

RESEARCH ARTICLE

Bayesian hierarchical lasso Cox model: A 9-gene prognostic signature for overall survival in gastric cancer in an Asian population

Jiadong Chu¹✉, Na Sun¹✉, Wei Hu¹✉, Xuanli Chen¹✉, Nengjun Yi^{2‡}, Yueping Shen^{1*}

1 Department of Biostatistics, School of Public Health, Medical College of Soochow University, Suzhou, P.R. China, **2** Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, Alabama, United States of America

✉ These authors contributed equally to this work.

‡ NY also contributed equally to this work.

* shenyueping@suda.edu.cn



Abstract

OPEN ACCESS

Citation: Chu J, Sun N, Hu W, Chen X, Yi N, Shen Y (2022) Bayesian hierarchical lasso Cox model: A 9-gene prognostic signature for overall survival in gastric cancer in an Asian population. PLoS ONE 17(4): e0266805. <https://doi.org/10.1371/journal.pone.0266805>

Editor: Andy T. Y. Lau, Shantou University Medical College, CHINA

Received: June 28, 2021

Accepted: March 29, 2022

Published: April 14, 2022

Copyright: © 2022 Chu et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All GEO files are available from the Gene Expression Omnibus (GEO) database (Project GSE66229 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66229>) and GSE15459 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE15459>)).

Funding: Our study was supported by the National Natural Science Foundation of China (project number 81973143). The main role of the funder (Yueping Shen) in this study was to project administration, supervision, review & editing.

Objective

Gastric cancer (GC) is one of the most common tumour diseases worldwide and has poor survival, especially in the Asian population. Exploration based on biomarkers would be efficient for better diagnosis, prediction, and targeted therapy.

Methods

Expression profiles were downloaded from the Gene Expression Omnibus (GEO) database. Survival-related genes were identified by gene set enrichment analysis (GSEA) and univariate Cox. Then, we applied a Bayesian hierarchical lasso Cox model for prognostic signature screening. Protein-protein interaction and Spearman analysis were performed. Kaplan–Meier and receiver operating characteristic (ROC) curve analysis were applied to evaluate the prediction performance. Multivariate Cox regression was used to identify prognostic factors, and a prognostic nomogram was constructed for clinical application.

Results

With the Bayesian lasso Cox model, a 9-gene signature included *TNFRSF11A*, *NMNAT1*, *EIF5A*, *NOTCH3*, *TOR2A*, *E2F8*, *PSMA5*, *TPMT*, and *KIF11* was established to predict overall survival in GC. Protein-protein interaction analysis indicated that *E2F8* was likely related to *KIF11*. Kaplan–Meier analysis showed a significant difference between the high-risk and low-risk groups ($P < 0.001$). Multivariate analysis demonstrated that the 9-gene signature was an independent predictor (HR = 2.609, 95% CI 2.017–3.370), and the C-index of the integrative model reached 0.75. Function enrichment analysis for different risk groups revealed the most significant enrichment pathway/term, including pyrimidine metabolism and respiratory electron transport chain.

Competing interests: The authors have declared that no competing interests exist.

Conclusion

Our findings suggested that a novel prognostic model based on a 9-gene signature was developed to predict GC patients in high-risk and improve prediction performance. We hope our model could provide a reference for risk classification and clinical decision-making.

Introduction

Gastric cancer (GC) has become a pervasive cancer worldwide and was responsible for over 1 million new cases and approximately 0.78 million deaths in 2018, making it the third leading cause of cancer death [1]. In addition, the incidence and diagnosis rates of GC in Asia were higher than those in other regions, especially among men [1]. Although constant improvements in therapy have been made, the survival rate of GC is still unsatisfactory, especially in the advanced stage [2]. With the development of high-throughput genome sequencing techniques, the exploration and application of various molecular biomarkers in GC would be efficient for better diagnosis, prediction, and targeted therapy [3]. Therefore, it is necessary and significant to establish a robust model based on genomic information for predicting prognosis. However, the occurrence and development of tumours are so complicated that considering a single gene to predict the survival of GC patients may not be accurate enough. Thus, it is feasible to construct a model by combining multiple genes with clinical characteristics to predict the survival prognosis of GC patients.

Further, identifying prognostic genes from high-dimensional data is critical, and several methods have been fully discussed [4]. Recently, Bayesian approaches have increased, thereby providing another option for variable selection and modelling based on high-dimensional survival data [5–11]. In previous research, Tang and Yi *et al.* proposed a novel Bayesian hierarchical Cox proportional hazards model (*i.e.*, the spike-and-slab lasso Cox) for adapting high-dimensional molecular data [12]. Extensive simulation studies showed that the spike-and-slab lasso Cox outperformed other methods, such as lasso Cox. However, few studies apply the Bayesian model to specific tumours in practice.

This study aimed to apply the spike-and-slab lasso Cox for identifying potential prognostic genes using the GSE66229 dataset. We constructed a 9-gene signature and established an integrative prognostic model for predicting overall survival (OS) in GC patients. Model performance was assessed in an independent external validation set. Functional enrichment analysis for different risk groups was performed by GSEA. Then, an instructive nomogram was drawn for prediction in clinical application. This study not only enriches the practical application of Bayesian methods but also provides new ideas and references for clinical prognosis prediction modelling.

Materials and methods

Cancer database download and processing

The mRNA expression profile and clinical information were publicly downloaded from the Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>). The gene expression microarray dataset GSE66229 came from an Asian Cancer Research Group (ACRG) study that included 300 tumour tissue samples (mainly adenocarcinoma) and 100 normal tissue samples, and this was used as the training set in this study. Another Singapore cohort dataset, GSE15459, was an independent external validation set and included 182 tumour tissue

samples. Both datasets were generated by the Affymetrix Human Genome U133 Plus 2.0 platform, and GSE66229 was normalized by robust multiarray average with the Affymetrix Power Tools package. GSE15459 was handled by Microarray Suite version 5.0 using Affymetrix default analysis settings and global scaling as a normalization method. Log2 transformation was utilized in this study.

Gene Set Enrichment Analysis (GSEA)

In contrast to traditional analysis, GSEA was not limited to providing a clear threshold (e.g., log2FC) for differentially expressed genes but was focused on genes contributing to specific biological function gene sets [13]. It could avoid removing important genes with no statistically significant expression differences. In this study, samples from GSE66229 were divided into the tumour and non-tumour groups. Then, GSEA v4.1.0 was used to analyze gene data. Based on the Molecular Signatures Database, the gene sets of Kyoto Encyclopedia of Genes and Genomes (KEGG, c2.cp.kegg.v7.2.symbols.gmt) and Gene Ontology (GO, c5.go.v7.2.symbols.gmt) were set as reference datasets. Nom-P value < 0.05 and false discovery rate (FDR) < 0.25 were set as the cut-off values. In addition, function enrichment analysis based on the same settings was performed for different risk groups.

Establishment of the prognostic signature

To screen prognostic genes, we split this process into two steps. First, univariate Cox regression was used to choose survival-related genes from the leading edge analysis in GSEA. Then, the spike-and-slab lasso Cox proposed by Tang and Yi *et al.* [12] was used to further identify prognostic genes. The spike-and-slab mixture double-exponential prior applied in this model was the key part and was expressed as follows:

$$\beta_j | \gamma_j, s_0, s_1 \sim (1 - \gamma_j) DE(\beta_j | 0, s_0) + \gamma_j DE(\beta_j | 0, s_1) \quad (1)$$

It had two positive value parameters, s_1 and s_0 ($s_1 > s_0 > 0$), which need to be preset. s_0 was chosen to be small and regarded as a “spike scale” for giving strong shrinkage on coefficient estimation, while s_1 was set to be large so that it served as a “slab scale” for giving weak shrinkage on important variables. γ_j was the indicator variable that linked the scale parameters with the coefficients. The algorithm for fitting the spike-and-slab Cox model was called the expectation-maximization (EM) cyclic coordinate descent algorithm, which had a fast computing speed [12].

