Clinical and Translational Oncology (2021) 23:2368–2381
https://doi.org/10.1007/s12094-021-02638-1

RESEARCH ARTICLE

A novel 12-gene signature as independent prognostic model in stage IA and IB lung squamous cell carcinoma patients

K. Wang[1,2] · Y. Li[3] · J. Wang[4] · R. Chen[5] · J. Li[4]

Received: 25 March 2021 / Accepted: 6 May 2021 / Published online: 24 May 2021

Abstract

Background There is currently no formal consensus on the administration of adjuvant chemotherapy to stage I lung squamous cell carcinoma (LUSC) patients despite the poor prognosis. The side effects of adjuvant chemotherapy need to be balanced against the risk of tumour recurrence. Prognostic markers are thus needed to identify those at higher risks and recommend individualised treatment regimens.

Methods Clinical and sequencing data of stage I patients were retrieved from the Lung Squamous Cell Carcinoma project of the Cancer Genome Atlas (TCGA) and three tissue microarray datasets. In a novel K-resample gene selection algorithm, gene-wise Cox proportional hazard regressions were repeated for 50 iterations with random resamples from the TCGA training dataset. The top 200 genes with the best predictive power for survival were chosen to undergo an L1-penalised Cox regression for further gene selection.

Results A total of 602 samples of LUSC were included, of which 42.2% came from female patients, 45.3% were stage IA cancer. From an initial pool of 11,212 genes in the TCGA training dataset, a final set of 12 genes were selected to construct the multivariate Cox prognostic model. Among the 12 selected genes, 5 genes, STAU1, ADGRF1, ATF7IP2, MALL and KRT23, were adverse prognostic factors for patients, while seven genes, NDUFB1, CNPY2, ZNF394, PIN4, FZD8, NBPF26 and EPYC, were positive prognostic factors. An equation for risk score was thus constructed from the final multivariate Cox model. The model performance was tested in the sequestered TCGA testing dataset and validated in external tissue microarray datasets (GSE4573, GSE31210 and GSE50081), demonstrating its efficacy in stratifying patients into high- and low-risk groups with significant survival difference both in the whole set (including stage IA and IB) and in the stage IA only subgroup of each set. The prognostic power remains significant after adjusting for standard clinical factors. When benchmarked against other prominent gene-signature based prognostic models, the model outperformed the rest in the TCGA testing dataset and in predicting long-term risk at eight years in all three validation datasets.

Conclusion The 12-gene prognostic model may serve as a useful complementary clinical risk-stratification tool for stage I and especially stage IA lung squamous cell carcinoma patients to guide clinical decision making.

Keywords Lung squamous cell carcinoma · Risk stratification · Prognostic model · Gene signature

K. Wang and Y. Li contributed equally to the manuscript.

✉ R. Chen
chenruigood@126.com

✉ J. Li
aljxcr@suda.edu.cn

[1] School of Clinical Medicine, The University of Cambridge, Cambridge, UK

[2] School of Medicine, The University of Leeds, Leeds, UK

[3] School of Medicine, The University of Manchester, Manchester, UK

[4] School of Public Health, Medical College of Soochow University, 199 Renai Rd., Suzhou 215123, Jiangsu, China

[5] Respiratory Department, The Second Affiliated Hospital of the Soochow University, Suzhou 215004, China

Introduction

To date, lung cancer remains the most commonly diagnosed cancer and the leading cause of cancer death worldwide [15]. Lung squamous cell carcinoma (LUSC) is the second most common lung cancer subtype, comprising approximately 20% of primary lung neoplasms in the United States [3].

The gold standard treatment for squamous cell lung carcinomas treatment is surgical resection, sometimes combined with adjuvant platinum-based chemotherapy, and complemented by immunotherapy and targeted therapy where necessary [26]. However, when diagnosed in the earlier stages, current guidance is unclear on the use of adjuvant chemotherapy and requires case-by-case judgment based on individual risk profiles. The latest meta-analysis of several large-scale trials [2, 5, 10, 25, 30] suggests platinum-based adjuvant chemotherapy confers survival benefits to patients in stage IB(> 4 cm) to stage III, but nil or even detrimental survival effects for stage IA patients. Accordingly, the 8th edition of the Tumour, Node, Metastasis (TNM) staging system has since reclassified tumour size > 4 cm without lymph node involvement as stage IIA and recommended adjuvant chemotherapy [14]. However, for stage IB cancers with tumours measuring between 3 and 4 cm, the 5-year survival averages still at 68%, not significantly different from the 60% 5-year survival in stage IIA patients [12]. At stage IA, the 5-year survival ranges from 77 to 92%, and there is no conclusive evidence to guide the administration of adjuvant chemotherapy in this group. Therefore, to balance out the risk of tumour recurrence and the side effects of adjuvant chemotherapy, other prognostic markers are needed for further risk stratification within early-stage LUSC patients to identify those at higher risks to recommend individualised treatment regimens.

