scientific reports

OPEN

A novel 14-gene signature for overall survival in lung adenocarcinoma based on the Bayesian hierarchical Cox proportional hazards model

Na Sun[1], Jiadong Chu[1], Wei Hu[1], Xuanli Chen[1], Nengjun Yi[2] & Yueping Shen[1]✉

There have been few investigations of cancer prognosis models based on Bayesian hierarchical models. In this study, we used a novel Bayesian method to screen mRNAs and estimate the effects of mRNAs on the prognosis of patients with lung adenocarcinoma. Based on the identified mRNAs, we can build a prognostic model combining mRNAs and clinical features, allowing us to explore new molecules with the potential to predict the prognosis of lung adenocarcinoma. The mRNA data (*n* = 594) and clinical data (*n* = 470) for lung adenocarcinoma were obtained from the TCGA database. Gene set enrichment analysis (GSEA), univariate Cox proportional hazards regression, and the Bayesian hierarchical Cox proportional hazards model were used to explore the mRNAs related to the prognosis of lung adenocarcinoma. Multivariate Cox proportional hazard regression was used to identify independent markers. The prediction performance of the prognostic model was evaluated not only by the internal cross-validation but also by the external validation based on the GEO dataset (*n* = 437). With the Bayesian hierarchical Cox proportional hazards model, a 14-gene signature that included CPS1, CTPS2, DARS2, IGFBP3, MCM5, MCM7, NME4, NT5E, PLK1, POLR3G, PTTG1, SERPINB5, TXNRD1, and TYMS was established to predict overall survival in lung adenocarcinoma. Multivariate analysis demonstrated that the 14-gene signature (HR 3.960, 95% CI 2.710–5.786), T classification (T_1, reference; T_3, HR 1.925, 95% CI 1.104–3.355) and N classification (N_0, reference; N_1, HR 2.212, 95% CI 1.520–3.220; N_2, HR 2.260, 95% CI 1.499–3.409) were independent predictors. The *C*-index of the model was 0.733 and 0.735, respectively, after performing cross-validation and external validation, a nomogram was provided for better prediction in clinical application. Bayesian hierarchical Cox proportional hazards models can be used to integrate high-dimensional omics information into a prediction model for lung adenocarcinoma to improve the prognostic prediction and discover potential targets. This approach may be a powerful predictive tool for clinicians treating malignant tumours.

Lung cancer is one of the most common cancers in the world and is the leading cause of cancer-related deaths[1]. With the aging of the global population, lung cancer has a critical impact on health worldwide. Furthermore, lung adenocarcinoma is an important lung cancer subtype that has attracted increasing attention from researchers[2,3]. Due to the 5-year survival rate of lung adenocarcinoma being comparatively low, thus, improving its clinical prognosis is one of the main goals of clinical workers and medical researchers. Most of the previous prognostic models of lung adenocarcinoma focused on the clinical factors, such as treatment, tumour node metastasis (TNM) stage, and tumour grade[4,5]. These models may not be able to accurately predict the survival of patients with lung adenocarcinoma.

With the development of molecular technologies, we have the opportunity to integrate high-dimensional omics information into a prediction model of lung adenocarcinoma to improve its prognostic prediction ability, discover potential therapeutic targets and guide clinical treatment. This has become a new strategy to predict the prognosis of patients with lung adenocarcinoma[6–8]. In previous studies, the most common analysis strategy

[1]Department of Epidemiology and Biostatistics, School of Public Health, Medical College of Soochow University, Suzhou 215123, China. [2]Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL 35294, USA. ✉email: shenyueping@suda.edu.cn

focused on selecting the most significant differential expression genes first, performing least absolute shrinkage and selection operator (LASSO) regression to calculate a risk score from high-dimensional omics data, and using Cox regression analysis to combine the risk score with clinical factors to establish an effective prognosis model[9,10]. To a certain extent, these model has a higher *C*-index than the prognosis models that only contain clinical factors[11].

However, the gradual development of the Bayesian method provides new ideas for research in this field and is recognized by an increasing number of scholars. Bayesian statistics is a kind of statistical inference based on population, sample, and prior information. In this context, Yi et al. combined Bayesian statistics with the classical LASSO Cox regression model and constructed a new prediction model, the Bayesian hierarchical Cox proportional hazards model, which obtained a higher *C*-index and had better stability[12]. More importantly, the expectation–maximization (EM) cyclic coordinate descent algorithm is used to fit the model, which increases the speed of the analysis. Up to now, the Bayesian hierarchical Cox proportional hazards model has not been applied to the prognosis and prediction of high-dimensional omics in lung adenocarcinoma.

In this study, the Bayesian hierarchical Cox proportional hazards model was applied to reduce the dimensionality of the transcriptomics data and explore the mRNAs related to the prognosis of lung adenocarcinoma. An independent prognostic factor was constructed involving a 14-gene prognostic signature based on a data set from The Cancer Genome Atlas (TCGA). Multivariate Cox proportional hazard regression was then used to build the final prediction model, combined with the risk score and clinical characteristics, and a prognostic nomogram was constructed for clinical application. In addition, the stability of the model was verified using the Gene Expression Omnibus (GEO) data set.

Material and methods

Study cohort.

TCGA data sets. The mRNA data and clinical data for lung adenocarcinoma samples from the TCGA-LUAD data set were obtained from the TCGA database[13]. The mRNA data sets consisted of normal samples ($n = 59$) and lung adenocarcinoma samples ($n = 535$). Additionally, the following clinical information was obtained: age, gender, race, T classification, N classification, M classification, stage, treatment, smoking history, survival status, and overall survival (OS). After excluding the samples from patients with missing values, more than 10 years of follow-up, and an OS time of fewer than 15 days, samples from a total of 470 patients were selected for the study cohort.

GEO data sets. The GEO database provides the largest available set of microarray data with clinical annotation for lung adenocarcinoma. The gene expression profiling data sets for the GSE68465 cohort were downloaded from the GEO database for validation studies[14]. The genetic and clinical data for 443 patients with lung adenocarcinoma were obtained and taken into account the aforementioned inclusion and exclusion criteria, 437 patients were selected for the validation cohort.

TCGA and GEO belong to public databases. The patients involved in the database have obtained ethical approval. Users can download relevant data for free for research and publish relevant articles. Our study is based on open-source data, so there are no ethical issues and other conflicts of interest.

Gene set enrichment analysis (GSEA).

GSEA[15] mainly uses genomic and gene sequencing to detect biological differences in microarray data sets[16]. In this study, critical pathways and leading-edge mRNAs in lung adenocarcinoma versus normal control samples were identified by GSEA, using the Molecular Signatures Database (MSigDB) c2 (c2.cp.kegg.v7.2.symbols.gmt)[17]. The false discovery rate (FDR) < 0.25, nominal *P* value < 0.05, and |Normalized Enrichment Score (NES)| > 1 were regarded as the criteria for the identification of significant pathways[18].

