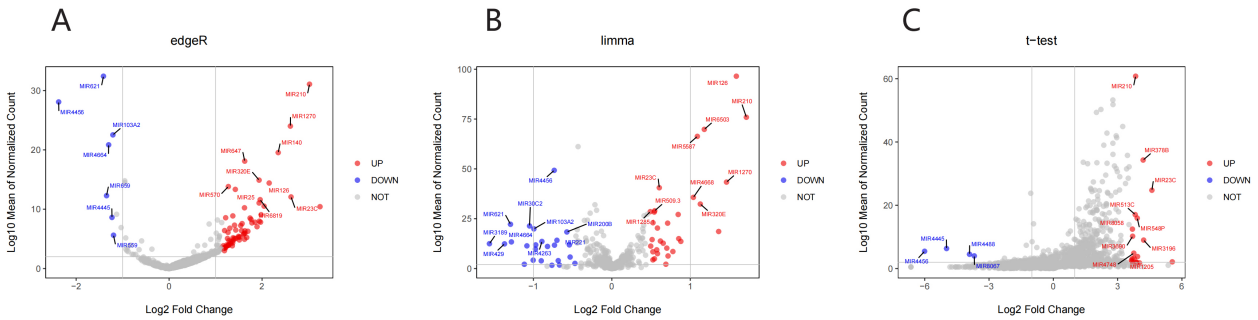

Fig. 5. Volcano plots of the differential genes. (A) The volcano plot of the differential genes searched by edgeR. (B) The volcano plot of the differential genes searched by limma. (C) The volcano plot of the differential genes searched by *t*-test. R packages were used to search for differential genes and draw corresponding volcano plots.

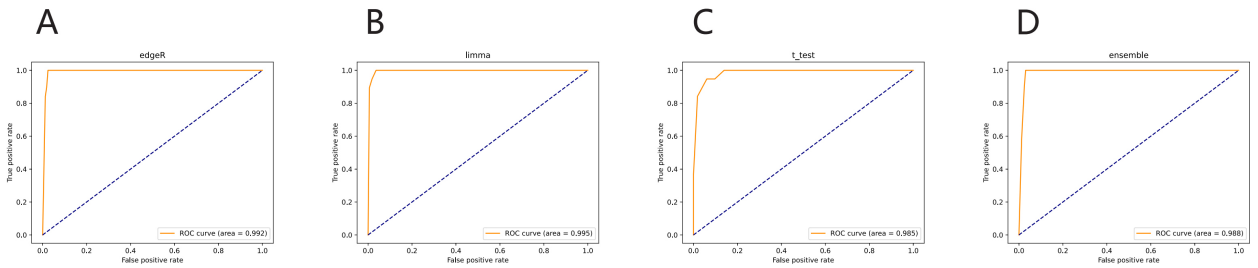

Fig. 6. ROC curves and AUCs. (A) ROC curves using edgeR searching for differential genes with AUC = 0.992. (B) ROC curves using limma searching for differential genes with AUC = 0.995. (C) ROC curves using two-sample *t*-test searching for differential genes with AUC = 0.985. (D) Receiver Operating Characteristic (ROC) curves using the ensemble method searching for important genes with Area Under Curve (AUC) = 0.988.

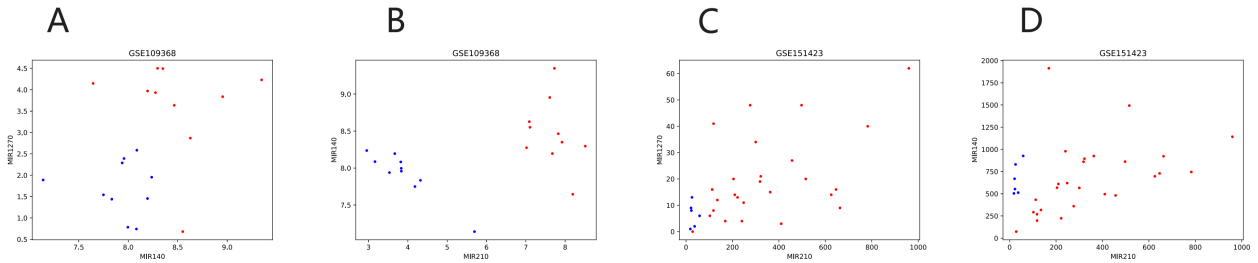

Fig. 7. The distribution of sample points. (A) The sample distribution on the GSE109368 data set using *miR-140* and *miR-1270* as a combination. (B) The sample distribution on the GSE109368 data set using *miR-210* and *miR-140* as a combination. (C) The sample distribution on the GSE151423 data set using *miR-210* and *miR-1270* as a combination. (D) The sample distribution on the GSE151423 data set using *miR-210* and *miR-140* as a combination. Samples in red represent cancer groups and samples in blue represent normal groups.

