Using a Machine Learning Approach to Identify Key Prognostic Molecules for Esophageal Squamous Cell Carcinoma

Meng-Xiang Li
School of Information Engineering, Henan University of Science and Technology, Henan Key Laboratory of Cancer Epigenetics, The First Affiliated Hospital of Henan University of Science and Technology

Xiao-Meng Sun
The Sixth People's Hospital of Luoyang, Oncology Department; Henan Key Laboratory of Cancer Epigenetics, The First Affiliated Hospital of Henan University of Science and Technology

Wei-Gang Cheng
Department of Thyroid and Breast Cancer Surgery, The First Affiliated Hospital, Henan University of Science and Technology

Hao-Jie Ruan
Henan Key Laboratory of Cancer Epigenetics, The First Affiliated Hospital of Henan University of Science and Technology

Ke Liu
School of Information Engineering of Henan University of Science and Technology

Pan Chen
Henan Key Laboratory of Cancer Epigenetics, The First Affiliated Hospital of Henan University of Science and Technology

Hai-Jun Xu
Henan Key Laboratory of Cancer Epigenetics, Medical College of Henan University of Science and Technology

She-Gan Gao
The First Affiliated Hospital of Henan University of Science and Technology

Xiao-Shan Feng
Henan Key Laboratory of Cancer Epigenetics, The First Affiliated Hospital of Henan University of Science and Technology

Yi-Jun Qi (qiyijun@haust.edu.cn)
Henan University of Science and Technology

Research article

Keywords: esophageal squamous cell carcinoma, stratifin, machine learning, support vector machine, random forest, logical regression, artificial neural network, eXtreme gradient boosting

Posted Date: March 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-310517/v1

Version of Record: A version of this preprint was published at BMC Cancer on August 9th, 2021. See the published version at https://doi.org/10.1186/s12885-021-08647-1.

Abstract

Objective

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 is 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 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 these 17 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 clinical proven molecules associated with ESCC progression were used to construct a molecular interaction network with 3 functional modules comprising 17 component molecules. The 131071 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.

Conclusion

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

Introduction

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, 10]. 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 [9].

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 [11] from Netherlands and Wang et al [12] 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 [13]. 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 [14, 15]. 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 [16].

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 [17]. 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 [18].

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 clinical 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 2 independent ESCC cohorts.

Results

Prognostic biomarkers of esophageal squamous cell carcinoma

We initially retrieved 49 articles, which reported a total of 48 molecules associated with the clinical survival, recurrence or therapeutic outcome of ESCC patients. Table 1 shows the characteristics of studies included in this study. 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.

Table 1
The total of 48 clinical proven molecules associated with ESCC.

Clinically verified mRNA					
ALDH1A1	CCND1	EGFR	MDM2	NOTCH1	REG1A
ALDH1A2	CD163	ERCC1	MKI67	PIK3CA	RRM2B
ALDH1A3	CD274	FAM84B	MLH1	PITX2	SFN
ALDH1B1	CD44	FDXR	MMS19	PROM1	SGTA
ALDH1L1	CD68	HOXC6	MT3	PTGS2	TGFB1
ALDH1L2	CDKN2A	HOXC8	MUC13	PTPN6	TP53
BRCA1	CEACAM5	IL6R	MUC20	RAC3	TRAM1
CCNA2	CRP	KRT19	MUC4	RAD51	VEGFA
ESCC: esophageal squamous cell carcinoma					

Identification of key prognostic molecules

Our approach for validating clinically selected molecules for 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 in NetBox software to define the module information. By inputting the Entrez ID of 48 molecules, three functional modules containing a total of 17 molecules as vertices and 19 edges were identified (Table 2). A subnetwork of these 17 molecules based on STRING database (https://string-db.org/) was built with 0.7 as the minimum interaction score (Fig. 2a).

Table 2
Molecules that construct functional modules by NetBox.

Molecules selected by NetBox					
CCNA2	CD44	MDM2	TRAM1	RRM2B	RAD51
CCND1	EGFR	MLH1	RAC3	SFN	TP53
BRCA1	CDKN2A	PTGS2	PIK3CA	VEGFA	

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. Figure 2c shows the top 5 important molecules identified by each machine learning algorithm and the intersecting molecule is SFN only, indicating that SFN may be the optimal prognostic biomarker for ESCC (Table 3).

Table 3
The molecule rank in 5 machine learning algorithms.

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

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 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 from The Cancer Genome Atlas (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).

