Li et al. BMC Cancer (2021) 21:906
https://doi.org/10.1186/s12885-021-08647-1

BMC Cancer

RESEARCH ARTICLE Open Access

Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma

Meng-Xiang Li[1,2†], Xiao-Meng Sun[2,3†], Wei-Gang Cheng[4], Hao-Jie Ruan[2], Ke Liu[1,2], Pan Chen[2], Hai-Jun Xu[2], She-Gan Gao[2], Xiao-Shan Feng[1,2*] and Yi-Jun Qi[2*]

Abstract

Background: A plethora of prognostic biomarkers for esophageal squamous cell carcinoma (ESCC) that have hitherto been reported are challenged with low reproducibility due to high molecular heterogeneity of ESCC. The purpose of this study was to identify the optimal biomarkers for ESCC using machine learning algorithms.

Methods: Biomarkers related to clinical survival, recurrence or therapeutic response of patients with ESCC were determined through literature database searching. Forty-eight biomarkers linked to recurrence or prognosis of ESCC were used to construct a molecular interaction network based on NetBox and then to identify the functional modules. Publicably available mRNA transcriptome data of ESCC downloaded from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) datasets included GSE53625 and TCGA-ESCC. Five machine learning algorithms, including logical regression (LR), support vector machine (SVM), artificial neural network (ANN), random forest (RF) and XGBoost, were used to develop classifiers for prognostic classification for feature selection. The area under ROC curve (AUC) was used to evaluate the performance of the prognostic classifiers. The importances of identified molecules were ranked by their occurrence frequencies in the prognostic classifiers. Kaplan-Meier survival analysis and log-rank test were performed to determine the statistical significance of overall survival.

Results: A total of 48 clinically proven molecules associated with ESCC progression were used to construct a molecular interaction network with 3 functional modules comprising 17 component molecules. The 131,071 prognostic classifiers using these 17 molecules were built for each machine learning algorithm. Using the occurrence frequencies in the prognostic classifiers with AUCs greater than the mean value of all 131,071 AUCs to rank importances of these 17 molecules, stratifin encoded by SFN was identified as the optimal prognostic biomarker for ESCC, whose performance was further validated in another 2 independent cohorts.

* Correspondence: samfeng137@hotmail.com; qiyijun@haust.edu.cn
†Meng-Xiang Li and Xiao-Meng Sun contributed equally to this work.
[1]School of Information Engineering of Henan University of Science and Technology, 263 Kaiyuan Road, Luolong Qu, Luoyang 471023, P. R. China
[2]Henan Key Laboratory of Microbiome and Esophageal Cancer Prevention and Treatment; Henan Key Laboratory of Cancer Epigenetics, Cancer Hospital, The First Affiliated Hospital, College of Clinical Medicine, Medical College of Henan University of Science and Technology, 24 Jinghua Road, Jianxi Qu, Luoyang 471003, P. R. China
Full list of author information is available at the end of the article

Conclusion: The occurrence frequencies across various feature selection approaches reflect the degree of clinical importance and stratifin is an optimal prognostic biomarker for ESCC.

Keywords: Esophageal squamous cell carcinoma, Stratifin, Machine learning, Support vector machine, Random forest, Logical regression, Artificial neural network, eXtreme gradient boosting

Background

There are approximate 572,000 new cases of esophageal cancer (EC) worldwide in 2018, half of which arise in China [1, 2]. EC ranks sixth and fourth in the incidence and mortality of malignant tumors in China, respectively [3, 4]. The predominant histological subtypes of EC comprise esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC), among which ESCC accounting for at least 90% of EC in China [5, 6]. Epidemiological studies show that the risk factors of ESCC implicate cigarette smoking, genetic family history, nutritional deficiencies, pickled vegetables intake, hot food and beverage, low socioeconomic status, etc. [7, 8]. In sharp contrast, the increasing risk for EAC is associated with excess body weight and gastroesophageal reflux disorders, which are prevalent in western countries. Furthermore, heavy smoking contributes to an elevated risk of both ESCC and EAC. In the case of alcohol consumption, however, modest to moderate consumption is linked to a reduced risk in ESCC in China, and in EAC in western countries [9]. Heavy alcohol consumption is a strong and well-established risk factor for ESCC in western settings, and cigarette smoking plays a negligible role in ESCC etiology in a high-incidence area of China [8].

As such, it is not possible to distinguish ESCC patients with disparate clinical outcomes under the same exposure conditions based on the risk factors alone. On the other hand, "omics" studies are characterized by poor reproducibility, which could be ascribed to molecular heterogeneity, sample source, tissue processing, detection technique, data analysis, etc. Van't Veer et al. [10] from Netherlands and Wang et al. [11] from USA analyzed the differentially expressed genes in 295 and 286 cases with breast cancer using gene chip technology, respectively, from which the 70- and 76-signature gene sets for prognostic prediction were developed but with only 3 overlapping genes. Each performed well on its own dataset but not on other datasets. This was also the case for colorectal cancer [12]. It is well-accepted that tumor heterogeneity increases the risk of recurrence and metastasis of tumor patients after treatment and even lead to the resistance to multimodality treatment [13, 14]. Recently, Lin et al. have revealed the molecular heterogeneity of ESCC and its biological significance for tumor development and metastasis from multiple cancers, and revealed the impacts of molecular heterogeneity on the occurrence, development, and prognosis of ESCC [15].

Machine learning is an important branch of artificial intelligence (AI), which provides a possible solution to the current problem of poor reproducibility in group learning. Generally, the machine learning algorithms are divided into weak classifier algorithm and strong classifier algorithm, such as logical regression (LR), support vector machine (SVM) and artificial neural network (ANN) as weak classifier algorithms, and random forest (RF) and eXtreme Gradient Boosting (XGBoost) as strong classifier algorithms. Machine learning algorithms have been widely used in medical science, especially in the diagnosis, prognostic prediction of patients with cancer. For example, Xu et al. identified 5 features among 31 features closely related to the prognosis of ESCC using the genetic algorithm, and established a new ESCC staging system MASAN, showing better prognostic prediction accuracy compared with the currently used TNM staging system [16]. In a prospective cohort study, four machine learning methods, including RF, LR, gradient lifting tree, and ANN, were employed to predict the risk of cardiovascular disease, and the performances were compared between machine learning algorithm and traditional method of ACC/AHA10 annual risk prediction model. The performance of the four machine learning algorithm models was superior [17].

Given the molecular heterogeneity of cancers, we hypothesized that key molecules could serve as genuine prognostic factors even in complicated interactions with other molecules. To further identify key prognostic biomarkers for ESCC, 48 clinically proven molecules associated with ESCC progression were used for subnetwork construction. Using all combinations of 17 component molecules from 3 functional modules, 5 different machine learning algorithms, including LR, SVM, ANN, RF and XGBoost, were used to develop prognostic classifiers. The importances of these 17 molecules were gauged according to the occurrence frequencies in the prognostic classifiers. The prognostic value of stratifin was validated in another 2 independent ESCC cohorts.