After filtering prognostic genes by the Bayesian hierarchical lasso Cox model, the risk score was calculated for each patient, and the formula was as follows:

$$\text{Risk Score} = \sum_{i=1}^n (\beta_i \times \text{expression}(\text{gene}_i)) \quad (2)$$

β_i represents the corresponding coefficient of a specific gene, and the expression () indicates the expression level of the corresponding gene. Next, samples were divided into high-risk or low-risk groups according to the median risk score. Kaplan-Meier and receiver operating characteristic (ROC) curve analysis were performed to assess the predictive effect. Also, we integrated the risk score and clinical factors into one model to build the final prognostic model and evaluated the performance by the C-index in the training set and validation set. Then, forest plots, nomogram plots, and calibration plots were used to demonstrate the main result fully.

Correlation and validation analysis

To further explore possible associations, we firstly applied the Search Tool for the Retrieval of Interacting Genes (STRING, <https://cn.string-db.org/>) [14] to construct the Protein-Protein interaction (PPI) network based on the above genes. Then Spearman correlation analysis was performed to investigate the association at expression levels. On the other hand, the protein expression levels of the 9 genes were verified using the publicly Human Protein Atlas (HPA) database (<https://www.proteinatlas.org/>) [15]. To better deal with the HPA data, we adopted the HPAanalyze package, a powerful tool for searching and analyzing the HPA database [16].

External validation of prognostic model

We downloaded GSE15459 as an independent external validation set, which comprised the gene expression and clinical data of 182 samples. The risk scores of the GC patients were calculated and divided into high-risk and low-risk groups based on the median value. The robustness of this model was tested by Kaplan-Meier and ROC curve analysis.

Implementation

Statistical analysis was performed based on R v4.0.2. `bmlasso()` was used for fitting the Bayesian hierarchical lasso Cox model, and `cv.bh()` was executed to select optimal s_0 based on the predictive performance of the model. These functions are from the freely available R package BhGLM [17].

Results

Identification of survival-related genes

To summarize this study more comprehensively, a schematic diagram was provided in Fig 1. GSE66229 had 400 samples (with a total of 20161 genes), including 300 cases and 100 normal samples. The GSEA results revealed that 331 genes involved in 15 pathways from KEGG and 2611 genes involved in 595 terms (biological process, molecular function, and cellular component) were filtered out with $FDR < 0.25$ and $Nom-P < 0.05$ (S1 Table). After removing duplicated genes, a total of 2641 genes were selected for the subsequent process. Among the above processes, pyrimidine metabolism ($Nom-p = 0$), spliceosome ($Nom-p = 0.024$), RNA export from the nucleus ($Nom-p = 0$), and viral gene expression ($Nom-p = 0$) from KEGG and GO based on the largest absolute normalized enrichment score (NES) were shown in Fig 2.

After excluding samples with missing values (i.e., survival time < 1 month, survival outcome, pathological stage, age and sex), univariate Cox proportional hazards regression analysis was used to identify survival-related genes. Basic clinical characteristics were described completely in Table 1, and the results showed that 1109 genes were significantly correlated with OS at $P < 0.05$.

Bayesian hierarchical lasso Cox for screening final prognostic genes

The selection criterion of two parameters, s_1 and s_0 , has been sufficiently discussed in a previous study [18]. The variety of the C-indexes of the survival model was sensitive to the change in s_0 but less susceptible to s_1 . Therefore, we decided to fix s_1 at 1 according to previous research. Regarding the value of s_0 , we first used the `glmNet()` function from BhGLM package to simulate 10-fold cross-validation repeated 10 times to obtain the stable penalty parameter λ (i.e., s_λ) and then adjusted the value with a limited range from -0.04 to 0.06 with intervals of

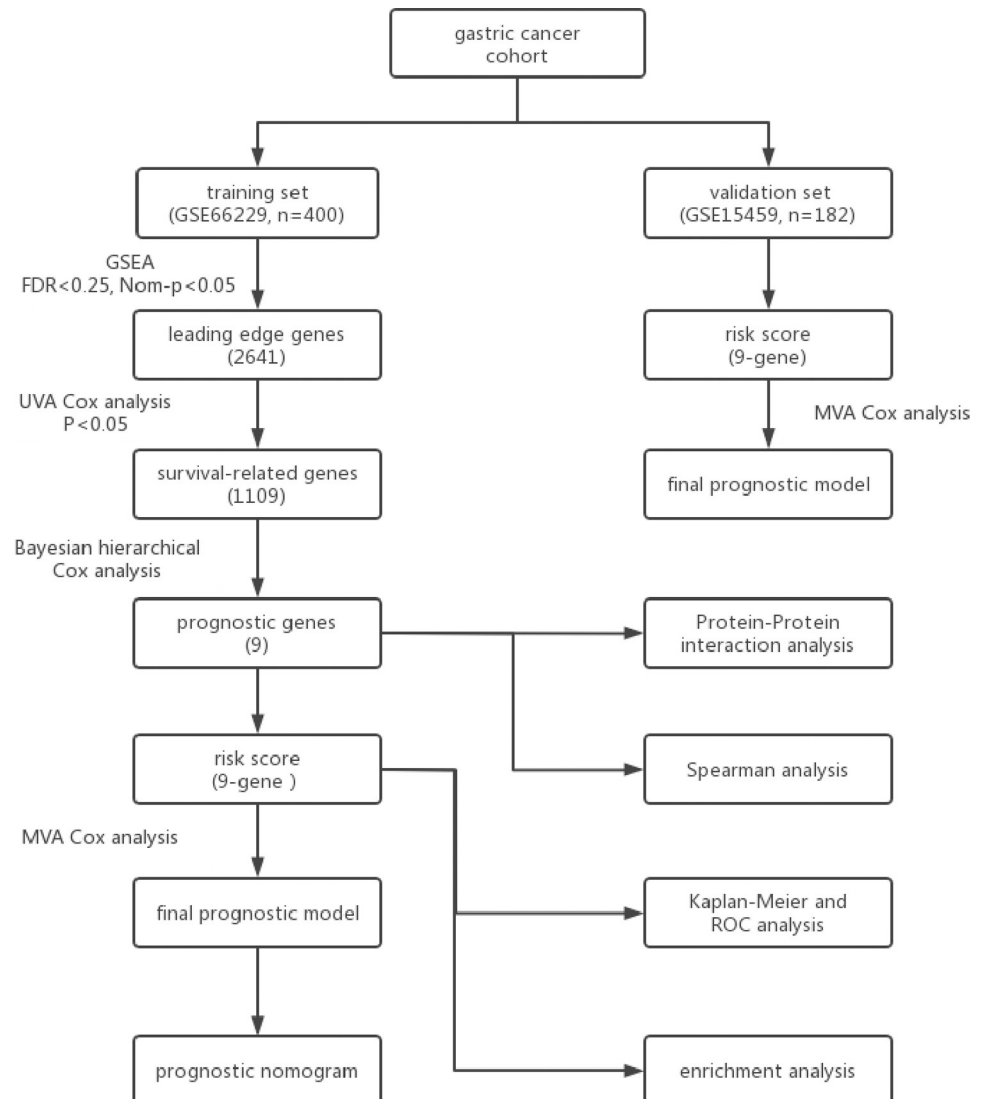


Fig 1. Schematic diagram of this study.