Statistical analysis.

Univariate Cox proportional hazards regression and Bayesian hierarchical Cox proportional hazards model. The univariate Cox proportional hazards regression was adopted for the initial dimension reduction of high-dimensional data. To explore the gene signatures potentially affecting the survival of lung adenocarcinoma patients, R version 4.0.2 software was used to analyze the data, and $P < 0.05$ was considered a statistically significant difference. The Bayesian hierarchical Cox proportional hazards model was used to establish the optimal multivariate model, and dimension reduction was realized by the bmlasso function through the R "BhGLM" package[19]. Moreover, the EM cyclic coordinate descent algorithm and spike-and-slab mixture double-exponential prior [formula (1)] were selected to fit the model[12].

$$\beta_j|\gamma_j, s_0, s_1 \sim DE(\beta_j|0, s_j) = \frac{1}{2s_j}exp(-\frac{|\beta_j|}{s_j})$$

$$s_j = (1-\gamma_j)s_0 + \gamma_j s_1 \tag{1}$$

The spike scale value s_0 and the slab scale value s_1 cause strong or weak shrinkage of β_j, respectively ($0 < s_0 < s_1$). Moreover, an initial value is required for the spike scale and the slab scale. Additionally, a previous study demonstrates that the spike scale value s_0 has a strong influence on the model effectiveness, while the slab scale has little effect on the model effectiveness[20]. Therefore, in this study, we set the initial values as follows: $s_0 = c$ ($s_\lambda - 0.05$, $s_\lambda - 0.04$, $s_\lambda - 0.03$, $s_\lambda - 0.02$, $s_\lambda - 0.01$, s_λ, $s_\lambda + 0.01$, $s_\lambda + 0.02$, $s_\lambda + 0.03$, $s_\lambda + 0.04$, $s_\lambda + 0.05$), $s_1 = 0.5$, where s_λ is the optimal penalty of the LASSO Cox model. The concordance index (*C*-index) and the validation deviance were used to select the optimal model through tenfold with 10 repeats cross-validation[21].

After building the optimal Bayesian hierarchical Cox proportional risk model, the genes with nonzero coefficients were selected to calculate the risk score [formula (2)].

$$risk\ scores = \sum\nolimits_{j=1}^{n} coefj * Xj \tag{2}$$

where *coefj* is the coefficient, *Xj* is the standardized gene expression in the optimal model. After calculating the risk score for each patient, the median risk score was regarded as the cut-off value that stratified lung adenocarcinoma patients into low-risk and high-risk groups to compare the survival. The area under the curve (AUC) of each data set was calculated for detailed evaluations.

Multivariate Cox proportional hazards regression. Finally, we combined the risk score with clinical characteristics to construct the prognostic model. The results were sequentially displayed by a forest plot using the R package "forestplot". In addition, the nomogram provided information on the relationship between the total points, risk score, and clinical characteristics to predict the 3-year, 5-year, 10-year overall survival rates for new patients. To ensure the stability of the results, the *C*-index obtained from 1000 bootstrap samples was used to measure the validity of the nomogram. Furthermore, we calculated the total point of each patient using the nomogram and divided the patients into two groups according to the median total point to compare the survival. Finally, calibration curves of the 3-year, 5-year, 10-year survival rates were drawn to verify the consistency of the overall survival rate data between the predicted values obtained using the nomogram and the actual values. The workflow of this study is shown in Fig. 1. I confirm that all methods were performed in accordance with the relevant guidelines and regulations.

Results

The clinical characteristics of the TCGA-LUAD cohort and the GEO cohort are shown in Table 1. The results showed that the distribution of clinical characteristics in the two cohorts was comparable.

Gene set enrichment analysis. GSEA revealed that 10 pathways were involved in the tumour group. After removing the repeated genes in the pathways, 165 genes were identified for subsequent analysis. The details are shown in Table S1 and Fig. 2. In addition, the expression of these 165 mRNAs was visualised by a heatmap (Fig. S1).

Prognosis-related mRNAs. Univariate Cox proportional hazards regression analysis showed that 87 genes were related to the prognosis of lung adenocarcinoma. The LASSO Cox model was used to show that the optimal penalty $s_\lambda = 0.0843$. According to the mean of the *C*-index, we found that the prediction model was optimal when $s_0 = 0.0743$ and $s_1 = 0.5$ (Table 2). The *C*-index of the Bayesian hierarchical Cox proportional hazards model was 0.651, slightly higher than the *C*-index of the LASSO Cox regression, which was 0.649. In this model, we found that the following 14 genes were significantly related to patient survival: CPS1, CTPS2, DARS2, IGFBP3, MCM5, MCM7, NME4, NT5E, PLK1, POLR3G, PTTG1, SERPINB5, TXNRD1 and TYMS (Fig. 3A). The distribution of these 14 genes in each pathway was visually displayed by a chord diagram (Fig. 3B).

We calculated the risk score for each patient and used the median risk score (median = − 0.068) to divide patients with lung adenocarcinoma into low-risk and high-risk groups. The Kaplan–Meier survival curve with log-rank test showed that patients with high-risk scores had shorter OS time than those with low-risk scores ($P < 0.001$, Fig. 4A), and the AUC of the risk score was 0.689 (Fig. 4C); similarly, the external validation set results were shown in Fig. 4B and D. Then, we analyzed the gene expression in lung adenocarcinoma and normal groups (CTPS2 and DARS2), which had not been fully explored. The results showed that the mRNA expression of CTPS2 was dramatically increased in lung adenocarcinoma samples compared with normal lung samples ($P < 0.001$, Fig. 5A). The mRNA level of DARS2 was significantly elevated in lung adenocarcinoma samples compared with normal lung samples ($P < 0.001$, Fig. 5B).

Prognostic model. Multivariate Cox proportional hazards regression showed that the risk score (HR 3.960, 95% CI 2.710–5.786), T classification (T_1, reference; T_3, HR 1.925, 95% CI 1.104–3.355) and N classification (N_0, reference; N_1, HR 2.212, 95% CI 1.520–3.220; N_2, HR 2.260, 95% CI 1.499–3.409) were independent predictors of lung adenocarcinoma patient survival (Fig. 6A). The *C*-indexes of the internal and external validation were 0.733 and 0.735, respectively. In addition, integrating the 14-gene signature and clinical factors, we generated a nomogram to predict the 3-year, 5-year and 10-year survival rates (Fig. 6B). Each factor was scored according to the proportion of its contribution to the survival rate. The Kaplan–Meier survival curve with log-rank test demonstrated that patients with high total points had shorter OS times than those with low total points ($P < 0.001$, Fig. S2). Calibration curves showed that there was consistency between the predicted and actual values (Fig. 6C–E), especially for the 3-year survival rate.

