Comparisons of forecasting for Survival Outcome for Head and Neck Squamous Cell Carcinoma by using Six Machine Learning Models Based on Multi-Omics

Liying Mo
Guangxi Medical University
Yuangang Su
Guangxi Medical University
Jianhui Yuan
Guangxi Medical University
Zhiwei Xiao
Guangxi Medical University
Ziyan Zhang
Guangxi Medical University
Xiuwan Lan
Guangxi Medical University
Daizheng Huang (huangdaizheng@gxmu.edu.cn)
Guangxi Medical University

Research Article

Keywords: Machine learning models, Multi-omics Integration, Head and Neck Squamous Cell Carcinoma, Survival prediction, Bayesian network

Posted Date: November 30th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1100398/v1

Version of Record: A version of this preprint was published at Current Genomics on February 4th, 2022. See the published version at https://doi.org/10.2174/1389202923666220204153744.

Abstract

Background: Machine learning methods showed excellent predictive ability in a wide range of fields. For the survival of head and neck squamous cell carcinoma (HNSC), its multi-omics influence is crucial. This study attempts to establish a variety of machine learning multi-omics models to predict the survival of HNSC and find the most suitable machine learning prediction method.

Results: For omics of HNSC, the results of the six models all showed that the performance of multi-omics was better than each single-omic alone. Results were presented which showed that the BN model played a good prediction performance (area under the curve [AUC] 0.8250) in HNSC multi-omics data. The other machine learning models RF (AUC = 0.8002), NN (AUC = 0.7200), and GLM (AUC = 0.7145) also showed high predictive performance except for DT(AUC = 0.5149) and SVM(AUC = 0.6981). And the results of a vitro qPCR were consistent with the Random forest algorithm.

Conclusion: Machine learning methods could better forecast the survival outcome of HNSC. Meanwhile, this study found that the Bayesian network was the most superior. Moreover, the forecast result of multi-omics was better than single-omic alone in HNSC.

1. Introduction

Cancer is a major public health problem and causes 1 in 6 deaths around the world[1]. Head and Neck Squamous Cell Carcinoma (HNSC), which arises from multiple anatomic subsites in the head and neck region is the seventh most common cancer worldwide. There is marked heterogeneity of tumors arising from the mucosal epithelium of the upper aerodigestivetract[2, 3]. The risk factors for the development of cancers of the oral cavity, oropharynx, hypopharynx, and larynx include tobacco exposure and alcohol dependence, and infection with oncogenic viruses is associated with cancers developing in the nasopharynx, palatine, and lingual tonsils of the oropharynx[4]. The high level of heterogeneity in HNSC along with the complex etiological factors makes the prognosis prediction deeply challenging. In the treatment of HNSC, a multispecialty team to evaluate the treatment choice is very important, since the head and neck cancers differ from the patients' statement, molecular change, and other environmental factors, such as alcohol and smoking. Tobacco smoking and alcohol drinking are used as all-cause mortality to diagnose HNSC[3, 5]. Surgery, radiation, and chemotherapy in various combinations are utilized for the treatment of HNSC[6]. But all of these treatments are associated with toxicity which can lead to different degrees of late organ dysfunction or other serious adverse reactions[7]. The main method of evaluating cancer development and providing survival estimation prognosis prediction is the main method of evaluating cancer development and providing survival estimation, which is mainly based on patients' clinical features and molecular profile[8].Some studies[9, 10]appliedpublic database such as TCGA and GEO datasets to identify biomarkers associated with the prognosis of cancer patients and predict the clinical outcome .Ghafouri-Fard Set al.[11]found the role of miRNA as prognostic biomarkers in HNSC. Although the single genomic analysis approaches have contributed towards the identification of cancer-specific mutations and molecular subtyping of tumors[12], single-omic only considers the role of single molecular biological information. And for HNSC, there was the little study of multi-omics data to its survival. We, therefore, attempted to apply multi-omics molecular biology information to predict the survival of HNSC, and at the same time, apply machine learning methods as a predictive tool.

Currently, many studies analyzed diseases not only from the level of gene expression alone but also from the multi-omics level. Multi-omics (mRNA, miRNA, DNA methylation, and copy number variation) is part of the prognostic effect. Multi-omics integration analysis and deep learning are used to predict high-grade patient survival and prognosis risk biomarkers[8, 13, 14]. HNSC is involved in a variety of complex mechanisms at the molecular level in vivo, and it is difficult to understand the development of cancer from gene-level alone, thereby making assessments to the patients. Integrative analyses that use information across the multi-omics profiling modalities promise to deliver more comprehensive insights into the survival prediction of cancer[15]. In the field of precision oncology, genomics approaches analyses have helped reveal several key mechanisms in cancer development, and several findings have been implemented in clinical oncology to help guide treatment decisions[16]. Moreover, multi-omics approaches can dissect the cellular response to chemo-/immunotherapy as well as discover molecular candidates with diagnostic/prognostic value[17]. In summary, multi-omics integration models driven by multi-omics data may help to overcome the chemo-/immunotherapy resistance phenotype of cancer cells rendering them vulnerable to targeted therapies and ultimately improving the quality of life of patients[17].

The classification of genomics data can be performed through machine learning (ML) algorithms to find significant features related to survival. Combined machine learning which is included random forest (RF), K-nearest neighbor (KNN), and artificial neural networks (ANN) is used to identify prognostic biomarkers in colorectal cancer and the performance of using RF is better[18]. Kaplan-Meier (KM), LASSO, and COX regression are performed to analyze the effect of CA9 on the survival of tongue squamous cell carcinoma (TSCC)[19]. Multi-omics data integration through machine learning (autoencoder and XGboost model) to construct an accurate and robust cancer prognosis prediction could cause abnormal C-index values fluctuations due to the neglect of tumor purity and known clinical data that affect the occurrence and development of tumors[20]. Therefore, the survival prediction that integrates multiple omics data and clinical data may acquire a robust and reliable prognostic prediction result. Yuri Fujinoetetal.[21]applied LASSO regression to predict the future visual field progression in glaucoma patients. A new algorithm based on LASSO called TG-LASSO was developed, which could predict clinical drug responseof cancer patients and identify genes related to drug response, including known targets genes and pathways related to the drug action mechanisms[22].Likewise, BN modeling has been used to develop decision-support tools in various oncologic diagnoses[23]. Thomas P Burghardtet al.[24]indicated that BN could explain the data set by defining the phenotype and pathogenicity of the given mutation position and the conditional probability of the residual substitution. And myypbc3 mutant disease was predicted via the neural/Bayesian network. Pau Llot et al.[25]used deep learning to do the performance about the genomic prediction of complex human traits.

In summary, a growing body of research has applied machine learning including LASSO algorithm, cox regression, BN, and neural network to analyze tumor data of integrated multi-omics. In this study, a combination of DNA methylation, a gene expression data analysis, a copy number variation, and a miRNA data analysis comprises a multi-omics integration. The aim of this study was to use machine learning modelto forecast the survival outcome of HNSC, and

compare the prediction performance in each omics. This comparison may be helpful in describing the hierarchal relationships between prognostic and outcome variables. Meanwhile, we aim to determine which machine learning model would be suitable for clinical use with decision curve analysis. Six machine learning methods: LASSO and BN combined model, RF, Neural Networks(NN), Generalized linear model(GLM), Decision Tree(DT), and Support Vector Machine(SVM), were used for prediction and their performances were compared.

2. Materials And Methods

2.1 The Datasets Source and Data Pre-processing

RNA-sequencing data (IlluminaHiSeq_RNASeqV2;Level 3), miRNA-seq data (IlluminaHiSeq_miRNASeq; Level 3), DNA methylation data (HumanMethylation450; Level 3), copy number variation(CNV) data (Affymetrix SNP 6.0 array; Level 3) and corresponding clinical information from HNSC were obtained from The Cancer Genome Atlas (TCGA) database (https://portal.gdc.cancer.gov/), in which the method of acquisition and application complied with the guidelines and policies. The clinical information of HNSC samples were downloaded as well.Meanwhile, the tumor samples ofthe multi-omicswere selected by filtering out the samples according to the nomenclature of TCGA sample IDs.

For the downloaded dataset, the data pre-processing and dimensionality reduction were required. The gene expression data and miRNA expression data were identified by comparison with tumor tissues and normal tissues expression and using the edgeR and DeSeq2 packages in R. Meanwhile, the chi-square test and Kruskal-Wallis test was used to reduce the number of genes in order to obtain DEGs in CNV data of HNSC. And for methylation data of HNSC, the limma package in R was used to filter DEGs.

Unless otherwise specified in the analysis of this paper, the programming language used is R (version 4.0.1).

2.2 Machine learning

Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO) is a regression-based algorithm that permits a large number of covariates in the model and has the function of penalizing the absolute value of the regression coefficient[26]. It is a linear regression method that uses L1 regularization, which can achieve the purpose of sparseness and feature selection. The LASSO regression is applied to data dimensionality reduction and feature selection due to its outstanding feature extraction and robust cancer prognosis[27].Formula 1 described the representation method of the minimum residual sum of squares of the LASSO algorithm.

$$\arg\min_{\beta}\left\{\sum_{i=1}^{q}\left(y_i-\sum_{j=1}^{m}x_{ij}\beta_j\right)^2\right\}\text{subject}\sum_{j=1}^{m}|\beta_j|=\gamma \qquad (1)$$

Bayesian Network

Bayesian Network (BN) is a multi-layered network of connections between clinical factors in a multi-omics data set that provides a multivariate mapping of complex data[28]. BN is a directed acyclic graph. Its nodes represent some random variables (Figure 1). Some of these random variables are observable and some are unobservable. Meanwhile, BN is a probabilistic graph model with a clear and transparent representation of the causal relationship between variables. Importantly, because the BN uses the posterior information of the data sets itself, it protects against over interpretation of the data. Survival predictions based on BN models have been developed for a number of tumors in order to improve prognostic estimates and to guide clinical decision making for appropriate treatment[29, 30].The BNs is one of the deep learning model methods, which also has the deep learning model's advantages. BNs with proper external validation could be useful as clinical decision support tools and provide clinicians and patients with information germane to the treatment of HNSC.

