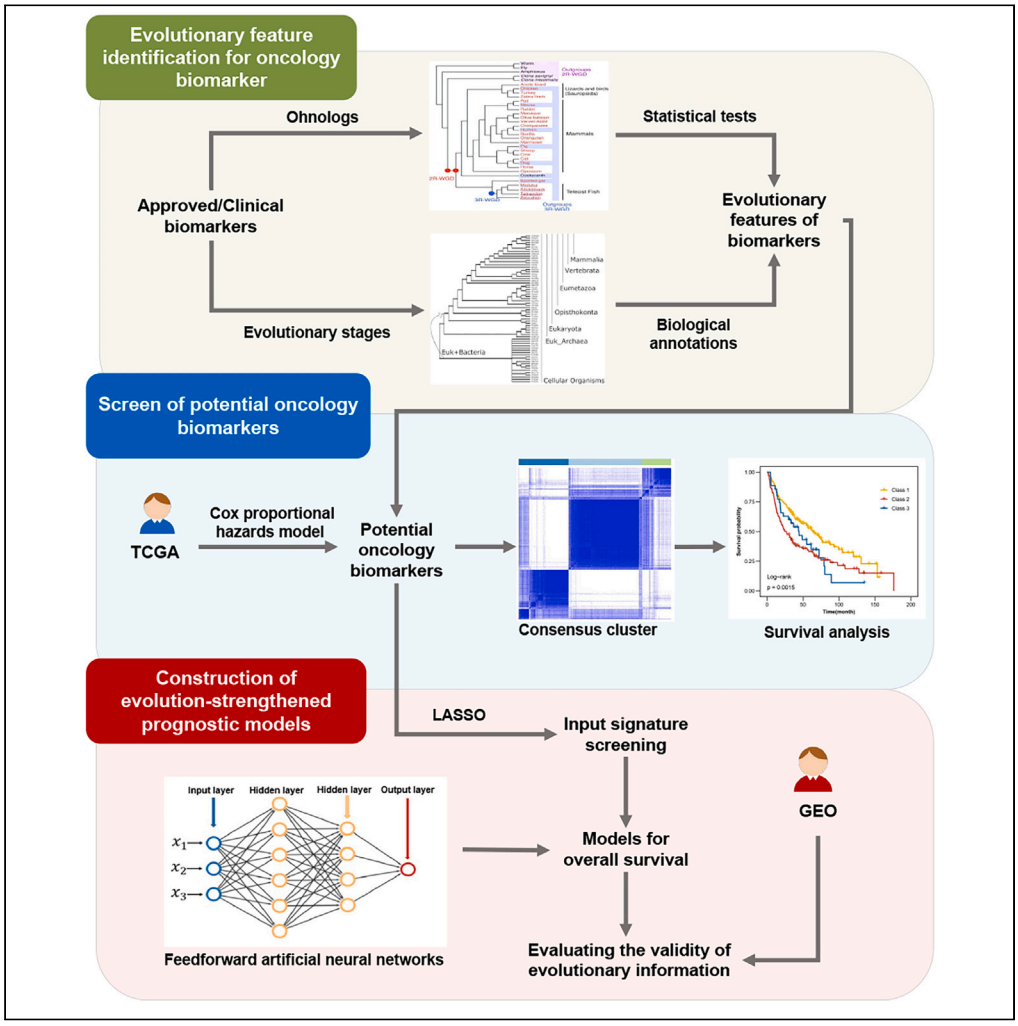
iScience

CellPress
OPEN ACCESS

Article

Evolutionary screening of precision oncology biomarkers and its applications in prognostic model construction



Zhi-Wen Zhang, Ke-Xin Zhang, Xuan Liao, Yuan Quan, Hong-Yu Zhang

quanyuan@mail.hzau.edu.cn (Y.Q.)
zhy630@mail.hzau.edu.cn (H.-Y.Z.)

Highlights

Clinically approved oncology biomarkers exhibit some evolutionary features

The evolutionary information of genes contributes to screening oncology biomarkers

Biomarkers with evolutionary features are more predictive of cancer prognosis

Zhang et al., iScience 27, 109859
June 21, 2024
https://doi.org/10.1016/j.isci.2024.109859

iScience

CellPress
OPEN ACCESS

Article

Evolutionary screening of precision oncology biomarkers and its applications in prognostic model construction

Zhi-Wen Zhang,[1] Ke-Xin Zhang,[1] Xuan Liao,[1] Yuan Quan,[1,*] and Hong-Yu Zhang[1,2,*]

SUMMARY

Biomarker screening is critical for precision oncology. However, one of the main challenges in precision oncology is that the screened biomarkers often fail to achieve the expected clinical effects and are rarely approved by regulatory authorities. Considering the close association between cancer pathogenesis and the evolutionary events of organisms, we first explored the evolutionary feature underlying clinically approved biomarkers, and two evolutionary features of approved biomarkers (Ohnologs and specific evolutionary stages of genes) were identified. Subsequently, we utilized evolutionary features for screening potential prognostic biomarkers in four common cancers: head and neck squamous cell carcinoma, liver hepatocellular carcinoma, lung adenocarcinoma, and lung squamous cell carcinoma. Finally, we constructed an evolution-strengthened prognostic model (ESPM) for cancers. These models can predict cancer patients' survival time across different cancer cohorts effectively and perform better than conventional models. In summary, our study highlights the application potentials of evolutionary information in precision oncology biomarker screening.

INTRODUCTION

Cancer is still the leading cause of premature death worldwide, and its prominence as a death cause is increasingly rising.[1–3] The lack of early diagnosis and appropriate treatment strategies may substantially result in a high cancer mortality rate.[4–6] Precision oncology is one of the most critical fields of modern medicine, and its dominant therapeutic paradigm is to personalize each patient's treatment based on oncology biomarkers.[7] Oncology biomarkers are essential in optimizing cancer prevention, diagnosis, and treatment.[8] Therefore, the effective identification of biomarkers will promote the progress of precision oncology. The rapid accumulation of multi-omics data over the past decades has facilitated the identification of cancer-related genes and screening oncology biomarkers.[9–11] However, biomarker screening has faced a dilemma: the screened biomarkers often do not achieve the expected clinical effects and are rarely approved by regulatory authorities.[12]

High-throughput genomics, proteomics, and metabolomics approaches allow the characterization and quantification of thousands of epigenetic markers, transcripts, proteins, and metabolites.[13] Precision oncology specialists are increasingly using computer technology to help explain complex cancer mechanisms and guide clinical decision-making.[14] Constructing a prognostic model using biomarkers can guide clinicians to stratify patients based on their pathological conditions and tailor personalized treatment plans, thereby improving survival rates for cancer patients.[15–17] Wang et al. used bioinformatics to explore the potential of *glypican 2* as a biomarker for pan-cancer.[18] Yang et al. screened long non-coding RNAs associated with prognosis in ovarian cancer and constructed prognostic prediction models. Their model showed better predictive precision than traditional clinical factors.[19] Based on the results of multivariate Cox regression analysis of patients with liver cancer, Zhang et al. constructed a model for survival in high-risk and low-risk groups that could be differentiated.[20] These results suggest that combining machine learning and multi-omics data is an effective way to screen biomarkers.

Carcinogenesis depends on key driver changes (mutations and epigenetic alterations in cancer cells), making it easily associated with evolution.[21] The various distinguishing hallmarks of cancer did not evolve independently with the advent of the organism but are an ordered and effective response to survival pressures. With the development of evolutionary medicine, accumulated evolutionary knowledge has been successfully used to interpret the pathogenesis of many diseases, including cancer, and to identify causative genes.[22,23] Cancer evolution is a process in which tumor cells adapt to the external environment, which can suppress the immune system's ability to recognize and attack tumors.[24] Based on the hallmarks of cancer cells, they can de- and *trans*-differentiate[25] and have unlimited replication potential, similar to

[1]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, P.R. China
[2]Lead contact
*Correspondence: quanyuan@mail.hzau.edu.cn (Y.Q.), zhy630@mail.hzau.edu.cn (H.-Y.Z.)
https://doi.org/10.1016/j.isci.2024.109859

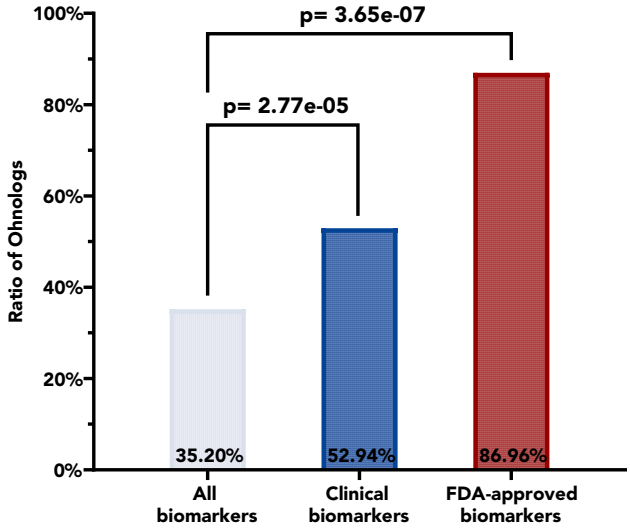


Figure 1. The proportion of Ohnologs in different types of biomarkers

Hypergeometric distribution tests confirmed a significant enrichment of Ohnologs in clinical biomarkers (p = 2.77e−05) and FDA-approved biomarkers (p = 3.65e−07) compared to all TTD biomarkers.

unicellular organisms.[26,27] Furthermore, cancer shares many characteristics with organisms at certain stages of evolution.[28,29] This provides an association between cancer and the evolutionary stage.

The vertebrate genome has undergone two important evolutionary events, namely whole-genome duplications (WGDs), which occurred in their jawless ancestor about 500 million years ago.[30] WGD gene duplicates, which have been uniformly termed "Ohnologs," contained approximately 30% of the human protein-coding genes.[31] Ohnologs are characterized by relatively high dosage sensitivity and tend to bring phenotypic variety upon changes in gene expression.[32] Ohnologs have been found to harbor mutations that are strongly associated with both cancer and genetic disorders.[30,31,33] The evolutionary stages of genes have also shown an association with cancer. Trigos et al. compared gene expression levels between cancer and normal tissue, and they found that overexpressed oncogenes were attributed to older age groups but that low-expressed genes were attributed to younger age groups.[34] Liebeskind et al. made inferences about the age of genes based on 13 popular homology inference algorithms, dividing human genes into eight evolutionary stages.[35] Then, it was demonstrated that the cancer driver genes were significantly enriched in genes that originated from the evolutionary stages of eukaryota, opisthokonta, and eumetazoa.[21,36] These results suggest that the evolutionary stages of genes may represent a valuable feature for screening oncology biomarkers. Incorporating evolutionary stages into systems-level analyses and retracing the cancer history may facilitate the identification of critical vulnerabilities in cancer, thereby identifying new oncology biomarkers and simplifying therapeutic strategies.[37]

In this study, we first investigated whether clinically approved biomarkers share some evolutionary features. Then, we used these evolutionary features to screen potential oncology biomarkers and validate the biological function of these potential oncology biomarkers. Finally, we constructed evolution-strengthened prognostic models (ESPMs) to predict the overall survival for four cancers and further validate their portability in different cancer cohorts. In summary, our study indicates the potential application of evolutionary information in biomedicine and provides a paradigm for how evolutionary information can be used for screening oncology biomarkers.

RESULTS

Ohnolog feature of approved biomarkers

Considering the close association between the Ohnologs with human diseases, we examined whether Ohnologs information can facilitate biomarker screening. We extracted 9,057 Ohnolog gene pairs from previous studies, encompassing 7,090 human genes.[31] The 1,514 biomarkers obtained from the TTD database include 23 Food and Drug Administration (FDA)-approved biomarkers and 119 biomarkers currently in clinical research. Notably, none of the 23 FDA-approved biomarkers were considered oncology biomarkers. This result further suggested the urgent need to improve the efficiency of oncology biomarker screening.