Methods

Literature search

Literatures related to the prognosis and treatment response of ESCC were retrieved from NCBI PubMed, Web of Science and Embase databases, published up to 31 December 2018, by two independent researchers. The key words for literature searching included "esophageal

squamous cell cancer", "prognosis or recurrence or resistance or sensitivity" and "chemotherapy or chemoradiotherapy". All relevant studies were retrieved.

Inclusion and exclusion criteria

We selected the studies using the following criteria: (1) clinical prognosis of patients with ESCC; (2) prediction of clinical response to chemotherapy or chemoradiotherapy; (3) clinical recurrence of ESCC; (4) retrospective and prospective cohort studies; (5) studies published in English. When disagreements occurred between reviewers, a third reviewer was invited for discussion of the eligibility of related studies.

Datasets downloads

Publicably available mRNA transcriptome data of ESCC from Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA) datasets included GSE53625 and TCGA-ESCC. GSE53625 included 179 patients with ESCC that were randomly divided into a training cohort of 134 patients and a test cohort of 45 patients. Since the GSE53625 data had been normalized in the original study [18] and all samples in the data set were paired samples, the difference between the expression values of cancer tissue and corresponding adjacent tissue was taken as the input data for all subsequent calculations. TCGA-ESCC contained 82 patients with ESCC, of which 37 Vietnamese patients with ESCC were used for an independent validation.

Patients and clinical samples

Eighty-six fresh-frozen ESCC with matched noncancerous mucosa samples were collected from the First Affiliated Hospital of Henan University of Science and Technology between 2012 and 2017. All ESCC patients received curative esophagectomy without preoperative neoadjuvant chemoradiotherapy.

Subnetwork construction

In this study, 48 molecules related to prognosis of ESCC were mapped and imported to NetBox (https://cbio.mskcc.org/tools/netbox/) to establish a molecular interaction subnetwork for network analysis [19]. NetBox, a java-based software tool, integrates four databases including the Human Protein Reference Database (HPRD), Reactome, NCI-Nature Pathway Interaction (PID) Database, and the MSKCC Cancer Cell Map. The shortest path between molecules in the network was defined as 1, denoting that molecules with direct interaction were selected as nodes of the subnetwork. Functional modules in the network were identified and degree of nodes were calculated by igraph R package.

Introduction of machine learning algorithms

This study used 5 machine learning algorithms, including LR, SVM, ANN, RF and XGBoost, to develop classifiers for prognostic classification.

The LR model is a generalized linear model, which is based on linear regression with a layer of Sigmoid function mapping. LR regression model is one of the most commonly used methods in medical research [20, 21].

SVM is a supervised learning method developed by Cortes and Vapnik in 1995 [22]. The support vectors are used to find the best hyperplane and then classify samples with different labels. The nonlinear features are mapped to the new high dimensional space by constructing a mapping function, and the inner product operation in the mapping space is simplified by kernel function to ensure that the results were equivalent, to achieve the linear separability of the samples. In this study, the Radial Basis Function (RBF) kernel function was used, and the RBF's transformation method was as follows:

$K(x, x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$, where σ is the hyperparameter controlled in accordance with deviation and error of variance.

Neural networks are an important machine learning technology and have widespread applications with advances of scientific computing capabilities such as supercomputers and quantum computing. In general, a neural network consists of an input layer, multiple hidden layers, and an output layer. The most important element in a neural network is the design of hidden layer and connection weight between neurons. Logistic regression belongs to the neural network with zero hidden layers.

RF and XGBoost are two integrated learning algorithms based on bagging and boosting algorithms, respectively. Integrated learning uses a certain method to learn multiple weak classifiers with some differences followed by combination of these classifiers. If the error rate of weak classifier is less than 0.5, the combination of multiple weak classifiers will gradually increase predictive ability and reduce classification error to achieve classification.

Development of classifiers

For 179 patients with ESCC samples, labels were assigned according to the survival time. Label 1 denotes the ESCC cases with survival times of more than 3 years and the remaining cases were labeled as 0. In the training cohort, cross-validation and parameter optimization were used to develop the models, and the test cohort was used for validation. Receiver operating characteristic (ROC) curve analysis was used to estimate predictive values of machine learning classifiers and the area under the curve AUC (area under ROC Curve) was calculated.

For each machine learning algorithm, 131,071 models representing various combinations of 17 selected features were established, and AUCs of the models in training and test cohort were calculated. During the development of classifiers, candidate classifiers were those classifiers with AUCs greater than the average of AUCs across all classifiers. Among all candidate classifiers, top 1000 models with the highest AUC values in test cohort were selected, and the occurrence frequencies of each molecule were counted in these 1000 classifiers. Top 5 molecules with the highest occurrence frequency were regarded as the important molecules of the corresponding machine learning algorithm.

The construction and testing of the classifiers in this study were implemented by using R 3.6.3. The weak classifier uses R packages such as bestglm, e1071, and nnet, and the integrated learning algorithm uses random forest and xgboost.

RNA extraction and quantitative RT-PCR

Total RNA of 86 pairs of ESCC samples with matched noncancerous tissues were isolated using Trizol reagent (Invitrogen, Carisbad, CA), and reverse transcription was performed using 1 μg of total RNA (Promega, USA). The primer pair for stratifin was as follows: forward primer, 5′-GACTACTACCGCTACCTGGC-3′, and reverse primer, 5′-GTTGGCGATCTCGTAGTGGA-3′. GAPDH was used as an internal standard and its primer pair was as follows, forward primer, 5′- GCCACATCGCTCAG ACACC − 3′, and reverse primer, 5′- GATGGCAACA ATATCCACTTTACC − 3′. Quantitative RT-PCR was performed in triplicate on an Applied Biosystems 7900 quantitative PCR system (Foster City, CA, USA). The Ct values were used for comparison using $2^{-\Delta\Delta Ct}$ method with GAPDH as the internal standard.

Statistical analysis

Differences of the quantitative data between 2 groups were performed using the unpaired or paired Student t-test. The relationship between the abundance of western blot and the expression level of SFN was analyzed by using linear regression. Overall survival was calculated from the date of surgery to the date of last follow-up or death. The Kaplan-Meier survival curves and log-rank tests were performed to determine the statistical significance of overall survival. All tests were 2-tailed and $P < 0.05$ were designated as significantly different.

Results

Prognostic biomarkers of esophageal squamous cell carcinoma

We initially retrieved 38 articles, which reported a total of 48 molecules associated with the clinical survival, recurrence or therapeutic outcome of ESCC patients (Table 1). In addition, a long non-coding RNAs LOC285194 and 6 microRNAs, including miR-23a, miR-24, miR-382, miR-7, and a combination of miR-133a and miR-133b, were identified as well. Due to their low numbers, these microRNAs and long non-coding RNA were excluded from this study. Thus, 48 unique molecules were included for subsequent study.

Identification of key prognostic molecules

Our approach for validating clinically proven molecules associated with prognosis of ESCC is summarized in Fig. 1. All 48 molecules were used to construct a protein-protein interaction network using NetBox. The shortest path between the molecules in the network was defined as 1, indicating that those molecules with direct interaction were retained as nodes in the network. This study is based on the local version of Java and Python using NetBox algorithm to define the functional modules. By inputting the Entrez ID of 48 molecules, 3 functional modules containing a total of 17 molecules as vertices and 19 edges were identified. A subnetwork of 16 molecules among these 17 molecules based on STRING database (https://string-db.org/) was built with 0.7 as the minimum interaction score (Fig. 2a).