Previous studies have explored the potential of using gene-expression profiles for the prognosis and prediction of response to lung cancer therapies [33]. Zhu et al. constructed a 15-gene model from 133 samples in the JBR.10 trial which was able to predict response to adjuvant chemotherapy in early-stage non-small cell lung cancer (NSCLC) [42]. Kratz et al. developed a 14-gene-expression based model from 361 resected non-squamous NSCLC samples which offered prognostic value in small < 2 cm node-negative stage IA patients beyond standard clinical factors [17]. Nonetheless, as Kinoshita pointed out in his editorial, given the distinct carcinogenesis pathways and molecular phenotypes in LUSCs that is not common to all NSCLCs [13, 16], there remains a need for a more targeted assay in the early-stage LUSC patients for more precise risk stratification.

Methodologically, despite the various promising genetic prognostic models in development [37], no model has so far been incorporated into routine clinical practice due to issues such as overfitting on small training datasets, lack of sufficient validation, improper patient selection [31], thus conferring limited clinical utility. Acknowledging these potential pitfalls, this study sets out to keep the end goal of patient-centred applications in mind to address unmet clinical needs, specifically by (1) focusing on stage IA and IB cancers where clinical decisions on chemotherapy administration require additional risk-stratification; (2) ensuring the model offers additional predictive power over the standard prognostic factors such as tumour size, histological grading and margin status; and (3) validating the model against external datasets [31, 43]. As such, this study aims to develop a robust genetic prognostic model as a complementary tool to stratify stage IA and IB lung squamous cell carcinomas, for potential clinical applications in risk stratification and individualised treatment.

Method

Study design and data curation

A retrospective data analysis was performed on expression profiles of frozen tissue samples retrieved from four publicly available datasets, including three tissue microarray datasets (GSE4573, GSE31210 and GSE50081, see each study for details on sample preparation [8, 22, 27]) and one RNA-Seq dataset from the Cancer Genome Atlas (TCGA) Lung Squamous Cell Carcinoma (LUSC) project. The GSE4573 dataset was retrieved directly from the Gene Expression Omnibus (GEO) online database, while samples from the GSE31210 and GSE50081 were acquired from a large published merged dataset [20], which have been pre-processed, normalised and corrected for batch effects. Genes were matched by entrez IDs cross datasets, and gene expressions were all log2-transformed. Only patients with stage IA or IB non-recurrent primary lung squamous cell carcinomas and complete survival information were included in the analysis. The process of the study is summarised in Fig. 1.

K-Cox gene selection and model construction

The TCGA–LUSC dataset, the largest of the four, was chosen for gene selection and model training. Only two-thirds of the samples ($N = 161$) in the TCGA dataset were used to train the model, while the remaining one-third of the samples ($N = 81$) was reserved for testing. The partitioning took into account the survival information, with approximately 40% of patients right-censored in both the training and testing dataset.

Genes with counts-per-million smaller than 1 in more than 80% of all samples were excluded from the analysis. *K*