4. Discussion

It has been suggested that *miR210* is associated with the characteristics of gene expressions in the presence of cellular hypoxia [32–34]. As to *miR210*, the overexpression in tumors is a direct consequence of reduced oxygen tension in the microenvironment [35]. Moreover, the cellular response to hypoxia is partly a transcriptional process orchestrated through an oxygen detection mechanism cen-

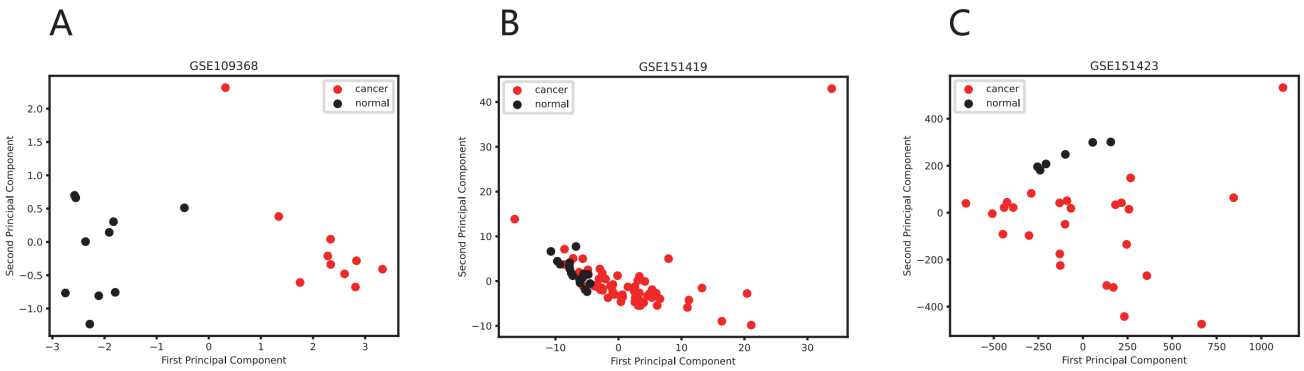

Fig. 8. The distribution of sample points using whole important genes. (A) The sample distribution on GSE109368 data set. (B) The sample distribution on GSE151419 data set. (C) The sample distribution on GSE151423 data set. Samples in red represent cancer groups and samples in black represent normal groups.