Discussion

In this study, the Bayesian hierarchical Cox proportional hazards model was adopted to reduce the dimensionality of the omics data as part of the research strategy. Through internal and external validation, the prediction of the prognosis model for lung adenocarcinoma performed well and its performance was better than that of models reported by others[22,23]. The clinical factors and 14-gene signature we identified through the prediction model are basically consistent with previous reports. Interestingly, we also found that CTPS2 and DARS2 which never be reported were associated with the increasing death risk of lung adenocarcinoma.

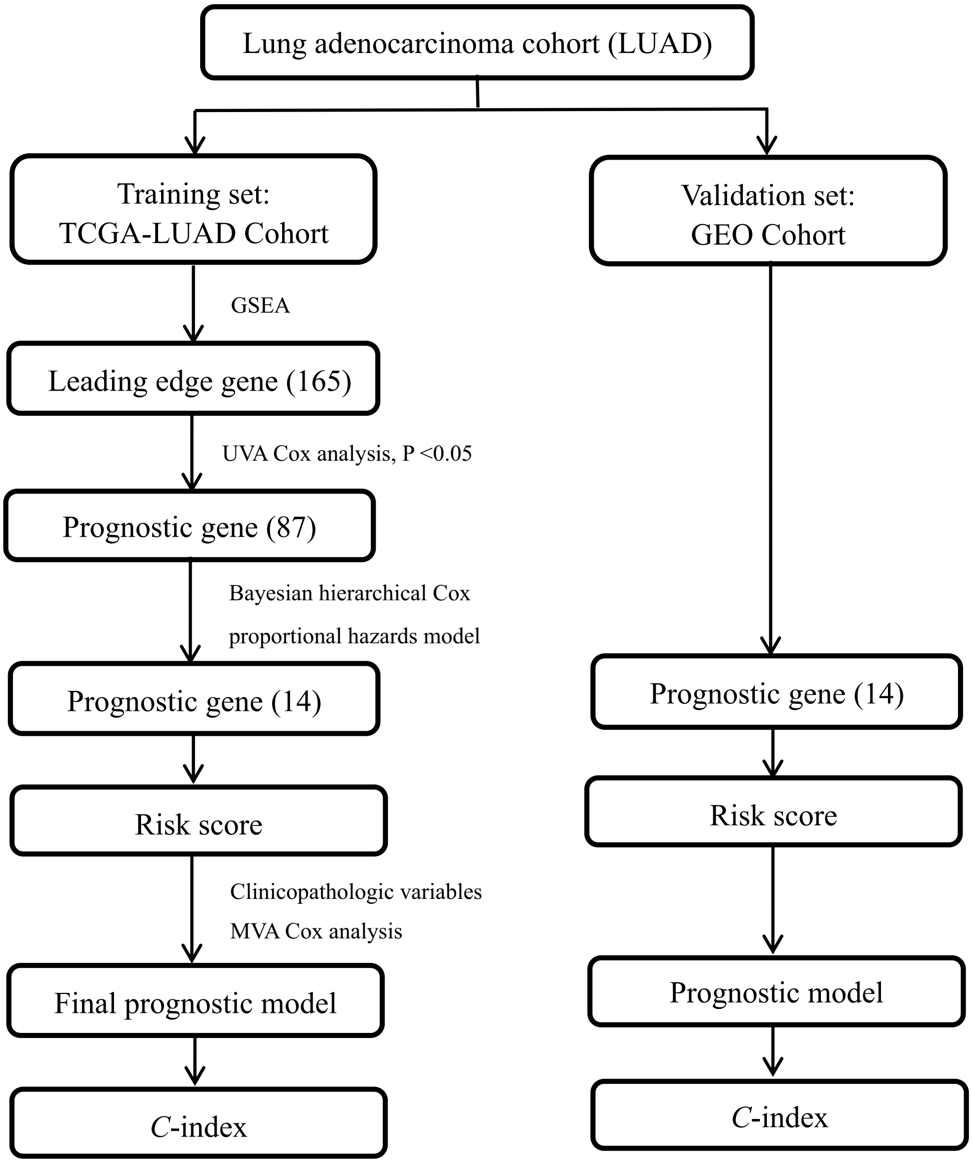
Lung adenocarcinoma cohort (LUAD)
Training set: TCGA-LUAD Cohort
Validation set: GEO Cohort
GSEA
Leading edge gene (165)
UVA Cox analysis, P <0.05
Prognostic gene (87)
Bayesian hierarchical Cox proportional hazards model
Prognostic gene (14)
Risk score
Clinicopathologic variables MVA Cox analysis
Final prognostic model
C-index
Prognostic gene (14)
Risk score
Prognostic model
C-index

Figure 1. The workflow of this study.

In this study, 14 prognostic genes were combined with clinical factors, and the final prognosis model for lung adenocarcinoma was constructed. In the training set and validation set, the *C*-index of the model reached 0.733 and 0.735, respectively, which indicates that the performance of the model is reliable. In a previous study of lung adenocarcinoma based on mRNA data from the TCGA database, Hugo Gómez-Rueda et al. constructed a prognostic model through LASSO regression and reported a lower *C*-index (*C*-index = 0.72)[22]. In another study, even with the combination of four omics datasets (mRNA, miRNA, DNA methylation and copy number variations) analysed by deep learning, the performance of the model was not as good as ours (*C*-index = 0.65)[23]. Although a study on early-stage lung adenocarcinoma further improved the *C*-index from 0.728 to 0.756 by adding BRCA1 and ERBB3 into the model, this method has not been verified internally and externally[24].

Our model was developed with a combination of LASSO Cox and Bayesian methods, which has several advantages over LASSO Cox. This method was also reported to be more accurate than the LASSO Cox regression model for coefficient estimation and prognosis prediction[25]. Additionally, the spike-and-slab prior used in the fitting of the Bayesian hierarchical Cox proportional hazards model can produce different shrinkages for different predictors, reduce the noise from irrelevant predictors and improve the accuracy of coefficient estimation and prediction[12]. The EM cyclic coordinate descent algorithm can make the convergence speed of the model faster on the premise of identifying important factors, which is an important element affecting the generalization of the model[26].