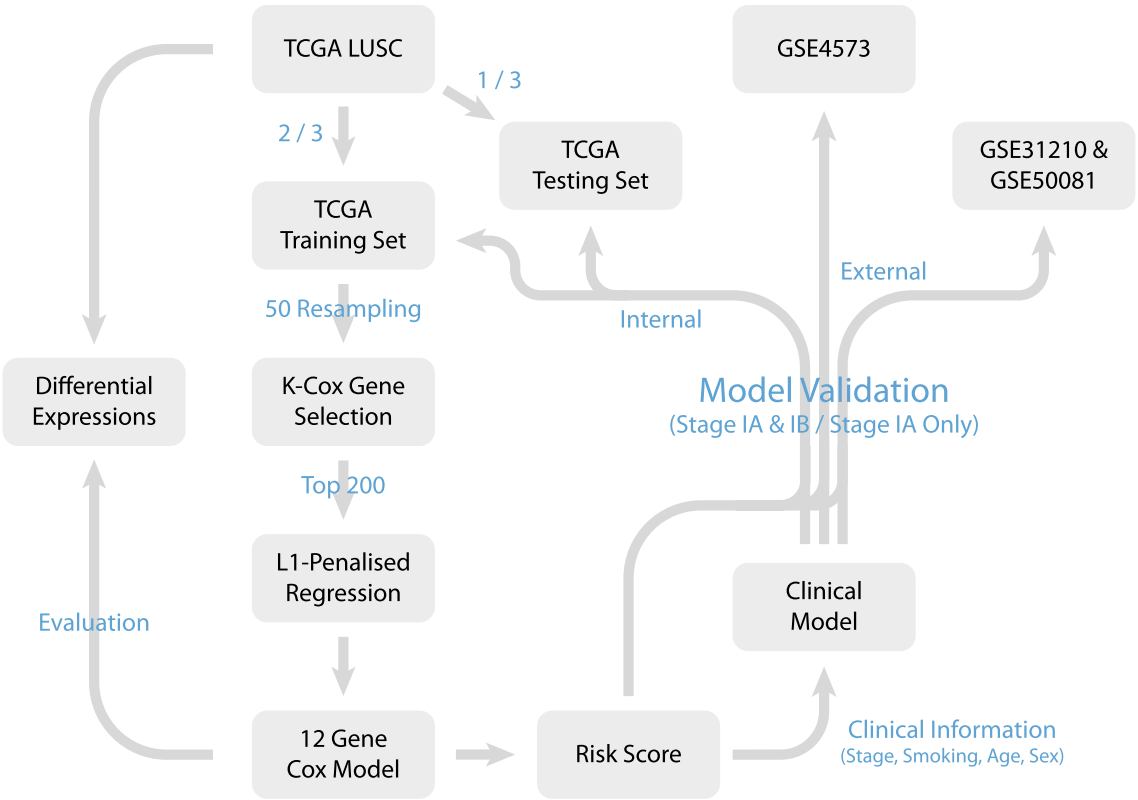

Fig. 1 Study flow diagram. *TCGA* the Cancer Genome Atlas, *LUCS* lung squamous cell carcinoma, *Cox* Cox proportional hazard model, *K-Cox* K-resampling Cox regression

random resamples ($K = 50$ in this study) were drawn from within the training dataset ($N = 161$), each consisting of two-thirds of the training set (i.e. 107 samples in each resample). In each resample, a separate Cox proportional hazard model was fitted for each gene, with the age of diagnosis included as a covariate. The results of the 50 runs of gene screening were combined by averaging the p values of each gene's predictive power in the Cox regressions. The top 200 genes with the smallest average p values were selected to obtain a stable list of genes most predictive of survival.

The expressions of these selected genes were then scaled and further selected by an L1-penalised Cox proportional hazard regression on the training set, with the penalty coefficient lambda chosen by a tenfold cross-validation process. A final multivariate Cox proportional hazard model was constructed using the final set of selected genes and other available clinical parameters to predict overall survival.

Model validation and evaluation

The final set of selected genes were validated internally against the reserved TCGA testing set, and externally against two sets of microarray samples, namely the GSE4573 dataset ($N = 73$), as well as a larger merged set consisting of all stage I lung squamous cell carcinoma (LUSC) and adenocarcinoma (LUAD) patients retrieved from the GSE50081 and GSE31210 repositories ($N = 287$). In the internal validation, a hazard score was computed for each patient in the training and the testing set using the coefficients associated with each predictor in the Cox model that was trained exclusively using the training dataset. The risk score was dichotomised at the medium to categorise patients into high-risk and low-risk groups, whose survival probabilities were calculated using the Kaplan–Meier estimator and compared using the log-rank test. A multivariate Cox regression on the overall survival was run using the derived risk score along with other available clinical parameters to assess the models' independent prognostic power. The same workflow was then performed in the two external datasets.

To further benchmark the model's performance, we evaluated our model against three promising, clinically ready NSCLC prognostic models and two best performing model in the LUSC population identified by existing literature [18, 23, 33, 34, 42]. All models selected contained similar numbers of gene signatures to avoid the problem of falsely elevated prognostic power due to overfitting using comparatively larger numbers of predictors in multivariate Cox regression. The set of genes from each study was taken

Table 1 Clinical characteristics of patients in datasets

	TCGA training $N=161$	TCGA testing $N=81$	GSE4573 $N=73$	GSE31210 and GSE50081 $N=287$
Age (years)	70 (57–78)	70 (60–78)	71 (59–81)	64 (52–76)
Gender				
Female	47 (29%)	27 (33%)	27 (37%)	153 (53%)
Male	114 (71%)	54 (67%)	46 (63%)	134 (47%)
Stage				
Stage IA	64 (40%)	25 (31%)	27 (37%)	158 (55%)
Stage IB	95 (59%)	55 (68%)	46 (63%)	129 (45%)
Stage I*	2 (1.2%)	1 (1.2%)	–	
Follow-up (years)	1.75 (0.23–6.49)	1.9 (0.2–7.2)	2.93 (0.85–7.37)	5.00 (2.24–7.40)
Number deceased	62	37	33	61

Numerical values are reported as medium (10th percentile–90th percentile); Categorical variables are reported by count (percentage in brackets)

*Stage I unspecified, no classification information available regarding the sample's substage

to fit a multivariate Cox model, from which the risk scores were calculated (We did not use their reported coefficients, if any, to allow a fairer comparison to our model). Genes that could not be mapped into our datasets were left out. Hazard ratio, concordance index (*C*-index) and AUC at 3, 5 and 8 years were used as benchmarking parameters, as was used in a previous meta-analysis evaluating the performance of various genetic prognostic models in lung cancer [33].

Lastly, to evaluate our methodology in comparison with the mainstream DEG-based prognostic gene selection, a differential expression analysis was carried out in the TCGA dataset ($N=271$, stage IA and IB) using the *edgeR* package, comparing the primary tumour samples ($N=242$) with normal solid tissue samples ($N=29$).

Statistical analysis

All statistical analyses were performed using R (version 3.6.3, https://www.r-project.org/). Data partitioning was conducted using the *caret* package, survival calculations using the *survival* and *survminer* packages, penalised regressions using the *glmnet* package and diagrams and illustrations using the *ggplot2* and *ggpubr* package. Statistical significance was defined as $p<0.05$ unless specified otherwise.

Results

Subject characteristics

Overall, data from a total of 602 samples of LUSC were included in the study, of which 42.2% came from female patients, 45.3% were stage IA cancer. The median age was 67 years (interquartile range, IQR = 61.0–73.0), and the median length of follow-up was 3.72 years (IQR = 1.72–5.40) (see Table 1).