Decision Tree

The Decision Tree (DT) is a basic classification and regression method. The DT is composed of nodes and directed edges. The DT reflected the mapping relationship between features and tags as well. DT learning is a process of recursively selecting the optimal feature, and segmenting the training data according to the feature so that each sub-data set has the best classification process.

Generalized linear model

The Generalized linear model (GLM) is based on the exponential distribution family, and the prototype of the exponential distribution family is as Formula 2.

$$P(y;\eta)=b(y)\cdot\exp\left(\eta^{T}T(y)-a(\eta)\right) \qquad (2)$$

Where η is a natural parameter, it may be a vector. *T(y)* is called a sufficient statistic.

Random forest

Random forest (RF) is an integrated learning method based on decision trees. At the same time, RF is also an improvement to the bagging algorithm. The process of RF was shown in the Figure2.

Neural Networks

Neural Networks (NN) is a two-stage regression or classification model. NN is a complex network system formed by a large number of simple processing units (called neurons) widely connected to each other. It is a highly complex nonlinear dynamic learning system. The network diagram was shown in Figure 3.

Support Vector Machine

Support Vector Machine (SVM) is a generalized linear classifier that classifies data binary in a supervised learning manner, and its decision boundary is the maximum-margin hyperplane that solves the learning sample. SVM can perform non-linear classification through the kernel method and is a classifier with sparsity and robustness.

Evaluation index of performance: AUC

Since the accuracy rate cannot fully evaluate the performance of the models, this study considered another evaluation indicator, namely AUC. AUC is the performance indicator to measure the pros and cons of machine learning. AUC is the abbreviation of Area Under roc Curve. As the name implies, the value of AUC is the size of the area under the ROC curve. The definition of AUC is given below:

$$\text{AUC value: } sensitive = \frac{\text{TP}}{\text{P}};\ \ secitificity = \frac{\text{TN}}{\text{FP+TN}};$$

The ROC curve is drawn by two variables. The abscissa is *1-specificity*, and the ordinate is *sensitivity*.

The meaning of these characters was shown here: *TP* represented the actual number of positive samples predicted as positive samples, *TN* represented the actual number of negative samples predicted as negative samples, *FP* represent edactually negative samples were predicted to be the number of positive samples, and *FN* represented the actual positive samples were predicted to be the number of negative samples.

2.3 Survival prediction process

By preprocessing the downloaded TCGA clinical data and omics data, the 490 HNSC samples shared by multi-omics were obtained. Likewise, DEGs were also obtained separately from each single-omic through preprocessing.

After the data pre-processing, the Lasso algorithm was used to select important variables for the survival outcome of HNSC from mRNA data, miRNA data, DNA methylation data, and CNV data. Random forest was used to calculate the ratio of each screened important variable. Integrated the four single omics and then six machine learning models (namely BN, RF, NN, GLM, DT, SVM) were performed to predict the survival outcome for HNSC. Likewise, using single-omic data as model input was performed to predict survival outcomes as well. Among them, the 490 HNSC samples were randomly divided into 3 groups, of which 2/3 were used as the training set and 1/3 were used as the test set. All mentioned models were operated 10 times.

Measure and compare the test results with performance indicators to find out which machine learning algorithms were effective and which omics were the most accurate for predicting HNSC survival. The flowchart for the main process of the study is presented in Figure 4.

2.4. In vitro experimental

Cell lines and culture

A normal human immortalized keratinocytes (Hacat) cell line and three HNSC cell lines (Cal-27, SCC-9 and FaDu) were used in the present study. All cell lines were obtained from the Cell Bank of the Chinese Academy of Sciences. Hacat, Cal-27 and FaDu cell lines were cultured in Dulbecco's Modified Eagle Medium (DMEM) and SCC-9 was cultured in Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12(DMEM/F12) in 5% CO2 at 37°C. All media was supplemented with 10% fetal bovine serum (FBS) and 1% penicillin streptomycin. All cell culture reagents were purchased from Gibco, Thermo Fisher Scientific company.

Quantitative Real-time PCR(qPCR) assay

Cells were seeded at a density of 10^5 cells per well in a 6-well plate and cultured overnight. Total RNA was extracted from cultured cells using TRIzol reagent (Invitrogen). Complementary DNA (cDNA) was synthesized using Transcriptor First Strand cDNA Synthesis Kit (Roche), in accordance with the manufacturer's instructions. Quantitative reverse-transcription PCR was performed with Fast Start Essential DNA Green Master (Roche) and special primer sequences (Table1). Relative mRNA expression was quantified by the comparative Ct (ΔCt) method and normalized to the internal control gene, ACTB.

Table 1

Primers sequences

Gene	Primer forward (5'→3')	Primer reverse (5'→3')
AQP5	GCCACCTTGTCGGAATCTACT	CCTTTGATGATGGCCACACG
ACTN3	GCCCGATCGAGATGATGATGG	GGCAGTGAAGGTTTTCCGCT
TAC1	GGGACTGTCCGTCGCAAAAT	ACAGGGCCACTTGTTTTTCA
ZFR2	ATGGCTACCTACCAGGACAGT	GTATCCCGAGGACAAGGTGC
MMP11	GATCGACTTCGCCAGGTACT	CAGTGGGTAGCGAAAGGTGT
ACTB	TCACCATGGATGATGATATCGC	ATAGGAATCCTTCTGACCCATGC

3. Results

3.1 The Datasets Source and Data Pre-processing

The HNSC multi-omics data downloaded by TCGA included mRNA expression data, miRNA expression data, DNA methylation data, CNV data, and 528 clinical data containing clinical information. The multi-omics samples downloaded from TCGA were screened and compared, and 490 tumor samples that the multi-omics both shared were obtained (Figure 5). From data pre-processing, we obtained 299 mRNA genes, 62 miRNA genes, 40 CNV genes, and 299 DNA methylation genes. Figure 6 showed the 40 top genes from the four single-omic data after DEGs.

3.2 Machine learning results

The parameter settings of the machine learning method used were shown in Table 2.

Table 2

Parameter in machine learning model.

Method	parameter
Bayesian Networks, BN	Hill Climbing, maximum likelihood estimation method
Random forest, RF	Select the number of trees corresponding to the smallest OOB error, mtry=sqrt(M),where M is the total number of features
Neural Networks, NN	linout=F, size=10, decay=0.001, iteration ordinal number=1000, The hidden node n2 and the input node n1 in the three-layer NNET were related by n2=2n1+1
Generalized linear model, GLM	family = "binomial", eliminate variables with p<0.05
Decision Tree, DT	method=class, parms=default
Support Vector Machine, SVM	method="C-classification", kernel="radial", cost=10，gamma=0.1

Before utilizing the machine Learning model to predict the survival outcome of HNSC, the LASSO algorithm was applied to select core genes in each single-omic. The LASSO algorithm took each single-omic data as the model input and took the event data of the clinical information as the model output. The results of each single-omics core gene obtained through the LASSO algorithm were shown in Table 3.And core genes that LASSO selected were integrated.

Table 3

Single-omiccore genes selected by LASSO algorithm.

Data type	Number of genes	Gene name
mRNA	21	*CST4,MSTN,ADH4,ACTN3,SLC13A2,TMEM210,MUC19,TAC1,METTL21C,AQP5,ADIPOQ,KCNJ16,CKM,DYNAP,GPRC6A,C14orf180,H*
miRNA	7	hsa-mir-378c,hsa-mir-411,hsa-mir-375,hsa-mir-499a,hsa-mir-503,hsa-mir-301a,hsa-mir-4776
Methylation	11	*RP11-266E6.3,AC007906.1,RP11-24M17.4,LRRC34,AC133644.2* *RNU2-37P,TCF24,CTD-2540M10.1,SLITRK1,AC093787.1,HIST1H4A*
Copy number variation, CNV	4	*FGF19,MMP11,PLIN1,HOXD13*

The data of each single-omic core gene selected from the LASSO algorithm were performed by the six machine learning models. The machine learning models took each single-omic core gene obtained from the LASSO algorithm as the model input and took the event data of the clinical information as the model output. Figure 7 provided the results. A comparison of Figure 7 reveals among the model predictions among HNSC omics, the prediction effect of mRNA was the best. The BN model had high predictive capacity (area under the curve [AUC] 0.7687), which was superior to that of other machine learning models. Besides, the CNV showed the worst performance in the survival outcome of HNSC. The accuracy for predicting the survival outcome was highest for NN (AUC 0.6220), followed by BN (AUC 5980).And overall, the prediction performance of BN in the four omic showed the best with 0.7687 on mRNA, 0.6418 on miRNA, 0.6325 on methylation and 0.5980 on CNV. Meanwhile, the predictive performance of miRNA data and methylation data were average, with AUC greater than 0.6 but less than 0.7.For the prediction performance of CNV, it could be inferred that CNV may have little effect on the survival outcome of HNSC.