<https://doi.org/10.1371/journal.pone.0266805.g001>

0.01. Our goal was to find an optimal value that simultaneously made the C-index larger and deviance smaller. According to the results of 10-fold cross-validation with repeated 10 times, we ultimately decided to choose $s_0 = s_\lambda - 0.04$ as an optimal value (Table 2). At the same time, the C-index and deviance of the constructed Bayesian hierarchical lasso Cox model were 0.684 (sd = 0.004) and 1582.454 (sd = 3.734), respectively. The C-index of traditional lasso Cox regression was 0.643 (sd = 0.011), which was lower than our Bayesian hierarchical lasso Cox model. Afterwards, we chose 9 prognostic genes whose coefficients were not zero, including tumour necrosis factor receptor superfamily member 11A (*TNFRSF11A*), nicotinamide nucleotide adenylyltransferase 1 (*NMNAT1*), eukaryotic translation initiation factor 5A (*EIF5A*), notch receptor 3 (*NOTCH3*), torsin family 2 member A (*TOR2A*), E2F transcription factor 8 (*E2F8*), proteasome 20S subunit alpha 5 (*PSMA5*), thiopurine S-methyltransferase (*TPMT*), and kinesin family member 11 (*KIF11*) (Table 3).

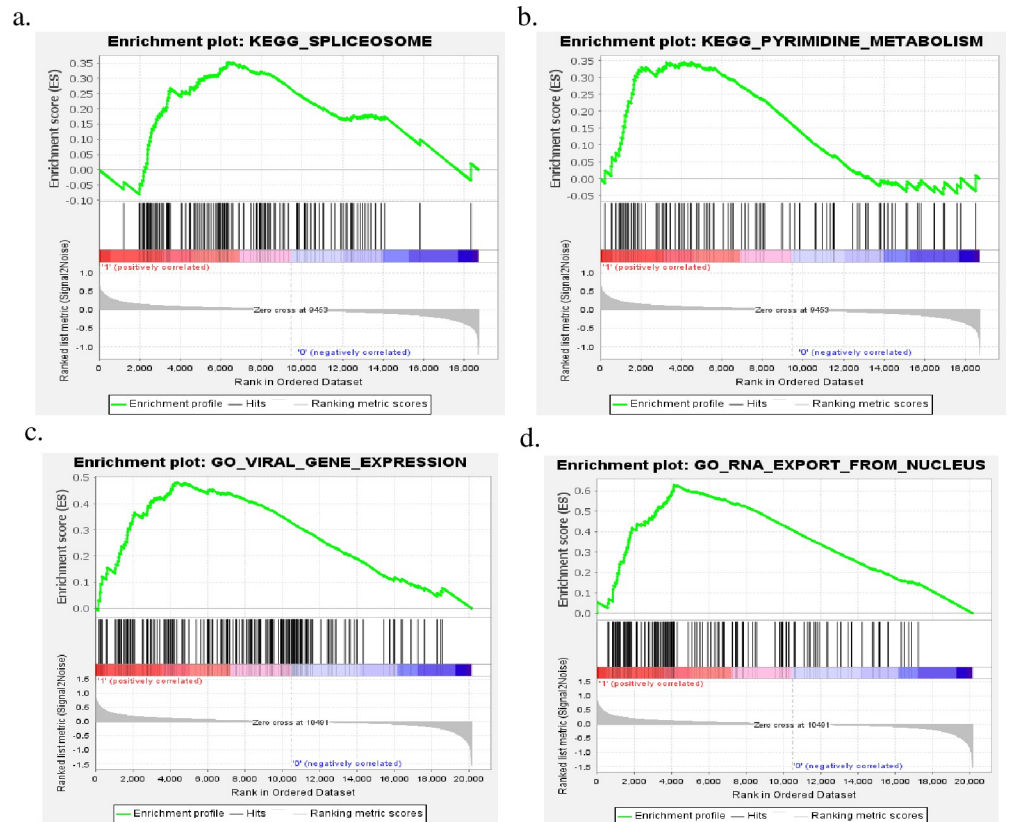


Fig 2. GSEA outcome of KEGG (a-b) and GO (c-d) with the largest absolute normalized enrichment size (NES).

<https://doi.org/10.1371/journal.pone.0266805.g002>

PPI, Spearman correlation and validation analysis

PPI analysis was carried out to investigate possible inter-relationships of 9 genes (Fig 3A). According to the STRING database, there existed an interaction between *E2F8* and *KIF11*. Except for the association between *NMNAT1* and *EIF5A* ($P > 0.05$), other pairs were statistically significant ($P < 0.05$) (S2 Table). Notably, the expression of *NOTCH3* was negatively associated with other genes, and the strongest association came from the expression of *E2F8* and

Table 1. Clinical characteristics of patients with gastric cancer.

		Training set		Validation set	
		count	%	count	%
Survival status	Alive, 0	148	49.33	87	47.80
	Dead, 1	152	50.67	95	52.20
Age	<65	161	53.67	75	41.21
	>= 65	139	46.33	107	58.79
Sex	female	101	33.67	66	36.26
	Male	199	66.33	116	63.74
Stage	I	30	10.00	31	17.03
	II	97	32.33	28	15.39
	III	96	32.00	66	36.26
	IV	77	25.67	57	31.32

<https://doi.org/10.1371/journal.pone.0266805.t001>

Table 2. Outcome of 10 times 10-fold cross-validation.

Method	C-index		Deviance	
	mean	sd	mean	sd
lasso	0.643	0.011	1602.125	5.641
$s_{\lambda}-0.04,1$	0.684	0.004	1582.454	3.734
$s_{\lambda}-0.03,1$	0.643	0.001	1601.508	1.258
$s_{\lambda}-0.02,1$	0.642	0.001	1602.134	1.720
$s_{\lambda}-0.01,1$	0.642	0.004	1603.394	5.236
$s_{\lambda},1$	0.636	0.009	1608.835	8.323
$s_{\lambda}+0.01,1$	0.648	0.012	1600.793	6.695
$s_{\lambda}+0.02,1$	0.653	0.012	1600.796	8.380
$s_{\lambda}+0.03,1$	0.654	0.011	1605.169	11.075
$s_{\lambda}+0.04,1$	0.657	0.011	1611.989	14.075
$s_{\lambda}+0.05,1$	0.657	0.011	1623.413	16.959
$s_{\lambda}+0.06,1$	0.655	0.012	1642.332	20.961

$s_{\lambda} = 0.0400874379965632$

<https://doi.org/10.1371/journal.pone.0266805.t002>

KIF11 ($r = 0.61$) (Fig 3B). Further, the protein levels of these genes were also explored through the HPA database (S2 Fig). The result showed that three genes (*E2F8*, *KIF11*, *TPMT*) had been found to be highly expressed in GC tissue.

Prognostic analysis of the 9-gene signature

We developed a prognostic signature based on the risk score constructed by the above 9 genes, and the formula of the risk score was given as follows: $\text{Risk score} = (-0.2277) * \text{TNFRSF11A} + (-0.1997) * \text{NMNAT1} + (-0.1693) * \text{EIF5A} + 0.1525 * \text{NOTCH3} + (-0.1423) * \text{TOR2A} + (-0.0393) * \text{E2F8} + (-0.0363) * \text{PSMA5} + (-0.0214) * \text{TPMT} + (-0.0060) * \text{KIF11}$.