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.

A variety of computational methods, such as dimensionality reduction [17], Cox multivariate regression [19], and subnetworks construction [20], 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 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 Random Forest-based and Support Vector Machine-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 [21]. 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, is a p53-inducible gene in response to DNA damage. 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 [22–29], lung [30], colon [31], liver [32], prostate [33–35], ovary [36–38], nasopharynx [39], and oral cancers [40]. In addition, downregulation of stratifin in ESCC has been reported in several studies, which showed a negative correlation between SFN and clinical outcome [41–43]. Collectively, the present study provided further evidence supporting stratifin as a reliable prognostic biomarker for ESCC.

In conclusion, 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.

Materials And 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 download

Publicably available mRNA transcriptome data of ESCC from Gene Expression Omnibus (GEO) and 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 [44], 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. Written informed consent was obtained from all patients. This study was approved by the Ethics Committee of the First Affiliated Hospital of Henan University of Science and Technology.

Subnetwork construction

In this study, 48 molecules related to prognosis of ESCC were used to establish a molecular interaction subnetwork by NetBox [45]. 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. 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.

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 [46, 47].

SVM is a supervised learning method developed by Cortes and Vapnik in 1995 [48]. 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 hyper-parameter 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 strong classifier 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, 131071 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 randomForest 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'- GCCACATCGCTCAGACACC -3', and reverse primer, 5'- GATGGCAACAATATCCACTTTACC -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. The Kaplan-Meier survival curves and log-rank tests were performed to determine the statistical significance of overall survival. All *P* values were 2-tailed and *P* values <0.05 were designated as significantly different.

Abreviation List

EC: esophageal cancer; ESCC: esophageal squamous cell carcinoma; EAC: esophageal adenocarcinoma; GEO: Gene Expression Omnibus; TCGA: The Cancer Genome Atlas; 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

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.

Consent for publication: All authors consent for publication.

Availability of data and materials: All data and materials are available on request.

Conflicts of interest: All authors declare no competing interests.

Funding support: This study was supported by grants from the National Natural Science Foundation of China (U1604191, 81872037).

Author contributions

Conception and design, Yi-Jun Qi; data curation, Xiao-Meng Sun, Wei-Gang Cheng; Methodology, Ke Liu, Pan Chen, Hao-Jie Ruan, Hai-Jun Xu; Writing original draft, Meng-Xiang Li; Writing-reviewing & editing, Yi-Jun Qi; Supervision, Xiao-Shan Feng, She-Gan Gao; Final approval of manuscript: All authors.

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.
2. Ferlay J, Shin H, Bray F, Forman D, Mathers C, Parkin D: **Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008**. *Int J Cancer* 2010, **127**(12):2893-2917.
3. Siegel RL, Miller KD, Jemal A: **Cancer statistics, 2019**. *CA Cancer J Clin* 2019, **69**(1):7-34.
4. Rongshou Z, Kexin S, Siwei Z, Hongmei Z, Xiaonong Z, Ru C, Xiuying G, Wenqiang W, Jie H: **Report of cancer epidemiology in China, 2015**. *Chinese Journal of Oncology* 2019, **41**(1).
5. Abnet C, Arnold M, Wei W: **Epidemiology of Esophageal Squamous Cell Carcinoma**. *Gastroenterology* 2018, **154**(2):360-373.
6. Song Y, Li L, Ou Y, Gao Z, Li E, Li X, Zhang W, Wang J, Xu L, Zhou Y *et al*: **Identification of genomic alterations in oesophageal squamous cell cancer**. *Nature* 2014, **509**(7498):91-95.
7. LS E, WH C, TL V, MD G, HA R, JL S, JB S, ST M, R D, H R *et al*: **Population attributable risks of esophageal and gastric cancers**. *J Natl Cancer Inst* 2003, **95**(18):1404-1413.
8. GD T, XD S, CC A, JH F, SM D, ZW D, SD M, YL Q, PR T: **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):456-463.
9. Tran G, Sun X, Abnet C, Fan J, Dawsey S, Dong Z, Mark S, Qiao Y, Taylor P: **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):456-463.
10. Freedman N, Murray L, Kamangar F, Abnet C, Cook M, Nyrén O, Ye W, Wu A, Bernstein L, Brown L *et al*: **Alcohol intake and risk of oesophageal adenocarcinoma: a pooled analysis from the BEACON Consortium**. *Gut* 2011, **60**(8):1029-1037.