The integrated core genes were selected again by the LASSO algorithm and then the 36 multi-omics genes that affect the occurrence of HNSC could be obtained (Figure 8). The 36 core genes that secondary selected were calculated to calculate the contribution of each gene. Moreover, both the results in Figure 7 and Figure 8 implied that CNV may have little effect on the survival outcome of HNSC.

The 36 genes that were secondary screened by LASSO were integrated. Meanwhile, the integrated multi-omics was used as the input of machine learning, and the event data of the clinical information were used as the output to predict the survival outcome of HNSC. Figure 9 presented the results of machine learning methods performed in multi-omics. Compared the performance with each single-omic in Figure 7 the multi-omics played the best performance. The AUC of the six machine learning models in multi-omics were 0.8250 on BN, 0.8002on RF, 0.7200on NN, 0.5149on DT, 0.7145on GLM and 0.6981 on SVM. Except for the machine learning performance in DT, the multi-omics data had the best forecast of HNSC survival outcome. Furthermore, whether it was the result of single-omic or multi-omics, BN played the best predictive effect. The prediction performance of RF ranked behind BN. Overall, these results suggested that the applying of multi-omics data to predict the survival outcome of HNSC was better than the applying of single-omic data alone. Likewise, the prediction performance of the BN model was better than other machine learning models as well. Together these results provide important insights that applying the LASSO algorithm to select the contributing variables and BN model to multi-omics data to predict the survival outcome may get a good performance.

3.3 qPCR results

Through the LASSO algorithm, we selected 5genes to verify *in vitro*, we detected their mRNA expression levels in a normal Hacat cell line and three HNSC cell lines. Cells were seeded with a number of 105 per well at 6-well plates overnight and detected via RT-qPCR. The mRNA expression levels of *AQP5*, *ACTN3*, *TAC1*, *ZFR2*, and *MMP11* were evaluated. The results were presented as mean ± SEM. *p< 0.05, **p< 0.01, and ***p< 0.001 as compared with the Hacat cell line group (Figure 10). We found that the mRNA expression of *MMP11* and *ZFR2* was significantly increased, meanwhile, the mRNA expression of the three genes *AQP5*, *ACTN3*, and *TAC1* was significantly decreased in the HNSCC cell line (Figure10). This is consistent with our model prediction of gene expression levels in HNSC.

4. Discussion

Although HNSC's recent advances have brought substantial improvements in outcome, it is still cancer with poor long-term survival due to the lack of specific therapeutic targets to predict its survival outcome[2]. Therefore, it is crucial to identify a robust method to predict the survival outcome of HNSC to evaluate its development and provide the survival estimation.

This study set out to assess the application of six machine learning models in multi-omics integration data to predict the survival outcome in HNSC. The obvious finding to emerge from the analysis was that the multi-omics prediction shows good performance compared with each single-omic data prediction effect. The prediction accuracy of multi-omics integrated data was better than that of single-omics prediction alone in general. Furthermore, the prediction effect of mRNA in six machine learning models was better than other single-omic, namely miRNA, methylation and CNV. The result may partly be explained by the mRNAs playing a key role in the development of HNSC related pathways and protein expression. And why the multi-omics integration data made the best prediction performance may probably be due to the multi-omics data integrate the molecular level information of each single-omic data in HNSC. The results indirectly showed multi-omics data could more comprehensively reflect the association between molecular level and HNSC survival outcome than single-omics data. According to the results, we can infer that integrating more data related to HNSC survival outcome can get better prediction performance.

In addition, by comparing the performance metrics with other machine learning models, as shown in Figure 7 and Figure 9,the results found that the prediction efficiency of the BN model was better than that of other machine learning models. Moreover, in all types of data, the BN model played the best predictive effect

among the six machine learning models except the miRNA expression data. In the integrated multi-omics data, the AUC of the BN model was 0.8250.These results suggested that the BN model could be a suitable model for survival outcome in HNSC. Comprehensive the performance indicator, the BN model made the best performance in predicting the survival outcome of HNSC. Although the RF model prediction performance was lower than the BN model, we found that the prediction performance of the RF model ranked second among these models. This result may be explained the fact that why current studies were fond of applying the RF as a model for predicting disease[31-33].The results indicated that the BN model could be a robust model to apply to omics data in predicting survival outcome. In summary, these results implied that the machine learning models especially the BN model could predict the survival outcome of HNSC. In other words, the BN model was more suitable for HNSC survival prediction, especially in HNSC data that combined with multiple omics.

The current study found that the Cox proportional hazards regression was widely used to predict survival and the prediction results were reliable[34, 35].In general, therefore, it seems that the Cox proportional hazards regression could be used to compare the performance with the previously mentioned models to further confirm the predictive performance of BN in HNSC(Figure 11).Surprisingly, the 0.82 C-index value in the Cox proportional hazard regression indicated that the performance of the Cox proportional hazard regression was consistent with BN. But the unexpected finding was the weight values that represented the influencing factors of multi-omics genes on overall HNSC's event in the last column of Figure 11 were different from the results predicted by the LASSO algorithm shown in Figure 8.For the inconsistent result, the importance of core genes was calculated again by RF(Figure 12).This finding was unexpected and suggested that the cox model and the BN model may have the same prediction of survival results, but the cox model and the LASSO regression model may have inconsistent results in the variable screening. This disappointing result might be explained by the fact that different model parameters cannot be ruled out. And we may need to perform further experiments *in vitro* to verify the result. In general, this intriguing result implied the BN model was suitable for applying multi-omics data and suitable for predicting the survival outcome.

The results of *in vitro* verification of the core genes selected by the random forest suggested that these genes could serve as therapeutic targets and poor prognostic factors for HNSC. At the same time, in the Cox proportional hazard regression and BN models, these genes indicated that they have a greater impact on the survival outcome of HNSC. Importantly, prior studies have noted *TAC1*wasa powerful epigenetic biomarker in HNSC[36]. Meanwhile, *ZFR2*,*AQP5*, and *ACTN3* were found on the association between tumors such as cervical cancer, prostate cancer, acute myeloid leukemia, colorectal cancer, and breast cancer[37-41]. However, the *MMP11* gene has not been studied in HNSC. The result of this study may suggest that *MMP11*could serve as a novel biomarker for the diagnosis in HNSC, while the matrix metalloproteases (MMP) family were related to pan-cancer, especially with HNSC[42].

These findings may be somewhat limited firstly by the lack of predicting the combination of other personal factors such as smoking condition and alcohol condition with multi-omics data, and the lack of combined predictions of multi-omics data and data that fully describe the prognosis of cancer, such as TMN stage, radiotherapy, and chemotherapy. Moreover, at the molecular level, we neglected to combine proteomics data with multi-omics to predict the survival outcome of HNSC. This may be one of the reasons that the AUC that describes the prediction performance was lower than 0.8. Secondly, this paper did not conduct experimental verification of the core genes screened. Despite these flaws, these results further support the idea of applying LASSO and BN combined model to multi-omics integration data to predict the survival outcome. And these results show that the use of machine learning methods, especially BN methods, is robust and accurate in predicting the survival outcome of HNSC.

5. Conclusion

Applying multi-omics integration data to machine learning has important implications for predicting survival outcomes. One of the strengths of this study is that it is the multi-omics integration data machine learning analysis to date. It applied the LASSO and six machine learning models across four HNSC single-omics types including mRNA, miRNA, methylation, and CNV to predict the affecting HNSC variables and survival outcome. This study set out to explore whether the LASSO and six machine learning models based on multi-omics integrated data could be robust in predicting the survival outcome of HNSC. However, this study had some limitations as it didn't combine other variables that may affect the survival progress of HNSC. And the further experiment of verifying the genes predicted by machine learning that affect the survival and development of HNSC didn't perform. Despite these limitations, the findings of this study are still valuable. The machine learning models especially the BN model are expected to become a practical prediction model for tumor survival and prognosis. The multi-omics integration data could bring more information about the molecular level to better predict survival outcome. And better clinical services may bring new ideas to the precise prognosis and treatment of tumors.

Abbreviations

HNSC: head and neck squamous cell carcinoma

ML: machine learning

KNN: K-nearest neighbor

ANN: artificial neural networks

KM: Kaplan-Meier

TSCC: tongue squamous cell carcinoma

LASSO: least absolute shrinkage and selection operator

BN: Bayesian network

RF: random forest

NN: neural networks

DT: decision tree

GLM: generalized linear model

SVM: support vector machine

CNV: copy number variation

MMP: matrix metalloproteases

Declarations

Author contributions

Daizheng Huang contributed to the study's conception and design. Data collection and analysis were performed by Liying Mo. Cell culture and qPCR experiments were performed by Yuangang Su. The first draft of the manuscript was written by Liying Mo and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability

All data generated or analyzed during this study are included in this article.

Acknowledgements

The authors would like to thank the BaGui scholar program of Guangxi Province, Guangxi Natural Science Foundation Innovation Research Team under Grant 2019GXNSFGA245002, Guangxi Natural Science Foundation under Grant 2018GXNSFAA281133, the National Natural Science Foundation of China under Grant 81860604.

Conflict of interest

The authors declare that they have no competing interests.