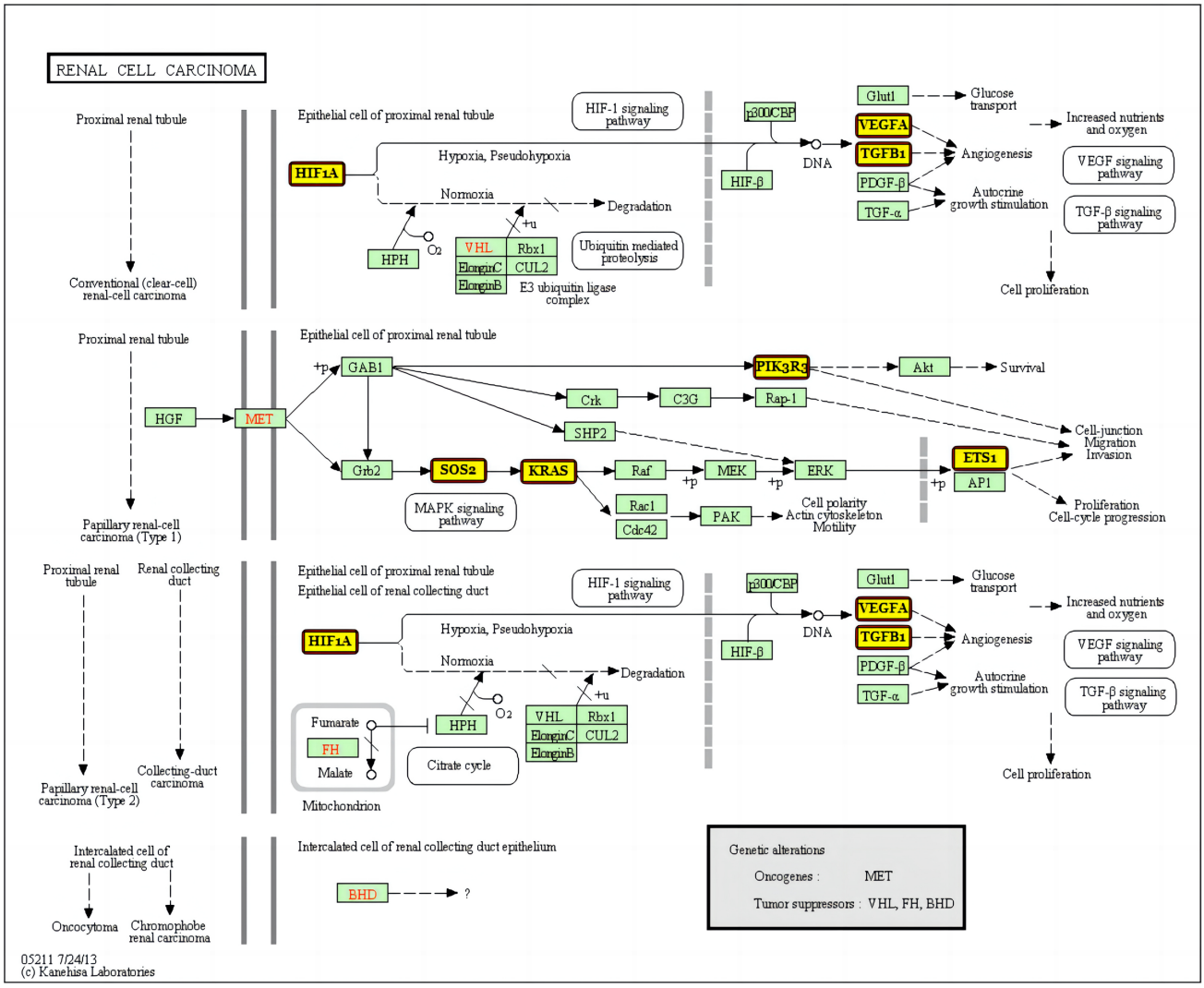

Fig. 9. KEGG Pathway of Renal cell carcinoma.

tered on the hypoxia-inducible factors (*HIFs*). Under conditions of normal oxygen content, *HIF* targets it for proteasomal destruction mediated by an E3 ubiquitin ligase containing Von Hippel-Lindau (*VHL*) protein.

And a specific disease closely related to the *HIF* pathway is KIRC, which is usually associated with the inactivation of the *VHL* tumor suppressor gene [36]. Mutations and heterozygous deletions of the *VHL* gene have been found in

57% and 98% of sporadic KIRC cases [37]. The *VHL* tumor suppressor gene product functions as an articulated subunit of the E3 ubiquitin ligase complex, which targets hydroxylated *HIF-1*α and *HIF-2*α for ubiquitination and subsequent degradation by the 26S proteasome [38,39]. Given its close relationship with *HIF*, it is not surprising that *miR-210* is specifically overexpressed in KIRC [40–42]. In addition, circulating *miR-210* levels are elevated in KIRC patients compared to healthy controls [43].

Besides, eight genes such as *HIF1A*, *VEGFA*, and *TGFB1* were found to interact with *miR-140* in the KEGG Pathway of 'Renal cell carcinoma'.

The eight genes are marked with yellow background and red border in Fig. 9.

miR-140 was also found to be associated with cell proliferation, migration and invasion by *in vitro* experiments [44]. *In vivo* knockdown of *miR-140* in mice revealed significant inhibition of renal cell carcinoma (RCC) tumor growth. Moreover, by analyzing the pathway of *miR-140*, it was found that this gene inhibited the expression of *KLF9* by binding to the 3′-UTR of *KLF9*, which could upregulate the expression of *KCNQ1* and thus reduce the growth and metastasis of RCC.

The functions of the other miRNAs were obtained by querying David's tool, the results of which are shown in **Supplementary Material**.