12-gene model construction and testing in TCGA dataset

After filtering out 45,281 genes with low expressions, all of the remaining 11,212 genes underwent the 50-resample K-Cox selection to screen for genes associated with overall survival. The 50 *p* values obtained for each gene were averaged, by which the genes were ranked (see Fig. 2a). The top 200 selected genes underwent an L1-penalised regression, and the cross-validated penalty coefficient lambda was chosen at 0.115, resulting in 12 genes with non-zero coefficients (see Fig. 2b). Among the 12 selected genes, 5 genes, STAU1, ADGRF1, ATF7IP2, MALL and KRT23, were adverse prognostic factors for patients, while 7 genes, NDUFB1, CNPY2, ZNF394, PIN4, FZD8, NBPF26 and EPYC, were positive prognostic factors (see Table 2 for hazard ratios in the univariate and multivariate Cox regressions).

A final multivariate Cox model was then constructed using the 12 selected genes (see Table 2), with 6 out of 12 genes being independent predictors of survival in the training dataset (see Fig. 2c). An equation for risk score was thus constructed from the final multivariate Cox model:

$$\begin{aligned}\text{Risk Score} = {} & 0.633 \times \text{STAU1} - 0.379 \times \text{NDUFB1} - 0.761 \times \text{CNPY2} - 0.454 \times \text{ZNF394} \\ & - 0.541 \times \text{PIN4} - 0.293 \times \text{FZD8} + 0.181 \times \text{ADGRF1} + 0.301 \times \text{ATF7IP2} \\ & + 0.032 \times \text{MALL} + 0.017 \times \text{KRT23} - 0.326 \times \text{NBPF26} - 0.226 \times \text{EPYC},\end{aligned}$$

where the log2-transformed gene expressions are multiplied by coefficients computed using data from the TCGA training set. Individual risk scores were thus calculated for the patients in both the TCGA training set and the TCGA