References

1. Siegel RL, Miller KD and Jemal A. Cancer statistics, 2020. *CA Cancer J Clin* 2020; 70: 7-30. 2020/01/09. DOI: 10.3322/caac.21590.
2. Chow LQM. Head and Neck Cancer. *N Engl J Med* 2020; 382: 60-72. 2020/01/02. DOI: 10.1056/NEJMra1715715.
3. Beynon RA, Lang S, Schimansky S, et al. Tobacco smoking and alcohol drinking at diagnosis of head and neck cancer and all-cause mortality: Results from head and neck 5000, a prospective observational cohort of people with head and neck cancer. *Int J Cancer* 2018; 143: 1114-1127. 2018/04/03. DOI: 10.1002/ijc.31416.
4. Torre LA, Bray F, Siegel RL, et al. Global cancer statistics, 2012. *CA Cancer J Clin* 2015; 65: 87-108. 2015/02/06. DOI: 10.3322/caac.21262.
5. Mourad M, Jetmore T, Jategaonkar AA, et al. Epidemiological Trends of Head and Neck Cancer in the United States: A SEER Population Study. *J Oral Maxillofac Surg* 2017; 75: 2562-2572. 2017/06/16. DOI: 10.1016/j.joms.2017.05.008.
6. Colevas AD, Yom SS, Pfister DG, et al. NCCN Guidelines Insights: Head and Neck Cancers, Version 1.2018. *J Natl Compr Canc Netw* 2018; 16: 479-490. 2018/05/13. DOI: 10.6004/jnccn.2018.0026.
7. Marur S and Forastiere AA. Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment. *Mayo Clin Proc* 2016; 91: 386-396. 2016/03/06. DOI: 10.1016/j.mayocp.2015.12.017.
8. Chaudhary K, Poirion OB, Lu L, et al. Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clin Cancer Res* 2018; 24: 1248-1259. 2017/10/07. DOI: 10.1158/1078-0432.CCR-17-0853.
9. Chen L, Lu D, Sun K, et al. Identification of biomarkers associated with diagnosis and prognosis of colorectal cancer patients based on integrated bioinformatics analysis. *Gene* 2019; 692: 119-125. 2019/01/18. DOI: 10.1016/j.gene.2019.01.001.
10. Ren N, Liang B and Li Y. Identification of prognosis-related genes in the tumor microenvironment of stomach adenocarcinoma by TCGA and GEO datasets. *Biosci Rep* 2020; 40 2020/10/06. DOI: 10.1042/BSR20200980.
11. Ghafouri-Fard S, Gholipour M, Taheri M, et al. MicroRNA profile in the squamous cell carcinoma: prognostic and diagnostic roles. *Heliyon* 2020; 6: e05436. 2020/11/19. DOI: 10.1016/j.heliyon.2020.e05436.
12. Hu F, Zeng W and Liu X. A Gene Signature of Survival Prediction for Kidney Renal Cell Carcinoma by Multi-Omic Data Analysis. *Int J Mol Sci* 2019; 20 2019/11/20. DOI: 10.3390/ijms20225720.

13. Vantaku V, Dong J, Ambati CR, et al. Multi-omics Integration Analysis Robustly Predicts High-Grade Patient Survival and Identifies CPT1B Effect on Fatty Acid Metabolism in Bladder Cancer. *Clin Cancer Res* 2019; 25: 3689-3701. 2019/03/09. DOI: 10.1158/1078-0432.CCR-18-1515.
14. Yin Z, Yan X, Wang Q, et al. Detecting Prognosis Risk Biomarkers for Colon Cancer Through Multi-Omics-Based Prognostic Analysis and Target Regulation Simulation Modeling. *Front Genet* 2020; 11: 524. 2020/06/13. DOI: 10.3389/fgene.2020.00524.
15. Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* 2018; 14: e8124. 2018/06/22. DOI: 10.15252/msb.20178124.
16. Olivier M, Asmis R, Hawkins GA, et al. The Need for Multi-Omics Biomarker Signatures in Precision Medicine. *Int J Mol Sci* 2019; 20 2019/09/29. DOI: 10.3390/ijms20194781.
17. Chakraborty S, Hosen MI, Ahmed M, et al. Onco-Multi-OMICS Approach: A New Frontier in Cancer Research. *Biomed Res Int* 2018; 2018: 9836256. 2018/11/08. DOI: 10.1155/2018/9836256.
18. Maurya NS, Kushwaha S, Chawade A, et al. Transcriptome profiling by combined machine learning and statistical R analysis identifies TMEM236 as a potential novel diagnostic biomarker for colorectal cancer. *Sci Rep* 2021; 11: 14304. 2021/07/14. DOI: 10.1038/s41598-021-92692-0.
19. Guan C, Ouyang D, Qiao Y, et al. CA9 transcriptional expression determines prognosis and tumour grade in tongue squamous cell carcinoma patients. *J Cell Mol Med* 2020; 24: 5832-5841. 2020/04/17. DOI: 10.1111/jcmm.15252.
20. Chai H, Zhou X, Zhang Z, et al. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. *Comput Biol Med* 2021: 134-115. 2021 May 9. DOI: 10.1101/807214.
21. Fujino Y, Murata H, Mayama C, et al. Applying "Lasso" Regression to Predict Future Visual Field Progression in Glaucoma Patients. *Invest Ophthalmol Vis Sci* 2015; 56: 2334-2339. 2015/02/24. DOI: 10.1167/iovs.15-16445.
22. Huang EW, Bhope A, Lim J, et al. Tissue-guided LASSO for prediction of clinical drug response using preclinical samples. *PLoS Comput Biol* 2020; 16: e1007607. 2020/01/23. DOI: 10.1371/journal.pcbi.1007607.
23. Nandra R, Parry M, Forsberg J, et al. Can a Bayesian Belief Network Be Used to Estimate 1-year Survival in Patients With Bone Sarcomas? *Clin Orthop Relat Res* 2017; 475: 1681-1689. 2017/04/12. DOI: 10.1007/s11999-017-5346-1.
24. Burghardt TP and Ajtai K. Neural/Bayes network predictor for inheritable cardiac disease pathogenicity and phenotype. *J Mol Cell Cardiol* 2018; 119: 19-27. 2018/04/15. DOI: 10.1016/j.yjmcc.2018.04.006.
25. P B, G dLC and M P-E. Can deep learning improve genomic prediction of complex human traits? *Genetics* 2018; 210: 809-819. 2018 Aug 31. DOI: 10.1534/genetics.118.301298.
26. McEligot AJ, Poynor V, Sharma R, et al. Logistic LASSO Regression for Dietary Intakes and Breast Cancer. *Nutrients* 2020; 12 2020/09/04. DOI: 10.3390/nu12092652.
27. Ying Q, Chen C and Xiaoyi L. Multi-Feature Fusion Combined with Machine Learning Algorithms to Quickly Screen Uveitis. *Journal of Xinjiang University(Natural Science Edition in Chinese and English)* 2021; 38: 439-449. DOI: 10.13568/j.cnki.651094.651316.
28. Stojadinovic A, Bilchik A, Smith D, et al. Clinical decision support and individualized prediction of survival in colon cancer: bayesian belief network model. *Ann Surg Oncol* 2013; 20: 161-174. 2012/08/18. DOI: 10.1245/s10434-012-2555-4.
29. A S, C E, L H, et al. Development of a Bayesian classifier for breast cancer risk stratification: a feasibility study. *Eplasty* 2010: 10-25.
30. Wang Q, Chen R, Cheng F, et al. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat Neurosci* 2019; 22: 691-699. 2019/04/17. DOI: 10.1038/s41593-019-0382-7.
31. Ni S, Qian Z, Yuan Y, et al. Schisandrin A restrains osteoclastogenesis by inhibiting reactive oxygen species and activating Nrf2 signalling. *Cell Prolif* 2020; 53: e12882. 2020/09/02. DOI: 10.1111/cpr.12882.
32. Ubels J, Schaefers T, Punt C, et al. RAINFOREST: a random forest approach to predict treatment benefit in data from (failed) clinical drug trials. *Bioinformatics* 2020; 36: i601-i609. 2021/01/01. DOI: 10.1093/bioinformatics/btaa799.
33. Yang L, Wu H, Jin X, et al. Study of cardiovascular disease prediction model based on random forest in eastern China. *Sci Rep* 2020; 10: 5245. 2020/04/07. DOI: 10.1038/s41598-020-62133-5.
34. Matsuo K, Purushotham S, Jiang B, et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. *Am J Obstet Gynecol* 2019; 220: 381 e381-381 e314. 2018/12/26. DOI: 10.1016/j.ajog.2018.12.030.
35. Shen Y, Peng X and Shen C. Identification and validation of immune-related lncRNA prognostic signature for breast cancer. *Genomics* 2020; 112: 2640-2646. 2020/02/23. DOI: 10.1016/j.ygeno.2020.02.015.
36. Misawa K, Mima M, Imai A, et al. The neuropeptide genes SST, TAC1, HCRT, NPY, and GAL are powerful epigenetic biomarkers in head and neck cancer: a site-specific analysis. *Clin Epigenetics* 2018; 10: 52. 2018/04/24. DOI: 10.1186/s13148-018-0485-0.
37. Direito I, Madeira A, Brito MA, et al. Aquaporin-5: from structure to function and dysfunction in cancer. *Cell Mol Life Sci* 2016; 73: 1623-1640. 2016/02/04. DOI: 10.1007/s00018-016-2142-0.
38. Rubicz R, Zhao S, Geybels M, et al. DNA methylation profiles in African American prostate cancer patients in relation to disease progression. *Genomics* 2019; 111: 10-16. 2016/02/24. DOI: 10.1016/j.ygeno.2016.02.004.
39. Yang X, Pang Y, Zhang J, et al. High Expression Levels of ACTN1 and ACTN3 Indicate Unfavorable Prognosis in Acute Myeloid Leukemia. *J Cancer* 2019; 10: 4286-4292. 2019/08/16. DOI: 10.7150/jca.31766.
40. Zhang L, Jiang Y, Lu X, et al. Genomic characterization of cervical cancer based on human papillomavirus status. *Gynecol Oncol* 2019; 152: 629-637. 2018/12/26. DOI: 10.1016/j.ygyno.2018.12.017.