Prognostic classification using 5 machine learning algorithms

Seeking to improve the predicative accuracy of ESCC prognosis, 5 different machine learning algorithms, including LR, SVM, ANN, RF and XGBoost, were leveraged for prognostic classification using the 17 prognostic molecules. Among the prognostic models with AUCs greater than the mean value of all AUCs of 131,071 models for each algorithm, the importances of those 17 prognostic molecules were weighted by their occurrence frequencies. Table 2 shows the top 5 important molecules identified by each machine learning algorithm and the intersecting molecule is SFN only (Fig. 2c), indicating that SFN may be the optimal prognostic biomarker for ESCC.

Correlation of stratifin mRNA and protein expression

Because we have reported that stratifin protein encoded by SFN by immunohistochemical assay was reduced significantly in ESCC compared with normal esophageal mucosa and intraepithelial neoplasia, the present study, however, revealed that stratifin mRNA expression was downregulated in ESCC compared with noncancerous tissues using an ESCC cohort of GSE53625. We assessed the correlation between stratifin protein and mRNA expression. Figure 2d shows that stratifin protein levels strongly correlate with its mRNA levels in ESCC tissues, detected by Western blot and by RT-PCR, respectively, suggesting that both the protein and mRNA expression

Table 1 Thirty-eight studies reporting 48 molecules that were associated with clinical survival, recurrence or therapeutic outcome of ESCC patients

Biomarkers	Prognosis/ recurrence /therapeutic outcome	PMID	Sample sizes	Conclusions
BRCA1	Prognosis and chemoradiotherapy	23,326,344	144	Low BRCA1 expression was an independent prognostic factor in cisplatin-based chemotherapy (HR 0.29, 95% CI 0.12–0.71; $P = 0.007$) or chemoradiotherapy (HR 0.12, 95% CI 0.04–0.37; $P < 0.001$) group.
CCNA2	Chemotherapy	23,205,070	48	The expression of cyclin A was an independent prognosis factor in patients with ESCC following paclitaxel-based chemotherapy.
CCND1	Prognosis and radiochemotherapy	9,988,238	172	Patients with cyclin D1-positive carcinomas showed significantly worse overall survival than patients with cyclin D1-negative carcinomas (HR 2.14, 95% CI 1.134–3.42; $P = 0.0038$)
CD163 & CD68	Prognosis and neoadjuvant chemotherapy	25,752,960	210	High infiltration of CD68+ macrophages and CD163+ macrophages was significantly associated with poor prognosis for patients undergoing neoadjuvant chemotherapy ($P = 0.057$, $P = 0.003$).
CD274	Prognosis and chemoradiotherapy	26,623,522	45	The higher PD-L1 H-scores had poorer overall survival (median 16.7 versus 32.9 months, P = 0.02) than those with lower H-scores. (HR 2.29, 95% CI 1.12–4.69; $P = 0.023$)
CD44 & PROM1	Prognosis	27,748,881	47	Patients with strong expression of CD44 or CD133 and those with a high ratio of CD133-positive tumor cells showed significantly poor prognosis regardless of the effect of chemotherapy. PROM1 (HR 5.05, 95% CI 1.12–4.69; P = 0.023)
CDKN2A	neoadjuvant chemotherapy	26,514,506	101	ESCC tumors that were found positive for p16 expression appeared to fall into responders group rather than non responders ($P = 0.008$) and reported with less mortality ($P = 0.048$).
CEA & KRT19	Chemoradiotherapy	19,863,186	84	CEA may be helpful in predicting the responsiveness in ESCC of primary lesions to CRT, with the effective rates (CR + PR) in CEA high and low groups of 58.3% (14/24) and 93.3% (56/60), respectively ($P =$ 0.013 and 0.013).
EGFR	Chemoradiotherapy	17,940,077	62	The difference in the CR rate between EGFR positive and -negative groups was significant (CR rate: 62% vs. 34%; $P = 0.037$).
CRP	Chemoradiotherapy	21,224,533	36	Serum CRP can predict of CRT response with an accuracy of 75%.
ERCC1	Chemotherapy	23,263,828	46	Patients with ERCC1 negative tumors had a higher treatment response than the ERCC1 positive group (radiological response rates; 92.3% vs.50%, $P = 0.013$).
FAM84B	Neoadjuvant chemoradiation	25,980,316	21	The fold-change of circulating FAM84B mRNA expression can predict the pCR with an AUC of 0.73.
FDXR	Prognosis and Neoadjuvant chemoradiation	26,637,858	50	Fdxr was significantly correlated with postoperative outcomes and an independent prognostic factor (HR 4.950, 95% CI 1.603–15.38; 0.012).
HOXC6 & HOXC8	Prognosis	24,525,058	274	HOXC6 and HOXC8 were independent prognostic factors in patients with ESCC. HOXC6: (HR 1.341, 95% CI 0.895–2.010; $P = 0.045$); HOXC8: (HR 1.657, 95% CI 1.146–2.395; $P = 0.007$).
IL6R	Prognosis	23,648,090	218	The sIL6R level was one of several significant independent predictors of an unfavorable outcome. (HR 3.20, 95% CI 1.34–7.53; $P = 0.008$)
MDM2 & MKI67	Prognosis and chemoradiotherapy	25,880,782	79	MDM2 and p16 are predictive markers for chemoradioresistance in cStageIII ESCC and Ki-67 is a prognostic marker following dCRT in cStageIII ESCC.
MLH1	Prognosis	18,053,639	51	The expression of hMLH1 is a potential marker of tumor response and survival.
MMS19	Chemoradiotherapy	25,892,874	103	High cytoplasmic MMS19 expression was associated with a good response to chemoradiotherapy (OR 11.5, 95% CI: 3.0–44.5; P < 0.001).
MT3	Prognosis	16,351,731	64	Esophageal squamous cell carcinomas with negative p53, positive CDC25B, and negative MT expressions respond well to CRT.
MUC13 & MUC20	Prognosis and neoadjuvant chemotherapy	26,323,930	186	The median survival time of patients with low MUC13/high MUC20 expression was significantly shorter than that of patients with high MUC13/low MUC20 expression (27.7 months vs. 59.5 months, $P =$

Table 1 Thirty-eight studies reporting 48 molecules that were associated with clinical survival, recurrence or therapeutic outcome of ESCC patients *(Continued)*