Of all TTD biomarkers, 533 were identified as Ohnolog genes, accounting for 35.20%. (Figure 1). Notably, 52.94% (63/119) of clinical research biomarkers were Ohnolog genes. Comparing the proportion of Ohnologs in clinical study biomarkers and all TTD biomarkers, the p value for the hypergeometric distribution test is 2.77e−05. This result indicated significant enrichment of Ohnologs in clinical biomarkers. Among FDA-approved biomarkers, Ohnologs accounted for an even higher 86.96% (20/23). Similarly, comparing this proportion with the proportion of Ohnologs in all TTD biomarkers, the hypergeometric distribution test had a p value of 3.65e−07. This result suggested

Table 1. The evolutionary stage distribution of FDA oncology pharmacogenomic biomarkers in drug labeling

Evolutionary stage	Number of oncology biomarker
Cellular organisms	1
Euk+bac	8
Euk_archaea	–
Eukaryota	9
Opisthokonta	6
Eumetazoa	12
Vertebrata	16
Mammalia	4
All	56

Table 1 shows the number of FDA oncology pharmacogenomic biomarkers in drug labeling for different evolutionary stages.

that Ohnologs were equally enriched in FDA-approved biomarkers. We extracted oncology biomarkers from the clinical trial biomarkers and also conducted a statistical analysis. Of 40 oncology biomarkers, 21 were Ohnologs, accounting for 52.5%. The hypergeometric test yielded a *p* value of 0.028. This finding suggested that Ohnologs are similarly enriched among tumor biomarkers. Therefore, Ohnologs may have the potential as an evolutionary feature that facilitates oncology biomarker screening.

Evolutionary stage feature of approved biomarkers

Previous studies have shown that cancer driver genes are significantly enriched in genes originating from the evolutionary stages of eukaryota, opisthokonta, and eumetazoa.[36] To determine the effectiveness of evolutionary stages in identifying oncology biomarkers, we obtained the list of FDA oncology pharmacogenomic biomarkers in drug labeling from the official website (Table 1). We found that FDA oncology pharmacogenomic biomarkers originating from the evolutionary stages of eukaryota, opisthokonta, and eumetazoa contain mutated genes commonly found in tumors.

Kirsten rat sarcoma viral oncogene homolog (*KRAS*), one of the most commonly mutated oncogenes, originated from the eukaryota stage. It usually acts as a molecular switch that, when turned on, activates some signaling pathways related to cell proliferation.[38] Wild-type *KRAS* amplification may be involved in the progression of tumors in the esophagogastric, colorectal, ovarian, and endometrial tissues.[39] Mutations in the *KRAS* are most commonly found in gastrointestinal and lung cancers, particularly pancreatic cancer.[40,41]

Another classical tumor target, *epidermal growth factor receptor* (*EGFR*), is a gene originating from the opisthokonta stage. EGFR is a transmembrane glycoprotein involved in cell proliferation, differentiation, and various regulatory mechanisms.[42] It is often overexpressed out of control in tumors, which makes it one of the proto-oncogenes.[43,44] *EGFR* is widely considered to be an important therapeutic target for non-small cell lung cancer, breast cancer, and esophageal-gastric cancer.[45,46]

Rearranged during transfection (*RET*) originates from the eumetazoa stage and is the transforming proto-oncogene that encodes a receptor tyrosine kinase.[47] The activated RET can initiate signaling pathways and promote cell proliferation and growth.[48] When *RET* undergoes oncogenic mutations, typically fusions or point mutations, the protein can become abnormally activated in a ligand-independent manner.[49] *RET* fusion is the main variant observed in non-small cell lung cancer and papillary thyroid cancer, and its point mutations are primarily associated with the development of sporadic medullary thyroid carcinoma.[50]

Furthermore, we analyzed and quantified the proportion of tumor biomarkers at various evolutionary stages within the MarkerDB database (Table 2). The analysis indicated that, compared to human genes, genes originating from eukaryota, opisthokonta, and eumetazoa were more enriched with oncology biomarkers. This enrichment, validated by Fisher's exact test with top three odds ratios and exceeding 1, highlights the potential of using evolutionary insights to identify oncology biomarkers. The aforementioned results suggested that biomarkers originating from three stages—eukaryota, opisthokonta, and eumetazoa—occupy pivotal positions in oncology research, providing evidence that the evolutionary stage of genes can serve as an evolutionary feature to screen potential precision oncology biomarkers.

Identification of potential oncology biomarkers

We downloaded RNA sequencing and clinical data from The Cancer Genome Atlas (TCGA) for four cancers: head and neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC) for further research (Table 3). To comprehensively evaluate the impact of clinicopathological characteristics on our subsequent research, we conducted a detailed statistical analysis of clinical data from cancer patients (Tables 4 and S1–S3). Specifically, we analyzed the age of patients, AJCC pathological stage, radiotherapy and chemotherapy received, and the distribution characteristics of high-frequency mutation genes in two groups of samples (positive and negative). Considering the significant intertumoral heterogeneity, three mutations with the highest frequency in each cancer were purposely selected for detailed statistics. The statistical results indicated that only the pathological stage in cancer types other than HNSC and *tumor protein p53* (*TP53*) gene mutations in LIHC and HNSC showed significant statistical differences. This

Table 2. Proportion of each evolutionary stage in oncology biomarkers of MarkedDB

Evolutionary stage	All human genes	MarkerDB	Odds ratio
Cellular organisms	4.54%	4.84%	1.06
Euk+bac	7.79%	8.06%	1.03
Euk_archaea	1.12%	0.00%	0
Eukaryota	29.26%	40.32%	1.63
Opisthokonta	5.75%	9.68%	1.76
Eumetazoa	25.50%	29.03%	1.19
Vertebrata	13.87%	6.45%	0.43
Mammalia	12.17%	1.61%	0.12

The table compares the proportion of genes from various evolutionary stages within all human genes and those identified as oncology biomarkers in the MarkerDB database. Fisher's exact test was employed to calculate the odds ratios. The results indicate that the odds ratios for the eukaryota, opisthokonta, and eumetazoa stages are more than 1, suggesting an enrichment of oncology biomarkers from these evolutionary stages compared to the overall distribution of human genes.

finding supported our grouping method, and the interference of clinicopathological features was excluded as much as possible when constructing the prognostic models.

Then, we performed survival analysis to identify the genes whose expression levels are significantly associated with survival time. We conducted Cox proportional hazards regression models for each gene and filtered for survival-related genes. We used a smaller threshold for some cancers to ensure their gene numbers were similar because of the remarkable heterogeneity between cancers.[51] 1,703, 1,896, 1,386, and 1,896 significant survival-related genes were obtained for LUSC, LUAD, LIHC, and HNSC, respectively. Additionally, we calculated the proportion of survival-related genes across various evolutionary stages. The outcomes of the Fisher's exact test highlighted a notable augmentation in the representation of these genes within the domains of eukaryota, opisthokonta, and eumetazoa, thereby substantiating our argument (Table S4). We screened genes being Ohnolog and originating from the evolutionary stages of eukaryota, opisthokonta, and eumetazoa, from survival-related genes as potential biomarkers. Finally, we identified 480, 545, 271, and 471 potential oncology prognostic biomarkers for LUSC, LUAD, LIHC, and HNSC, respectively (Table S5). These potential biomarkers would be further analyzed for biological enrichment analysis to evaluate their associations with cancer.

Biological enrichment analyses of potential oncology biomarkers

To explore the biology of potential oncology biomarkers, we performed Gene Ontology (GO) (Figure 2) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (Figure 3) enrichment analyses. For HNSC, GO terms showed that its potential biomarkers were mainly enriched in cellular growth regulation, axonogenesis, and axon development. KEGG analysis showed they were majorly involved in axon guidance, calcium signal pathway, and focal adhesion, which have been demonstrated to be associated with the growth and invasion of cancers.[52–54] The GO enrichment analysis of LIHC indicated that its potential biomarkers were related to biomolecules and cell metabolism regulation. These genes are enriched in pathways associated with bacteria, viral infections, and cancer. Both the GO and KEGG results are relevant to the cell cycle. For LUAD, the potential biomarkers are mostly signal transduction and protein synthesis. Notably, these genes are enriched in the Wnt pathway, the activation of which promotes the development, progression, and metastasis of cancers, including LUAD,[55,56] and some drugs targeting the Wnt pathway are already available for clinical use.[57–59] KEGG for LUSC was also enriched for some pathways associated with signal transduction, including the MAPK signaling pathway. MAPK pathway represents ubiquitous signal transduction pathways and is often altered in human tumors and cancer cell lines.[60–62] GO terms showed that the biological functions of its biomarkers were more related to neuronal development and membrane potential. Overall, these oncology biomarkers are involved in cell growth, neurological development, energy metabolism, protein expression, and signal transmission, and therefore associated with cancer.

Furthermore, we performed consensus clustering to classify prognostically stratified subgroups based on the expression levels of potential oncology biomarkers and found 3, 3, 4, and 4 molecular classifications for LUAD, LIHC, LUSC, and HNSC based on cumulative distribution function plot and delta area plot, respectively (Figures S1–S4). Finally, the Kaplan-Meier survival analysis showed significant differences in overall survival between classes for each cancer (Figure 4). These results indicate that the potential oncology biomarkers screened based on the evolutionary features of genes are biologically significant and can be further used to construct prognostic models.

Construction of ESPMs

LASSO regression was utilized to screen gene signatures for cancer. As a result, 36, 57, 21, and 50 gene signatures were obtained for LUSC, LUAD, LIHC, and HNSC, respectively (Table S6). The magnitude of the LASSO coefficients, whether positive, negative, or absolute, signifies the impact of gene expression levels on patient survival. We found associations between the genes with the highest LASSO coefficients and the cancer prognosis (Tables S7, S8, S9, and S10).

Table 3. Sample compositions of cancer cohorts

Cancer	All samples	Available samples	Mean survival time (months)	Negative sample[a]	Positive sample[a]
HNSC[b]	520	378	25.50	155	223
LIHC[b]	370	234	22.95	81	153
LUAD[b]	508	307	27.01	109	198
LUSC[b]	495	314	30.58	127	187

This study's sample sizes for each of the four cancer types are detailed in Table 1. Exclusions were made for samples with incomplete data or follow-up periods less than the calculated mean survival time, which was derived exclusively from the samples that experienced mortality. These criteria ensured the selection of positive and negative samples to construct future predictive models.

[a]Positive and negative samples are divided according to mean survival time.

[b]HNSC, head and neck squamous cell carcinoma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma.

Fatty acid-binding protein 6 (*FABP6*) had the highest coefficient for LUSC. It has been demonstrated to be overexpressed in bladder cancer, colorectal cancer, and glioblastoma.[63–66] The negative coefficient for *FABP6* suggested that overexpression of *FABP6* may be detrimental to patient survival, consistent with current findings.[65,66] *FABP6* also has been used to construct prognostic prediction models for many cancers.[67,68] For LUAD, *protein kinase C delta* (*PRKCD*) had the highest positive coefficient. *PRKCD* is a regulator of mitochondrial autophagy[69,70] and is associated with resistance to some cancer treatment regimens.[71,72] This gene's high expression benefits patient survival.[73] For LIHC, two members of the annexin family, *annexin A10* (*ANXA10*) and *annexin A2* (*ANXA2*), were the top contributors with positive coefficients. They have been reported to inhibit the progression of certain types of cancer.[74–76] *ANXA10* may be a prognostic marker and an inhibitor of LIHC.[77] Conversely, *ANXA10* and *ANXA2* would impede survival in LUAD, which has been recognized to some extent.[78] This result suggested that the effects of the same biomarker may vary considerably in different cancers. *Muscleblind-like splicing regulator 1*, the primary contributor in the HNSC model, would inhibit tumor progression,[79] as demonstrated again in our research. The aforementioned results suggest that these genes are associated with patient prognosis, and their effects align with LASSO coefficients, offering new insights into the impact of altered gene expression on patient prognosis. Although these genes may not have received widespread attention in the field of cancer, our results demonstrate that their potential roles in cancer prognosis research still warrant further exploration for a deeper understanding of their biological significance.