41. Zhu Z, Jiao L, Li T, et al. Expression of AQP3 and AQP5 as a prognostic marker in triple-negative breast cancer. *Oncol Lett* 2018; 16: 2661-2667. 2018/07/18. DOI: 10.3892/ol.2018.8955.
42. Gobin E, Bagwell K, Wagner J, et al. A pan-cancer perspective of matrix metalloproteases (MMP) gene expression profile and their diagnostic/prognostic potential. *BMC Cancer* 2019; 19: 581. 2019/06/16. DOI: 10.1186/s12885-019-5768-0.

Figures

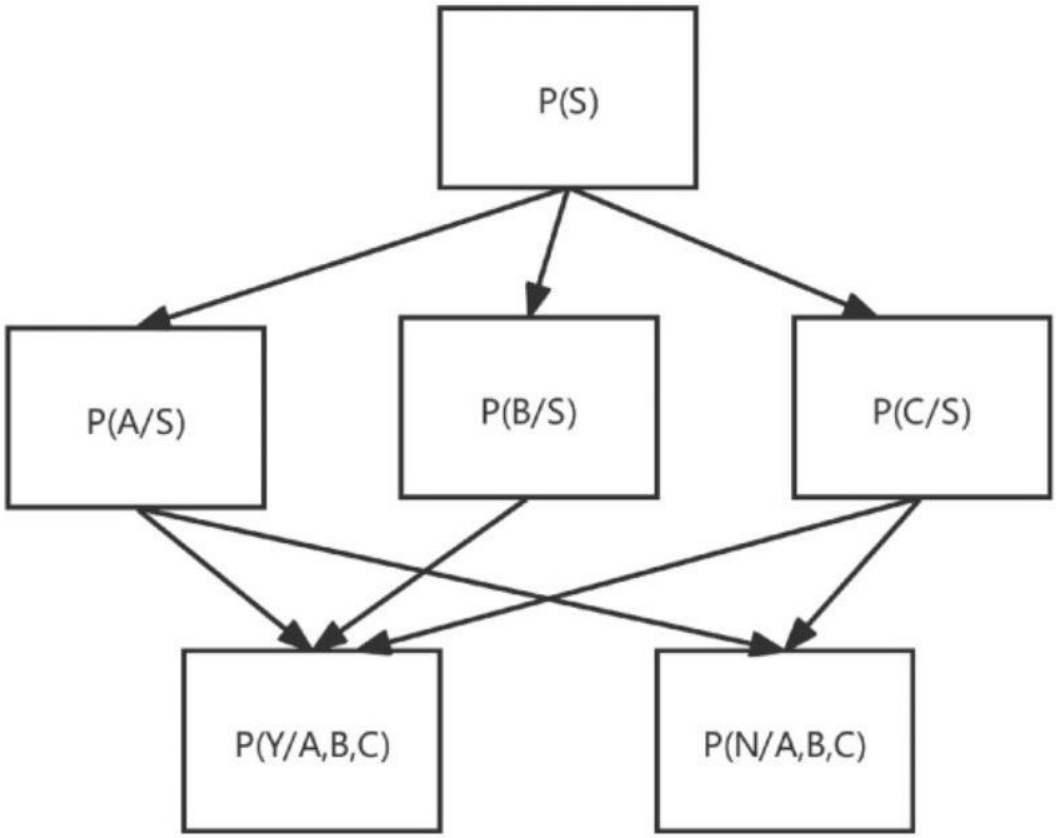

Figure 1

A simple BN network diagram.

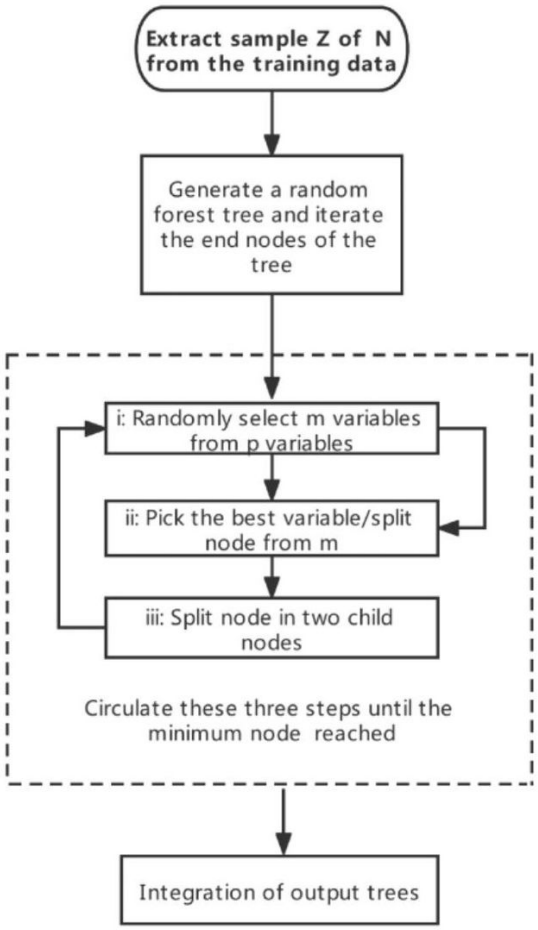

Figure 2

The process of RF.

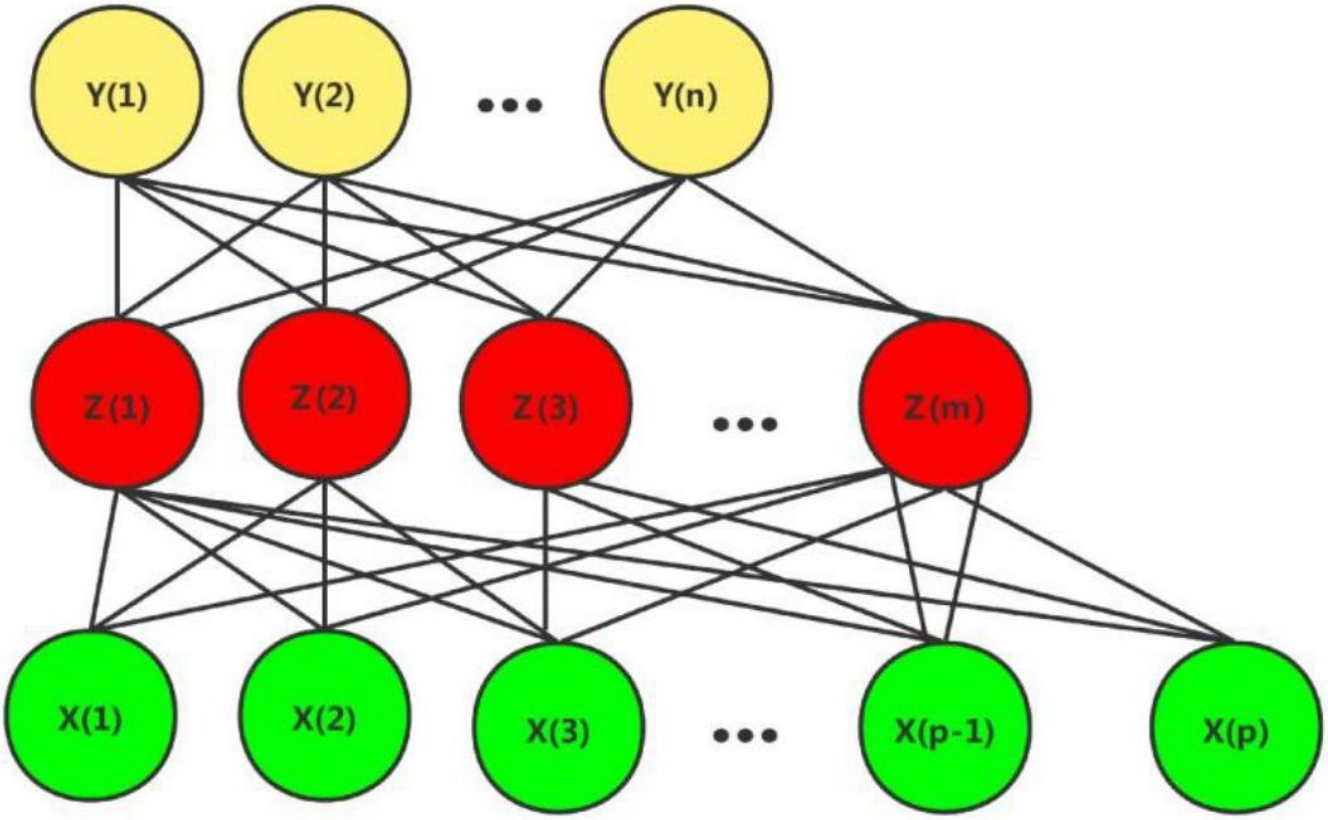

Figure 3

The single hidden layer neural network diagram.