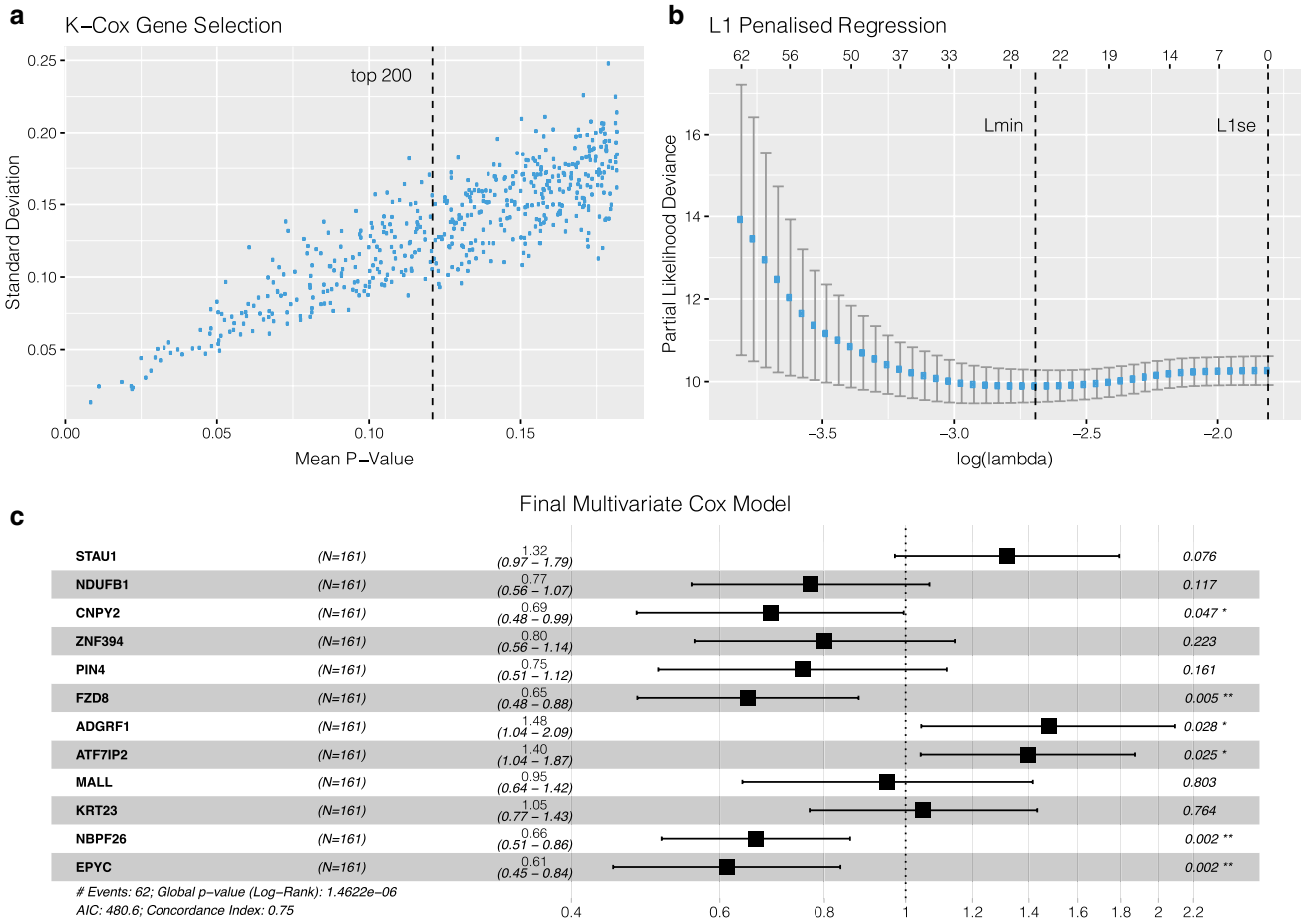

Fig. 2 Gene selection processes. **a** The top 200 genes with the lowest mean *p* values (left of the dotted line) were selected for L1-penalised regression. **b** L1-penalised regression were conducted on the 200 genes, the lambda was selected through a tenfold cross-validation process. **c** A multivariate Cox model was constructed with the 12 selected genes, and the forest plot illustrates the adjusted hazard ratio corresponding to each gene

testing set. Patients in either set were separately dichotomised by their median risk score and categorised into the high-risk and low-risk groups. There was a statistically significant distinction in survival between the high-risk and low-risk groups in both the training and testing datasets (see Fig. 3a). We then investigated if the same stratification survival difference could be extended to the stage IA subgroup. Notably, the computed risk score was sensitive enough to differentiate patients into high-risk and low-risk groups even within the stage IA cohorts in both sets. A multivariate Cox proportional hazard model was fitted with the computed risk score and clinical variables including patient age, gender, tumour stage (either stage IA or IB) and pack-years smoked (see Fig. 3b and c). The model showed that the computed risk score remained an independent predictor of survival in TCGA training set (Hazard ratio = 2.98, 95% CI 2.19–4.0, $p < 0.001$) and testing set (Hazard ratio = 2.43, 95% CI 1.69–3.5, $p < 0.001$). It is noteworthy that no clear prognostic differences were observed using tumour stage between patients in stage IA and those in stage IB in both TCGA training (Hazard ratio = 1.33, 95% CI 0.74–2.4, $p = 0.344$) and testing set (Hazard ratio = 1.24, 95% CI 0.49–3.2, $p = 0.648$).

Model validation in external datasets

In the two validation datasets, similarly, multivariate Cox models were constructed using the 12-gene signature and computed risk scores were dichotomised to separate the patients into high-risk and low-risk groups. In both the GSE4573 dataset (LUSC-only) and the merged GSE31210 and GSE50081 dataset (both a mixture of LUSC and LUAD), the high-risk groups exhibited significantly worse prognosis than the low-risk groups. This risk-stratification power was again sensitive enough even in the subset of stage IA patients (see Fig. 4). Multivariate Cox proportional hazard model suggested computed risk score is again an independent prognostic factor in GSE4573 (Hazard ratio = 2.77,

Table 2 Characteristics of the final 12-gene set

Symbol	K-Cox	Univariate Cox regression		Multivariate Cox regression		Differential Expr	
	Mean *p* (SD)	HR (95% CI)	p	HR (95% CI)	*p*	logFC	FDR
ADGRF1	0.030 (0.045)	1.179 (1.065–1.305)	0.001	1.199 (1.020–1.408)	0.028	0.215	0.639
NDUFB1	0.039 (0.047)	0.585 (0.416–0.824)	0.002	0.684 (0.426–1.100)	0.117	0.104	0.566
FZD8	0.052 (0.075)	0.833 (0.729–0.953)	0.008	0.746 (0.607–0.916)	0.005	−0.687	0.008
CNPY2	0.053 (0.097)	0.441 (0.274–0.712)	0.001	0.467 (0.221–0.989)	0.047	0.578	0.000
KRT23	0.056 (0.064)	1.117 (1.040–1.201)	0.003	1.017 (0.909–1.140)	0.764	4.265	0.000
MALL	0.073 (0.138)	1.225 (1.059–1.418)	0.006	0.968 (0.752–1.247)	0.803	−1.287	0.000·
STAU1	0.081 (0.095)	1.616 (1.021–2.557)	0.041	1.883 (0.935–3.790)	0.076	0.288	0.003
PIN4	0.091 (0.121)	0.540 (0.353–0.827)	0.005	0.582 (0.273–1.241)	0.161	0.220	0.097
ATF7IP2	0.105 (0.131)	1.266 (1.050–1.528)	0.014	1.351 (1.038–1.758)	0.025	−0.059	0.784
NBPF26	0.106 (0.098)	0.816 (0.700–0.951)	0.009	0.722 (0.588–0.886)	0.002	−0.249	0.265
EPYC	0.107 (0.136)	0.875 (0.789–0.971)	0.012	0.797 (0.691–0.921)	0.002	4.650	0.000
ZNF394	0.113 (0.137)	0.426 (0.253–0.717)	0.001	0.635 (0.306–1.318)	0.223	0.012	0.937