We then sought to leverage the prognostic capability of the gene signatures into a clinical tool capable of estimating the five-year overall survival probability of cancer patients. Neural networks have been widely used for cancer diagnosis, prognosis, and treatment selection with favorable predictive results.[80,81] Therefore, we constructed a binary classifier, ESPM, to predict the survival time of patients. The classification labels depended on the mean survival time of patients (Table 3). It should be noted that the parameters and steps for the four cancers are identical when constructing the models. We employed 5-fold cross-validation by dividing the data 100 times at random and using the mean values of the model outputs to evaluate the model's performance (Figure 5). The mean area under the curve (AUC) values for LUSC, LUAD, LIHC, and HNSC were 0.70, 0.77, 0.77, and 0.69, respectively.

Portability of ESPMs

We further verify the portability of ESPMs across different cancer cohorts. We used genes without the evolutionary feature screening of what we consider to be oncology biomarkers to construct conventional models. For each cancer, equal numbers of genes with oncology biomarker evolutionary features were randomly selected. These genes were followed in the same steps and parameters when constructing models. The mean AUC values for conventional models were 0.65 (LUSC), 0.72 (LUAD), 0.76 (LIHC), and 0.70 (HNSC). The results indicated that the ESPMs performed better than the conventional models in three cancers: LIHC (from 0.76 to 0.77), LUAD (from 0.72 to 0.77), and LUSC (from 0.65 to 0.70). However, the performance of the ESPM model for HNSC showed a slight decrease (from 0.70 to 0.69) (Figure 6). Although the AUC of the ESPM model for HNSC showed a slight decrease, its higher accuracy and specificity suggested a better classification of negative samples, which would be equally relevant in guiding clinical treatment.

To demonstrate the prognostic value of these potential biomarkers, the dataset from the GEO database was analyzed (key resources table). The overall survival time of cancer patients in the GEO dataset was predicted based on the model features obtained from TCGA. The AUCs for LUSC, LUAD, LIHC, and HNSC were 0.69, 0.75, 0.72, and 0.62, respectively. In addition, similar to the results of the TCGA cohorts, the ESPMs outperformed the conventional models (Figure 6). Our results confirmed that the portability of the biomarkers screened using evolutionary features facilitated the construction of cancer prognostic models.

DISCUSSION

Precision oncology aims to provide individualized treatment for cancer patients to more appropriately meet the specific treatment needs of different patients. The identification of predictive biomarkers has the potential to significantly enhance treatment selection and improve patient outcomes, as well as reduce side effects associated with cancer treatment.[82] With the availability of multi-omics data, an increasing number of novel markers are being proposed for the diagnosis, treatment, and prognostic survival assessment of tumors, driving personalized therapy. However, the high background noise in omics data poses a challenge in distinguishing between clinically meaningful results and

Table 4. Clinicopathological characteristics of lung adenocarcinoma patients

Clinicopathological characteristics	*n*	Positive sample	Negative sample	*p*
Mean age	64.66			
≥	159	99	60	0.7184
<	138	89	49	
AJCC pathologic stage				
I	148	118	30	9.011e−08
II	81	46	35	
III	55	22	33	
IV	18	8	10	
Pharmaceutical therapy				
Yes	88	59	29	0.5992
No	219	139	80	
Radiation therapy				
Yes	46	25	21	0.1338
No	261	173	88	
Mutation				
TP53				
Yes	134	85	49	0.8100
No	173	113	60	
MUC16				
Yes	116	74	42	0.9022
No	191	124	67	
CSMD3				
Yes	115	76	39	0.7121
No	192	122	70	

Table 4 evaluates the impact of pathological characteristics on the positive and negative group samples among LUAD (lung adenocarcinoma) patients. These characteristics encompass age, pathological stage, received treatment regimens, and the top three mutations identified in these patients. *p* values were derived using the Fisher's exact test to determine statistical significance.

merely noisy results.[83] Evolutionary information has shown potential in identifying oncology biomarkers and improving the accuracy of cancer diagnosis and treatment selection.[84,85] Incorporating additional biological information can improve prediction accuracy and reduce false positives.

Ohnologs play a crucial role in the development and regulation of organisms.[30,31,86] In this study, the statistical analysis results demonstrated the association between Ohnologs and biomarkers. Indeed, the positive selection that genes have undergone during evolution makes these genes often more meaningful for vertebrates, which could explain their associations with biomarkers. Moreover, the ability of cancer cells to evolve rapidly and escape the control of cell division and programmed cell death allows them to spread rapidly, similar to the characteristics in organisms of specific evolutionary stages.[28,87,88] Many of the FDA oncology pharmacogenomic biomarkers in drug labeling originated from the evolutionary stages of eukaryota, opisthokonta, and eumetazoa. Furthermore, cancer driver genes are mainly enriched in these three stages,[36] suggesting that genes originating from these stages are more likely to be involved in critical biological processes that cause cancer occurrence and progression.

Considering the close association between cancer and evolution, exploring biomarker features from an evolutionary perspective and combining them with omics data may be an effective approach to screening high-quality oncology biomarkers. We screened genes with needed evolutionary features as potential oncology biomarkers. To minimize the impact of tumor heterogeneity, we specifically selected the mutations with the highest frequency in each tumor cohort for our high-frequency mutation statistics and separately identified potential tumor biomarkers for each cancer type. In LUAD and LUSC, we observed a substantial degree of similarity. *TP53* exhibited the highest mutation frequency in all four cancer cohorts (Table 3) and showed significant differences in the distribution of positive and negative samples in HNSC and LIHC. This suggests that its impact on the prognosis of these two cancers is worth further investigation.[89,90] Although cancer patients respond differently to treatments, this usually prolongs survival, thereby affecting prognosis.[91] According to TCGA data, the number of patients who had undergone prior treatment is seldom, making its impact on patient prognosis negligible. Furthermore, our grouping method appears to have excluded the influence of subsequent treatment modalities on the positive and negative sample

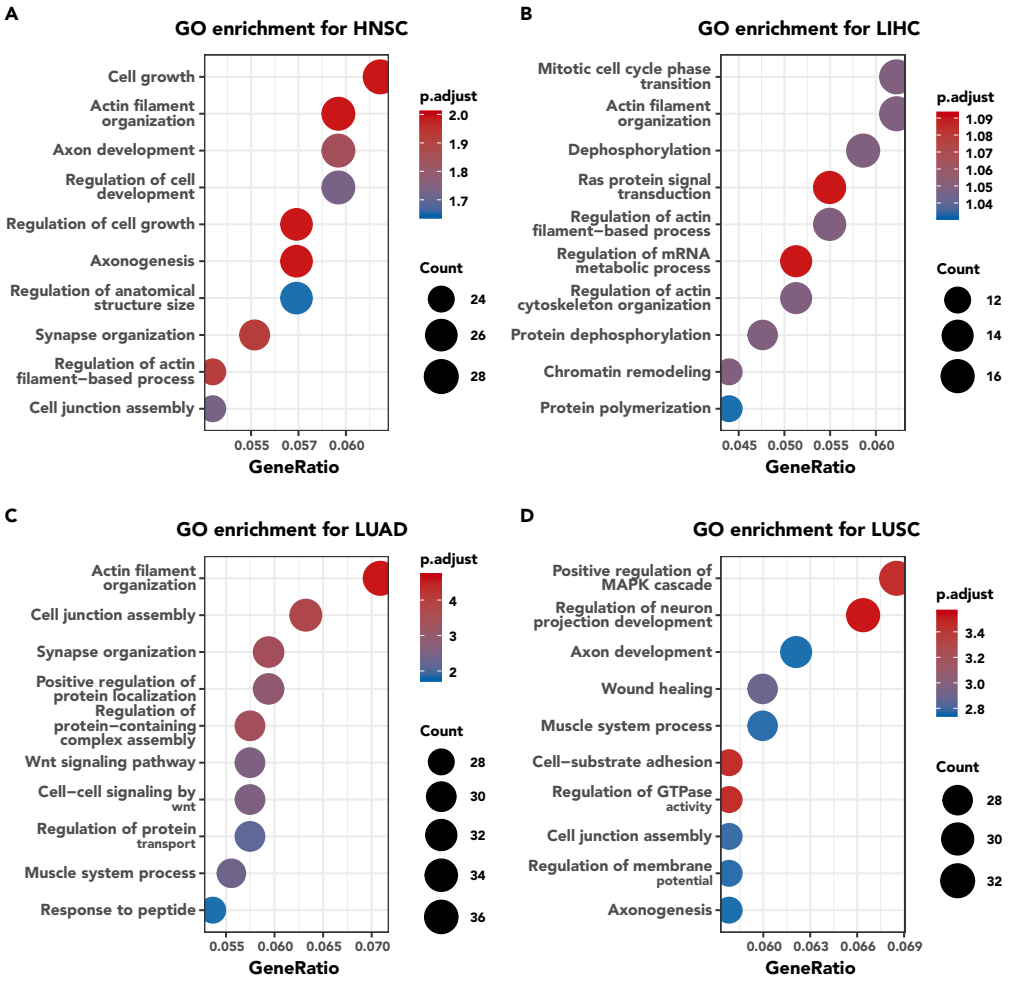


Figure 2. GO enrichment analyses of potential oncology biomarkers

(A) GO enrichment analysis results for HNSC.

(B) GO enrichment analysis results for LIHC.

(C) GO enrichment analysis results for LUAD.

(D) GO enrichment analysis results for LUSC. The vertical coordinate represents the biological function of gene enrichment, the bubble size represents the number of genes enriched in the biological function, and the bubble colors represent the corrected *p* value.

categorization. This result not only enhanced the reliability of our prognostic model but also further demonstrated the great potential of using evolutionary information in the selection of oncology biomarkers. GO and KEGG analyses showed that these oncology biomarkers were involved in biological processes that are consistent with the characteristics of cancer.[39,41,92] We also noted that the enriched terms and pathways vary considerably between different cancers (even LUAD and LUSC). This result may represent specific hallmarks of different cancers. We generated molecular subtyping schemes based on the expression of these oncology biomarkers for each cancer and found that these subtypes were associated with overall survival. Therefore, these signatures may provide clinically important information for prognostication.

We constructed classifiers to predict patient survival time in different cancer cohorts. Comparing the results of the ESPMs with the conventional models, higher AUCs were obtained because of the addition of evolutionary information. Furthermore, the stability of ESPMs is greater than the conventional models, suggesting that ESPMs have a better predictive capability. The improved performance of ESPMs highlighted the value of evolutionary information in oncology biomarker screening. Heterogeneity manifests across all levels of tumor organization, its quantification thus requiring measurements at multiple scales.[93] Although transcriptomic data are widely used in oncology because of the marked differences in gene expression between cancer cells and normal cells, it cannot provide comprehensive information on other biological processes.[94,95] Therefore, each histological platform's specific limitations and noise may diminish the contribution of evolutionary information in oncology biomarker screening.[96–98] While the expression levels of potential biomarkers could significantly differentiate samples into subtypes with varying survival periods, this heterogeneity might magnify the differences in expression among samples when constructing