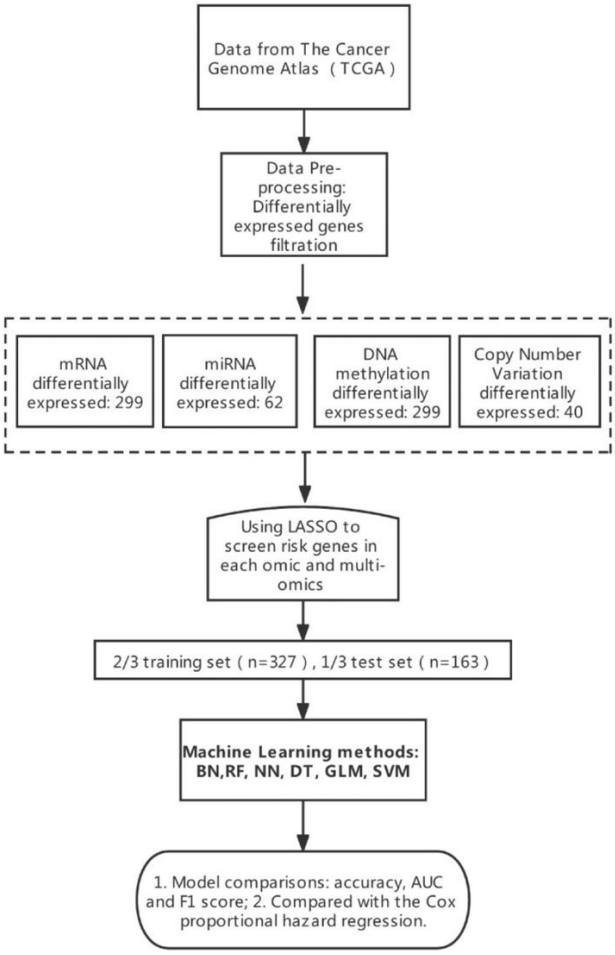

Figure 4

The main process of the research.

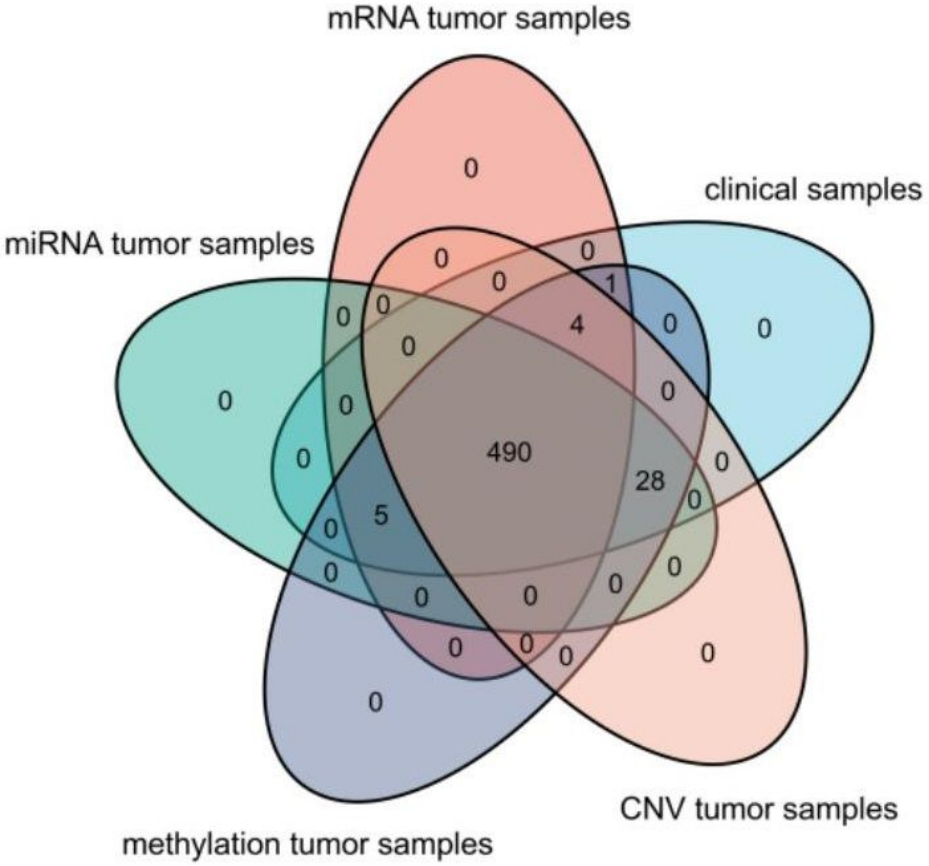

Figure 5

The screened out of the shared tumor samples from TCGA.

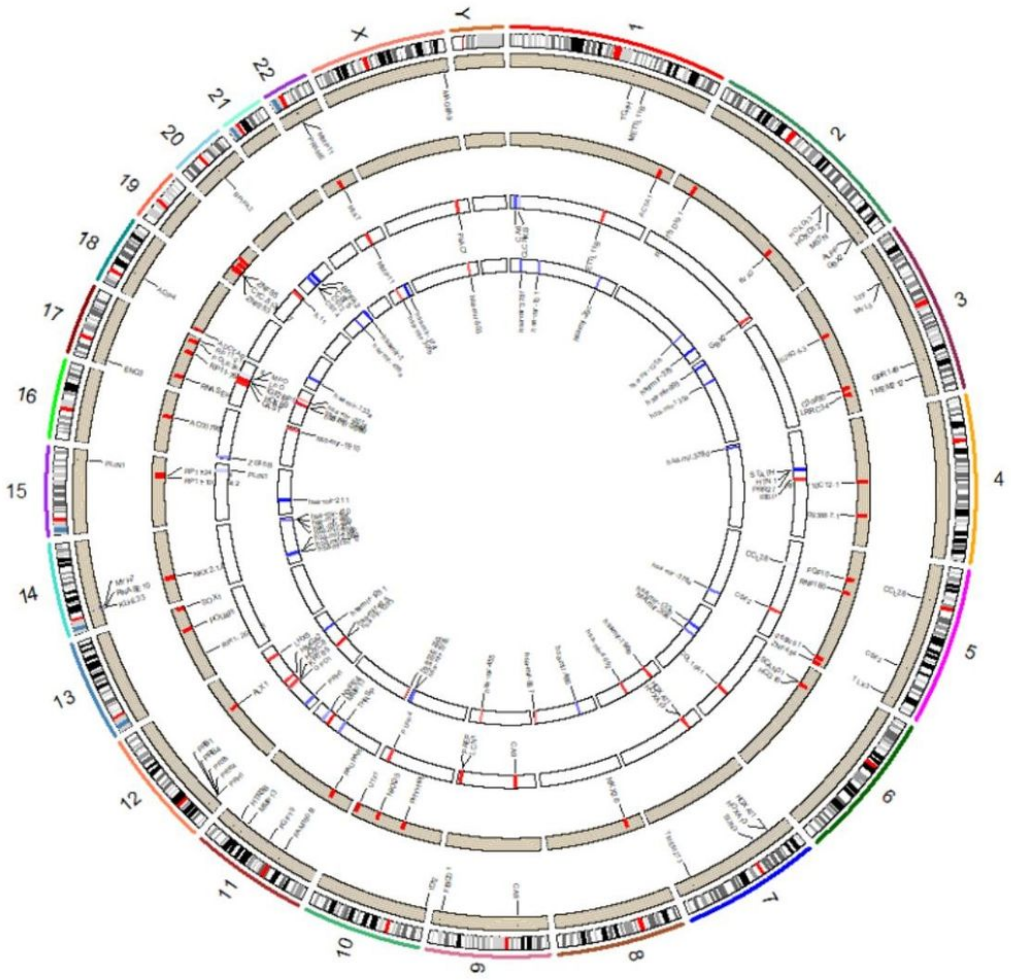

Figure 6

The results of DEGs in TCGA. Legend: The most inner circle was the top 40 differentially expressed miRNAs in HNSC, the second inner circle was the top 20 up-regulated mRNAs and top 20 down-regulated mRNAs in HNSC, the third inner circle was top 40 differentially expressed methylation genes in HNSC, and the outer circle was CNV identification differentially genes.