After calculating the risk score for each patient, the median risk score (median = -0.077) was regarded as the cut-off value that stratified GC patients into low-risk and high-risk groups. Among them, the range of risk scores was [-1.569, 2.104]. The low-risk group was defined as [-1.569, -0.077], and the high-risk group was defined as [-0.077, 2.104]. The Kaplan-Meier survival analysis demonstrated a statistically significant difference between the low-risk and high-risk groups ($P < 0.0001$, Fig 4A), and the AUC of the risk score was 0.765 (Fig 4B). The distribution of survival status is also shown in Fig 5A. As the risk score of the GC patients increased,

Table 3. Bayesian hierarchical lasso Cox model of 9 genes associated with OS in GC patients.

Gene	Coefficient	HR
<i>TNFRSF11A</i>	-0.2277	0.7964
<i>NMNAT1</i>	-0.1997	0.8190
<i>EIF5A</i>	-0.1693	0.8443
<i>NOTCH3</i>	0.1525	1.1648
<i>TOR2A</i>	-0.1423	0.8673
<i>E2F8</i>	-0.0393	0.9615
<i>PSMA5</i>	-0.0363	0.9644
<i>TPMT</i>	-0.0214	0.9788
<i>KIF11</i>	-0.0060	0.9941

<https://doi.org/10.1371/journal.pone.0266805.t003>

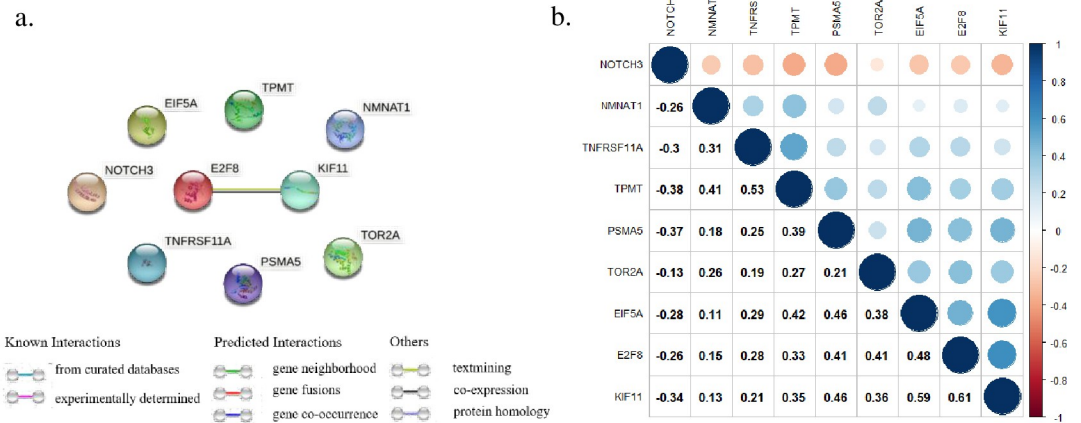


Fig 3. Interaction and correlation analysis. a) possible interactions in 9 genes based on STRING analysis. b) Spearman correlation plot of 9 genes based on expression profile.

<https://doi.org/10.1371/journal.pone.0266805.g003>

the expression of the mRNAs in the high-risk group (*TNFRSF11A*, *NMNAT1*, *EIF5A*, *TOR2A*, *E2F8*, *PSMA5*, *TPMT*, *KIF11*) showed obvious downregulation, whereas the expression of mRNAs in the low-risk group (*NOTCH3*) was upregulated. Additionally, the Kaplan-Meier analysis of the expression of each gene is shown in [S1 Fig](#).

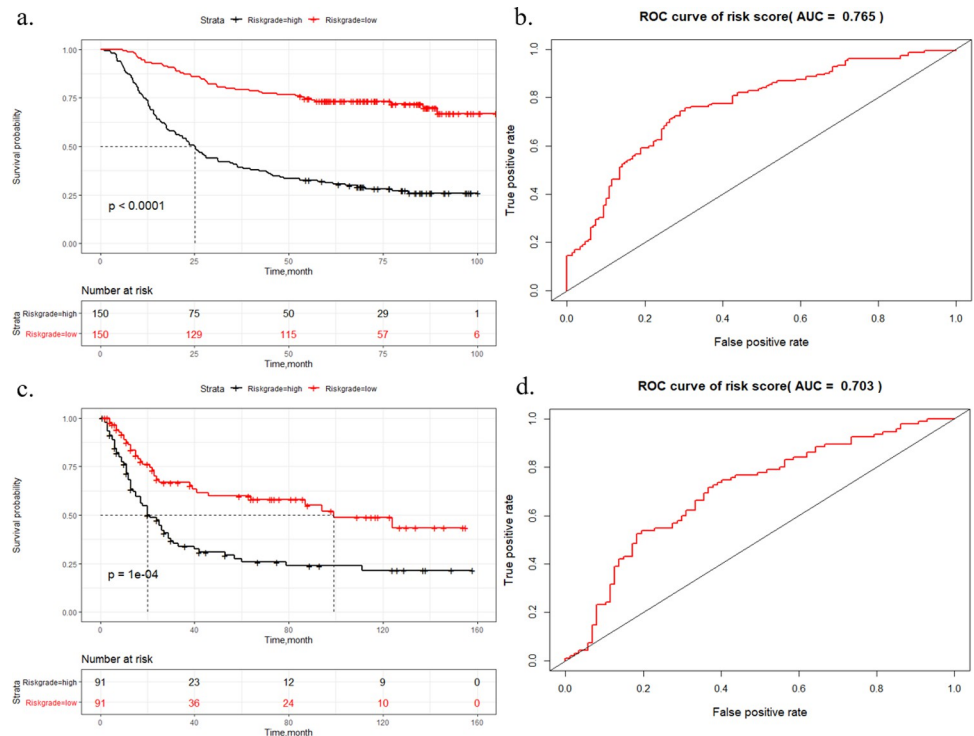


Fig 4. a) Kaplan-Meier curve of GSE66229 survival data for high-risk and low-risk groups with $p < 0.0001$. b) The ROC curve of the risk score for predicting survival in the GSE66229 dataset. c) Kaplan-Meier curve of GSE15459 survival data for high-risk and low-risk groups with $p = 0.0001$. d) The ROC curve of the risk score for predicting survival in the GSE15459 dataset.