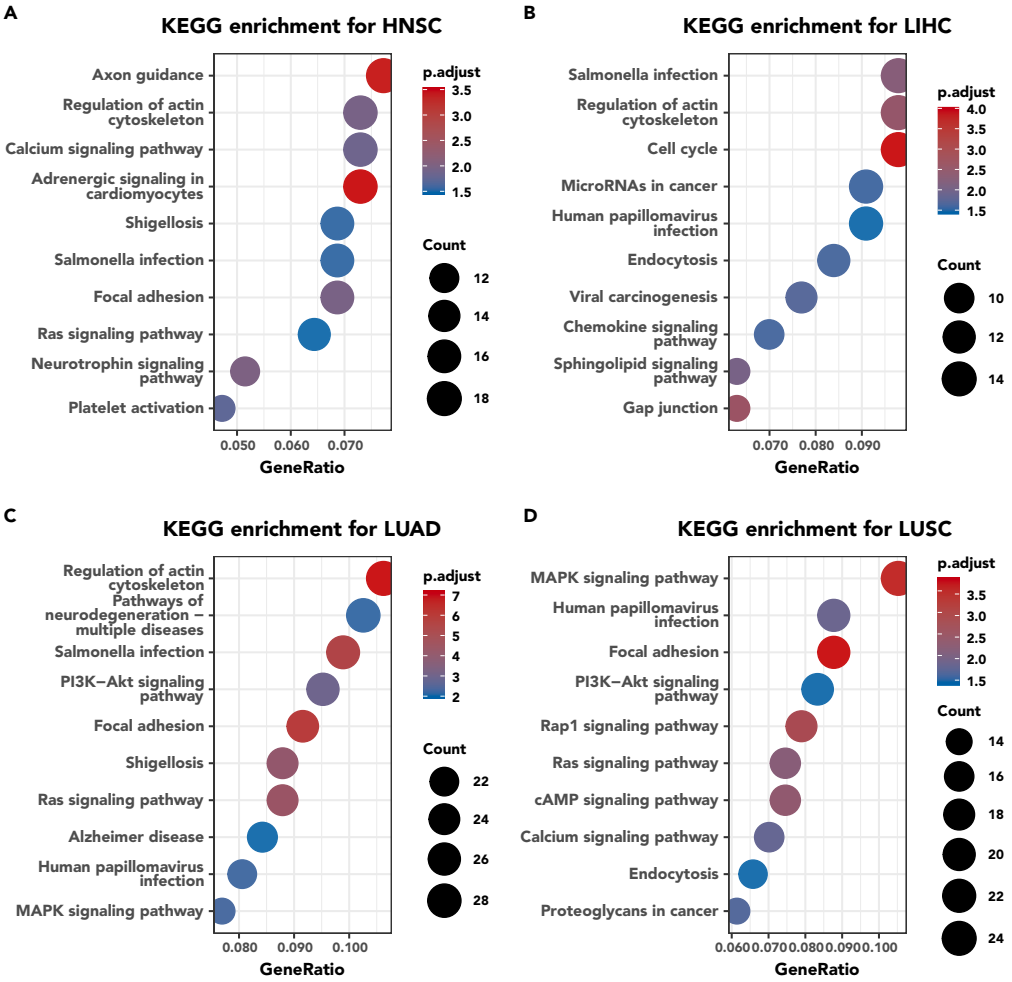


Figure 3. KEGG enrichment analyses of potential oncology biomarkers

(A) KEGG enrichment analysis results for HNSC.

(B) KEGG enrichment analysis results for LIHC.

(C) KEGG enrichment analysis results for LUAD.

(D) KEGG enrichment analysis results for LUSC. The vertical coordinate represents the pathway of gene enrichment, the bubble size represents the number of genes enriched in the pathway, and the bubble colors represent the corrected *p* value.

more refined binary prognostic models, which affected the predictive effectiveness of the model. In the validation set (GEO dataset), the model's predictive performance was further compromised by the heterogeneity among different cohort samples. This accounted for the less favorable outcomes observed in the validation set. However, within the same data context, the results of the ESPM were still superior to traditional models, indicating that incorporating evolutionary information into tumor biomarker selection demonstrates a certain level of stability.

In this study, we validated the potential application of evolutionary information in biomedicine and provided a paradigm for the application of evolutionary information in precision oncology. Our approach is rooted in a biological perspective, aiming to uncover the significant connections between cancer and genes that have been preserved with greater significance over the vast expanse of evolutionary history. We posit that variations in these genes are likely to exert substantial impacts, thereby possessing the potential to serve as biomarkers. Subsequently, we systematically validated our hypotheses using statistical and machine learning techniques to ensure the reliability of our findings. The results showed that the performance of ESPM was better than other conventional models in different cancer prognostic models, suggesting that incorporating evolutionary information may improve the efficiency of oncology biomarker screening. Additionally, with the continuous development of new technologies and the deepening of clinical research, we can foresee that more oncology biomarkers will be discovered and approved, which will help to explore the evolutionary features of oncology biomarkers and assess the application potentials of evolutionary information in precision oncology more comprehensively.

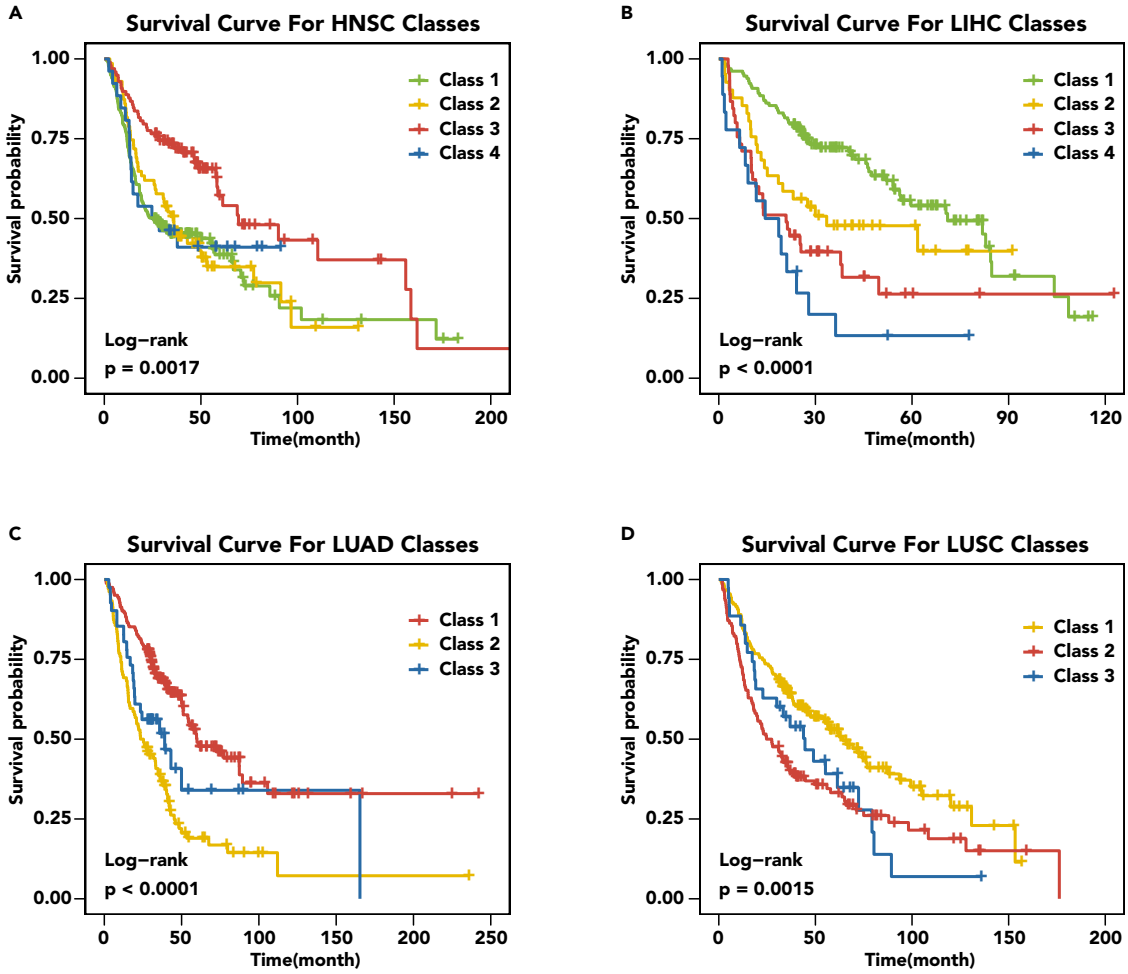


Figure 4. Kaplan-Meier survival analysis of the molecular subtypes in different cancers

(A) Survival analysis of different subtypes for LUAD.

(B) Survival analysis of different subtypes for LIHC.

(C) Survival analysis of different subtypes for LUSC.

(D) Survival analysis of different subtypes for HNSC.

Limitations of the study

Notably, there are few FDA-approved biomarkers, limiting us to exploring their evolutionary features from limited perspectives. Tumor heterogeneity is also reflected in the pathological stages of patients, which can alter patient's gene expression patterns. In the TCGA-HNSC dataset only, there was no significant difference in the stage distribution of patients. This lack of distinction may blur the boundaries between positive and negative sample expression patterns during the model training process, leading to instability in the results. Single-cell sequencing data, with its unique characteristics, offer a promising avenue for identifying biomarkers. However, single-cell technologies are advancements made in recent years. There is a lack of sufficient clinical follow-up information and the accumulated data on cancer patient survival rates. The pronounced batch effects can impact the integration of datasets, making the use of these technologies for predicting oncology prognosis biomarkers a challenging endeavor. Moreover, neural networks require more sample data to improve accuracy, but the currently available cancer databases do not adequately fulfill this requirement.[99,100] In future research endeavors, we aim to examine the applicability of evolutionary information across diverse datasets using varied methodologies. This will enable us to corroborate and refine our hypotheses from multiple perspectives.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY

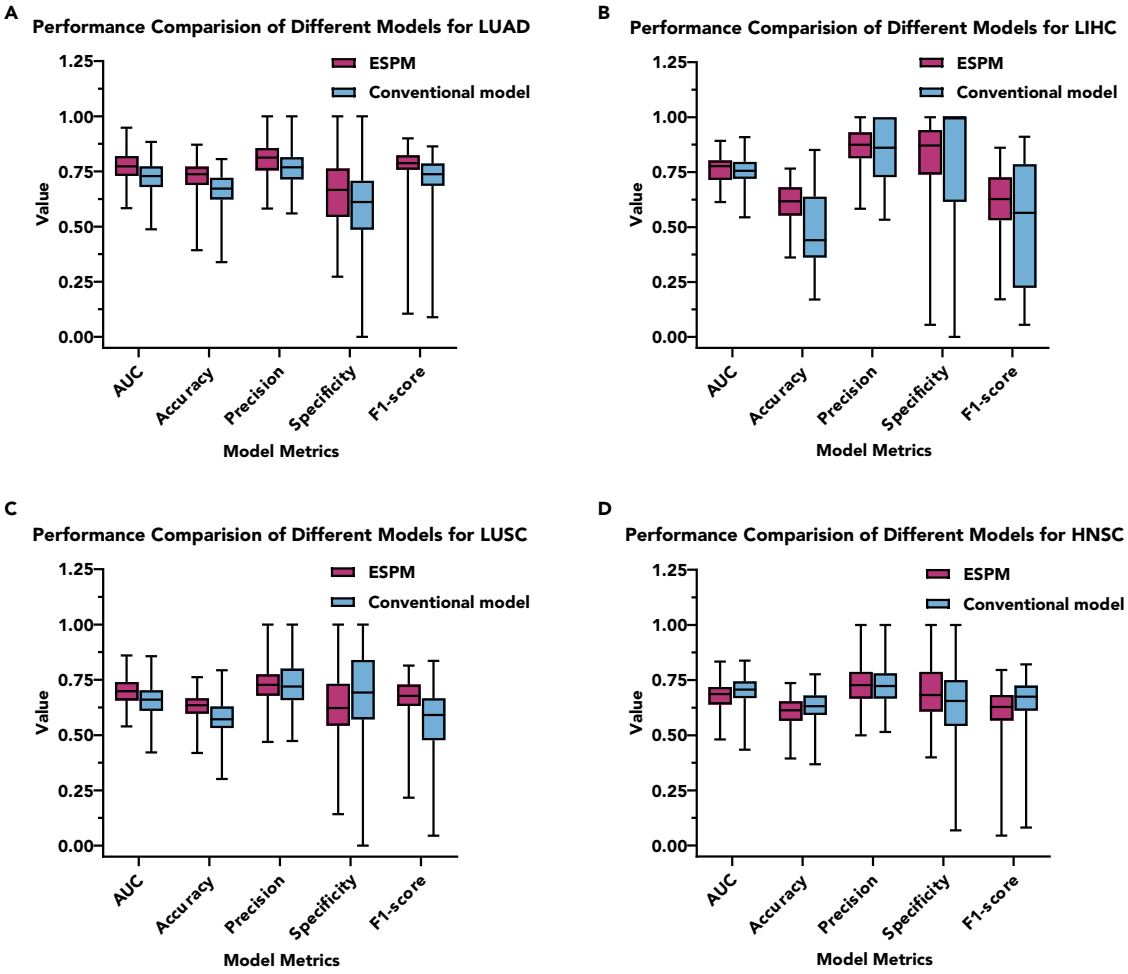


Figure 5. Performance comparison between ESPM and conventional model for each type of cancer

(A) Comparison of the performance between LUAD's ESPM and the conventional model.