Biomarkers	Prognosis/ recurrence /therapeutic outcome	PMID	Sample sizes	Conclusions
				0.021; HR 0.531, 95% CI: 0.299–0.944; $P = 0.031$).
MUC4	neoadjuvant chemotherapy	26,673,820	186	Low expression of MUC4 and MUC20 in resection samples was significantly correlated with better TRG (tumor regression grade). MUC4 and MUC20 were identified as potential biomarkers for predicting the efficacy of neoadjuvant chemotherapy in ESCC patients.
NOTCH1 & PIK3CA	Prognosis and Chemotherapy	26,528,858	104	NOTCH1 mutations was correlated with shorter survival times and failed to respond to chemotherapy, whereas PIK3CA mutations pointed to better responses to chemotherapy and longer survival times than patients without PIK3CA mutations.
PTGS2	Prognosis and Chemoradiotherapy	21,437,756	58	Negative or weak expression of PTGS2 was correlated significantly with CRT response (OR 6.296, 95% CI 1.58–25.096; $P = 0.010$).
PTPN6	Prognosis	32,536,826	184	Elevated PTPN6 expression indicated longer OS (HR 1.123, 95% CI: 0.565–2.230; $P = 0.741$).
RAD51	Prognosis and recurrence	24,065,387	89	Rad51 expression in ESCC was associated with poor survival ($P = 0.0324$) and recurrence ($P = 0.0171$).
REG1A	Prognosis	23,645,481	177	REG1A expression was a significant prognostic factor (HR 3.095, 95% CI: 1.569–5.943; $P = 0.0015$).
SFN	Chemoradiation therapy	15,999,354	62	SFN-positive expressions were closely related to the response to CRT.
	Prognosis	24,743,601	278	Downregulation of 14–3-3σpredicts poor survival, suggesting that 14–3-3σmay be a biomarker for early detection of high-risk subjects and diagnosis of ESCC. (HR 0.466, 95% CI 0.251–0.866; $P = 0.016$).
	Prognosis	20,108,042	148	Reduced stratifin expression, T4 stage, lymph node metastasis, and distant metastasis were independent risk factors for worse prognosis in ESCC patients.
SGTA	Prognosis	23,939,810	120	SGTA expression indicated poor prognosis (RR 3.513, 95% CI: 2.161–9.791; $P = 0.016$).
TGFB1 & VEGFA	Prognosis	24,623,035	79	VEGFA and TGFB1 were significantly associated with pathological response and/or DFS, and may be used to predict pathological response and survivals for ESCC patients receiving combined modality therapy.
TP53 & RRM2B	Prognosis and chemoradiotherapy	15,655,547	62	p53 or p53R2 (RRM2B) expression was correlated with a favorable response to CRT ($P = 0.0001$ or 0.041 clinical, $P = 0.016$ or 0.0018 histological, respectively; TP53, RR 2.688, 95% CI: 1.157–6.250; $P = 0.0011$. RRM2B, RR 2.469, 95% CI: 1.164–5.235; $P = 0.0057$).
	Prognosis	25,135,238	36	The median tumor associated survival was 34.2 months for patients with normal TP53, compared with 8.9 months for those with mutant TP53. The latter had a 3-fold higher risk of death (HR 3.01, 95% CI 1.359–6.86; $P = 0.005$).
	Prognosis	10,414,702	42	The current study indicated that p53 mutation of tumor tissues might be a prognostic factor for esophageal squamous cell carcinoma cases and one of the risk factors for its recurrence.
	Chemotherapy	19,941,080	97	Patients with mutations in p53 therefore showed significantly poorer prognosis than those without mutant p53.
RAC3 & TRAM1	Prognosis and chemoradiotherapy	19,552,757	98	Overexpression of AIB1/RAC3/ TRAM1 is a useful predictor of CRT resistance and an independent molecular marker of poor prognosis for ESCC patients.
ALDH1A1, ALDH1A2, ALDH1A3, ALDH1B1, ALDH1L1, ALDH1L2	Prognosis and recurrence	22,847,125	152	ALDH1 was a predictor of postoperative recurrence and prognosis in ESCC, and CD44 might be a predictor of recurrence and prognosis.
PITX2	Prognosis and chemoradiotherapy	23,132,660	454	High expression of PITX2 was associated with poor disease-specific survival (HR 1.732, 95% CI 1.133–2.646; $P = 0.011$) in ESCC.

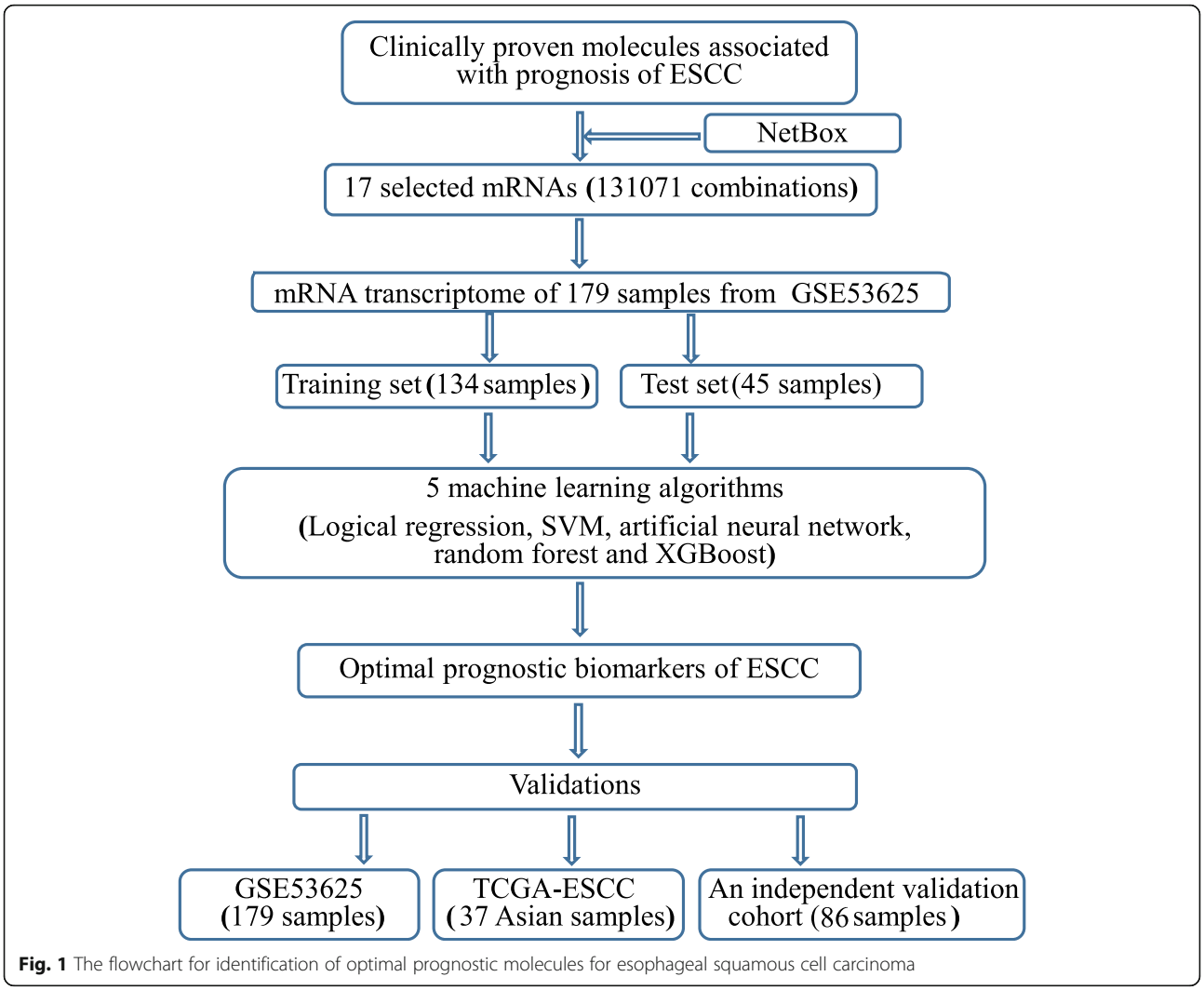

Fig. 1 The flowchart for identification of optimal prognostic molecules for esophageal squamous cell carcinoma

patterns of stratifin may have prognostic implication in ESCC.