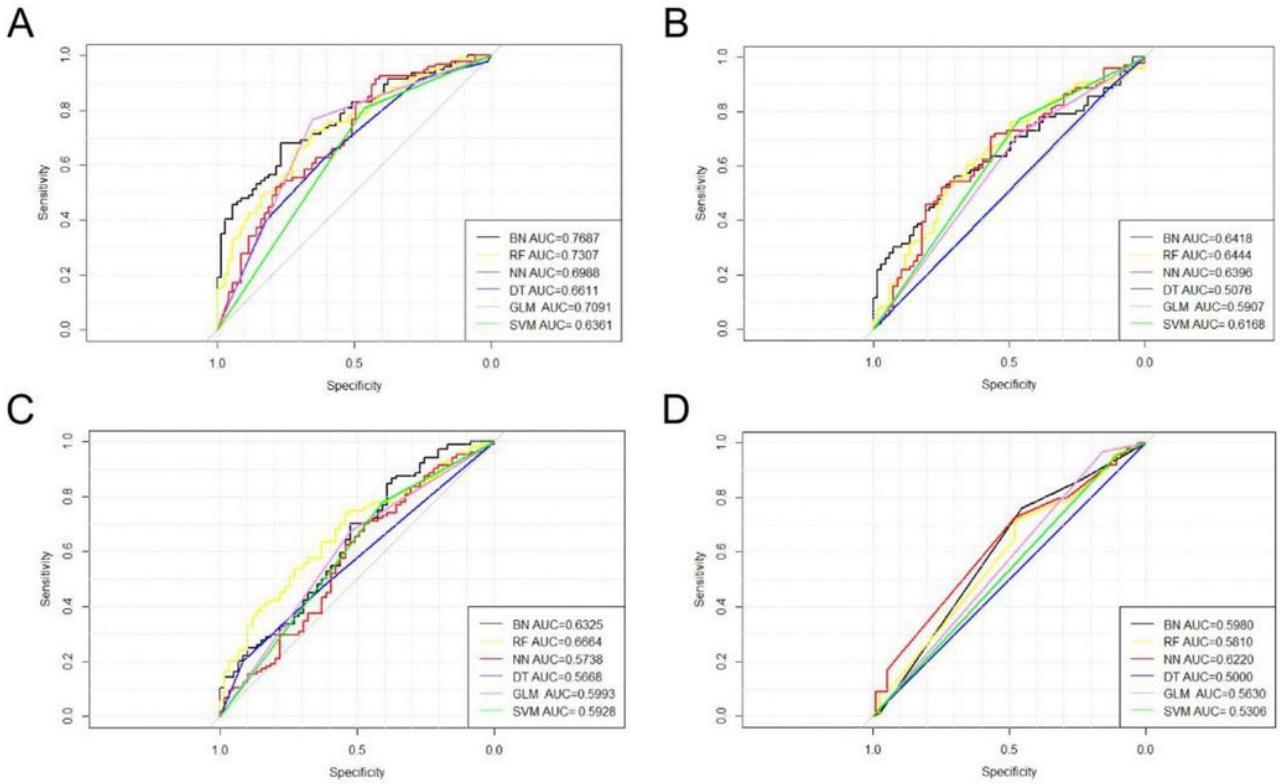

Figure 7

Comparison of single-omic prediction results applying by the six machine learning. (A)the ROC curve of mRNA prediction results. (B) the ROC curve of miRNA prediction results. (C)the ROC curve of methylation prediction results. (D) the ROC curve of CNV prediction results.

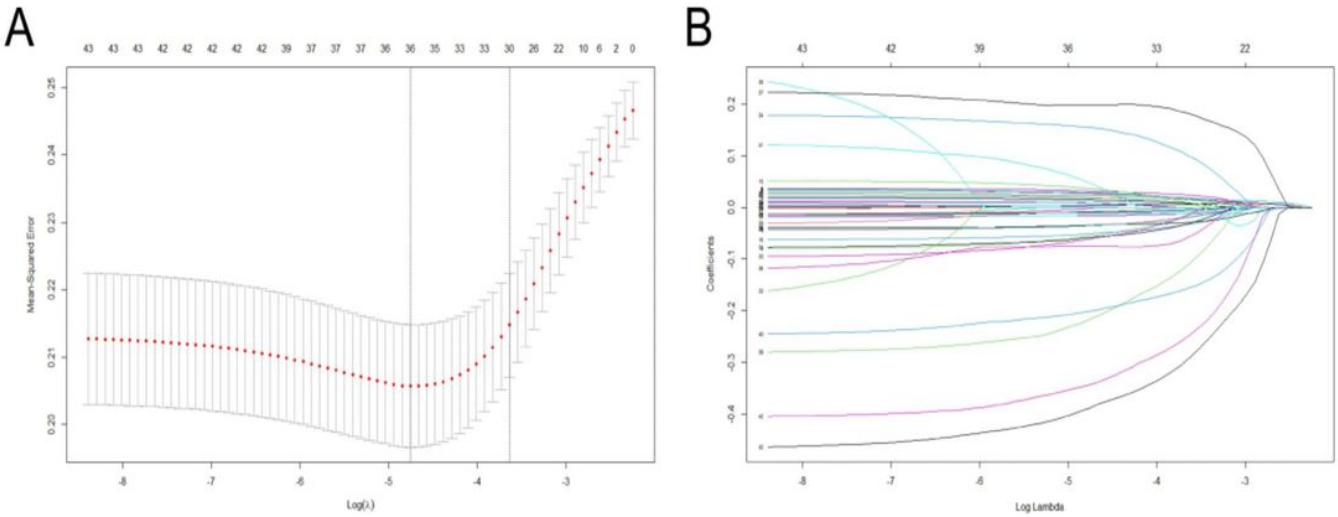

Figure 8

The secondary selected results through LASSO.(A) To determine the penalty value at the lowest point of between lines determines. (B) Showed the contribution of variables.

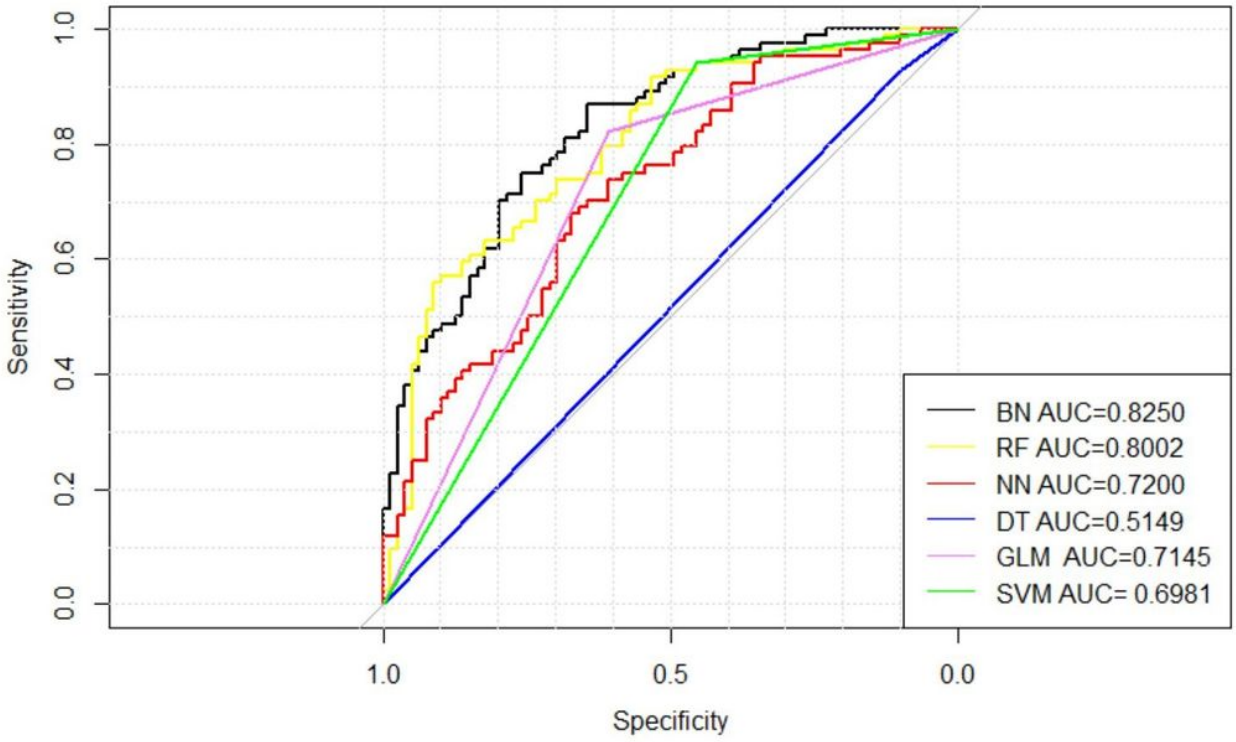

Figure 9

Comparison of the machine learning models prediction results in multi-omics.