(B) Comparison of the performance between LIHC's ESPM and the conventional model.

(C) Comparison of the performance between LUSC's ESPM and the conventional model.

(D) Comparison of the performance between HNSC's ESPM and the conventional model. The evaluation metrics of the model, including AUC, accuracy, precision, specificity, and F1-score, are presented in the boxplots. ESPM, evolution-strengthened prognostic model; conventional model, the model constructed by genes without evolutionary feature screening; AUC, area under the curve.

- Lead contact
- Materials availability
- Data and code availability

- METHOD DETAILS
  - Biomarker information
  - Evolutionary information of genes
  - RNA-seq and clinical data of cancer cohorts
  - Identification of evolutionary features of biomarkers
  - GO and KEGG enrichment analysis
  - Identification of potential prognostic biomarkers
  - Unsupervised consensus clustering analysis
  - Construction of the evolution-strengthened prognostic model

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.109859.

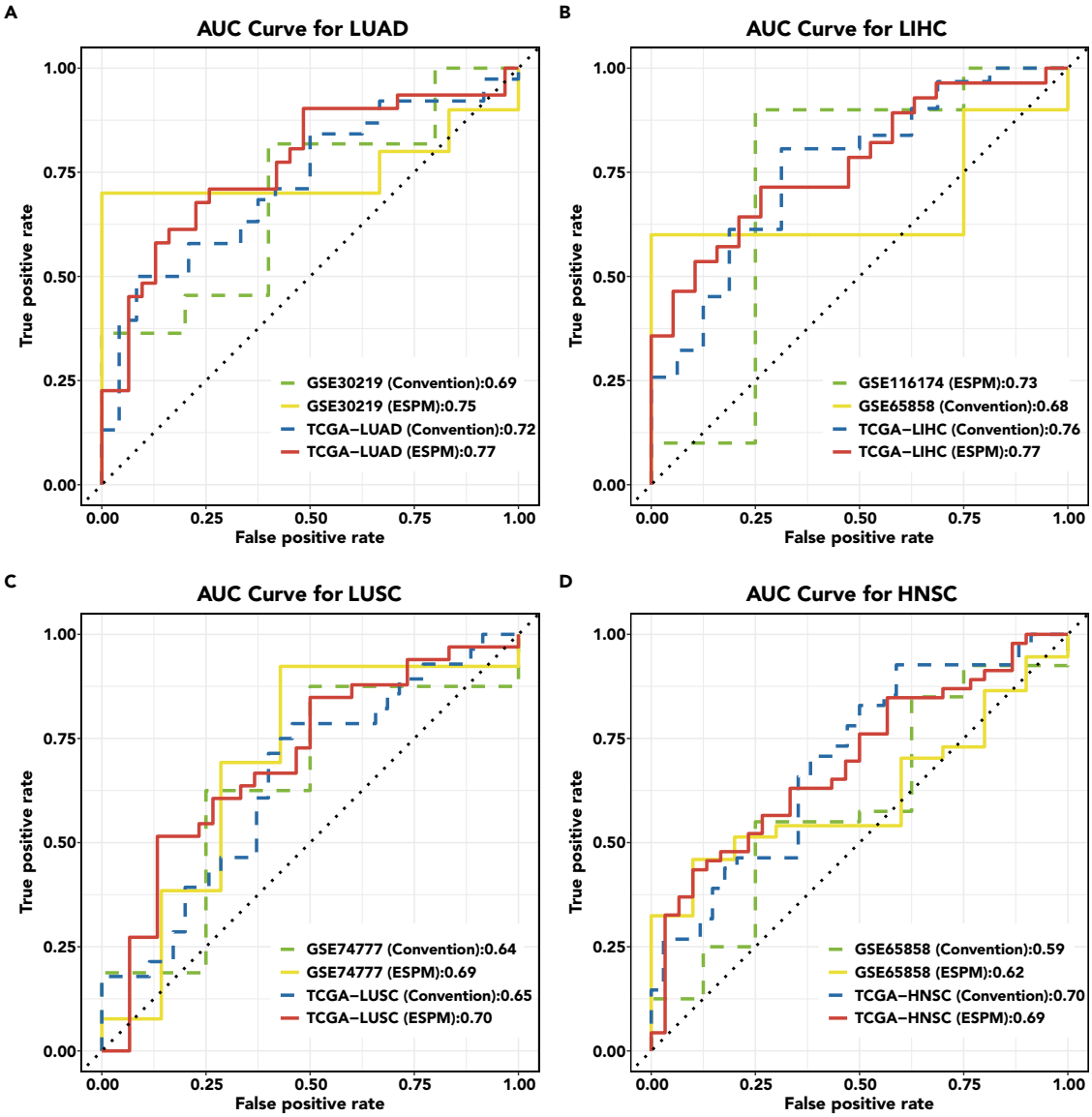


Figure 6. AUC curves for prognostic models in different cancer cohorts

(A) AUC curves for LUAD prognostic models in different cohorts.

(B) AUC curves for LIHC prognostic models in different cohorts.

(C) AUC curves for LUSC prognostic models in different cohorts.

(D) AUC curves for HNSC prognostic models in different cohorts. AUC, area under the curve.

ACKNOWLEDGMENTS

We thank the National Natural Science Foundation of China (grant number 32300545), Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), and the National Key R&D Program of China (2022YFA1304104) for grants.

AUTHOR CONTRIBUTIONS

Z.-W.Z.: writing – original draft, investigation, formal analysis, data curation, validation, and conceptualization. K.-X.Z.: writing – original draft, validation, and conceptualization. X.L.: formal analysis, data curation, and conceptualization. Y.Q.: writing – review and editing, validation, supervision, software, resources, project administration, methodology, and conceptualization. H.-Y.Z.: writing – review and editing, supervision, resources, project administration, investigation, and funding acquisition.

DECLARATION OF INTERESTS

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Received: October 10, 2023
Revised: March 15, 2024
Accepted: April 27, 2024
Published: April 30, 2024

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA A Cancer J. Clin. *71*, 209–249. https://doi.org/10.3322/caac.21660.
2. Siegel, R.L., Miller, K.D., Fuchs, H.E., and Jemal, A. (2022). Cancer statistics, 2022. CA A Cancer J. Clin. *72*, 7–33. https://doi.org/10.3322/caac.21708.
3. Bray, F., Laversanne, M., Weiderpass, E., and Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer *127*, 3029–3030. https://doi.org/10.1002/cncr.33587.
4. Ciardiello, F., Ciardiello, D., Martini, G., Napolitano, S., Tabernero, J., and Cervantes, A. (2022). Clinical management of metastatic colorectal cancer in the era of precision medicine. CA A Cancer J. Clin. *72*, 372–401. https://doi.org/10.3322/caac.21728.
5. Di, Z., Zhou, S., Xu, G., Ren, L., Li, C., Ding, Z., Huang, K., Liang, L., and Yuan, Y. (2022). Single-cell and WGCNA uncover a prognostic model and potential oncogenes in colorectal cancer. Biol. Proced. Online *24*, 13. https://doi.org/10.1186/s12575-022-00175-x.
6. Hanna, T.P., King, W.D., Thibodeau, S., Jalink, M., Paulin, G.A., Harvey-Jones, E., O'Sullivan, D.E., Booth, C.M., Sullivan, R., and Aggarwal, A. (2020). Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ *371*, m4087. https://doi.org/10.1136/bmj.m4087.
7. Tsimberidou, A.M., Fountzilas, E., Nikanjam, M., and Kurzrock, R. (2020). Review of precision cancer medicine: Evolution of the treatment paradigm. Cancer Treat Rev. *86*, 102019. https://doi.org/10.1016/j.ctrv.2020.102019.
8. Sarhadi, V.K., and Armengol, G. (2022). Molecular Biomarkers in Cancer. Biomolecules *12*, 1021. https://doi.org/10.3390/biom12081021.
9. Picard, M., Scott-Boyer, M.P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. Comput. Struct. Biotechnol. J. *19*, 3735–3746. https://doi.org/10.1016/j.csbj.2021.06.030.
10. Tsimberidou, A.M. (2015). Targeted therapy in cancer. Cancer Chemother. Pharmacol. *76*, 1113–1132. https://doi.org/10.1007/s00280-015-2861-1.
11. Xiao, Y., Bi, M., Guo, H., and Li, M. (2022). Multi-omics approaches for biomarker discovery in early ovarian cancer diagnosis. EBioMedicine *79*, 104001. https://doi.org/10.1016/j.ebiom.2022.104001.
12. Goossens, N., Nakagawa, S., Sun, X., and Hoshida, Y. (2015). Cancer biomarker discovery and validation. Transl. Cancer Res. *4*, 256–269. https://doi.org/10.3978/j.issn.2218-676X.2015.06.04.
13. Rutledge, J., Oh, H., and Wyss-Coray, T. (2022). Measuring biological age using omics data. Nat. Rev. Genet. *23*, 715–727. https://doi.org/10.1038/s41576-022-00511-7.
14. Swanson, K., Wu, E., Zhang, A., Alizadeh, A.A., and Zou, J. (2023). From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell *186*, 1772–1791. https://doi.org/10.1016/j.cell.2023.01.035.
15. Luo, X.J., Zhao, Q., Liu, J., Zheng, J.B., Qiu, M.Z., Ju, H.Q., and Xu, R.H. (2021). Novel Genetic and Epigenetic Biomarkers of Prognostic and Predictive Significance in Stage II/III Colorectal Cancer. Mol. Ther. *29*, 587–596. https://doi.org/10.1016/j.ymthe.2020.12.017.
16. Cui, G., Cai, F., Ding, Z., and Gao, L. (2019). MMP14 predicts a poor prognosis in patients with colorectal cancer. Hum. Pathol. *83*, 36–42. https://doi.org/10.1016/j.humpath.2018.03.030.
17. Tang, X., Pang, T., Yan, W.F., Qian, W.L., Gong, Y.L., and Yang, Z.G. (2020). A novel prognostic model predicting the long-term cancer-specific survival for patients with hypopharyngeal squamous cell carcinoma. BMC Cancer *20*, 1095. https://doi.org/10.1186/s12885-020-07599-2.
18. Chen, G., Luo, D., Zhong, N., Li, D., Zheng, J., Liao, H., Li, Z., Lin, X., Chen, Q., Zhang, C., et al. (2022). GPC2 Is a Potential Diagnostic, Immunological, and Prognostic Biomarker in Pan-Cancer. Front. Immunol. *13*, 857308. https://doi.org/10.3389/fimmu.2022.857308.
19. Yang, S., Ji, J., Wang, M., Nie, J., and Wang, S. (2023). Construction of Ovarian Cancer Prognostic Model Based on the Investigation of Ferroptosis-Related lncRNA. Biomolecules *13*, 306. https://doi.org/10.3390/biom13020306.
20. Zhang, H., Xia, P., Liu, J., Chen, Z., Ma, W., and Yuan, Y. (2021). ATIC inhibits autophagy in hepatocellular cancer through the AKT/FOXO3 pathway and serves as a prognostic signature for modeling patient survival. Int. J. Biol. Sci. *17*, 4442–4458. https://doi.org/10.7150/ijbs.65669.
21. Graham, T.A., and Sottoriva, A. (2017). Measuring cancer evolution from the genome. J. Pathol. *241*, 183–191. https://doi.org/10.1002/path.4821.
22. Stearns, S.C., Nesse, R.M., Govindaraju, D.R., and Ellison, P.T. (2010). Evolution in health and medicine Sackler colloquium: Evolutionary perspectives on health and medicine. Proc. Natl. Acad. Sci. USA *107*, 1691–1695. https://doi.org/10.1073/pnas.0914475107.
23. Greaves, M. (2015). Evolutionary determinants of cancer. Cancer Discov. *5*, 806–820. https://doi.org/10.1158/2159-8290.CD-15-0439.
24. Zhu, X., Li, S., Xu, B., and Luo, H. (2021). Cancer evolution: A means by which tumors evade treatment. Biomed. Pharma *133*, 111016. https://doi.org/10.1016/j.biopha.2020.111016.
25. Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discov. *12*, 31–46. https://doi.org/10.1158/2159-8290.CD-21-1059.
26. Cisneros, L., Bussey, K.J., Orr, A.J., Miočević, M., Lineweaver, C.H., and Davies, P. (2017). Ancient genes establish stress-induced mutation as a hallmark of cancer. PLoS One *12*, e0176258. https://doi.org/10.1371/journal.pone.0176258.
27. Lineweaver, C.H., Bussey, K.J., Blackburn, A.C., and Davies, P.C.W. (2021). Cancer progression as a sequence of atavistic reversions. Bioessays *43*, e2000305. https://doi.org/10.1002/bies.202000305.
28. Domazet-Loso, T., and Tautz, D. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. BMC Biol. *8*, 66. https://doi.org/10.1186/1741-7007-8-66.
29. Jacques, F., Baratchart, E., Pienta, K.J., and Hammarlund, E.U. (2022). Origin and evolution of animal multicellularity in the light of phylogenomics and cancer genetics. Med. Oncol. *39*, 160. https://doi.org/10.1007/s12032-022-01740-w.
30. Singh, P.P., Arora, J., and Isambert, H. (2015). Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. PLoS Comput. Biol. *11*, e1004394. https://doi.org/10.1371/journal.pcbi.1004394.
31. Makino, T., and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc. Natl. Acad. Sci. USA *107*, 9270–9274. https://doi.org/10.1073/pnas.0914697107.
32. Xie, T., Yang, Q.Y., Wang, X.T., McLysaght, A., and Zhang, H.Y. (2016). Spatial Colocalization of Human Ohnolog Pairs Acts to Maintain Dosage-Balance. Mol. Biol. Evol. *33*, 2368–2375. https://doi.org/10.1093/molbev/msw108.
33. Singh, P.P., Affeldt, S., Cascone, I., Selimoglu, R., Camonis, J., and Isambert, H. (2012). On the expansion of "dangerous" gene repertoires by whole-genome duplications in early vertebrates. Cell Rep. *2*,