Prognostic validation of stratifin

Using the dataset of GSE53625, 125 and 54 patients with ESCC were dichotomized into high-risk and low-risk subgroups according to optimal expression threshold of stratifin. The Kaplan-Meier survival analysis showed that the median survival times of the high-risk and low-risk subgroups were 25.5 months and > 60 months, respectively (Fig. 3a). Moreover, log-rank test showed that the survival times of two groups were significantly different, with a hazard ratio of 0.49 for patients with high stratifin expression (95% CI, 0.31 to 0.78, $P = 0.002$). The 3-year survival rates for these 2 subgroups were 42.4 and 63.1%, respectively. These results indicate that high expression of gene SFN is favorable to long-term survival of ESCC patients. In the 37 cases of ESCC with Asian ancestry from TCGA database, there was a trend for a favorable prognosis in ESCC patients with high mRNA levels of stratifin ($P = 0.094$, Fig. 3b).

We then validated the prognostic value of stratifin mRNA in another independent 86 ESCC cases. Using the median of stratifin mRNA levels as a cut-off value, 40 patients with ESCC were assigned to the high-risk subgroup and the other 46 patients to the low-risk subgroup. In consistent with previous results, ESCC patients in the high-risk subgroup had a significantly poorer survival than those in the low-risk subgroup. The median survival time for patients in the high-risk group was 37.5 months, while that for ESCC patients in the low-risk group was 60 months. The 3-year survival rates for the high-risk and low-risk subgroups were 53.6 and 73.5%, respectively. The log-rank test showed that the survival times of two groups were significantly different, with hazard ratio of 0.44 (95% CI, 0.26 to 0.75, $P = 0.0018$, Fig. 3c).

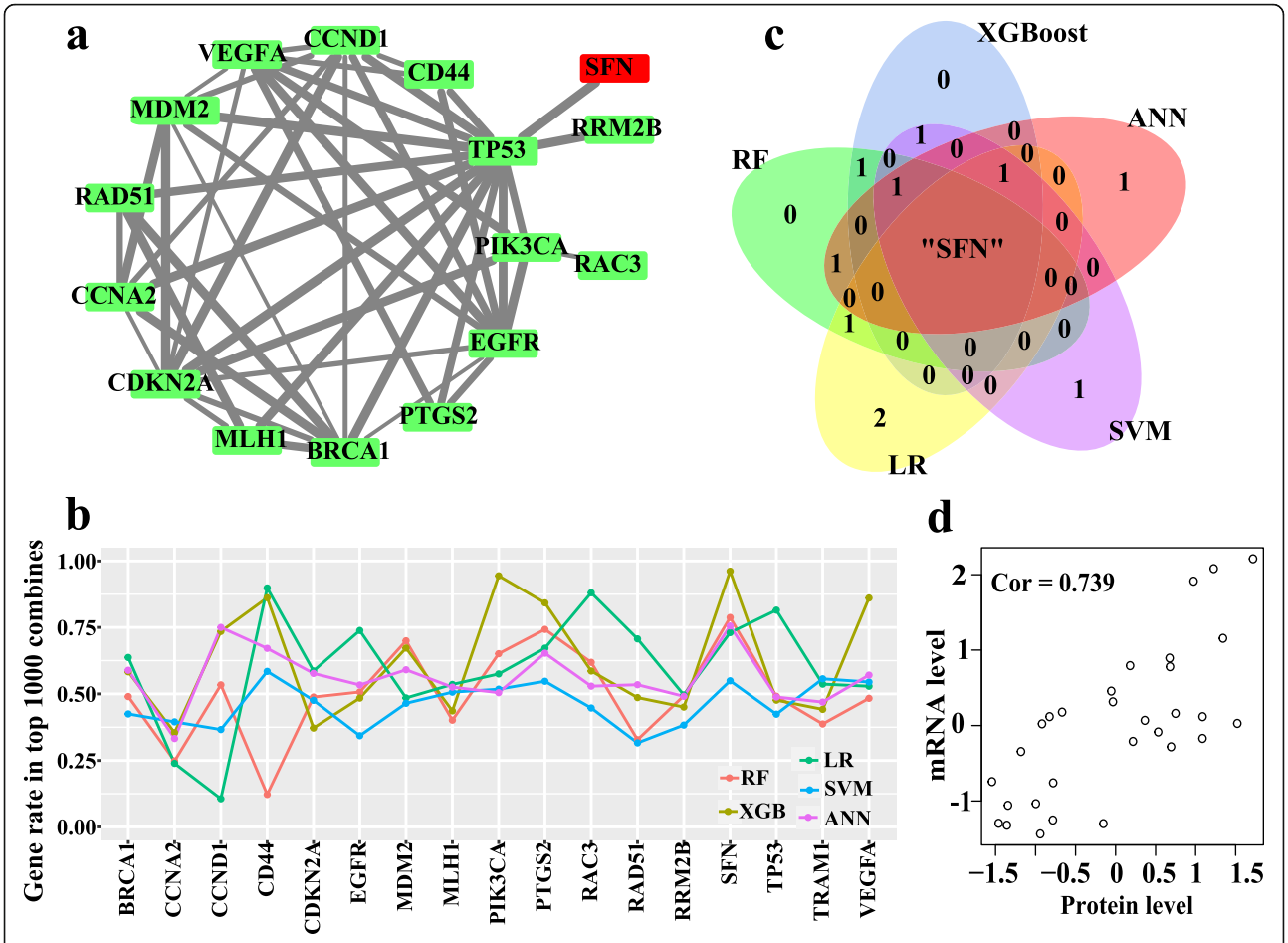

Fig. 2 NetBox and Machine learning model. **a** Molecular interaction network constructed by NetBox and STRING; **b** The occurrence frequencies of each molecule in top 1000 classifiers across 5 machine learning algorithms; **c** The intersection of the optimal genes from five machine learning algorithms; **d** Pearson correlation analysis between protein and mRNA expression levels of SFN detected by WB and RT-PCR, respectively. LR: logical regression; SVM: support vector machine; ANN: artificial neural network; RF: random forest; XGBoost: eXtreme gradient boosting; WB: Western bolt

Discussion

In this study, 48 molecules associated with clinical outcome of ESCC were used for construction of a molecular interaction network and subsequent identification of functional modules. Afterwards, all combinations of 17 component molecules from 3 modules were used to develop prognostic classifiers with 5 machine learning algorithms. Stratifin encoded by SFN was identified as the key prognostic biomarker for ESCC because it was the top overlapping molecule across the 5 prognostic methods used in this study. The down-regulation of stratifin mRNA and protein expression was associated with an overall poor survival of ESCC patients in 3 independent cohorts. Therefore, stratifin encoded by SFN was a robust biomarker for prognostic prediction of ESCC patients.