<https://doi.org/10.1371/journal.pone.0266805.g004>

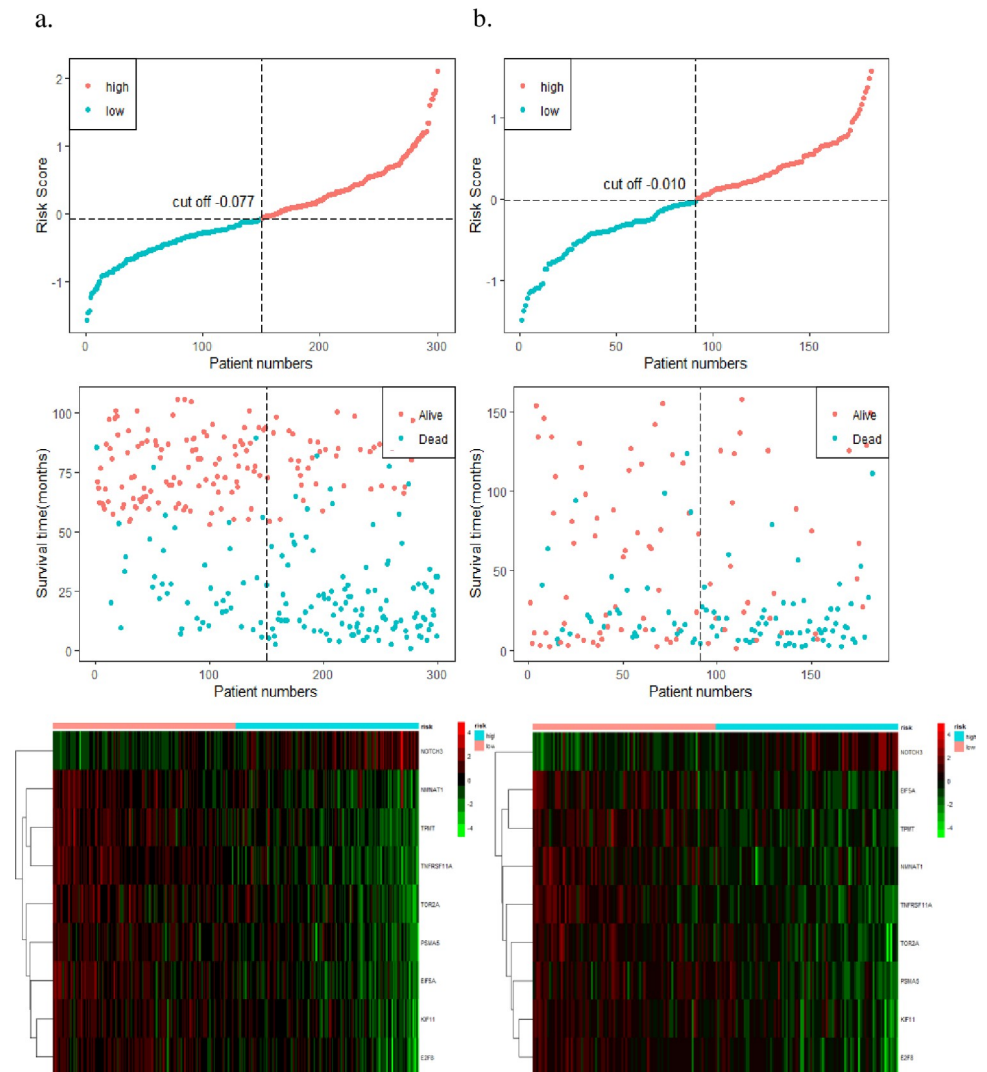


Fig 5. Risk score model based on 9 genes in the training and validation sets. a) The training set. b) The validation set. The top row shows the tendency of the risk score with the cut-off value in each dataset. The middle row shows the distribution of the survival status of patients in each dataset. The bottom row shows the mRNA expression of 9 genes based on the median risk score in each dataset.

<https://doi.org/10.1371/journal.pone.0266805.g005>

Function enrichment analysis

To elucidate the possible biological terms or pathways associated with the risk score, we also utilized GSEA to perform GO and KEGG pathway analyses based on differentially expressed risk genes (DRGs) between the low-risk and high-risk groups. As shown in the chart, we intuitively discovered that DRGs were enriched mainly in particular molecular terms/pathways, such as pyrimidine metabolism and respiratory electron transport chain (Fig 6A and 6B).

Combining 9-gene signature with clinical characteristics

Next, we integrated the gene signature plus several clinical factors, including age, sex and stage, into a super prognostic model to completely predict the OS of GC patients. After integration, the C-index of the final prognostic model reached 0.75, in contrast to the model that

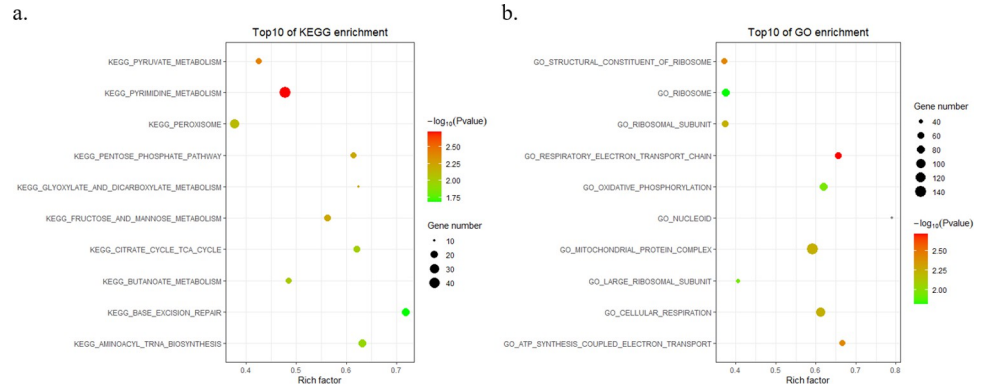


Fig 6. Functional enrichment analysis based on low-risk and high-risk groups. a) Top 10 enrichment pathways in KEGG based on the largest absolute NES. b) Top 10 enriched GO pathways based on the largest absolute NES.

<https://doi.org/10.1371/journal.pone.0266805.g006>

considered only clinical factors (C-index = 0.686) (Fig 7A). The HR of the risk score was 2.609, 95% CI: 2.017–3.370. Finally, considering the application of clinical practice, we utilized a nomogram to predict the survival probability of GC patients (Fig 8A). Moreover, calibration plots were used to illustrate the stability of the nomogram in predicting 1-year, 3-year, or 5-year OS (Fig 8B–8D).

Independent external validation in GSE15459

To assess the robustness of the 9-gene prognostic signature, we selected GSE15459 as an independent external validation set. Similar to the training set, Kaplan-Meier analysis indicated that low-risk patients had longer survival times than high-risk patients ($P = 0.0001$, Fig 4C), and the overall AUC of the risk score was 0.703 (Fig 4D). The distribution between the risk score and survival status is displayed in Fig 5B. We observed that the expression levels between different risk groups were similar to those in the training set, which further verified the accuracy of our results. After integrating clinical characteristics, the C-index of the final prognostic model also increased to 0.75, in contrast to the model with only clinical characteristics (C-index = 0.696). Multivariate analysis demonstrated that the risk score could be a stable prognostic factor for the prediction of OS (HR = 1.815, 95% CI: 1.288–2.560, $P < 0.001$) (Fig 7B).

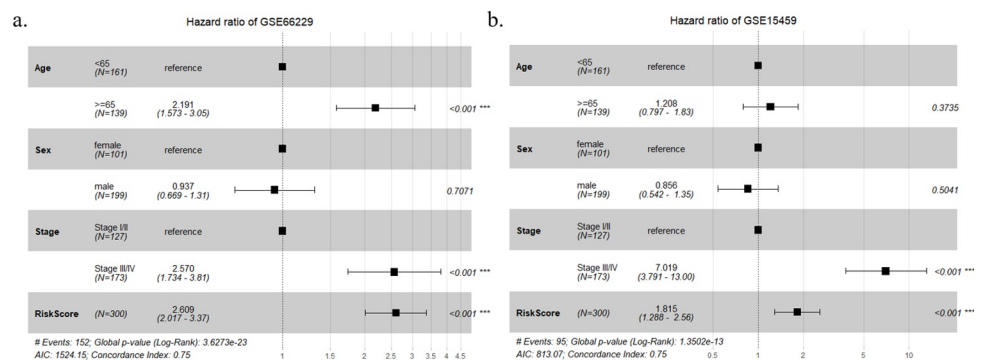


Fig 7. Integrative prognostic models combining the risk score and clinical factors in GC patients. a) Forest plot combining the risk score with clinical factors (age, sex, stage) in the training set. b) Forest plot combining the risk score with clinical factors (age, sex, stage) in the validation set.