The mean *p* values in K-Cox analysis and the coefficients for both univariate and multivariate Cox regression are calculated in the TCGA training set. Differential expression is calculated using all stage I patients in the TCGA dataset, comparing gene expressions in normal solid tissues with that in primary tumours

K-Cox K-resampling Cox regression gene selection, *SD* standard deviation, *HR* hazard ratio, *logFC* log fold change in differential gene-expression analysis, *FDR* false discovery rate, *Expr* expression

95% CI 1.66–4.6, $p < 0.001$) and interestingly also in the merged dataset of GSE31210 and GSE50081 (Hazard ratio = 1.91, 95% CI 1.04–3.5, $p = 0.038$). Again, tumour stage alone was not prognostic between patients in stage IA and those in stage IB in GSE4573 (Hazard ratio = 1.82, 95% CI 0.82–4.0, $p = 0.139$). However, in the merged dataset of GSE31210 and GSE50081, tumour stage exhibited a significant prognostic difference between stage IA and IB patients (Hazard ratio = 2.45, 95% CI 1.28–4.7, $p = 0.007$).

Benchmarking against other prominent gene-based prognostic models

The summary of each model's final gene signatures in every dataset is outlined in Table 3. When compared against the five other promising clinical models, our model outperformed the rest in the TCGA training and testing set across all three metrics (see Fig. 5). Looking at hazard ratio and *C*-index, our model achieved a decent performance in external validation dataset GSE4573 and the merged set GSE31210 and 50,081 containing LUAD samples. When looking at AUC, our model performed the best at predicting longer term survival, i.e. at 8 years, while its predictive power is on par with other models at shorter time frames.

Differential expression analysis of the 12 genes

Finally, a differential expression analysis was carried out in the TCGA dataset, comparing the gene expressions of stage I tumour tissue samples and the corresponding normal solid lung tissue surgically removed from patients with stage I cancers. Only two out of the 12 genes exhibited differential expression patterns in tumour samples (see Fig. 6c). A combination of differential expression level and hazard ratio data revealed EPYC as a prognostically beneficial gene that is significantly up-regulated in tumour tissues and KRT23 as a prognostically detrimental gene that is also significantly up-regulated (Fig. 6d).

Discussion

In this study, we used the K-Cox gene selection method based on RNASeq data from TCGA to construct a reliable prognostic model that was able to stratify stage IA and IB LUSC patients into different risk groups. The risk category derived from the model was able to better predict mortality risks than using standard clinical prognostic factors alone. The independent prognostic power of the model was further validated in two external validation sets including one with LUAD patients. We also benchmarked our model against several of the most promising gene expression-based prognostic models and demonstrated that our model outperformed the rest in the RNA-seq based TCGA dataset as well as in predicting long-term survival at 8 years across all datasets. This could be possibly due to the lack of training data with long-term survival information and RNASeq-based datasets for the other models at the time of their development. This model was developed with a particular focus on stage IA and IB LUSC patients, while most previous gene expression-based prognostic models were trained to be deployed in the general NSCLC cohorts [43].