Table 2 The top 5 important molecules identified by each machine learning algorithm

Molecule rank	Weak classifiers			Strong classifiers	
	LR	SVM	ANN	RF	XGBoost
1	CD44	CD44	**SFN**	**SFN**	**SFN**
2	RAC3	TRAM1	CCND1	PTGS2	PIK3CA
3	TP53	**SFN**	CD44	MDM2	CD44
4	EGFR	PTGS2	PTGS2	PIK3CA	VEGFA
5	**SFN**	VEGFA	MDM2	RAC3	PTGS2

A variety of computational methods, such as dimensionality reduction [16], Cox multivariate regression [23], and subnetworks construction [24], have been used to identify biomarkers for detection, diagnosis and prognosis of patients suffering from cancers. In most cases, these methods were applied independently. As a result, distinct sets of molecules are identified by using various algorithms. It is conceivable, however, that the key molecules exerting crucial biological functions in cancer

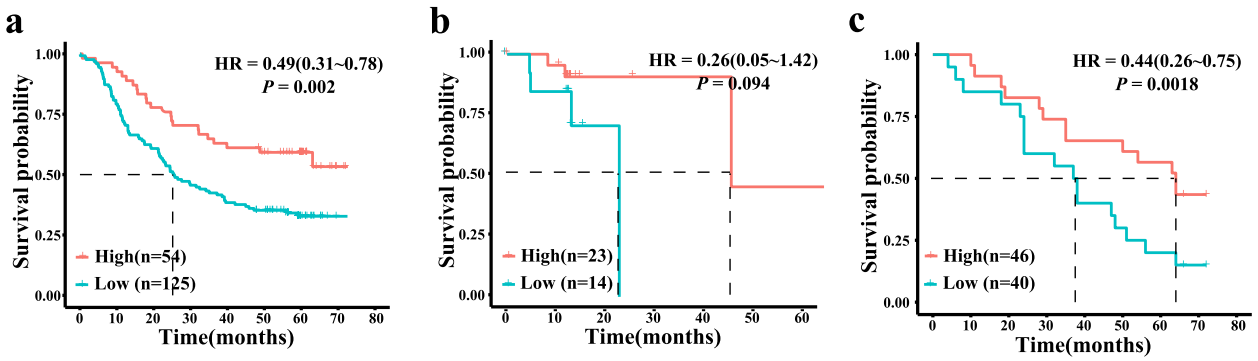

Fig. 3 Kaplan-Meier survival curves of ESCC patients in each dataset. **a** Kaplan-Meier survival curves of ESCC patients in GSE53625 dataset; **b** Kaplan-Meier survival curves of ESCC patients in TCGA-ESCC dataset (37 Asian cases); **c** Kaplan-Meier survival curves of ESCC patients in an independent cohort of 86 patients with ESCC

progression might be identified by these different computational analyses. The frequencies of overlapping molecules identified across these computational algorithms represent the degrees of functional importance. Using a subset of 38 miRNAs with experimental evidence associated with breast cancer, Oneeb et al. employed 3 feature selection methods, including Information Gain, Chi Squared, and Least Absolute Shrinkage and Selection Operation, to rank the importances of miRNAs. The top 10 important miRNAs were utilized to build optimal classifiers for discrimination between breast cancer cases and healthy subjects using RF-based and SVM-based algorithms. A 3-miRNA signature showed the best performance for diagnosis of breast cancer, indicating that not all miRNAs are equally important as cancer biomarkers [25]. Notably, these results demonstrate that the machine learning is a useful tool for feature selection without transformation of original features. In the present study, 48 biomarkers with clinical evidence for prognosis of ESCC were used to construct a subnetwork with 3 functional modules, including 17 component molecules. To rank the importances of these 17 molecule features, 5 machine learning algorithms were used for feature selection with SFN as the top overlapping gene, suggesting that SFN might be the optimal prognostic biomarker for ESCC.

In line with our previous findings, the expression pattern of stratifin mRNA resembled its protein expression, both of which were downregulated in ESCC compared with adjacent noncancerous mucosa. In the ESCC cohort of GSE53625, stratifin mRNA was an independent prognostic biomarker. This was also the case in another independent 86 ESCC cohort. Furthermore, a strong positive correlation between mRNA and protein expression of stratifin was found as well. Stratifin, one of the seven isoforms of 14–3-3 proteins in mammals, form homodimers and heterodimers that could bind to a number of target proteins in native state. Through association, stratifin regulates the functions of its ligands, including cytoskeletal dynamics, cell cycle regulation, polarity, adhesion, motility, mitogenic signaling and oncogenic signaling. In response to DNA damage, p53 can induce stratifin expression. In this manner, upregulation of stratifin causes G_2 arrest through sequestration of cdc2-cyclin B1 complex in cytoplasm and allows the repair of damaged DNA before further cell cycle progression. Thus, stratifin has been suggested to be a potential tumor suppressor. Decreased expression levels of stratifin occur frequently in many human cancers including breast [26–33], lung [34], colon [35], liver [36], prostate [37–39], ovary [40–42], nasopharynx [43], and oral cancers [44]. In addition, downregulation of stratifin in ESCC has been reported in several studies, which showed a negative correlation between SFN and clinical outcome [45–47]. Collectively, the present study provided further evidence supporting stratifin as a reliable prognostic biomarker for ESCC.

There are certain limitations to our study. Firstly, the present study only validated the clinical significance of stratifin in ESCC. Due to tumor heterogeneity, a composite biomarker comprising multiple functional molecules could represent the biology of ESCC much better than single molecule, and thus is able to improve the overall prediction of ESCC outcome. Secondly, liquid biopsy, in particular a simple blood test, offers a less-invasive approach to real-time monitor metastatic progression and therapeutic outcome of ESCC compared with tissue biopsy. The profile of stratifin in blood of ESCC patients should be characterized in future studies.

Conclusions

The present study presents stratifin as an optimal prognostic biomarker for ESCC using machine learning algorithms. In 3 independent cohorts of ESCC, stratifin can discriminate between ESCC patients with different clinical outcomes. Further prospective studies from different

institutions are needed to validate the robustness of stratifin in prognostic prediction of ESCC patients. Thus, our study demonstrates that the overlapping frequencies across different feature selection approaches represent the degree of importance, with top one as the key molecule with clinical implication. This method of mining key molecules that stably affect the prognosis of ESCC could be applied to the other relevant research.

Abbreviations

EC: Esophageal cancer; ESCC: Esophageal squamous cell carcinoma; EAC: Esophageal adenocarcinoma; GEO: Gene Expression Omnibus; TCGA: The Cancer Genome Atlas; AI: Artificial intelligence; LR: Logical regression; SVM: Support vector machine; ANN: Artificial neural network; RF: Random forest; XGBoost: eXtreme Gradient Boosting; WB: Western blot; PCR: Polymerase chain reaction; ROC: Receiver operating characteristic; AUC: Area under ROC curve

Acknowledgements

Not applicable.