Furthermore, using the novel Bayesian hierarchical Cox proportional hazards model, most of the 14 prognostic genes we found can be explained in terms of basic study and population study. It is reported that CPS1, IGFBP3, MCM5, MCM7, NT5E, PLK1, PTTG1, SERPINB5, TXNRD1 and TYMS were associated with the

Factor	TCGA	GEO
No. of patients	470	437
Age, years, mean (SD)	65.2 (10.01)	64.4 (10.10)
Gender, no. (%)		
Female	251 (53.40)	218 (49.89)
Male	219 (46.60)	219 (50.11)
Race, no. (%)		
White	371 (78.94)	289 (66.13)
Other	57 (12.13)	19 (4.35)
Unknown	42 (8.94)	129 (29.52)
T classification, no. (%)		
T_1	160 (34.04)	149 (34.10)
T_2	251 (53.40)	249 (56.98)
T_3	42 (8.94)	28 (6.41)
T_4	17 (3.62)	11 (2.52)
N classification, no. (%)		
N_0	312 (66.38)	297 (67.96)
N_1	91 (19.36)	87 (19.91)
N_2	67 (14.26)	53 (12.13)
M classification, no. (%)		
M_0	314 (66.81)	437 (100.00)
M_1	21 (4.47)	0 (0.00)
MX	135 (28.72)	0 (0.00)
Stage, no. (%)		
I	253 (53.83)	–
II	114 (24.26)	–
III	76 (16.17)	–
IV	21 (4.47)	–
Unknown	6 (1.28)	–
Neither	0 (0.00)	316 (72.31)
Treatment, no. (%)		
Chemotherapy	242 (51.49)	43 (9.84)
Radiotherapy	228 (48.51)	20 (4.58)
Chemotherapy and Radiotherapy	0 (0.00)	44 (10.07)
Unknown	0 (0.00)	14 (3.20)
Smoking history, no. (%)		
No	63 (13.40)	48 (10.98)
Yes	389 (82.77)	296 (67.73)
Unknown	18 (3.83)	93 (21.28)

Table 1. Clinical characteristics of the lung adenocarcinoma cohort in the study.

prognosis of lung adenocarcinoma or non-small cell lung cancer, which was also included in our 14 genetic findings[27–36]. Basic studies found that NME4 and POLR3G were related to tumorigenesis and the progression of lung adenocarcinoma[37,38]. Our study also revealed that high expression of NME4 and POLR3G may adversely affect the poor prognosis of lung adenocarcinoma. To the best of our knowledge, there were no biological mechanism studies about CTPS2 and DARS2 that affect the tumorigenesis and progression of lung adenocarcinoma. The relationship between CTPS2 and DARS2 and the prognosis of lung adenocarcinoma has not been studied.

The protein encoded by CTPS2 is an important enzyme belonging to the CTP synthase family, which regulates cytosine nucleotide synthesis and provides the necessary precursors for RNA and DNA synthesis[39]. As early as 1978, researchers discovered that cancer cells with increased cell proliferation capabilities also showed increased CTP synthase activity, especially hepatocellular carcinoma cells[40]. Another study also reported that CTPS2 is a key gene that affects the prognosis of osteosarcoma[39]. Here, our study also showed that CTPS2 is an important gene for the prognosis of lung adenocarcinoma, it is highly expressed in patients with lung adenocarcinoma, and the prognosis is poor. Based on the above evidence, it is reasonable to suggest that the CTPS2 gene may be a new potential target for selective chemotherapy of lung adenocarcinoma. However, the mechanism of CTPS2 in lung adenocarcinoma is not clear, and more research is necessary.

The protein encoded by DARS2 is a critical mitochondrial enzyme belonging to the class-II aminoacyl-tRNA synthetase family, which is important for the mitochondrial unfolded protein response[41]. The relationship

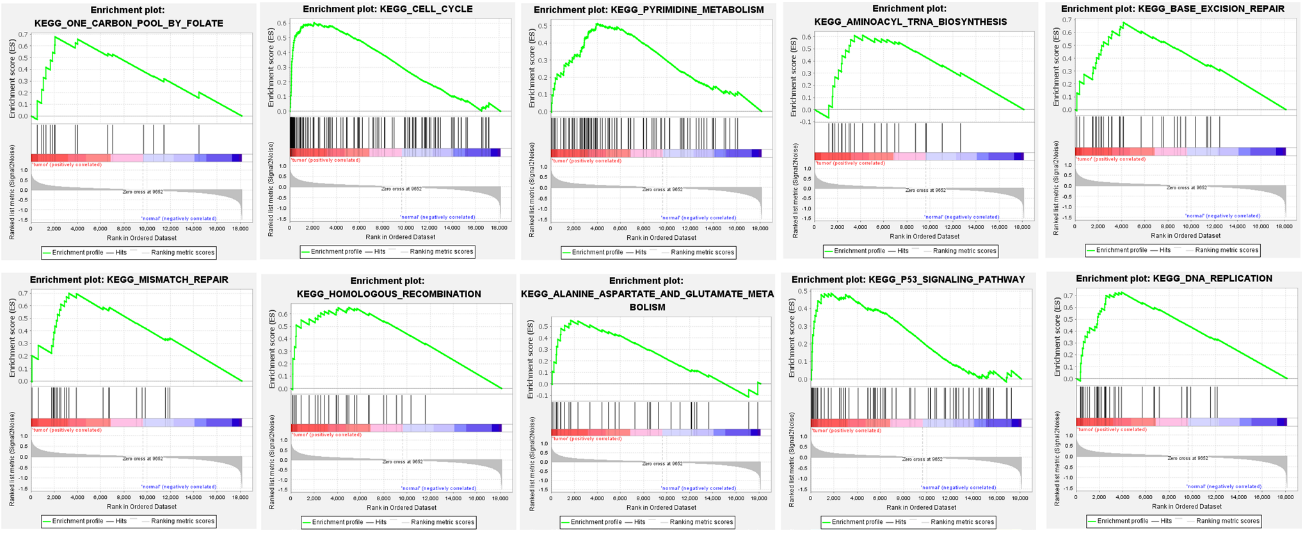

Figure 2. GSEA results from the c2 reference gene sets of the tumour group.

	C-index		Deviance	
Method	Mean	SD	Mean	SD
LASSO Cox	0.649	0.007	1779.056	4.590
$s_\lambda - 0.05$, 0.5	0.626	0.013	1787.719	6.311
$s_\lambda - 0.04$, 0.5	0.637	0.006	1786.711	3.539
$s_\lambda - 0.03$, 0.5	0.645	0.006	1781.919	2.940
$s_\lambda - 0.02$, 0.5	0.649	0.006	1779.648	3.111
$\mathrm{s}_\lambda - 0.01$, 0.5	0.651	0.006	1779.030	3.328
s_λ, 0.5	0.650	0.006	1779.095	3.984
$s_\lambda + 0.01$, 0.5	0.649	0.007	1779.092	4.659
$s_\lambda + 0.02$, 0.5	0.648	0.007	1779.666	5.335
$s_\lambda + 0.03$, 0.5	0.646	0.007	1780.872	5.847
$s_\lambda + 0.04$, 0.5	0.645	0.007	1782.788	6.259
$s_\lambda + 0.05$, 0.5	0.643	0.007	1785.146	6.803

Table 2. The measurements of the optimal models for the TCGA lung adenocarcinoma (LUAD) dataset mRNAs by tenfold with 10 repeats cross-validation. Significant values are in bold. $s_0 = c$ ($s_\lambda - 0.05$, $s_\lambda - 0.04$, $s_\lambda - 0.03$, $s_\lambda - 0.02$, $s_\lambda - 0.01$, s_λ, $s_\lambda + 0.01$, $s_\lambda + 0.02$, $s_\lambda + 0.03$, $s_\lambda + 0.04$, $s_\lambda + 0.05$), $s_\lambda = 0.0843$.

between the DARS2 gene and leukoencephalopathy with brain stem and spinal cord involvement and lactate elevation has been studied most frequently[42]. The first report on the relationship between DARS2 and cancer was in 2017, in which it was reported that DARS2 can promote the development of hepatocellular carcinoma by accelerating the cell cycle and reducing apoptosis[43]. Our study also showed that with an increased expression of DARS2, the death risk of patients with lung adenocarcinoma gradually elevated. We infer that DARS2 also affects the prognosis of lung adenocarcinoma by accelerating cell cycle progression and attenuating cell apoptosis, but further research is necessary to verify its function.