<https://doi.org/10.1371/journal.pone.0266805.g007>

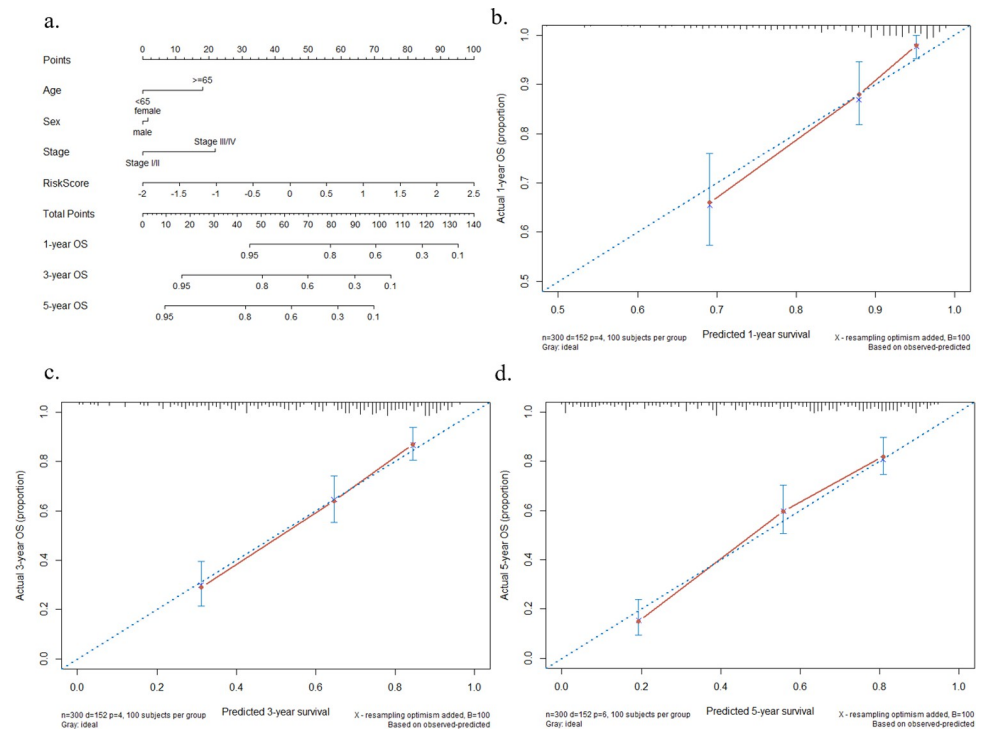


Fig 8. Nomogram and calibration plots for the prognostic model. a) Nomogram plotted by gene signature and clinical factors. b-d) Calibration plots demonstrating the consistency between predicted and observed 1-year, 3-year, and 5-year survival outcomes.

<https://doi.org/10.1371/journal.pone.0266805.g008>

Discussion

As one of the most common tumour diseases, GC has a relatively high incidence and mortality rate, especially in Asia [1]. Reliable markers could be applied as the vital indicator for GC risk stratification and following treatment. Compared to the traditional model with clinical characteristics, the prognostic model combining biomarkers would be more powerful for prediction performance, especially in the current era of precision medicine. However, the high-dimensional feature of genetic data makes it difficult to model directly, and some effective statistical methods are needed, such as PCA, ridge, and lasso [4]. In this study, we used a Bayesian approach (also known as spike-and-slab lasso Cox) to screen target genes [12]. This method integrated mainly the penalized lasso and Bayesian variable selection, which was superior to lasso. First, this Bayesian model could achieve the same function of variable selection as lasso Cox. Second, compared with lasso Cox giving the same penalty parameter to all coefficients, spike-and-slab lasso Cox could achieve more flexibility. In other words, it could selectively shrink different predictors based on different scales from the data (*i.e.*, giving relatively small shrinkage to those predictors with large effect and giving strong shrinkage to irrelevant or weak predictors at the same time). Therefore, to some extent, this Bayesian model could reduce the estimation bias of lasso Cox. Our study also demonstrated that the cross-validated C-index of the Bayesian model was better than the traditional lasso (0.684 vs 0.643).

After choosing the optimal model, 9 prognostic genes with non-zero coefficients were selected. Next, we established the risk score index based on these genes, and Kaplan-Meier analysis demonstrated that GC patients in the high-risk group (≥ -0.077) had a shorter survival time than those in the low-risk group (< -0.077). Additionally, we tested the predictive power of the 9-gene signature alone, and the results showed that it was a great predictor for OS

(AUC = 0.765 in the training set; AUC = 0.703 in the validation set). Compared to a 4-gene signature reported by a previous study, our signature acquired a higher level (AUC for OS, 0.765 vs 0.684) [19]. Furthermore, we integrated the risk score with clinical factors (age, sex, pathological stage) to construct a multivariate Cox regression model. The results proved that this risk signature could be an independent prediction marker for GC. The C-index of the integrative model reached 0.75 in both the training set and validation set, which showed the reliable performance of our model. Finally, considering the possible clinical application, a nomogram was provided to visually predict the GC patients' survival. By evaluating 1-year, 3-year, and 5-year OS calibration curves, the drawn nomogram had a relatively high accuracy of prediction.

Through the Bayesian hierarchical lasso Cox model, the 9 genes we found were also the focus of other basic researches or population studies. *NOTCH3*, a member of the NOTCH family, is involved in the NOTCH signalling pathway, which is regarded as one of the key pathways constituting the stem cell signalling network [20]. Our results demonstrated that the increased expression of *NOTCH3* was associated with poor OS among GC patients, which was consistent with previous studies [21, 22]. *TNFRSF11A*, also regarded as the receptor activator of NF- κ B (RANK), can activate several pathways, such as NF- κ B, JNK, ERK, p38, and Akt/PKB, and was reported to be a novel and frequent target for de novo methylation in gliomas [23]. Our study showed that the expression of *TNFRSF11A* was positively associated with survival (HR = 0.7964 < 1), which was consistent with another study [24]. In population-based studies, *TPMT*, *TOR2A*, *KIF11*, and *EIF5A* were reported to be associated with prognosis in other tumours, such as childhood acute lymphoblastic leukaemia [25, 26], ovarian cancer [27, 28], breast cancer [29], and colorectal cancer [30].

NMNAT1 is involved in the NAD⁺ salvage/recycling pathway, which is crucial for maintaining the functions of a wide variety of NAD⁺-dependent enzymes in the cytoplasm and nucleus [31]. Knockdown of *NMNAT1* enhanced rRNA transcription, which might facilitate increased ribosome biogenesis and tumour development [32]. However, to our knowledge, for the *NMNAT1* signature, there have been no relevant studies based on populations. Our study revealed that high expression of *NMNAT1* was associated with lower mortality risk and higher OS than low expression. *E2F8* is a member of the E2F family of transcription factors that regulates various cellular functions related to the cell cycle and apoptosis [33]. *E2F8* is considered to be a kind of transcriptional repressor that is similar to *E2F7* in that it can inhibit E2F-driven promoters [34]. A previous study found that the increased expression of E2F family members (*E2F2*, *E2F5*, *E2F6*, and *E2F7*) was significantly associated with favourable OS in GC [35]. Our research further revealed that increased expression of *E2F8* was also associated with favourable prognosis.