Authors' contributions

Conception and design, YJQ; data curation, XMS, WGC; Methodology, KL, PC, HJR, HJX; Writing original draft, MXL; Writing-reviewing & editing, YJQ; Supervision, XSF, SGG. All authors read and approved the final manuscript.

Funding

This study was supported by grants from the National Natural Science Foundation of China (U1604191, 81872037). The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The data sets of GSE53625 from GEO (https://www.ncbi.nlm.nih.gov/geo/) and 37 ESCC cases with Asian ancestry from TCGA (UCSC Xena, https://xena.ucsc.edu/) are public open and available.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of The First Affiliated Hospital of Henan University of Science and Technology. Written informed consent was obtained from all patients. This study was conducted in accordance with the Declaration of Helsinki and the ethical standards of the Committee on human experimentation of the institution.

Consent for publication

Not applicable.

Competing interests

All authors declare no competing interests.

Author details

[1]School of Information Engineering of Henan University of Science and Technology, 263 Kaiyuan Road, Luolong Qu, Luoyang 471023, P. R. China. [2]Henan Key Laboratory of Microbiome and Esophageal Cancer Prevention and Treatment; Henan Key Laboratory of Cancer Epigenetics, Cancer Hospital, The First Affiliated Hospital, College of Clinical Medicine, Medical College of Henan University of Science and Technology, 24 Jinghua Road, Jianxi Qu, Luoyang 471003, P. R. China. [3]The Sixth People's Hospital of Luoyang, Oncology Department, 14 Xiyuan Road, Jianxi Qu, Luoyang 471003, P. R. China. [4]Department of Thyroid and Breast Cancer Surgery, The First Affiliated Hospital, College of Clinical Medicine, Medical College of Henan University of Science and Technology, 24 Jinghua Road, Jianxi Qu, Luoyang 471003, P. R. China.

Received: 1 December 2020 Accepted: 19 July 2021
Published online: 09 August 2021

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68(6):394–424. https://doi.org/10.3322/caac.21492.
2. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer. 2010; 127(12):2893–917. https://doi.org/10.1002/ijc.25516.
3. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019; 69(1):7–34. https://doi.org/10.3322/caac.21551.
4. Zheng RS, Sun KX, Zhang SW, Zeng HM, Zou XN, Chen R, et al. He J: [report of cancer epidemiology in China, 2015]. Zhonghua Zhong Liu Za Zhi. 2019; 41(1):19–28. https://doi.org/10.3760/cma.j.issn.0253-3766.2019.01.005.
5. Abnet CC, Arnold M, Wei WQ. Epidemiology of esophageal squamous cell carcinoma. Gastroenterology. 2018;154(2):360–73. https://doi.org/10.1053/j.gastro.2017.08.023.
6. Song Y, Li L, Ou Y, Gao Z, Li E, Li X, et al. Identification of genomic alterations in oesophageal squamous cell cancer. Nature. 2014;509(7498):91–5. https://doi.org/10.1038/nature13176.
7. Engel LS, Chow WH, Vaughan TL, Gammon MD, Risch HA, Stanford JL, et al. Population attributable risks of esophageal and gastric cancers. J Natl Cancer Inst. 2003;95(18):1404–13. https://doi.org/10.1093/jnci/djg047.
8. Tran GD, Sun XD, Abnet CC, Fan JH, Dawsey SM, Dong ZW, et al. Prospective study of risk factors for esophageal and gastric cancers in the Linxian general population trial cohort in China. Int J Cancer. 2005;113(3): 456–63. https://doi.org/10.1002/ijc.20616.
9. Freedman ND, Murray LJ, Kamangar F, Abnet CC, Cook MB, Nyrén O, et al. Alcohol intake and risk of oesophageal adenocarcinoma: a pooled analysis from the BEACON consortium. Gut. 2011;60(8):1029–37. https://doi.org/10.1136/gut.2010.233866.
10. Van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002; 415(6871):530–6.
11. Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, Yang F, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet. 2005;365(9460):671–9. https://doi.org/10.1016/S0140-6736(05)17947-1.
12. Tsuji S, Midorikawa Y, Takahashi T, Yagi K, Takayama T, Yoshida K, et al. Potential responders to FOLFOX therapy for colorectal cancer by random forests analysis. Br J Cancer. 2012;106(1):126–32. https://doi.org/10.1038/bjc.2011.505.
13. Gao YB, Chen ZL, Li JG, Hu XD, Shi XJ, Sun ZM, et al. Genetic landscape of esophageal squamous cell carcinoma. Nat Genet. 2014;46(10):1097–102. https://doi.org/10.1038/ng.3076.
14. Liu W, Snell JM, Jeck WR, Hoadley KA, Wilkerson MD, Parker JS, et al. Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis. JCI Insight. 2016;1(16):e88755. https://doi.org/10.1172/jci.insight.88755.
15. Lin L, Lin DC. Biological Significance of Tumor Heterogeneity in Esophageal Squamous Cell Carcinoma. Cancers (Basel). 2019;11(8):1156.
16. Liu W, He JZ, Wang SH, Liu DK, Bai XF, Xu XE, et al. MASAN: a novel staging system for prognosis of patients with oesophageal squamous cell carcinoma. Br J Cancer. 2018;118(11):1476–84. https://doi.org/10.1038/s41416-018-0094-x.
17. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944. https://doi.org/10.1371/journal.pone.0174944.
18. Li J, Chen Z, Tian L, Zhou C, He MY, Gao Y, et al. LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. Gut. 2014;63(11):1700–10. https://doi.org/10.1136/gutjnl-2013-305806.
19. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS One. 2010;5(2):e8918. https://doi.org/10.1371/journal.pone.0008918.
20. López-Martínez F, Schwarcz A, Núñez-Valdez ER, García-Díaz V. Machine learning classification analysis for a hypertensive population as a function of several risk factors. Expert Syst Appl. 2018;110:206–15. https://doi.org/10.1016/j.eswa.2018.06.006.