In summary, we constructed a prognosis prediction model of lung adenocarcinoma that the performance of the model is well and drew a nomogram, which provided a powerful tool for clinicians to predict the prognosis of lung adenocarcinoma patients. What's more, the main innovation of our study is the application of the Bayesian hierarchical Cox proportional hazards model for the reduction of omics data dimensionality to screen for prognostic genes. However, there are some limitations to the study. First, though our study adopted a new strategy of combining omics data with clinical characteristics, there are many possible research strategies in this field. It is a major challenge to determine which procedure is the best for model construction. To solve these problems,

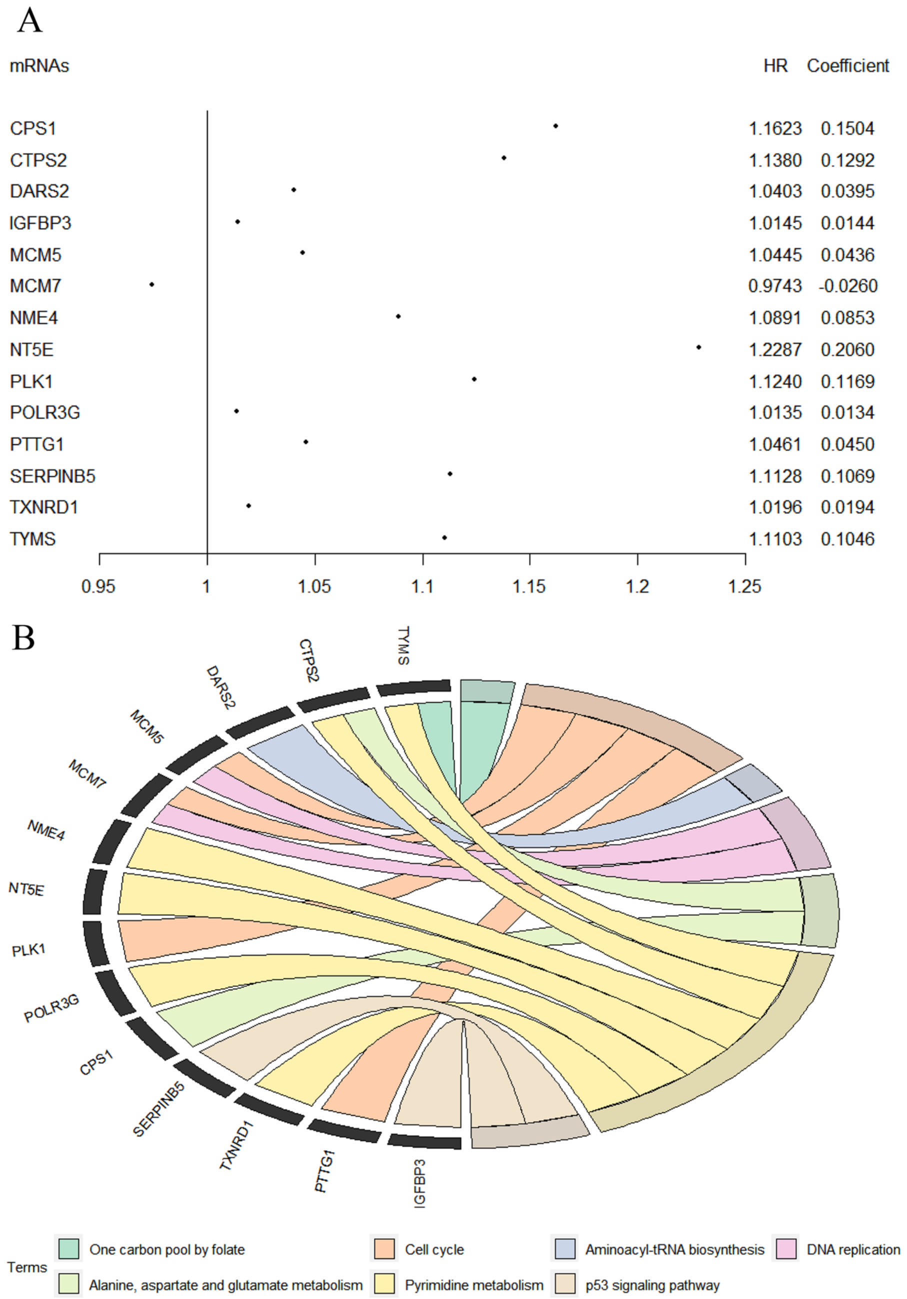

Figure 3. 14-gene prognostic signature. (**A**) Estimate of HR for 14 genes using the Bayesian hierarchical Cox proportional hazards model with a spike-and-slab prior. (**B**) The chord diagram of prognosis-related mRNAs. Genes are represented on the left, and pathways are represented on the right. Different pathways are differentiated by different colours.