5. Conclusions

In this paper, we proposed a machine learning-based miRNA importance calculation method to find several subsets of miRNAs most associated with KIRC by analyzing the miRNA expressions of KIRC. Traditional differential analysis was used to find the differential genes. The classifier trained with the optimal feature subset, compared to the former, had close or better classification accuracy in a relatively small much feature dimension, which reflected the superiority of this method. Results on the independent testing set also verified the effectiveness of the found important miRNAs. The obtained important miRNAs were investigated and *miR-210* was found to be associated with cellular hypoxia response and *VHL* tumor suppressor genes. Besides, *miR140* was also involved in the biochemical processes of cell proliferation, migration and invasion. The identified important miRNAs all played a role in key signaling pathways for KIRC progression and metastasis, and represented potential targets for KIRC therapy. Some miRNAs associated with KIRC (e.g., *miR4456*, *miR1270*, *miR647* and *miR4664*) but not documented in the literature were also identified for further experimental validation.

Availability of Data and Materials

Publicly available datasets were analyzed in this study. TCGA data can be found at: https://portal.gdc.cancer.gov/repository and GEO data can be found at: https://www.ncbi.nlm.nih.gov/geo/ .

Author Contributions

QW conceived the general project and supervised it. EYZ initiated the idea, conceived the whole process, and finalized the manuscript. EYZ, and ZQX were the principal developers and performed the experiments. QW helped to modify the manuscript. All authors contributed to editorial changes in the manuscript. All authors read and approved the final manuscript. All authors have participated sufficiently in the work to take public responsibility for appropriate portions of the content and agreed to be accountable for all aspects of the work in ensuring that questions related to its accuracy or integrity.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

Not applicable.

Funding

This work has been supported by the financial support of the Heilongjiang Health Commission Project [grant numbers 2020-067]. The results shown are in whole or part based upon data generated by TCGA Research Network.

Conflict of Interest

The authors declare no conflict of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/j.fbl2811283.

References

[1] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA: A Cancer Journal for Clinicians. 2016; 66: 7–30.

[2] Ljungberg B, Campbell SC, Choi HY, Jacqmin D, Lee JE, Weikert S, *et al*. The epidemiology of renal cell carcinoma. European Urology. 2011; 60: 615–621.

[3] Pal SK, Williams S, Josephson DY, Carmichael C, Vogelzang NJ, Quinn DI. Novel therapies for metastatic renal cell carcinoma: efforts to expand beyond the VEGF/mTOR signaling paradigm. Molecular Cancer Therapeutics. 2012; 11: 526–537.

[4] Song E, Ma X, An R, Zhang P, Zhang X, Wang B, *et al*. Retroperitoneal Laparoscopic Partial Nephrectomy for Tumors Larger than 7 cm in Renal Cell Carcinoma: Initial Experience of Single-Institution. Journal of Laparoendoscopic & Advanced Surgical Techniques. Part a. 2017; 27: 1127–1131.

[5] Dhillon A, Singh A. eBreCaP: extreme learning-based model for breast cancer survival prediction. IET Systems Biology. 2020; 14: 160–169.

[6] Sinha A, Singh C, Parmar D, Singh MP. Proteomics in clinical interventions: achievements and limitations in biomarker development. Life Sciences. 2007; 80: 1345–1354.

[7] Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nature Clinical Practice. Oncology. 2008; 5: 588–599.