21. Wang WT, Guo CQ, Cui GH, Zhao S. Correlation of plasma miR-21 and miR-93 with radiotherapy and chemotherapy efficacy and prognosis in patients with esophageal squamous cell carcinoma. World J Gastroenterol. 2019; 25(37):5604–18. https://doi.org/10.3748/wjg.v25.i37.5604.
22. CORTES C, VAPNIK V. Support-vector networks. Mach Learn. 1995;20(3):273–97. https://doi.org/10.1007/BF00994018.
23. Liu Y, Gu Y, Su M, Liu H, Zhang S, Zhang Y. An analysis about heterogeneity among cancers based on the DNA methylation patterns. BMC Cancer. 2019; 19(1):1259. https://doi.org/10.1186/s12885-019-6455-x.
24. Yu D, Ruan X, Huang J, Hu W, Chen C, Xu Y, et al. Comprehensive analysis of competitive endogenous RNAs network, Being Associated With Esophageal Squamous Cell Carcinoma and Its Emerging Role in Head and Neck Squamous Cell Carcinoma. Front Oncol. 2019;9:1474.
25. Rehman O, Zhuang H, Muhamed Ali A, Ibrahim A, Li Z. Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach. Cancers (Basel). 2019;11(3):431.
26. Ferguson AT, Evron E, Umbricht CB, Pandita TK, Chan TA, Hermeking H, et al. High frequency of hypermethylation at the 14-3-3 sigma locus leads to gene silencing in breast cancer. Proc Natl Acad Sci U S A. 2000;97(11): 6049–54. https://doi.org/10.1073/pnas.100566997.
27. Umbricht CB, Evron E, Gabrielson E, Ferguson A, Marks J, Sukumar S. Hypermethylation of 14-3-3 sigma (stratifin) is an early event in breast cancer. Oncogene. 2001;20(26):3348–53. https://doi.org/10.1038/sj.onc.1204438.
28. Moreira JM, Ohlsson G, Rank FE, Celis JE. Down-regulation of the tumor suppressor protein 14-3-3sigma is a sporadic event in cancer of the breast. Mol Cell Proteomics. 2005;4(4):555–69. https://doi.org/10.1074/mcp.M400205-MCP200.
29. Wilker EW, van Vugt MA, Artim SA, Huang PH, Petersen CP, Reinhardt HC, et al. 14-3-3sigma controls mitotic translation to facilitate cytokinesis. Nature. 2007;446(7133):329–32. https://doi.org/10.1038/nature05584.
30. Feng W, Shen L, Wen S, Rosen DG, Jelinek J, Hu X, et al. Correlation between CpG methylation profiles and hormone receptor status in breast cancers. Breast Cancer Res. 2007;9(4):R57. https://doi.org/10.1186/bcr1762.
31. Urano T, Saito T, Tsukui T, Fujita M, Hosoi T, Muramatsu M, et al. Efp targets 14-3-3 sigma for proteolysis and promotes breast tumour growth. Nature. 2002;417(6891):871–5. https://doi.org/10.1038/nature00826.
32. Ling C, Zuo D, Xue B, Muthuswamy S, Muller WJ. A novel role for 14-3-3sigma in regulating epithelial cell polarity. Genes Dev. 2010;24(9):947–56. https://doi.org/10.1101/gad.1896810.
33. Zurita M, Lara PC, del Moral R, Torres B, Linares-Fernández JL, Arrabal SR, et al. Hypermethylated 14-3-3-sigma and ESR1 gene promoters in serum as candidate biomarkers for the diagnosis and treatment efficacy of breast cancer metastasis. BMC Cancer. 2010;10(1):217. https://doi.org/10.1186/1471-2407-10-217.
34. Osada H, Tatematsu Y, Yatabe Y, Nakagawa T, Konishi H, Harano T, et al. Frequent and histological type-specific inactivation of 14-3-3sigma in human lung cancers. Oncogene. 2002;21(15):2418–24. https://doi.org/10.1038/sj.onc.1205303.
35. Suzuki H, Itoh F, Toyota M, Kikuchi T, Kakiuchi H, Imai K. Inactivation of the 14-3-3 sigma gene is associated with 5′ CpG island hypermethylation in human cancers. Cancer Res. 2000;60(16):4353–7.
36. Iwata N, Yamamoto H, Sasaki S, Itoh F, Suzuki H, Kikuchi T, et al. Frequent hypermethylation of CpG islands and loss of expression of the 14-3-3 sigma gene in human hepatocellular carcinoma. Oncogene. 2000;19(46):5298–302. https://doi.org/10.1038/sj.onc.1203898.
37. Lodygin D, Diebold J, Hermeking H. Prostate cancer is characterized by epigenetic silencing of 14-3-3sigma expression. Oncogene. 2004;23(56): 9034–41. https://doi.org/10.1038/sj.onc.1208004.
38. Cheng L, Pan CX, Zhang JT, Zhang S, Kinch MS, Li L, et al. Loss of 14-3-3sigma in prostate cancer and its precursors. Clin Cancer Res. 2004;10(9): 3064–8. https://doi.org/10.1158/1078-0432.CCR-03-0652.
39. Pulukuri SM, Rao JS. CpG island promoter methylation and silencing of 14-3-3sigma gene expression in LNCaP and tramp-C1 prostate cancer cell lines is associated with methyl-CpG-binding protein MBD2. Oncogene. 2006; 25(33):4559–72. https://doi.org/10.1038/sj.onc.1209462.
40. Akahira J, Sugihashi Y, Suzuki T, Ito K, Niikura H, Moriya T, et al. Decreased expression of 14-3-3 sigma is associated with advanced disease in human epithelial ovarian cancer: its correlation with aberrant DNA methylation. Clin Cancer Res. 2004;10(8):2687–93. https://doi.org/10.1158/1078-0432.CCR-03-0510.
41. Kaneuchi M, Sasaki M, Tanaka Y, Shiina H, Verma M, Ebina Y, et al. Expression and methylation status of 14-3-3 sigma gene can characterize the different histological features of ovarian cancer. Biochem Biophys Res Commun. 2004;316(4):1156–62. https://doi.org/10.1016/j.bbrc.2004.02.171.
42. Mhawech P, Benz A, Cerato C, Greloz V, Assaly M, Desmond JC, et al. Downregulation of 14-3-3sigma in ovary, prostate and endometrial carcinomas is associated with CpG island methylation. Mod Pathol. 2005; 18(3):340–8. https://doi.org/10.1038/modpathol.3800240.
43. Yi B, Tan SX, Tang CE, Huang WG, Cheng AL, Li C, et al. Inactivation of 14-3-3 sigma by promoter methylation correlates with metastasis in nasopharyngeal carcinoma. J Cell Biochem. 2009;106(5):858–66. https://doi.org/10.1002/jcb.22051.
44. Gasco M, Bell AK, Heath V, Sullivan A, Smith P, Hiller L, et al. Epigenetic inactivation of 14-3-3 sigma in oral carcinoma: association with p16(INK4a) silencing and human papillomavirus negativity. Cancer Res. 2002;62(7):2072–6.
45. Qi YJ, Wang M, Liu RM, Wei H, Chao WX, Zhang T, et al. Downregulation of 14-3-3σ correlates with multistage carcinogenesis and poor prognosis of esophageal squamous cell carcinoma. PLoS One. 2014;9(4):e95386. https://doi.org/10.1371/journal.pone.0095386.
46. Ren HZ, Pan GQ, Wang JS, Wen JF, Wang KS, Luo GQ, et al. Reduced stratifin expression can serve as an independent prognostic factor for poor survival in patients with esophageal squamous cell carcinoma. Dig Dis Sci. 2010;55(9):2552–60. https://doi.org/10.1007/s10620-009-1065-0.
47. Lai KK, Chan KT, Choi MY, Wang HK, Fung EY, Lam HY, et al. 14-3-3σ confers cisplatin resistance in esophageal squamous cell carcinoma cells via regulating DNA repair molecules. Tumour Biol. 2016;37(2):2127–36. https://doi.org/10.1007/s13277-015-4018-6.

Publisher's Note