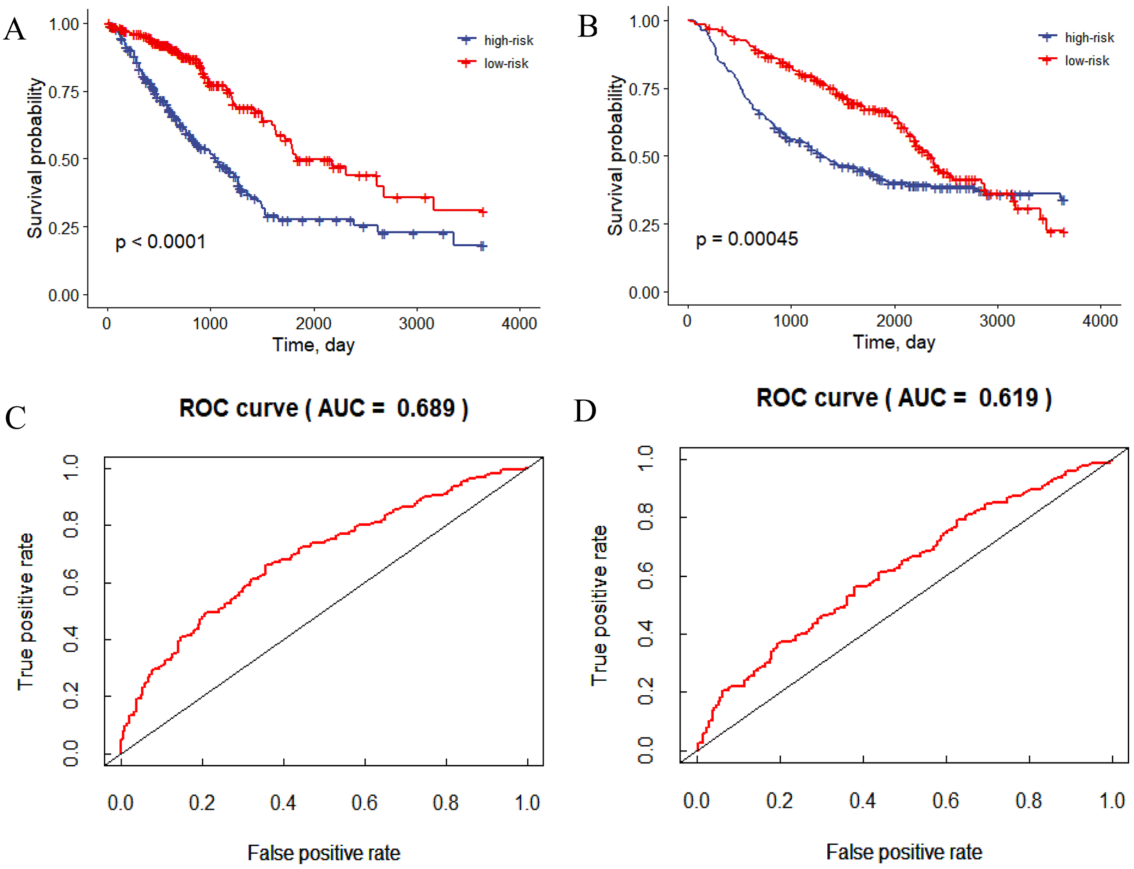

Figure 4. (**A**) Kaplan–Meier curve of TCGA-LUAD survival data for high-risk and low-risk groups with $P<0.001$. (**B**) Kaplan–Meier curve of GEO survival data for high-risk and low-risk groups with $P<0.001$. (**C**) The ROC curve of the risk score for predicting survival in the TCGA-LUAD Cohort. (**D**) The ROC curve of the risk score for predicting survival in the GEO Cohort.

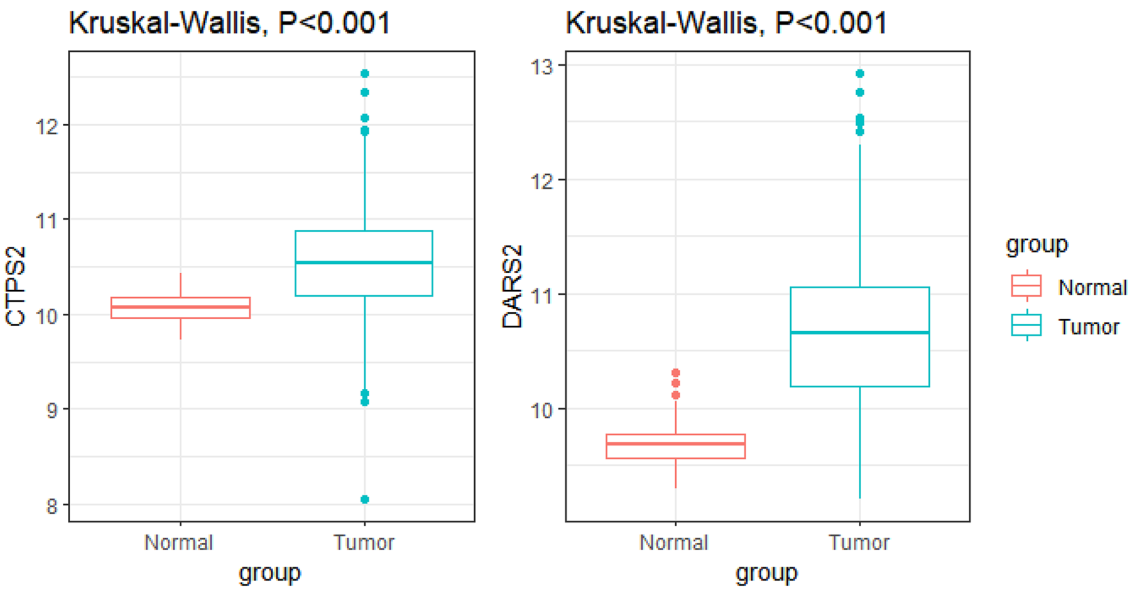

Figure 5. The CTPS2 and DARS2 expression in lung adenocarcinoma and normal groups in TCGA data. (**A**) The mRNA level of CTPS2 was dramatically increased in lung adenocarcinoma samples compared with normal lung samples. (**B**) The mRNA level of DARS2 was significantly higher in lung adenocarcinoma samples compared with normal lung samples.