Notably, some previously reported genes, such as *PSMA5* and *KIF11*, can be treated as potential therapeutic targets for tumours [36, 37]. Previous studies demonstrated that the upregulation of *PSMA5*, by activating the key Nrf2/ARE signalling pathway, played a critical role in the mechanism of inducing tumour cell apoptosis caused by combined chemotherapy regimens [38, 39]. *KIF11* silencing induced chromosome instability (CIN), which might contribute to cancer development and progression [40]. However, other studies showed that the suppression of *PSMA5* could strengthen the sensitivity of myeloma to bortezomib [41]. Apigenin induced apoptosis in prostate cancer cells, which was accompanied by the downregulated expression of *PSMA5* [42]. These studies suggested that the same gene might involve different mechanisms in different tumour types.

In general, our study used a Bayesian hierarchical lasso Cox model to screen a prognosis-related gene signature. It was the first to apply this Bayesian approach to construct prognosis-related models in GC. Notably, there might be some restrictions regarding generalizability to

populations in other regions because this model we built was based on the Asian population. Moreover, our analysis focused mainly on mRNA expression data, but other molecular types, such as microRNA, CNV, or methylation, might contain important prognostic information in GC. Therefore, in further research, we would consider establishing a more generalized model that combines different molecular data for better prediction based on our Bayesian approach. Finally, although the 9-gene signature was explored by statistical analysis or database validation, we expect to further verify these genes by in vitro/vivo experiments in the future study.

Conclusion

Our research confirmed that the Bayesian hierarchical lasso Cox model had great prediction power than the traditional Cox model. Based on this Bayesian approach, we proposed a 9-gene prognostic signature as an independent predictor for the overall survival of GC patients. Finally, combined with clinical characteristics, a comprehensive nomogram was provided for clinical application. Overall, our study offers certain reference significance for clinical prognosis prediction in GC.

Supporting information

S1 Table. Leading edge subset outcome of GSEA.
(XLSX)

S2 Table. The Spearman correlation test among 9 genes from GSE66229.
(DOCX)

S1 Fig. Kaplan-Meier analysis of each gene based on median expression level for overall survival.
(DOCX)

S2 Fig. The immunohistochemistry images of nine genes in normal and tumour tissues derived from the Human Protein Atlas (HPA) database.
(DOCX)

Author Contributions

Conceptualization: Jiadong Chu, Na Sun.

Data curation: Jiadong Chu.

Formal analysis: Jiadong Chu.

Funding acquisition: Yueping Shen.

Methodology: Jiadong Chu, Na Sun, Nengjun Yi, Yueping Shen.

Software: Jiadong Chu, Na Sun, Wei Hu, Xuanli Chen.

Supervision: Nengjun Yi, Yueping Shen.

Validation: Jiadong Chu, Xuanli Chen.

Visualization: Wei Hu, Xuanli Chen.

Writing – original draft: Jiadong Chu.

Writing – review & editing: Nengjun Yi, Yueping Shen.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018; 68(6):394–424. <https://doi.org/10.3322/caac.21492> PMID: 30207593
2. Digkila A, Wagner AD. Advanced gastric cancer: Current treatment landscape and future perspectives. *World J Gastroenterol*. 2016; 22(8):2403–14. <https://doi.org/10.3748/wjg.v22.i8.2403> PMID: 26937129
3. Abbas M, Faggian A, Sintali DN, Khan GJ, Naeem S, Shi M, et al. Current and future biomarkers in gastric cancer. *Biomed Pharmacother*. 2018; 103:1688–700. <https://doi.org/10.1016/j.biopha.2018.04.178> PMID: 29864959
4. Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, et al. Predicting survival from microarray data—a comparative study. *Bioinformatics*. 2007; 23(16):2080–7. <https://doi.org/10.1093/bioinformatics/btm305> PMID: 17553857
5. Bonato V, Baladandayuthapani V, Broom BM, Sulman EP, Aldape KD, Do KA. Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*. 2011; 27(3):359–67. <https://doi.org/10.1093/bioinformatics/btq660> PMID: 21148161
6. Kaderali L, Zander T, Faigle U, Wolf J, Schultze JL, Schrader R. CASPAR: a hierarchical bayesian approach to predict survival times in cancer from gene expression data. *Bioinformatics*. 2006; 22(12):1495–502. <https://doi.org/10.1093/bioinformatics/btl103> PMID: 16554338
7. Chekouo T, Stingo FC, Doecke JD, Do KA. miRNA-target gene regulatory networks: A Bayesian integrative approach to biomarker selection with application to kidney cancer. *Biometrics*. 2015; 71(2):428–38. <https://doi.org/10.1111/biom.12266> PMID: 25639276
8. Treppmann T, Ickstadt K, Zucknick M. Integration of Multiple Genomic Data Sources in a Bayesian Cox Model for Variable Selection and Prediction. *Comput Math Methods Med*. 2017; 2017:7340565. <https://doi.org/10.1155/2017/7340565> PMID: 28828032
9. Annest A, Bumgarner RE, Raftery AE, Yeung KY. Iterative Bayesian Model Averaging: a method for the application of survival analysis to high-dimensional microarray data. *BMC Bioinformatics*. 2009; 10:72. <https://doi.org/10.1186/1471-2105-10-72> PMID: 19245714
10. Nikoienjad A, Wang W, Johnson VE. Bayesian variable selection for survival data using inverse moment priors. *Ann Appl Stat*. 2020; 14(2):809–28. <https://doi.org/10.1214/20-AOAS1325> PMID: 33456641
11. Lee KH, Chakraborty S, Sun J. Bayesian Variable Selection in Semiparametric Proportional Hazards Model for High Dimensional Survival Data. *The International Journal of Biostatistics*. 2011; 7(1):1–32. <https://doi.org/10.2202/1557-4679.1301>
12. Tang Z, Shen Y, Zhang X, Yi N. The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics*. 2017; 33(18):2799–807. <https://doi.org/10.1093/bioinformatics/btx300> PMID: 28472220
13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005; 102(43):15545–50. <https://doi.org/10.1073/pnas.0506580102> PMID: 16199517
14. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015; 43(Database issue):D447–52. <https://doi.org/10.1093/nar/gku1003> PMID: 25352553
15. Asplund A, Edqvist PH, Schwenk JM, Pontén F. Antibodies for profiling the human proteome—The Human Protein Atlas as a resource for cancer research. *Proteomics*. 2012; 12(13):2067–77. <https://doi.org/10.1002/pmic.201100504> PMID: 22623277
16. Tran AN, Dussaq AM, Kennell T Jr., Willey CD, Hjelmeland AB. HPAanalyze: an R package that facilitates the retrieval and analysis of the Human Protein Atlas data. *BMC Bioinformatics*. 2019; 20(1):463. <https://doi.org/10.1186/s12859-019-3059-z> PMID: 31500569
17. Yi N, Tang Z, Zhang X, Guo B. BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology. *Bioinformatics*. 2019; 35(8):1419–21. <https://doi.org/10.1093/bioinformatics/bty803> PMID: 30219850
18. Tang Z, Shen Y, Zhang X, Yi N. The Spike-and-Slab Lasso Generalized Linear Models for Prediction and Associated Genes Detection. *Genetics*. 2017; 205(1):77–88. <https://doi.org/10.1534/genetics.116.192195> PMID: 27799277
19. Chen J, Wang A, Ji J, Zhou K, Bu Z, Lyu G, et al. An Innovative Prognostic Model Based on Four Genes in Asian Patient with Gastric Cancer. *Cancer Res Treat*. 2021; 53(1):148–61. <https://doi.org/10.4143/crt.2020.424> PMID: 32878427
20. Katoh M, Katoh M. Notch signaling in gastrointestinal tract (review). *Int J Oncol*. 2007; 30(1):247–51. PMID: 17143535