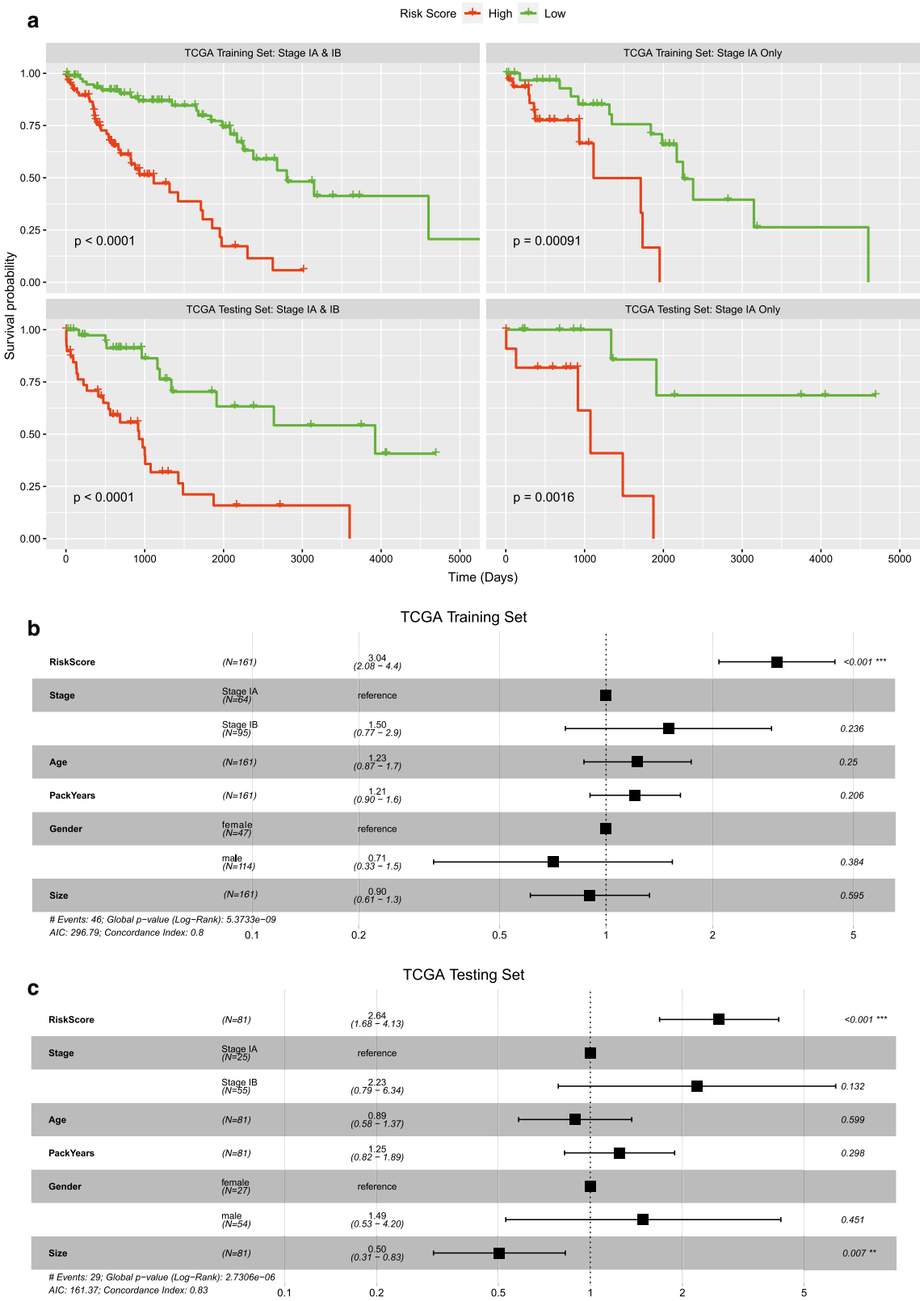

Fig. 3 Internal validation in the TCGA training dataset and testing dataset. **a** Kaplan–Meier survival curves for risk groups dichotomised at median risk score in stage IA and IB patients and stage IA only patients. **b** and **c** Multivariate Cox model shows the risk scores are independent prognostic factors