[8] Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010; 11: 422.
[9] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Research. 2008; 18: 1509–1517.
[10] Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007; 23: 2881–2887.
[11] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11: R106.
[12] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26: 139–140.
[13] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012; 40: 4288–4297.
[14] Chen Y, Lun ATL, Smyth GK. From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. F1000Research. 2016; 5: 1438.
[15] Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010; 11: 422.
[16] Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research. 2015; 43: e47.
[17] Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Research. 2011; 21: 2213–2223.
[18] Tarazona S, Furió-Tarí P, Turrà D, Pietro AD, Nueda MJ, Ferrer A, *et al*. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Research. 2015; 43: e140.
[19] Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, *et al*. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013; 29: 1035–1043.
[20] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology. 2013; 31: 46–53.
[21] Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American Journal of Botany. 2012; 99: 248–256.
[22] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods. 2011; 8: 469–477.
[23] Gao Y, Jia Y, Yu Z, Ji X, Liu X, Han L, *et al*. Analysis of the differential expression and prognostic relationship of DEGs in AML based on TCGA database. European Journal of Medical Research. 2023; 28: 103.
[24] Yuan L, Zeng G, Chen L, Wang G, Wang X, Cao X, *et al*. Identification of key genes and pathways in human clear cell renal cell carcinoma (ccRCC) by co-expression analysis. International Journal of Biological Sciences. 2018; 14: 266–279.
[25] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008; 9: 559.
[26] Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009; 136: 215–233.
[27] Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends in Molecular Medicine. 2014; 20: 460–469.
[28] Garzon R, Calin GA, Croce CM. MicroRNAs in Cancer. Annual Review of Medicine. 2009; 60: 167–179.
[29] Dhillon A, Kaur A, Singh A. Application of Machine Learning for Prediction of Lung Cancer using Omics Data. International Journal of Innovative Technology and Exploring Engineering (IJITEE). 2020; 9: 230–236.
[30] Singh A, Dhillon A, Kumar N, Hossain MS. eDiaPredict: an ensemble-based framework for diabetes prediction. ACM Transactions on Multimidia Computing Communications and Applications. 2021; 17: 1–26.
[31] Dhillon A, Singh A, Bhalla VK. A Systematic Review on Biomarker Identification for Cancer Diagnosis and Prognosis in Multi-omics: From Computational Needs to Machine Learning and Deep Learning. Archives of Computational Methods in Engineering. 2023; 30: 917–949.
[32] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research. 2011; 12: 2825–2830.
[33] Chen HY, Lin YM, Chung HC, Lang YD, Lin CJ, Huang J, *et al*. miR-103/107 promote metastasis of colorectal cancer by targeting the metastasis suppressors DAPK and KLF4. Cancer Research. 2012; 72: 3631–3641.
[34] Sarkar J, Gou D, Turaka P, Viktorova E, Ramchandran R, Raj JU. MicroRNA-21 plays a role in hypoxia-mediated pulmonary artery smooth muscle cell proliferation and migration. American Journal of Physiology. Lung Cellular and Molecular Physiology. 2010; 299: L861–L871.
[35] Huang X, Ding L, Bennewith KL, Tong RT, Welford SM, Ang KK, *et al*. Hypoxia-inducible mir-210 regulates normoxic gene expression involved in tumor initiation. Molecular Cell. 2009; 35: 856–867.
[36] Ivan M, Huang X. miR-210: fine-tuning the hypoxic response. Advances in Experimental Medicine and Biology. 2014; 772: 205–227.
[37] Brugarolas J. Renal-cell carcinoma–molecular pathways and therapies. The New England Journal of Medicine. 2007; 356: 185–187.
[38] Gnarra JR, Tory K, Weng Y, Schmidt L, Wei MH, Li H, *et al*. Mutations of the VHL tumour suppressor gene in renal carcinoma. Nature Genetics. 1994; 7: 85–90.
[39] Ivan M, Kondo K, Yang H, Kim W, Valiando J, Ohh M, *et al*. HIFalpha targeted for VHL-mediated destruction by proline hydroxylation: implications for O2 sensing. Science. 2001; 292: 464–468.
[40] Jaakkola P, Mole DR, Tian YM, Wilson MI, Gielbert J, Gaskell SJ, *et al*. Targeting of HIF-alpha to the von Hippel-Lindau ubiquitylation complex by O2-regulated prolyl hydroxylation. Science. 2001; 292: 468–472.
[41] White NMA, Bao TT, Grigull J, Youssef YM, Girgis A, Diamandis M, *et al*. miRNA profiling for clear cell renal cell carcinoma: biomarker discovery and identification of potential controls and consequences of miRNA dysregulation. The Journal of Urology. 2011; 186: 1077–1083.
[42] Juan D, Alexe G, Antes T, Liu H, Madabhushi A, Delisi C, *et al*. Identification of a microRNA panel for clear-cell kidney cancer. Urology. 2010; 75: 835–841.
[43] Redova M, Poprach A, Besse A, Iliev R, Nekvindova J, Lakomy R, *et al*. MiR-210 expression in tumor tissue and in vitro effects of its silencing in renal cell carcinoma. Tumour Biology. 2013; 34: 481–491.
[44] Zhao A, Li G, Péoc'h M, Genin C, Gigante M. Serum miR-210 as a novel biomarker for molecular diagnosis of clear cell renal cell carcinoma. Experimental and Molecular Pathology. 2013; 94: 115–120.