11. LJ vtV, H D, MJ vdV, YD H, AA H, M M, HL P, K vdK, MJ M, AT W *et al*: **Gene expression profiling predicts clinical outcome of breast cancer**. *Nature* 2002, **415**(6871):530-536.
12. Y W, JG K, Y Z, AM S, MP L, F Y, D T, M T, ME M-vG, J Y *et al*: **Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer**. *Lancet (London, England)* 2005, **365**(9460):671-679.
13. S T, Y M, T T, K Y, T T, K Y, Y S, H A: **Potential responders to FOLFOX therapy for colorectal cancer by Random Forests analysis**. *Br J Cancer* 2012, **106**(1):126-132.
14. YB G, ZL C, JG L, XD H, XJ S, ZM S, F Z, ZR Z, ZT L, ZY L *et al*: **Genetic landscape of esophageal squamous cell carcinoma**. *Nat Genet* 2014, **46**(10):1097-1102.
15. W L, JM S, WR J, KA H, MD W, JS P, N P, YB M, G M, NG L *et al*: **Subtyping sub-Saharan esophageal squamous cell carcinoma by comprehensive molecular analysis**. *JCI insight* 2016, **1**(16):e88755.
16. Le-hang L, De-Chen L: **Biological Significance of Tumor Heterogeneity in Esophageal Squamous Cell Carcinoma**. *Cancers (Basel)* 2019, **11**(8).
17. W L, JZ H, SH W, DK L, XF B, XE X, JY W, Y J, CQ L, LQ C *et al*: **MASAN: a novel staging system for prognosis of patients with oesophageal squamous cell carcinoma**. *Br J Cancer* 2018, **118**(11):1476-1484.
18. 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.
19. 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.
20. Yu D, Ruan X, Huang J, Hu W, Chen C, Xu Y, Hou J, Li S: **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.
21. O R, H Z, A MA, A I, Z L: **Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach**. *Cancers (Basel)* 2019, **11**(3).
22. Ferguson A, Evron E, Umbricht C, Pandita T, Chan T, Hermeking H, Marks J, Lambers A, Futreal P, Stampfer M *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):6049-6054.
23. Umbricht C, 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):3348-3353.
24. Moreira J, Ohlsson G, Rank F, Celis J: **Down-regulation of the tumor suppressor protein 14-3-3sigma is a sporadic event in cancer of the breast**. *Molecular & cellular proteomics : MCP* 2005, **4**(4):555-569.
25. Wilker E, van Vugt M, Artim S, Huang P, Petersen C, Reinhardt H, Feng Y, Sharp P, Sonenberg N, White F *et al*: **14-3-3sigma controls mitotic translation to facilitate cytokinesis**. *Nature* 2007, **446**(7133):329-332.

26. Feng W, Shen L, Wen S, Rosen D, Jelinek J, Hu X, Huan S, Huang M, Liu J, Sahin A *et al*: **Correlation between CpG methylation profiles and hormone receptor status in breast cancers**. *Breast cancer research : BCR* 2007, **9**(4):R57.
27. Urano T, Saito T, Tsukui T, Fujita M, Hosoi T, Muramatsu M, Ouchi Y, Inoue S: **Efp targets 14-3-3 sigma for proteolysis and promotes breast tumour growth**. *Nature* 2002, **417**(6891):871-875.
28. Ling C, Zuo D, Xue B, Muthuswamy S, Muller W: **A novel role for 14-3-3sigma in regulating epithelial cell polarity**. *Genes Dev* 2010, **24**(9):947-956.
29. Zurita M, Lara P, del Moral R, Torres B, Linares-Fernández J, Arrabal S, Martínez-Galán J, Oliver F, Ruiz de Almodóvar J: **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**:217.
30. Osada H, Tatematsu Y, Yatabe Y, Nakagawa T, Konishi H, Harano T, Tezel E, Takada M, Takahashi T: **Frequent and histological type-specific inactivation of 14-3-3sigma in human lung cancers**. *Oncogene* 2002, **21**(15):2418-2424.
31. 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):4353-4357.
32. Iwata N, Yamamoto H, Sasaki S, Itoh F, Suzuki H, Kikuchi T, Kaneto H, Iku S, Ozeki I, Karino Y *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):5298-5302.
33. Lodygin D, Diebold J, Hermeking H: **Prostate cancer is characterized by epigenetic silencing of 14-3-3sigma expression**. *Oncogene* 2004, **23**(56):9034-9041.
34. Cheng L, Pan C, Zhang J, Zhang S, Kinch M, Li L, Baldridge L, Wade C, Hu Z, Koch M *et al*: **Loss of 14-3-3sigma in prostate cancer and its precursors**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2004, **10**(9):3064-3068.
35. Pulukuri S, Rao J: **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):4559-4572.
36. Akahira J, Sugihashi Y, Suzuki T, Ito K, Niikura H, Moriya T, Nitta M, Okamura H, Inoue S, Sasano H *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**. *Clinical cancer research : an official journal of the American Association for Cancer Research* 2004, **10**(8):2687-2693.
37. Kaneuchi M, Sasaki M, Tanaka Y, Shiina H, Verma M, Ebina Y, Nomura E, Yamamoto R, Sakuragi N, Dahiya R: **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):1156-1162.
38. Mhawech P, Benz A, Cerato C, Greloz V, Assaly M, Desmond J, Koeffler H, Lodygin D, Hermeking H, Herrmann F *et al*: **Downregulation of 14-3-3sigma in ovary, prostate and endometrial carcinomas is**