1387–1398. https://doi.org/10.1016/j.celrep.2012.09.034.
34. Trigos, A.S., Pearson, R.B., Papenfuss, A.T., and Goode, D.L. (2017). Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. Proc. Natl. Acad. Sci. USA *114*, 6406–6411. https://doi.org/10.1073/pnas.1617743114.
35. Liebeskind, B.J., McWhite, C.D., and Marcotte, E.M. (2016). Towards Consensus Gene Ages. Genome Biol. Evol. *8*, 1812–1823. https://doi.org/10.1093/gbe/evw113.
36. Chu, X.Y., Jiang, L.H., Zhou, X.H., Cui, Z.J., and Zhang, H.Y. (2017). Evolutionary Origins of Cancer Driver Genes and Implications for Cancer Prognosis. Genes *8*, 182. https://doi.org/10.3390/genes8070182.
37. Trigos, A.S., Pearson, R.B., Papenfuss, A.T., and Goode, D.L. (2018). How the evolution of multicellularity set the stage for cancer. Br. J. Cancer *118*, 145–152. https://doi.org/10.1038/bjc.2017.398.
38. Awad, M.M., Liu, S., Rybkin, I.I., Arbour, K.C., Dilly, J., Zhu, V.W., Johnson, M.L., Heist, R.S., Patil, T., Riely, G.J., et al. (2021). Acquired Resistance to KRASG12C Inhibition in Cancer. N. Engl. J. Med. *384*, 2382–2393. https://doi.org/10.1056/NEJMoa2105281.
39. Uprety, D., and Adjei, A.A. (2020). KRAS: From undruggable to a druggable Cancer Target. Cancer Treat Rev. *89*, 102070. https://doi.org/10.1016/j.ctrv.2020.102070.
40. Wong, G.S., Zhou, J., Liu, J.B., Wu, Z., Xu, X., Li, T., Xu, D., Schumacher, S.E., Puschhof, J., McFarland, J., et al. (2018). Targeting wild-type KRAS-amplified gastroesophageal cancer through combined MEK and SHP2 inhibition. Nat. Med. *24*, 968–977. https://doi.org/10.1038/s41591-018-0022-x.
41. Punekar, S.R., Velcheti, V., Neel, B.G., and Wong, K.K. (2022). The current state of the art and future trends in RAS-targeted cancer therapies. Nat. Rev. Clin. Oncol. *19*, 637–655. https://doi.org/10.1038/s41571-022-00671-9.
42. Du, Z., and Lovly, C.M. (2018). Mechanisms of receptor tyrosine kinase activation in cancer. Mol. Cancer *17*, 58. https://doi.org/10.1186/s12943-018-0782-4.
43. Talukdar, S., Emdad, L., Das, S.K., and Fisher, P.B. (2020). EGFR: An essential receptor tyrosine kinase-regulator of cancer stem cells. Adv. Cancer Res. *147*, 161–188. https://doi.org/10.1016/bs.acr.2020.04.003.
44. Cheng, W.L., Feng, P.H., Lee, K.Y., Chen, K.Y., Sun, W.L., Van Hiep, N., Luo, C.S., and Wu, S.M. (2021). The Role of EREG/EGFR Pathway in Tumor Progression. Int. J. Mol. Sci. *22*, 12828. https://doi.org/10.3390/ijms222312828.
45. Levantini, E., Maroni, G., Del Re, M., and Tenen, D.G. (2022). EGFR signaling pathway as therapeutic target in human cancers. Semin. Cancer Biol. *85*, 253–275. https://doi.org/10.1016/j.semcancer.2022.04.002.
46. Friedlaender, A., Subbiah, V., Russo, A., Banna, G.L., Malapelle, U., Rolfo, C., and Addeo, A. (2022). EGFR and HER2 exon 20 insertions in solid tumours: from biology to treatment. Nat. Rev. Clin. Oncol. *19*, 51–69. https://doi.org/10.1038/s41571-021-00558-1.
47. Takahashi, M., Ritz, J., and Cooper, G.M. (1985). Activation of a novel human transforming gene, ret, by DNA rearrangement. Cell *42*, 581–588. https://doi.org/10.1016/0092-8674(85)90115-1.
48. Salvatore, D., Santoro, M., and Schlumberger, M. (2021). The importance of the RET gene in thyroid cancer and therapeutic implications. Nat. Rev. Endocrinol. *17*, 296–306. https://doi.org/10.1038/s41574-021-00470-9.
49. Ding, S., Wang, R., Peng, S., Luo, X., Zhong, L., Yang, H., Ma, Y., Chen, S., and Wang, W. (2020). Targeted therapies for RET-fusion cancer: Dilemmas and breakthrough. Biomed. Pharmacother. *132*, 110901. https://doi.org/10.1016/j.biopha.2020.110901.
50. Thein, K.Z., Velcheti, V., Mooers, B.H.M., Wu, J., and Subbiah, V. (2021). Precision therapy for RET-altered cancers with RET inhibitors. Trends Cancer *7*, 1074–1088. https://doi.org/10.1016/j.trecan.2021.07.003.
51. Wang, S., Xiong, Y., Zhang, Q., Su, D., Yu, C., Cao, Y., Pan, Y., Lu, Q., Zuo, Y., and Yang, L. (2021). Clinical significance and immunogenomic landscape analyses of the immune cell signature based prognostic model for patients with breast cancer. Briefings Bioinf. *22*, bbaa311. https://doi.org/10.1093/bib/bbaa311.
52. Jurcak, N.R., Rucki, A.A., Muth, S., Thompson, E., Sharma, R., Ding, D., Zhu, Q., Eshleman, J.R., Anders, R.A., Jaffee, E.M., et al. (2019). Axon Guidance Molecules Promote Perineural Invasion and Metastasis of Orthotopic Pancreatic Tumors in Mice. Gastroenterology *157*, 838–850.e6. https://doi.org/10.1053/j.gastro.2019.05.065.
53. Patergnani, S., Danese, A., Bouhamida, E., Aguiari, G., Previati, M., Pinton, P., and Giorgi, C. (2020). Various Aspects of Calcium Signaling in the Regulation of Apoptosis, Autophagy, Cell Proliferation, and. Int. J. Mol. Sci. *21*, 8323. https://doi.org/10.3390/ijms21218323.
54. Lin, X., Zhuang, S., Chen, X., Du, J., Zhong, L., Ding, J., Wang, L., Yi, J., Hu, G., Tang, G., et al. (2022). lncRNA ITGB8-AS1 functions as a ceRNA to promote colorectal cancer growth and migration through integrin-mediated focal adhesion signaling. Mol. Ther. *30*, 688–702. https://doi.org/10.1016/j.ymthe.2021.08.011.
55. Li, Y., Sheng, H., Ma, F., Wu, Q., Huang, J., Chen, Q., Sheng, L., Zhu, X., Zhu, X., and Xu, M. (2021). RNA m6A reader YTHDF2 facilitates lung adenocarcinoma cell proliferation and metastasis by targeting the AXIN1/Wnt/β-catenin signaling. Cell Death Dis. *12*, 479. https://doi.org/10.1038/s41419-021-03763-z.
56. Jiang, N., Zou, C., Zhu, Y., Luo, Y., Chen, L., Lei, Y., Tang, K., Sun, Y., Zhang, W., Li, S., et al. (2020). HIF-1α-regulated miR-1275 maintains stem cell-like phenotypes and promotes the progression of LUAD by simultaneously activating Wnt/β-catenin and Notch signaling. Theranostics *10*, 2553–2570. https://doi.org/10.7150/thno.41120.
57. Li, S., Yang, F., Wang, M., Cao, W., and Yang, Z. (2017). miR-378 functions as an onco-miRNA by targeting the ST7L/Wnt/β-catenin pathway in cervical cancer. Int. J. Mol. Med. *40*, 1047–1056. https://doi.org/10.3892/ijmm.2017.3116.
58. Duchartre, Y., Kim, Y.M., and Kahn, M. (2016). The Wnt signaling pathway in cancer. Crit. Rev. Oncol. Hematol. *99*, 141–149. https://doi.org/10.1016/j.critrevonc.2015.12.005.
59. Xu, X., Zhang, M., Xu, F., and Jiang, S. (2020). Wnt signaling in breast cancer: biological mechanisms, challenges and opportunities. Mol. Cancer *19*, 165. https://doi.org/10.1186/s12943-020-01276-5.
60. Lee, S., Rauch, J., and Kolch, W. (2020). Targeting MAPK Signaling in Cancer: Mechanisms of Drug Resistance and Sensitivity. Int. J. Mol. Sci. *21*, 1102. https://doi.org/10.3390/ijms21031102.
61. Wagner, E.F., and Nebreda, A.R. (2009). Signal integration by JNK and p38 MAPK pathways in cancer development. Nat. Rev. Cancer *9*, 537–549. https://doi.org/10.1038/nrc2694.
62. Zhu, H., Liu, Q., Yang, X., Ding, C., Wang, Q., and Xiong, Y. (2022). LncRNA LINC00649 recruits TAF15 and enhances MAPK6 expression to promote the development of lung squamous cell carcinoma via activating MAPK signaling pathway. Cancer Gene Ther. *29*, 1285–1295. https://doi.org/10.1038/s41417-021-00410-9.
63. Lian, W., Wang, Z., Ma, Y., Tong, Y., Zhang, X., Jin, H., Zhao, S., Yu, R., Ju, S., Zhang, X., et al. (2022). FABP6 Expression Correlates with Immune Infiltration and Immunogenicity in Colorectal Cancer Cells. J. Immunol. Res. *2022*, 3129765. https://doi.org/10.1155/2022/3129765.
64. Zhang, Y., Zhao, X., Deng, L., Li, X., Wang, G., Li, Y., and Chen, M. (2019). High expression of FABP4 and FABP6 in patients with colorectal cancer. World J. Surg. Oncol. *17*, 171. https://doi.org/10.1186/s12957-019-1714-5.
65. Pai, F.C., Huang, H.W., Tsai, Y.L., Tsai, W.C., Cheng, Y.C., Chang, H.H., and Chen, Y. (2021). Inhibition of FABP6 Reduces Tumor Cell Invasion and Angiogenesis through the Decrease in MMP-2 and VEGF in Human Glioblastoma Cells. Cells *10*, 2782. https://doi.org/10.3390/cells10102782.
66. Lin, C.H., Chang, H.H., Lai, C.R., Wang, H.H., Tsai, W.C., Tsai, Y.L., Changchien, C.Y., Cheng, Y.C., Wu, S.T., and Chen, Y. (2022). Fatty Acid Binding Protein 6 Inhibition Decreases Cell Cycle Progression, Migration and Autophagy in Bladder Cancers. Int. J. Mol. Sci. *23*, 2154. https://doi.org/10.3390/ijms23042154.
67. Lin, J., Yang, J., Xu, X., Wang, Y., Yu, M., and Zhu, Y. (2020). A robust 11-genes prognostic model can predict overall survival in bladder cancer patients based on five cohorts. Cancer Cell Int. *20*, 402. https://doi.org/10.1186/s12935-020-01491-6.
68. Hu, B., Yang, X.B., and Sang, X.T. (2020). Development of an immune-related prognostic index associated with hepatocellular carcinoma. Aging *12*, 5010–5030. https://doi.org/10.18632/aging.102926.
69. Munson, M.J., Mathai, B.J., Ng, M.Y.W., Trachsel-Moncho, L., de la Ballina, L.R., and Simonsen, A. (2022). GAK and PRKCD kinases regulate basal mitophagy. Autophagy *18*, 467–469. https://doi.org/10.1080/15548627.2021.2015154.
70. Munson, M.J., Mathai, B.J., Ng, M.Y.W., Trachsel-Moncho, L., de la Ballina, L.R., Schultz, S.W., Aman, Y., Lystad, A.H., Singh, S., Singh, S., et al. (2021). GAK and PRKCD are positive regulators of PRKN-independent mitophagy. Nat. Commun. *12*, 6101. https://doi.org/10.1038/s41467-021-26331-7.
71. Chen, Y., Ke, G., Han, D., Liang, S., Yang, G., and Wu, X. (2014). MicroRNA-181a enhances the chemoresistance of human cervical squamous cell carcinoma to