21. Cui Y, Li Q, Li W, Wang Y, Lv F, Shi X, et al. NOTCH3 is a Prognostic Factor and Is Correlated With Immune Tolerance in Gastric Cancer. *Front Oncol.* 2020; 10:574937. <https://doi.org/10.3389/fonc.2020.574937> PMID: 33479597
22. Kang W, Zhang J, Huang T, Zhou Y, Wong CC, Chan RCK, et al. NOTCH3, a crucial target of miR-491-5p/miR-875-5p, promotes gastric carcinogenesis by upregulating PHLDB2 expression and activating Akt pathway. *Oncogene.* 2021. <https://doi.org/10.1038/s41388-020-01579-3> PMID: 33452458
23. von dem Knesebeck A, Felsberg J, Waha A, Hartmann W, Scheffler B, Glas M, et al. RANK (TNFRSF11A) is epigenetically inactivated and induces apoptosis in gliomas. *Neoplasia.* 2012; 14(6):526–34. <https://doi.org/10.1596/neo.12360> PMID: 22787434
24. Hou JY, Wang YG, Ma SJ, Yang BY, Li QP. Identification of a prognostic 5-Gene expression signature for gastric cancer. *Journal of Cancer Research and Clinical Oncology.* 2017; 143(4):619–29. <https://doi.org/10.1007/s00432-016-2324-z> PMID: 28035468
25. Schmiegelow K, Forestier E, Kristinsson J, Söderhäll S, Vettenranta K, Weinshilboum R, et al. Thiopurine methyltransferase activity is related to the risk of relapse of childhood acute lymphoblastic leukemia: results from the NOPHO ALL-92 study. *Leukemia.* 2009; 23(3):557–64. <https://doi.org/10.1038/leu.2008.316> PMID: 18987654
26. Lennard L, Cartwright CS, Wade R, Vora A. Thiopurine dose intensity and treatment outcome in childhood lymphoblastic leukaemia: the influence of thiopurine methyltransferase pharmacogenetics. *Br J Haematol.* 2015; 169(2):228–40. <https://doi.org/10.1111/bjh.13240> PMID: 25441457
27. Zhang Q, Chen WM, Zhang XX, Zhang HX, Wang HC, Zheng FY, et al. Overexpression of salusin- β is associated with poor prognosis in ovarian cancer. *Oncol Rep.* 2017; 37(3):1826–32. <https://doi.org/10.3892/or.2017.5429> PMID: 28184918
28. Liu J, Meng H, Li S, Shen Y, Wang H, Shan W, et al. Identification of Potential Biomarkers in Association With Progression and Prognosis in Epithelial Ovarian Cancer by Integrated Bioinformatics Analysis. *Front Genet.* 2019; 10:1031. <https://doi.org/10.3389/fgene.2019.01031> PMID: 31708970
29. Deng JL, Xu YH, Wang G. Identification of Potential Crucial Genes and Key Pathways in Breast Cancer Using Bioinformatic Analysis. *Front Genet.* 2019; 10:695. <https://doi.org/10.3389/fgene.2019.00695> PMID: 31428132
30. Tunca B, Tezcan G, Cecener G, Egeli U, Zorluoglu A, Yilmazlar T, et al. Overexpression of CK20, MAP3K8 and EIF5A correlates with poor prognosis in early-onset colorectal cancer patients. *J Cancer Res Clin Oncol.* 2013; 139(4):691–702. <https://doi.org/10.1007/s00432-013-1372-x> PMID: 23322277
31. Zhang T, Berrocal JG, Frizzell KM, Gamble MJ, DuMond ME, Krishnakumar R, et al. Enzymes in the NAD⁺ salvage pathway regulate SIRT1 activity at target gene promoters. *J Biol Chem.* 2009; 284(30):20408–17. <https://doi.org/10.1074/jbc.M109.016469> PMID: 19478080
32. Song T, Yang L, Kabra N, Chen L, Koomen J, Haura EB, et al. The NAD⁺ synthesis enzyme nicotinamide mononucleotide adenyltransferase (NMNAT1) regulates ribosomal RNA transcription. *J Biol Chem.* 2013; 288(29):20908–17. <https://doi.org/10.1074/jbc.M113.470302> PMID: 23737528
33. Xanthoulis A, Tiniakos DG. E2F transcription factors and digestive system malignancies: how much do we know? *World J Gastroenterol.* 2013; 19(21):3189–98. <https://doi.org/10.3748/wjg.v19.i21.3189> PMID: 23745020
34. Christensen J, Cloos P, Toftegaard U, Klinkenberg D, Bracken AP, Trinh E, et al. Characterization of E2F8, a novel E2F-like cell-cycle regulated repressor of E2F-activated transcription. *Nucleic Acids Res.* 2005; 33(17):5458–70. <https://doi.org/10.1093/nar/gki855> PMID: 16179649
35. Manicum T, Ni F, Ye Y, Fan X, Chen BC. Prognostic values of E2F mRNA expression in human gastric cancer. *Biosci Rep.* 2018; 38(6). <https://doi.org/10.1042/BSR20181264> PMID: 30487158
36. Frankland-Searby S, Bhaumik SR. The 26S proteasome complex: an attractive target for cancer therapy. *Biochim Biophys Acta.* 2012; 1825(1):64–76. <https://doi.org/10.1016/j.bbcan.2011.10.003> PMID: 22037302
37. Garcia-Saez I, Skoufias DA. Eg5 targeting agents: From new anti-mitotic based inhibitor discovery to cancer therapy and resistance. *Biochem Pharmacol.* 2020; 184:114364. <https://doi.org/10.1016/j.bcp.2020.114364> PMID: 33310050
38. Seo SU, Cho HK, Min KJ, Woo SM, Kim S, Park JW, et al. Thioridazine enhances sensitivity to carboplatin in human head and neck cancer cells through downregulation of c-FLIP and Mcl-1 expression. *Cell Death Dis.* 2017; 8(2):e2599. <https://doi.org/10.1038/cddis.2017.8> PMID: 28182008
39. Seo SU, Kim TH, Kim DE, Min KJ, Kwon TK. NOX4-mediated ROS production induces apoptotic cell death via down-regulation of c-FLIP and Mcl-1 expression in combined treatment with thioridazine and curcumin. *Redox Biol.* 2017; 13:608–22. <https://doi.org/10.1016/j.redox.2017.07.017> PMID: 28806703

40. Asbaghi Y, Thompson LL, Lichtensztein Z, McManus KJ. KIF11 silencing and inhibition induces chromosome instability that may contribute to cancer. *Genes Chromosomes Cancer*. 2017; 56(9):668–80. <https://doi.org/10.1002/gcc.22471> PMID: 28510357
41. Zhu YX, Tiedemann R, Shi CX, Yin H, Schmidt JE, Bruins LA, et al. RNAi screen of the druggable genome identifies modulators of proteasome inhibitor sensitivity in myeloma including CDK5. *Blood*. 2011; 117(14):3847–57. <https://doi.org/10.1182/blood-2010-08-304022> PMID: 21289309
42. Singh V, Sharma V, Verma V, Pandey D, Yadav SK, Maikhuri JP, et al. Apigenin manipulates the ubiquitin-proteasome system to rescue estrogen receptor- β from degradation and induce apoptosis in prostate cancer cells. *Eur J Nutr*. 2015; 54(8):1255–67. <https://doi.org/10.1007/s00394-014-0803-z> PMID: 25408199