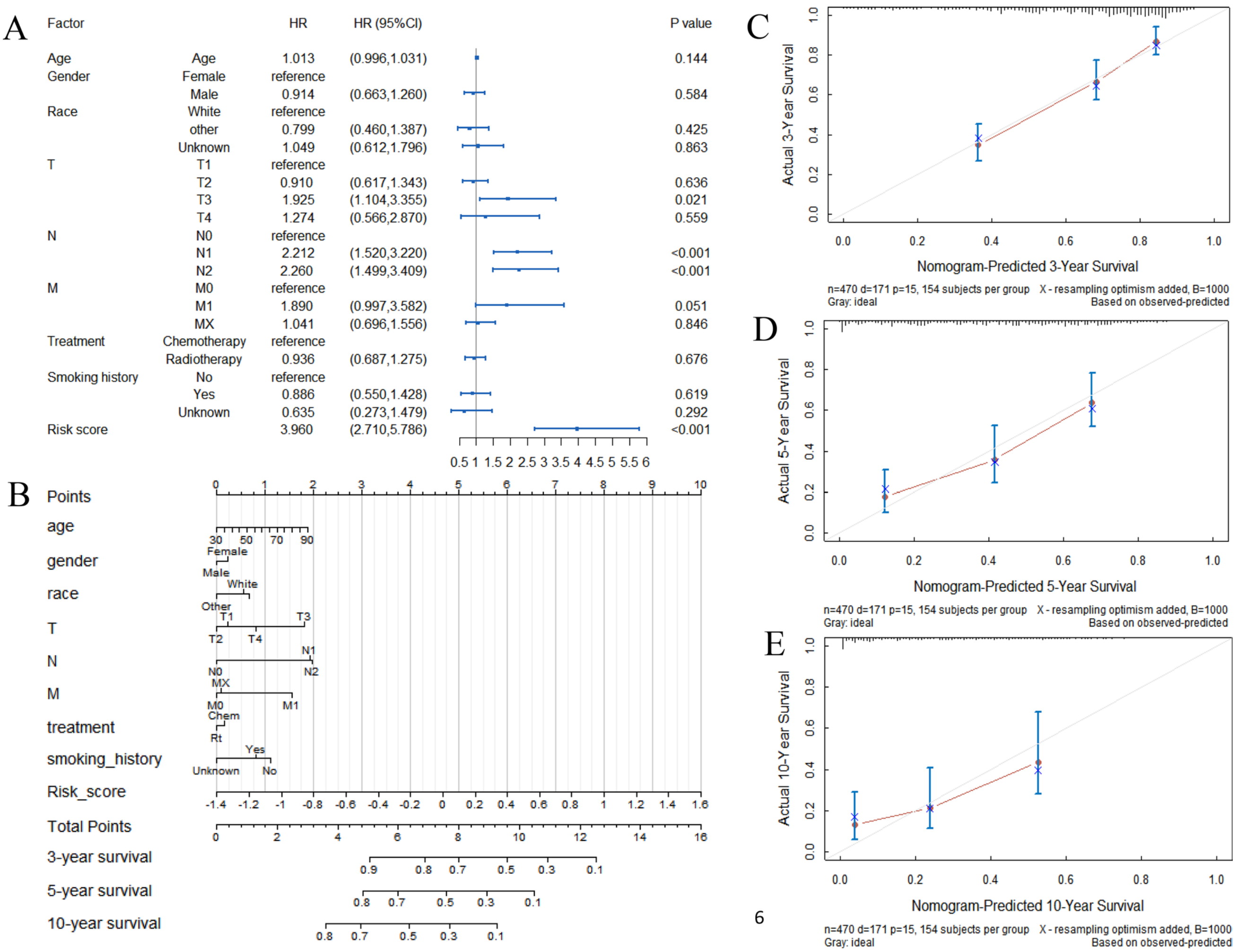

Figure 6. The final prognostic model. (**A**) Forest diagram of the risk score and clinical variables. (**B**) The nomogram for predicting the survival probability of lung adenocarcinoma patients at 3, 5 and 10 years. (**C**–**E**) Calibration curve of the nomogram for predicting 3-year, 5-year, and 10-year overall survival probability.

we should conduct some simulations and case studies in the future to explore the best research strategy for cancer prognosis prediction. In addition, there may be interactions and more complex nonlinear relationships between genes, which were unfortunately not analyzed in this study. Therefore, whether this method can be used to identify complex nonlinear relationships will be a focus of future research. Finally, although the statistical analysis was used to test the expression of genes that have not been fully explored in lung adenocarcinoma, we also expect to verify the expression of related genes by in vitro and in vivo experiments and explain the important role of CTPS2 and DARS2 in lung adenocarcinoma in further study.

Conclusions

The Bayesian hierarchical Cox proportional hazards model is a highly effective and alternative method for dealing with high-dimensional omics data when constructing cancer prediction and prognosis models. CTPS2 and DARS2 are new signatures affecting the prognosis of lung adenocarcinoma and may be potential new treatment targets.

Received: 15 July 2021; Accepted: 6 December 2021
Published online: 07 January 2022

References

1. Fitzmaurice, C. *et al.* Global, regional, and national cancer incidence, mortality, years of life lost, years lived with disability, and disability-adjusted life-years for 29 cancer groups, 1990 to 2017: A systematic analysis for the global burden of disease study. *JAMA Oncol.* **5**, 1749–1768. https://doi.org/10.1001/jamaoncol.2019.2996 (2019).
2. Song, Q. *et al.* Identification of an immune signature predicting prognosis risk of patients in lung adenocarcinoma. *J. Transl. Med.* **17**, 70. https://doi.org/10.1186/s12967-019-1824-4 (2019).