associated with CpG island methylation. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc* 2005, **18**(3):340-348.

39. Yi B, Tan S, Tang C, Huang W, Cheng A, Li C, Zhang P, Li M, Li J, Yi H *et al*: **Inactivation of 14-3-3 sigma by promoter methylation correlates with metastasis in nasopharyngeal carcinoma**. *J Cell Biochem* 2009, **106**(5):858-866.
40. Gasco M, Bell A, Heath V, Sullivan A, Smith P, Hiller L, Yulug I, Numico G, Merlano M, Farrell P *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):2072-2076.
41. Qi Y, Wang M, Liu R, Wei H, Chao W, Zhang T, Lou Q, Li X, Ma J, Zhu H *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.
42. Ren H, Pan G, Wang J, Wen J, Wang K, Luo G, Shan X: **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):2552-2560.
43. Lai K, Chan K, Choi M, Wang H, Fung E, Lam H, Tan W, Tung L, Tong D, Sun R *et al*: **14-3-3σ confers cisplatin resistance in esophageal squamous cell carcinoma cells via regulating DNA repair molecules**. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 2016, **37**(2):2127-2136.
44. Li J, Chen Z, Tian L, Zhou C, He MY, Gao Y, Wang S, Zhou F, Shi S, Feng X *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):1700-1710.
45. E C, E D, N S, BS T, C S: **Automated network analysis identifies core pathways in glioblastoma**. *PLoS One* 2010, **5**(2):e8918.
46. López-Martínez F, Schwarcz.Md A, Núñez-Valdez ER, García-Díaz V: **Machine learning classification analysis for a hypertensive population as a function of several risk factors**. *Expert Systems with Applications* 2018, **110**:206-215.
47. WT W, CQ G, GH C, S Z: **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):5604-5618.
48. CORTES C, VAPNIK V: **Support-Vector Networks**. *Machine Learning* 1995, **20**:273-297.