cisplatin by targeting PRKCD. Exp. Cell Res. *320*, 12–20. https://doi.org/10.1016/j.yexcr.2013.10.014.

72. Ke, G., Liang, L., Yang, J.M., Huang, X., Han, D., Huang, S., Zhao, Y., Zha, R., He, X., and Wu, X. (2013). MiR-181a confers resistance of cervical cancer to radiation therapy through targeting the pro-apoptotic PRKCD gene. Oncogene *32*, 3019–3027. https://doi.org/10.1038/onc.2012.323.
73. Yao, L., Wang, L., Li, F., Gao, X., Wei, X., and Liu, Z. (2015). MiR181c inhibits ovarian cancer metastasis and progression by targeting PRKCD expression. Int. J. Clin. Exp. Med. *8*, 15198–15205.
74. Munksgaard, P.P., Mansilla, F., Brems Eskildsen, A.S., Fristrup, N., Birkenkamp-Demtröder, K., Ulhøi, B.P., Borre, M., Agerbæk, M., Hermann, G.G., Orntoft, T.F., and Dyrskjøt, L. (2011). Low ANXA10 expression is associated with disease aggressiveness in bladder cancer. Br. J. Cancer *105*, 1379–1387. https://doi.org/10.1038/bjc.2011.404.
75. Seidlitz, T., Chen, Y.T., Uhlemann, H., Schölch, S., Kochall, S., Merker, S.R., Klimova, A., Hennig, A., Schweitzer, C., Pape, K., et al. (2019). Mouse Models of Human Gastric Cancer Subtypes With Stomach-Specific CreERT2-Mediated Pathway Alterations. Gastroenterology *157*, 1599–1614.e2. https://doi.org/10.1053/j.gastro.2019.09.026.
76. Xiong, M., Chen, L., Zhou, L., Ding, Y., Kazobinka, G., Chen, Z., and Hou, T. (2019). NUDT21 inhibits bladder cancer progression through ANXA2 and LIMK2 by alternative polyadenylation. Theranostics *9*, 7156–7167. https://doi.org/10.7150/thno.36030.
77. Zhang, C., Peng, L., Gu, H., Wang, J., Wang, Y., and Xu, Z. (2023). ANXA10 is a prognostic biomarker and suppressor of hepatocellular carcinoma: a bioinformatics analysis and experimental validation. Sci. Rep. *13*, 1583. https://doi.org/10.1038/s41598-023-28527-x.
78. Hung, M.S., Chen, Y.C., Lin, P., Li, Y.C., Hsu, C.C., Lung, J.H., You, L., Xu, Z., Mao, J.H., Jablons, D.M., and Yang, C.T. (2019). Cul4A Modulates Invasion and Metastasis of Lung Cancer Through Regulation of ANXA10. Cancers *11*, 618. https://doi.org/10.3390/cancers11050618.
79. Su, J., Chen, D., Ruan, Y., Tian, Y., Lv, K., Zhou, X., Ying, D., and Lu, Y. (2022). LncRNA MBNL1-AS1 represses gastric cancer progression via the TGF-β pathway by modulating miR-424-5p/Smad7 axis. Bioengineered *13*, 6978–6995. https://doi.org/10.1080/21655979.2022.2037921.
80. Tran, K.A., Kondrashova, O., Bradley, A., Williams, E.D., Pearson, J.V., and Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med. *13*, 152. https://doi.org/10.1186/s13073-021-00968-x.
81. Hu, X., Cammann, H., Meyer, H.A., Miller, K., Jung, K., and Stephan, C. (2013). Artificial neural networks and prostate cancer–tools for diagnosis and management. Nat. Rev. Urol. *10*, 174–182. https://doi.org/10.1038/nrurol.2013.9.
82. Batis, N., Brooks, J.M., Payne, K., Sharma, N., Nankivell, P., and Mehanna, H. (2021). Lack of predictive tools for conventional and targeted cancer therapy: Barriers to biomarker development and clinical translation. Adv. Drug Deliv. Rev. *176*, 113854. https://doi.org/10.1016/j.addr.2021.113854.
83. Poste, G. (2011). Bring on the biomarkers. Nature *469*, 156–157. https://doi.org/10.1038/469156a.
84. Takan, I., Karakülah, G., Louka, A., and Pavlopoulou, A. (2023). "In the light of evolution:" keratins as exceptional tumor biomarkers. PeerJ *11*, e15099. https://doi.org/10.7717/peerj.15099.
85. Gerhauser, C., Favero, F., Risch, T., Simon, R., Feuerbach, L., Assenov, Y., Heckmann, D., Sidiropoulos, N., Waszak, S.M., Hübschmann, D., et al. (2018). Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. Cancer Cell *34*, 996–1011.e8. https://doi.org/10.1016/j.ccell.2018.10.016.
86. Gillard, G.B., Grønvold, L., Røsæg, L.L., Holen, M.M., Monsen, Ø., Koop, B.F., Rondeau, E.B., Gundappa, M.K., Mendoza, J., Macqueen, D.J., et al. (2021). Comparative regulomics supports pervasive selection on gene dosage following whole genome duplication. Genome Biol. *22*, 103. https://doi.org/10.1186/s13059-021-02323-0.
87. Pajkos, M., Zeke, A., and Dosztányi, Z. (2020). Ancient Evolutionary Origin of Intrinsically Disordered Cancer Risk Regions. Biomolecules *10*, 1115. https://doi.org/10.3390/biom10081115.
88. He, Y., Sun, M.M., Zhang, G.G., Yang, J., Chen, K.S., Xu, W.W., and Li, B. (2021). Targeting PI3K/Akt signal transduction for cancer therapy. Signal Transduct. Targeted Ther. *6*, 425. https://doi.org/10.1038/s41392-021-00828-5.
89. Chen, Y., Li, Z.Y., Zhou, G.Q., and Sun, Y. (2021). An Immune-Related Gene Prognostic Index for Head and Neck Squamous Cell Carcinoma. Clin. Cancer Res. *27*, 330–341. https://doi.org/10.1158/1078-0432.CCR-20-2166.
90. Lam, Y.K., Yu, J., Huang, H., Ding, X., Wong, A.M., Leung, H.H., Chan, A.W., Ng, K.K., Xu, M., Wang, X., and Wong, N. (2023). TP53 R249S mutation in hepatic organoids captures the predisposing cancer risk. Hepatology *78*, 727–740. https://doi.org/10.1002/hep.32802.
91. LeSavage, B.L., Suhar, R.A., Broguiere, N., Lutolf, M.P., and Heilshorn, S.C. (2022). Next-generation cancer organoids. Nat. Mater. *21*, 143–159. https://doi.org/10.1038/s41563-021-01057-5.
92. Ahmed, M.B., Alghamdi, A.A.A., Islam, S.U., Lee, J.S., and Lee, Y.S. (2022). cAMP Signaling in Cancer: A PKA-CREB and EPAC-Centric Approach. Cells *11*, 2020. https://doi.org/10.3390/cells11132020.
93. Kashyap, A., Rapsomaniki, M.A., Barros, V., Fomitcheva-Khartchenko, A., Martinelli, A.L., Rodriguez, A.F., Gabrani, M., Rosen-Zvi, M., and Kaigala, G. (2022). Quantification of tumor heterogeneity: from data acquisition to metric generation. Trends Biotechnol. *40*, 647–676. https://doi.org/10.1016/j.tibtech.2021.11.006.
94. Wishart, D. (2022). Metabolomics and the Multi-Omics View of Cancer. Metabolites *12*, 154. https://doi.org/10.3390/metabo12020154.
95. Vasaikar, S.V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res. *46*, 956–963. https://doi.org/10.1093/nar/gkx1090.
96. Chai, H., Zhou, X., Zhang, Z., Rao, J., Zhao, H., and Yang, Y. (2021). Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput. Biol. Med. *134*, 104481. https://doi.org/10.1016/j.compbiomed.2021.104481.
97. Poirion, O.B., Jing, Z., Chaudhary, K., Huang, S., and Garmire, L.X. (2021). DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome Med. *13*, 112. https://doi.org/10.1186/s13073-021-00930-x.
98. Huang, W., Chen, J., Weng, W., Xiang, Y., Shi, H., and Shan, Y. (2020). Development of cancer prognostic signature based on pan-cancer proteomics. Bioengineered *11*, 1368–1381. https://doi.org/10.1080/21655979.2020.1847398.
99. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. *51*, 12–18. https://doi.org/10.1038/s41588-018-0295-5.
100. Kang, M., Ko, E., and Mersha, T.B. (2022). A roadmap for multi-omics data integration using deep learning. Briefings Bioinf. *23*, bbab454. https://doi.org/10.1093/bib/bbab454.
101. Zhou, Y., Zhang, Y., Lian, X., Li, F., Wang, C., Zhu, F., Qiu, Y., and Chen, Y. (2022). Therapeutic target database update 2022: facilitating drug discovery with enriched comparative data of targeted agents. Nucleic Acids Res. *50*, 1398–1407. https://doi.org/10.1093/nar/gkab953.
102. Wishart, D.S., Bartok, B., Oler, E., Liang, K.Y.H., Budinski, Z., Berjanskii, M., Guo, A., Cao, X., and Wilson, M. (2021). MarkerDB: an online database of molecular biomarkers. Nucleic Acids Res. *49*, 1259–D1267. https://doi.org/10.1093/nar/gkaa1067.
103. Shi, R., Bao, X., Unger, K., Sun, J., Lu, S., Manapov, F., Wang, X., Belka, C., and Li, M. (2021). Identification and validation of hypoxia-derived gene signatures to predict clinical outcomes and therapeutic responses in stage I lung adenocarcinoma patients. Theranostics *11*, 5061–5076. https://doi.org/10.7150/thno.56202.
104. Lee, J.H., Jung, S., Park, W.S., Choe, E.K., Kim, E., Shin, R., Heo, S.C., Lee, J.H., Kim, K., and Chai, Y.J. (2019). Prognostic nomogram of hypoxia-related genes predicting overall survival of colorectal cancer-Analysis of TCGA database. Sci. Rep. *9*, 1803. https://doi.org/10.1038/s41598-018-38116-y.
105. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R., et al. (2020). Ensembl 2020. Nucleic Acids Res. *48*, 682–688. https://doi.org/10.1093/nar/gkz966.
106. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinf. *12*, 323. https://doi.org/10.1186/1471-2105-12-323.
107. Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS *16*, 284–287. https://doi.org/10.1089/omi.2011.0118.
108. Pinto, J.A., Araujo, J., Cardenas, N.K., Morante, Z., Doimi, F., Vidaurre, T., Balko, J.M., and Gomez, H.L. (2016). A prognostic

signature based on three-genes expression in triple-negative breast tumours with residual disease. NPJ Genom. Med. *1*, 15015. https://doi.org/10.1038/npjgenmed.2015.15.