3. Sun, L., Zhang, Z., Yao, Y., Li, W. Y. & Gu, J. Analysis of expression differences of immune genes in non-small cell lung cancer based on TCGA and ImmPort data sets and the application of a prognostic model. *Ann. Transl. Med.* **8**, 550. https://doi.org/10.21037/atm.2020.04.38 (2020).
4. Tao, H. *et al.* Analysis of clinical characteristics and prognosis of patients with anaplastic lymphoma kinase-positive and surgically resected lung adenocarcinoma. *Thorac. Cancer* **8**, 8–15. https://doi.org/10.1111/1759-7714.12395 (2017).
5. Zhang, Y. *et al.* Real-world study of the incidence, risk factors, and prognostic factors associated with bone metastases in women with uterine cervical cancer using surveillance, epidemiology, and end results (SEER) data analysis. *Med. Sci. Monit.* **24**, 6387–6397. https://doi.org/10.12659/MSM.912071 (2018).
6. Shukla, S. *et al.* Development of a RNA-seq based prognostic signature in lung adenocarcinoma. *J. Natl. Cancer Inst.* **109**, 200. https://doi.org/10.1093/jnci/djw200 (2017).
7. Zhang, L., Zhang, Z. & Yu, Z. Identification of a novel glycolysis-related gene signature for predicting metastasis and survival in patients with lung adenocarcinoma. *J. Transl. Med.* **17**, 423. https://doi.org/10.1186/s12967-019-02173-2 (2019).
8. Xia, L. *et al.* Decreased expression of EFCC1 and its prognostic value in lung adenocarcinoma. *Ann. Transl. Med.* **7**, 672. https://doi.org/10.21037/atm.2019.10.41 (2019).
9. Sun, S. *et al.* Development and validation of an immune-related prognostic signature in lung adenocarcinoma. *Cancer Med.* https://doi.org/10.1002/cam4.3240 (2020).
10. Zhuang, Z. *et al.* Diagnostic, progressive and prognostic performance of m(6)A methylation RNA regulators in lung adenocarcinoma. *Int. J. Biol. Sci.* **16**, 1785–1797. https://doi.org/10.7150/ijbs.39046 (2020).
11. Mo, Z. *et al.* Identification of a hypoxia-associated signature for lung adenocarcinoma. *Front. Genet.* **11**, 647. https://doi.org/10.3389/fgene.2020.00647 (2020).
12. Tang, Z., Shen, Y., Zhang, X., Yi, N. & Hancock, J. The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics* **33**, 2799–2807. https://doi.org/10.1093/bioinformatics/btx300 (2017).
13. *TCGA-LUAD*. https://portal.gdc.cancer.gov/repository.
14. *GEO*. http://www.ncbi.nlm.nih.gov/geo/.
15. *GSEA*. http://software.broadinstitute.org/gsea/index.jsp.
16. He, W. *et al.* Gene set enrichment analysis and meta-analysis identified 12 key genes regulating and controlling the prognosis of lung adenocarcinoma. *Oncol. Lett.* **17**, 5608–5618. https://doi.org/10.3892/ol.2019.10236 (2019).
17. *MSigDB*. https://www.gsea-msigdb.org/gsea/msigdb/index.jsp.
18. Zhang, L. *et al.* Genome-wide investigation of the clinical significance and prospective molecular mechanisms of kinesin family member genes in patients with lung adenocarcinoma. *Oncol. Rep.* **42**, 1017–1034. https://doi.org/10.3892/or.2019.7236 (2019).
19. Yi, N., Tang, Z., Zhang, X. & Guo, B. BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology. *Bioinformatics* **35**, 1419–1421. https://doi.org/10.1093/bioinformatics/bty803 (2019).
20. Tang, Z., Shen, Y., Zhang, X. & Yi, N. The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics* **205**, 77–88. https://doi.org/10.1534/genetics.116.192195 (2017).
21. Li, R. *et al.* Fast Lasso method for large-scale and ultrahigh-dimensional Cox model with applications to UK Biobank. *Biostatistics* https://doi.org/10.1093/biostatistics/kxaa038 (2020).
22. Gomez-Rueda, H., Martinez-Ledesma, E., Martinez-Torteya, A., Palacios-Corona, R. & Trevino, V. Integration and comparison of different genomic data for outcome prediction in cancer. *BioData Min.* **8**, 32. https://doi.org/10.1186/s13040-015-0065-1 (2015).
23. Lee, T. Y., Huang, K. Y., Chuang, C. H., Lee, C. Y. & Chang, T. H. Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication. *Comput. Biol. Chem.* **87**, 107277. https://doi.org/10.1016/j.compbiolchem.2020.107277 (2020).
24. Sun, Y. *et al.* Two-gene signature improves the discriminatory power of IASLC/ATS/ERS classification to predict the survival of patients with early-stage lung adenocarcinoma. *Onco Targets Ther.* **9**, 4583–4591. https://doi.org/10.2147/OTT.S107272 (2016).
25. Mallick, H. & Yi, N. A new Bayesian lasso. *Stat. Interface* **7**, 571–582. https://doi.org/10.4310/SII.2014.v7.n4.a12 (2014).
26. Mallick, H. & Yi, N. Bayesian methods for high dimensional linear models. *J. Biometr. Biostat.* **1**, 005–005 (2013).
27. Wu, G. *et al.* CPS1 expression and its prognostic significance in lung adenocarcinoma. *Ann. Transl. Med.* **8**, 341. https://doi.org/10.21037/atm.2020.02.146 (2020).
28. Liu, Y. Z. *et al.* MCMs expression in lung cancer: implication of prognostic significance. *J. Cancer* **8**, 3641–3647. https://doi.org/10.7150/jca.20777 (2017).
29. Jiang, T. *et al.* Comprehensive evaluation of NT5E/CD73 expression and its prognostic significance in distinct types of cancers. *BMC Cancer* **18**, 267. https://doi.org/10.1186/s12885-018-4073-7 (2018).
30. Li, H. *et al.* The clinical and prognostic value of polo-like kinase 1 in lung squamous cell carcinoma patients: Immunohistochemical analysis. Biosci. Rep. https://doi.org/10.1042/BSR20170852 (2017).
31. Long, H. P., Liu, J. Q., Yu, Y. Y., Qiao, Q. & Li, G. PKMYT1 as a potential target to improve the radiosensitivity of lung adenocarcinoma. *Front. Genet.* **11**, 376. https://doi.org/10.3389/fgene.2020.00376 (2020).
32. Wang, X. F. *et al.* The roles of MASPIN expression and subcellular localization in non-small cell lung cancer. Biosci. Rep. https://doi.org/10.1042/BSR20200743 (2020).
33. Huang, J. *et al.* Identification of gene and microRNA changes in response to smoking in human airway epithelium by bioinformatics analyses. *Medicine* **98**, e17267. https://doi.org/10.1097/MD.0000000000017267 (2019).
34. Wang, H., Wang, X., Xu, L., Zhang, J. & Cao, H. High expression levels of pyrimidine metabolic rate-limiting enzymes are adverse prognostic factors in lung adenocarcinoma: A study based on The Cancer Genome Atlas and Gene Expression Omnibus datasets. *Purinergic Signal* **16**, 347–366. https://doi.org/10.1007/s11302-020-09711-4 (2020).
35. Yang, L. *et al.* Up-regulation of insulin-like growth factor binding protein-3 is associated with brain metastasis in lung adenocarcinoma. *Mol. Cells* **42**, 321–332. https://doi.org/10.14348/molcells.2019.2441 (2019).
36. Fan, X., Wang, Y. & Tang, X. Q. Extracting predictors for lung adenocarcinoma based on Granger causality test and stepwise character selection. *BMC Bioinform.* **20**, 197. https://doi.org/10.1186/s12859-019-2739-z (2019).
37. Wang, W. *et al.* NME4 may enhance nonsmall cell lung cancer progression by overcoming cell cycle arrest and promoting cellular proliferation. *Mol. Med. Rep.* **20**, 1629–1636. https://doi.org/10.3892/mmr.2019.10413 (2019).
38. Papadopoulos, A. *et al.* Cigarette smoking and lung cancer in women: Results of the French ICARE case-control study. *Lung Cancer* **74**, 369–377. https://doi.org/10.1016/j.lungcan.2011.04.013 (2011).
39. Fan, H., Lu, S., Wang, S. & Zhang, S. Identification of critical genes associated with human osteosarcoma metastasis based on integrated gene expression profiling. *Mol. Med. Rep.* **20**, 915–930. https://doi.org/10.3892/mmr.2019.10323 (2019).
40. Williams, J. C., Kizaki, H., Weber, G. & Morris, H. P. Increased ctp synthetase-activity in cancer-cells. *Nature* **271**, 71–73. https://doi.org/10.1038/271071a0 (1978).
41. Seiferling, D. *et al.* Loss of CLPP alleviates mitochondrial cardiomyopathy without affecting the mammalian UPRmt. *EMBO Rep.* **17**, 953–964. https://doi.org/10.15252/embr.201642077 (2016).
42. Rumyantseva, A., Motori, E. & Trifunovic, A. DARS2 is indispensable for Purkinje cell survival and protects against cerebellar ataxia. *Hum. Mol. Genet.* **29**, 2845–2854. https://doi.org/10.1093/hmg/ddaa176 (2020).
43. Qin, X. *et al.* Upregulation of DARS2 by HBV promotes hepatocarcinogenesis through the miR-30e5p/ MAPK/NFAT5 pathway. *J. Exp. Clin. Cancer Res.* https://doi.org/10.1186/s13046-017-0618-x (2017).

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Number 81973143. All of the authors have participated actively in this study, and agree to the content of the manuscript and they are being listed as an author on the paper.

Author contributions

N.S. conceived, designed, analyzed the data, and write the manuscript. J.C. conceptualized and developed an outline for the manuscript and revised the manuscript. W.H. and X.C. analyzed the data and generated the figures and tables. N.Y. and Y.S. revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-03645-6.

Correspondence and requests for materials should be addressed to Y.S.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.