Figures

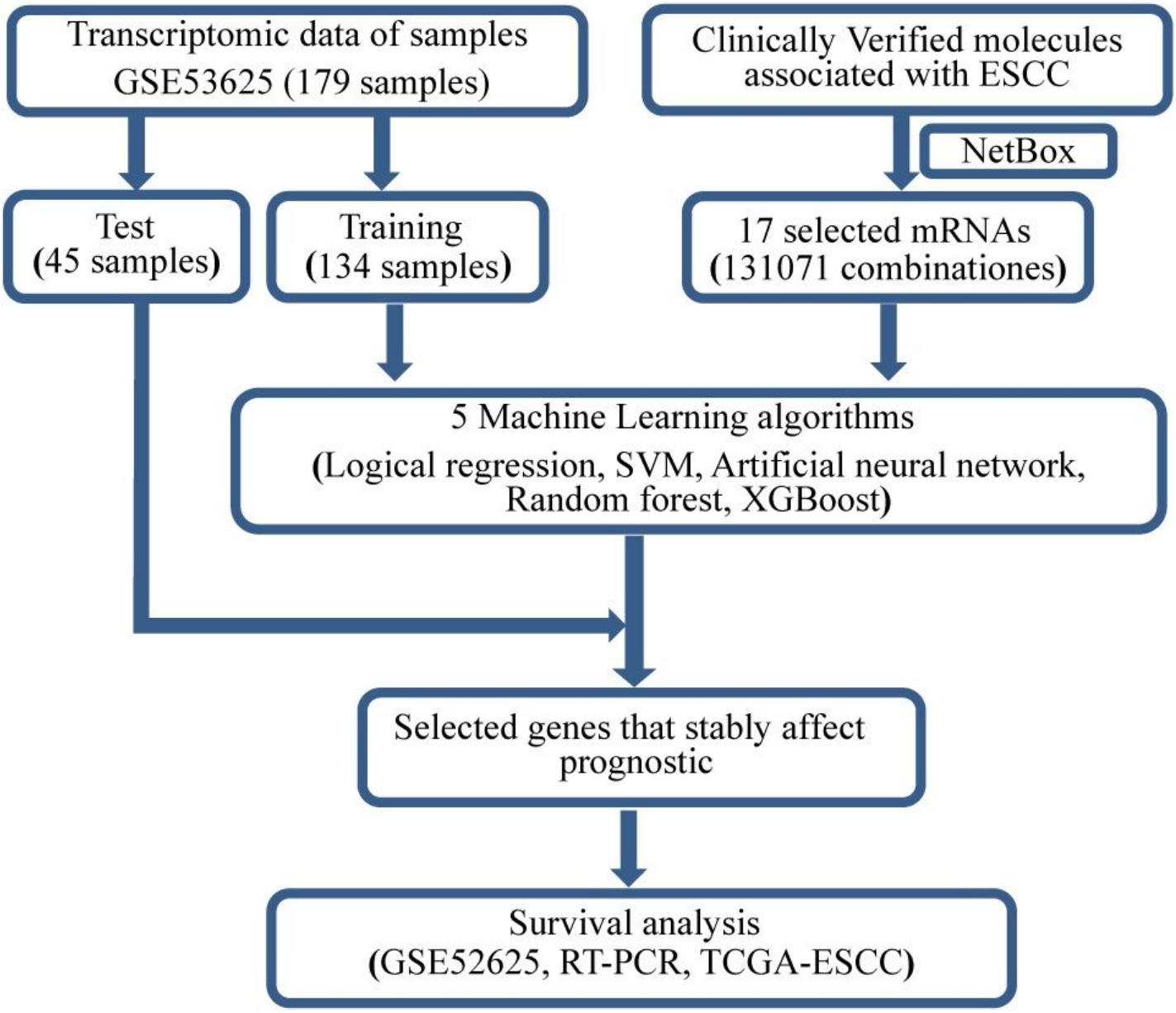

Figure 1

Study design of this study.