109. Cheong, J.H., Wang, S.C., Park, S., Porembka, M.R., Christie, A.L., Kim, H., Kim, H.S., Zhu, H., Hyung, W.J., Noh, S.H., et al. (2022). Development and validation of a prognostic and predictive 32-gene signature for gastric cancer. Nat. Commun. *13*, 774. https://doi.org/10.1038/s41467-022-28437-y.

110. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170.

111. Gao, F., Wang, W., Tan, M., Zhu, L., Zhang, Y., Fessler, E., Vermeulen, L., and Wang, X. (2019). DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. Oncogenesis *8*, 44. https://doi.org/10.1038/s41389-019-0157-8.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
TCGA-HNSC\LIHC\LUAD\LUSC	TCGA	https://portal.gdc.cancer.gov/
GSE30219	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30219
GSE65858	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65858
GSE74477	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE74477
GSE116174	GEO	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE116174
TTD-biomarkers	TTD	https://db.idrblab.org/ttd/
Pharmacogenomic biomarkers in drug labeling	FDA	https://www.fda.gov/
MarkerDB oncology biomarkers	MarkerDB	https://markerdb.ca/
Ohnologs	Makino et al.[31]	https://doi.org/10.1073/pnas.0914697107
Evolutionary stages of genes	Liebeskind et al.[35]	https://github.com/marcottelab/Gene-Ages/
Source Code	This paper	https://github.com/cdoebra/ESPMs
Software and algorithms		
BioMart	Ensembl	http://asia.ensembl.org/biomart/martview/
TPM	Li et al.[105]	https://doi.org/10.1186/1471-2105-12-323
R version 4.2.2	R software	https://www.r-project.org/
Python version 3.6.8	Python software	https://www.python.org/
clusterProfiler	Bioconductor	https://bioconductor.org/packages/clusterProfiler
Survival	R-project	https://cran.r-project.org/web/packages/survival/
Scikit-learn	Sklearn package	https://scikit-learn.org/stable/install.html
ConsensusClusterPlus	Bioconductor	https://bioconductor.org/packages/ConsensusClusterPlus
keras	Keras software	https://keras.io/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Hong-yu Zhang (e-mail: zhy630@mail.hzau.edu.cn).

Materials availability

This study did not generate new unique reagents or materials.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- Original codes have been deposited at Github, and are publicly accessible as of the date of publication. Open access link is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Biomarker information

The biomarker annotations utilized in this study were obtained from the Therapeutic Target Database version 7.1.01[101] and the MarkerDB database.[102] The TTD database contains 1514 biomarkers, including 119 clinical biomarkers and 23 Food and Drug Administration (FDA)-approved biomarkers. The types of biomarkers include classification, diagnosis, detection, prognosis, monitoring, theragnosis, and pharmacodynamics. The list of FDA oncology pharmacogenomic biomarkers in drug labeling was downloaded from the website (key resources table).

Evolutionary information of genes

Based on the work by Makino et al.,[31] we downloaded 9,057 Ohnolog pairs and used the BioMart of the Ensembl database to convert the Ensembl ID of the Ohnolog genes to corresponding gene symbol. After screening and de-duplication, 7,090 human genes were obtained. The evolutionary stages corresponding to human genes were downloaded from the work of Liebeskind et al.[35] They classified the origin of human genes into eight evolutionary stages: cellular organisms, euk_archaea, eukaryota, opisthokonta, eumetazoa, vertebrata, mammalia, and euk+bac, containing 813, 201, 5242, 1030, 4568, 2484, 2181, and 1396 genes, respectively.

RNA-seq and clinical data of cancer cohorts

The RNA-seq and clinical data used in this study were obtained from The Cancer Genome Atlas Project (TCGA) portal. We downloaded data from cancer cohorts with sample sizes >300. We extracted the Transcripts Per Kilobase Million (TPM) values of samples that met the following criteria: (1) tissues were collected from primary cancer and (2) complete overall survival information, including follow-up time and vital status,[103,104] and survival time over 30 days. For a gene that had multiple Ensembl IDs, we calculated the average level as the final expression level.[105] For each cancer, the mean survival time of the dead samples was calculated and used to classify the samples into positive and negative samples. Only two types of patients were used for subsequent analysis: (1) survival status was "Alive" and follow-up time longer than mean survival; or (2) survival status was "Dead." Prognostic predictions for rapidly progressing cancers may be more meaningful, and these types of cancers are relatively less influenced by environmental and external factors. Considering the reliability and practical significance of the results, we selected cancers with an average survival time of about 36 months or less and with a ratio that did not exceed 1 to 2 between positive and negative samples.

We obtained four cancers, head and neck squamous cell carcinoma (HNSC), liver hepatocellular carcinoma (LIHC), lung adenocarcinoma (LUAD), and lung squamous cell carcinoma (LUSC), for further analysis. Four independent cancer cohorts, namely, GSE65858 (HNSC, n = 226), GSE116174 (LIHC, n = 64), GSE30219 (LUAD, n = 83), and GSE74477 (LUSC, n = 107) were obtained from the Gene Expression Omnibus (GEO) database for portability validation of ESPMs. Low-quality and duplicate arrays were removed from each dataset based on the same criteria as the TCGA data, and the expression values were log2-transformed and normalized using RMA. We converted gene expression to TPM for model construction like TCGA datasets.[106]

$$\mathrm{TPM}(x) = \frac{C_x / L_x \times 10^6}{\sum_{i=1}^{N} C_i / L_i} \quad \text{(Equation 1)}$$

Where x can represent a gene, a transcript, or a specific region on the genome, C_x denotes the number of reads aligned to the exonic region of gene x, L_x represents the number of bases included in the exonic region of gene x, and N indicates the total number of genes in the sample.

Identification of evolutionary features of biomarkers

All biomarkers downloaded from the TTD database are divided into two groups according to their research stage: FDA-approved and those undergoing clinical trials. Correlations between biomarkers and Ohnolog genes were determined by hypergeometric distribution test. The p-value can be computed using the formula as follows:

$$Pvalue = 1 - \sum_{i=0}^{x-1} \frac{\left(C_K^i \times C_{M-K}^{N-i}\right)}{C_M^N} \quad \text{(Equation 2)}$$

The number of approved or research biomarkers is K. The total number of biomarkers is M and the number of Ohnologs in all biomarkers is N. x represents the number of Ohnologs in approved or in-research biomarkers. Results are considered significant when the p-value is less than 0.05.

GO and KEGG enrichment analysis

Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis compare the functional annotation information known from a species' genome with the gene list to determine which functions or pathways appear more frequently in these genes than in the randomly expected value. The enrichment analysis in this paper was implemented using the "clusterProfiler" package of R.[107] The terms with an adjusted p-value < 0.05 were considered significantly enriched genes, and only the top 10 GeneRatio were demonstrated for each cancer.

Identification of potential prognostic biomarkers

Patients were divided into two groups, high expression (top 25%) and low expression (bottom 25%), based on the expression levels of genes. The Cox proportional hazards model was utilized to evaluate overall survival (OS) and survival status.[108,109] The Cox regression analysis was performed using the "Survival" package of R. The hazard ratio (HR) was calculated to quantify this correlation, as expressed below:

$$HR(t) = \frac{h_1(t)}{h_0(t)} = e^{\gamma_1} \quad \text{(Equation 3)}$$

When only 1 exposure variable remains, the risk of the sample at time t is $h_1(t)$, and the baseline risk is $h_0(t)$. γ_1 is the coefficient by which the gene expression value affects patient survival. Taking into account the inevitable heterogeneity among tumors and to prevent the discrepancy between the number of input features and the sample size from adversely impacting the model's performance, we have established distinct significance thresholds for each type of cancer when selecting survival-significant genes (the p-values for LUSC, LUAD, LIHC, and HNSC were set to be 0.05, 0.025, 0.001, and 0.025, respectively). Then, we selected genes with evolutionary features as potential biomarkers.

Unsupervised consensus clustering analysis

The unsupervised consensus clustering ($k = 2 - 8$) algorithm and visualization were conducted using the "ConsensusClusterPlus" package in R[110] to explore cancer molecular classification based on the expression matrix of the potential oncology biomarkers. The optimal numbers of clusters for each cancer were identified by visual inspection and consideration of the cumulative distribution function (CDF) and the Delta areas for each k group. We then performed the Kaplan–Meier survival curves from the "survival" R packages to explore the prognosis among different cluster.

Construction of the evolution-strengthened prognostic model

To identify the most salient features, we used the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm to further filter the potential biomarkers. This process used a 10-fold cross-validation with 10000 iterations to determine the optimal α. The optimal α value was employed to construct the most appropriate model by capturing the optimal features and their corresponding LASSO coefficients. The entire process was executed through the 'sklearn' module of the Python software.

ESPM was a supervised and gene expression level-based model. The deep learning framework of ESPM was implemented in Python using the 'keras' module. The classifier is implemented by drawing references from DeepC.[111] We built a fully trainable connected multilayer artificial neural network (ANN) perceptron with seven hidden layers. The SELU activation function was utilized in each hidden layer, whereas the final output layer used SoftMax. The entire network is initialized using the Glorot uniform distribution method, and the optimizer can be selected from either SDG or Adam.

Each neuron in a layer is connected to every neuron in the subsequent layer. We assume the current layer t of the neural network has k neurons and layer $t - 1$ has l neurons. The information transfer between two neurons $i(i \in \{1, \cdots k\})$, $j(j \in \{1, \cdots k\})$ depends on the weight w_{ij} and the bias value b_j between them, which can be expressed as:

$$h_j^t = f\left(\sum_{i=1}^{l} w_{ij} h_i^{t-1} + b_j\right) \quad \text{(Equation 4)}$$

Here, h and f denote the outputs of the neuron and the activation function, respectively. The output layer then maps these values to probabilities for each category, enabling the model to make predictions about the data. To prevent overfitting, we employed 10-fold cross-validation and a callback function. Specifically, the 'auc' and 'val_loss' were chosen as monitoring metrics. In addition, 5-fold cross-validation was performed to promise the stability of the prediction model and was repeated for 100 cycles, with the data used for each cycle being extracted randomly. It should be noted that we have constructed two models – ESPM and a traditional model. These two models were the same in terms of steps and parameter settings, the only difference being that the input features for the ESPM model were selected from survival-significant genes that were Ohnologs and originated from eukaryota, opisthokonta, and eumetazoa. In contrast, the traditional model selected input features from the remaining survival-significant genes. This approach maximized the demonstration of the effectiveness of evolutionary information.