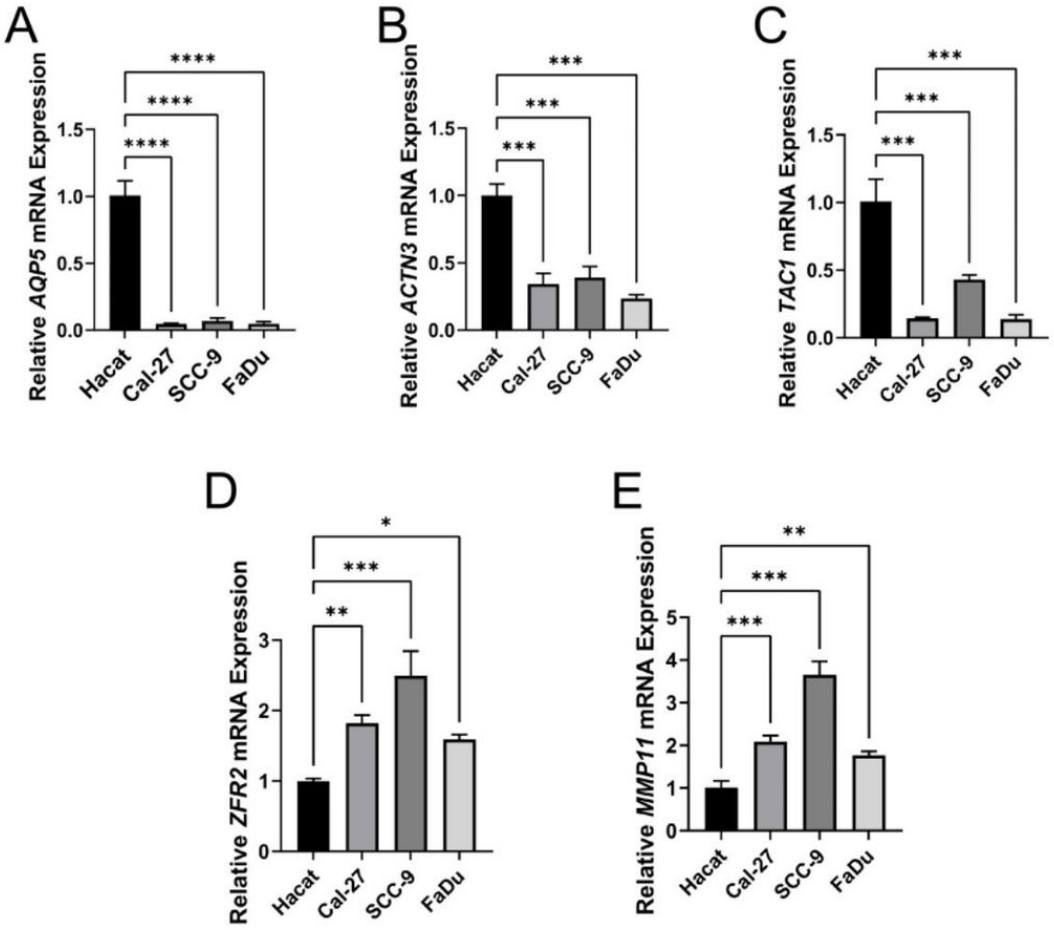

Figure 10

The mRNA expression of 5genes. (A)AQP5 expression;(B)ACTN3expression; (C)TAC1 expression; (D)ZFR2 expression; (E)MMP11 expression.

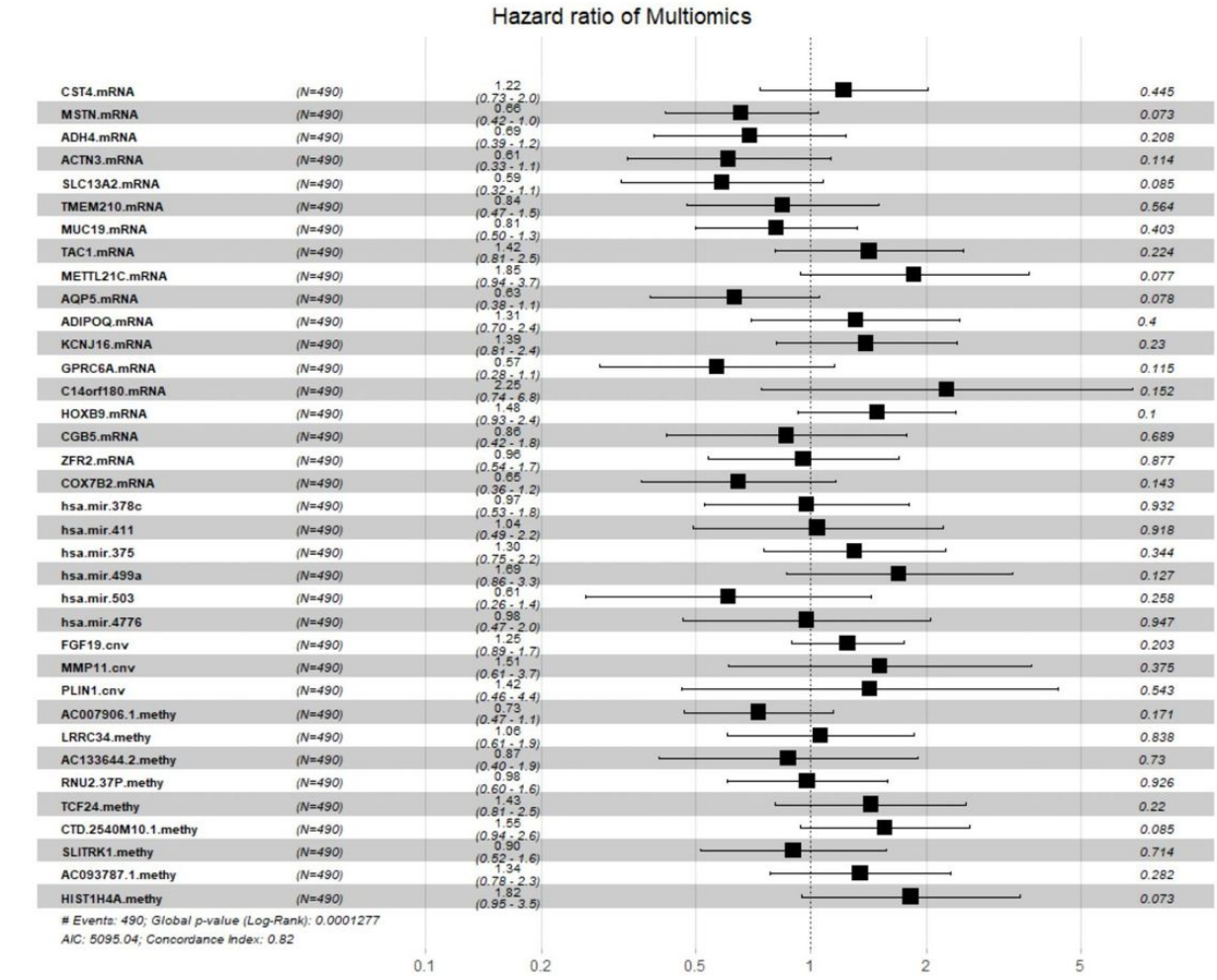

Figure 11

The prediction results in Cox proportional hazard regression.

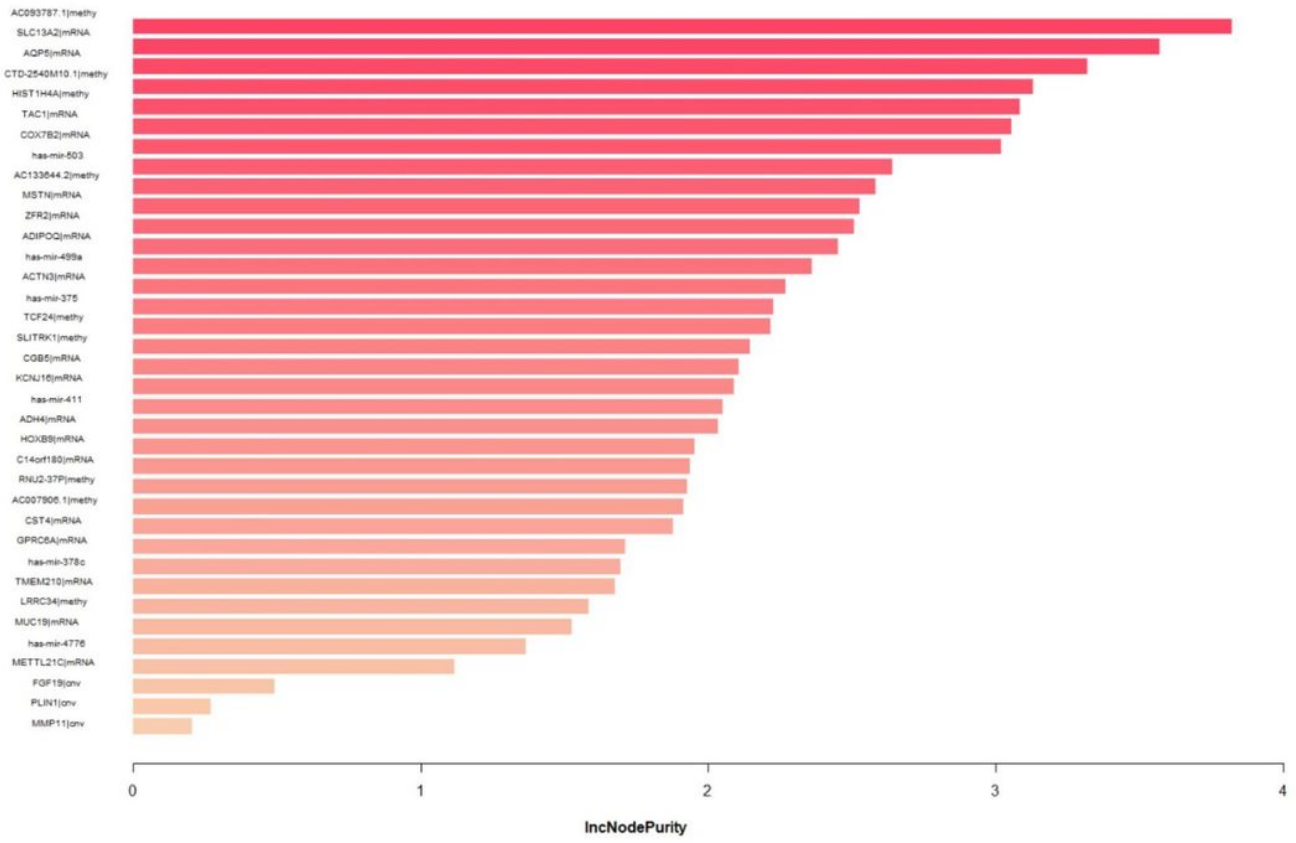

Figure 12

The feature of each core gene was obtained through Random forest.