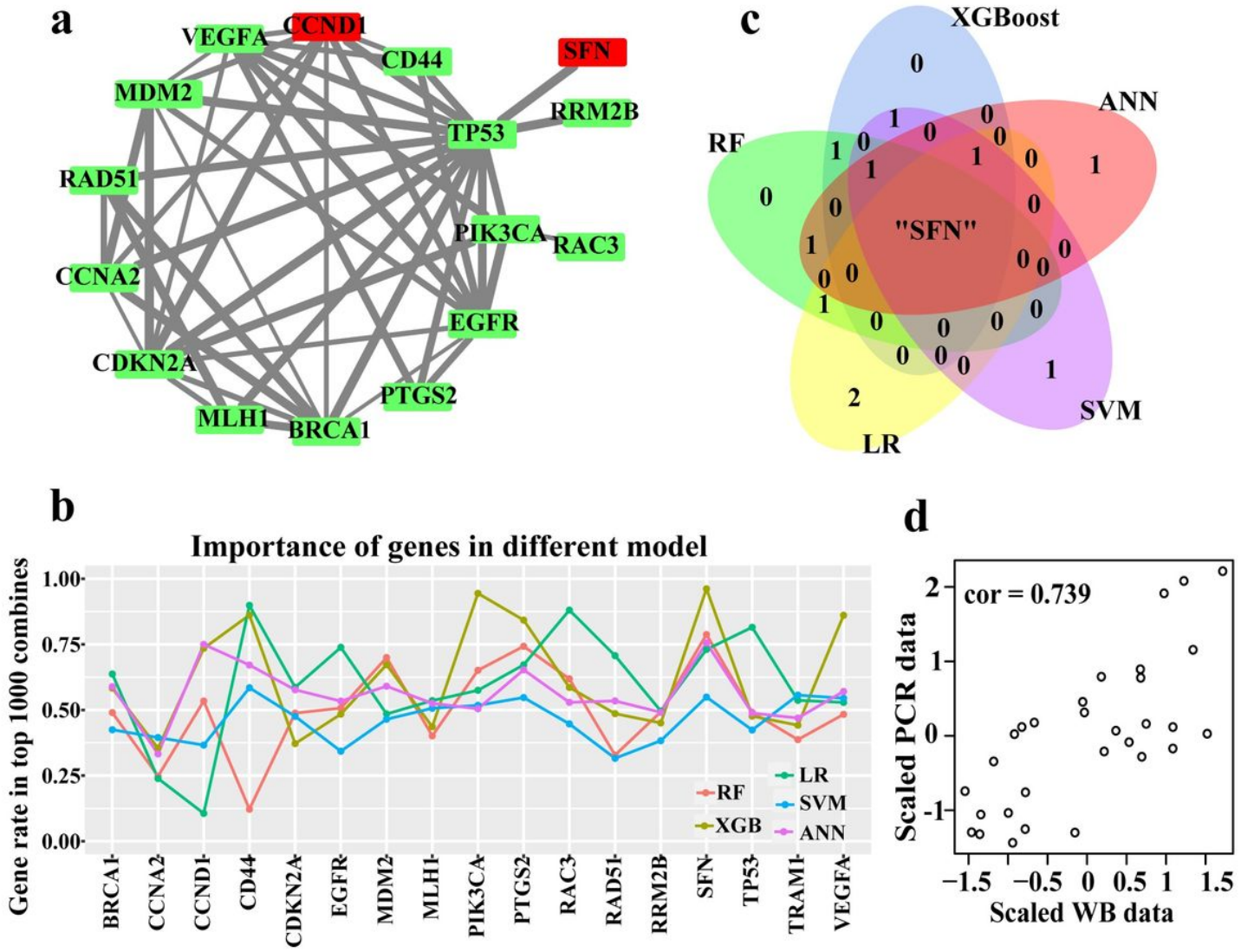

Figure 2

NetBox and Machine learning model results. (a) Molecular interaction network constructed by NetBox; (b) The occurrence frequencies of each molecular as optimal gene 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 detected by Western bolt and RT-PCR, respectively. LR: logical regression; SVM: support vector machine; ANN: artificial neural network; RF: random forest; XGBoost: eXtreme gradient boosting.

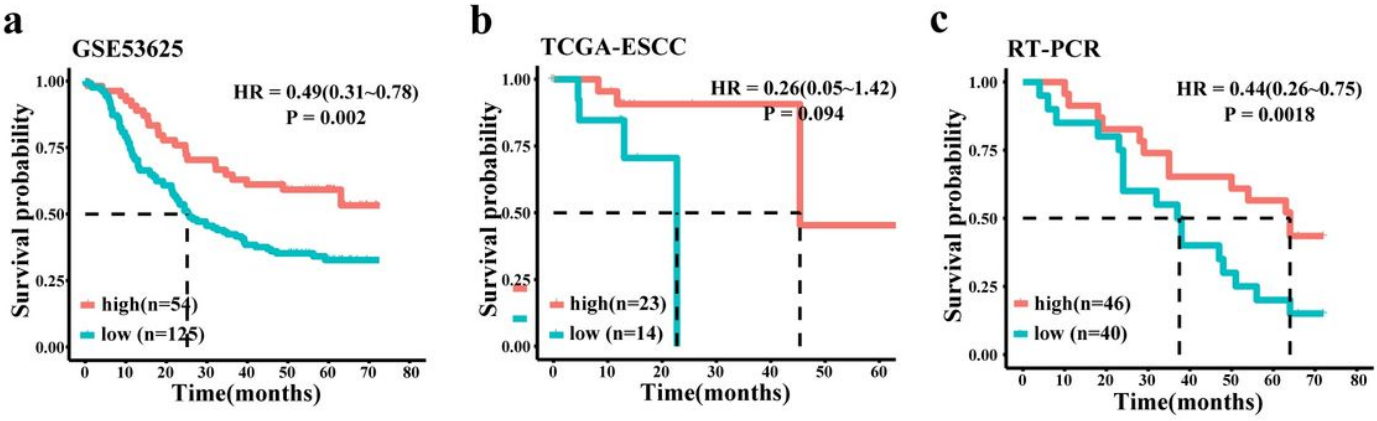

Figure 3

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; (c) Kaplan-Meier survival curves of ESCC patients in an independent cohort of 86 patients with ESCC.