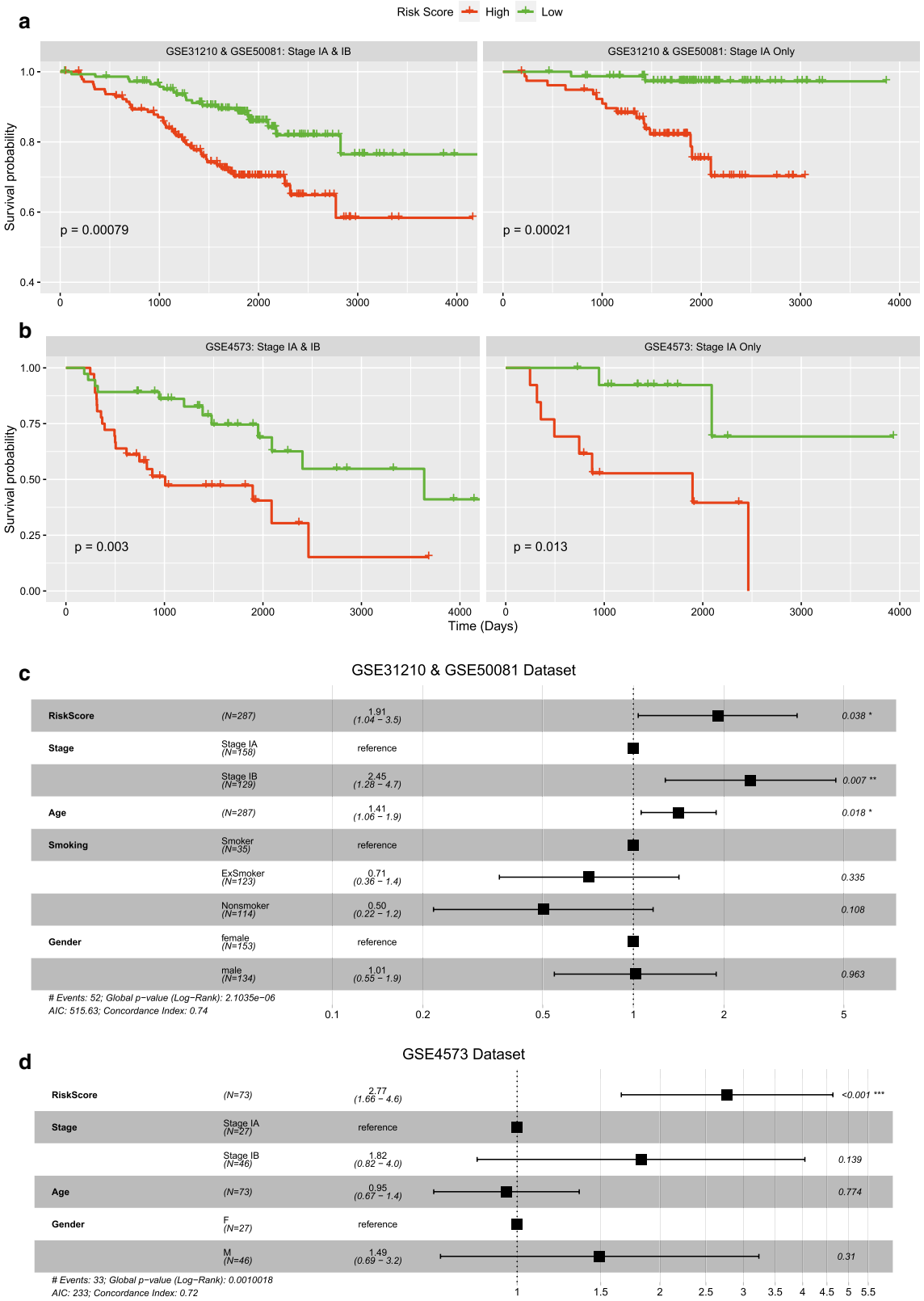

Fig. 4 External validation in the GSE31210 & GSE50081 and GSE4573 dataset. **a** and **b** Kaplan–Meier survival curves for risk groups dichotomised at median risk score in stage IA and IB patients and stage IA only patients. **c** and **d** Multivariate Cox model shows the risk scores are independent prognostic factors

Table 3 Summary of the current model used for the benchmark

Study	Raw gene	TCGA		GSE4573		GSE31210_50081	
		Ngenes	Missing genes	Ngenes	Missing genes	Ngenes	Missing genes
Kratz_2012	BAG1, BRCA1, CDC6, CDK2AP1, ERBB3, FUT3, IL11, LCK, RND3, SH3BGR, WNT3A, ESD, TBP, YAP1	14		13	WNT3A	13	WNT3A
Zhu_2010	MLANA, ATP1B1, L1CAM, STMN2, TRIM14, FAM64A, MB, EDN3, UMPS, MDM2, ZNF236, FOSL2, HEXIM1, MYT1L, IKBKAP, MDM2	13	FAM64A, IKBKAP	13	FAM64A, IKBKAP	16	
Tang_2013	DOCK9, RRM2, AURKA, HOPX, NKX2-1, TTC37, COL4A3, IFT57, C1orf116, HSD17B6, MBIP, ATP8A1	11	NKX2-1	11	NKX2-1	11	NKX2-1
Parmigiani_2004	GPC3, MALL, IRX5, FGFR2, FOLR1, TYRP1, STX1A, IGJ, MAD2L1, VEGFC, KIAA0101, IL6ST, SELE, ARHGDIB	12	IGJ, KIAA0101	12	IGJ, KIAA0101	14	
Kadara_2011	UBE2C, MCM2, MCM6, FEN1, TPX2	5		5		5	
This study	STAU1, ADGRF1, ATF7IP2, MALL, KRT23, NDUFB1, CNPY2, ZNF394, PIN4, FZD8, NBPF26, EPYC	12		12		10	ADGRF1, NBPF26

Raw gene the original genes reported in each model, *Ngenes* number of genes used for each model in respective datasets, *Missing gene* genes that are not found from each model in respective datasets

The performance in LUSC dataset GSE 4573 represents room for improvement in our prognostic signatures, likely since our model was only trained on RNASeq data. Mixing a portion of microarray data into the training set in the gene selection process may improve the general applicability of future prognostic models.

Methodologically, our gene selection approach presents a unique alternative to the mainstream differentially expressed genes (DEG) approach. A significant limitation of the DEG approach is that some of the healthy tissue annotated in many databases are in fact paracarcinomatous tissues rather than normal healthy tissue [1]. According to the recently proposed field cancerisation theory, paracarcinomatous tissues have already acquired certain genetic mutations in a stepwise manner [28] and thus carry genetic patterns different from true healthy tissues. Using the mRNA expression levels from paracarcinomatous tissues as a "healthy reference" in DEG analysis would overlook essential genes that are differentially expressed in both cancer and paracarcinomatous tissues compared to the truly healthy tissues. Furthermore, to screen for potential candidate genes among a significantly larger pool of genes than among only the DEGs, potential problems might arise from "false discoveries" due to the large number of individual regressions and the random characteristics of the particular sampling of the training dataset. In this study, by taking different subsets of the training dataset for K repetitions of Cox regressions, the K-Cox approach reduces the chances of Type I errors of falsely relying on a single significant p value in rejecting the null hypothesis, i.e. potentially including prognostically meaningless genes or excluding important ones. It is also worth noting that, the present algorithm started with an initial pool of 11,212 candidate genes before K-Cox selection, while typical DEG-based approaches select from a much smaller subset of genes, typically in the hundreds. This, therefore, demonstrates the sensitivity of our method in producing a set of prognostically significant genes that can be plausibly supported by mechanistic studies in the literature, as shown in the following section.

Even though only two genes exhibited a $\log_2$ fold change of more than 1.5, which is the conventional threshold for determining differentially expressed gene, the importance of the selected genes in tumourigenesis can be substantiated by previous mechanistic studies. Overexpression of ADGRF1 (Adhesion G Protein-Coupled Receptor F1) was shown to dampen mammosphere formation and anchorage-independent growth, a common feature of metastatic cell lines [4]. Another study found that ADGRF1 mRNA expression was positively correlated with E-cadherin (CDH1) and negatively correlated with vimentin and N-cadherin

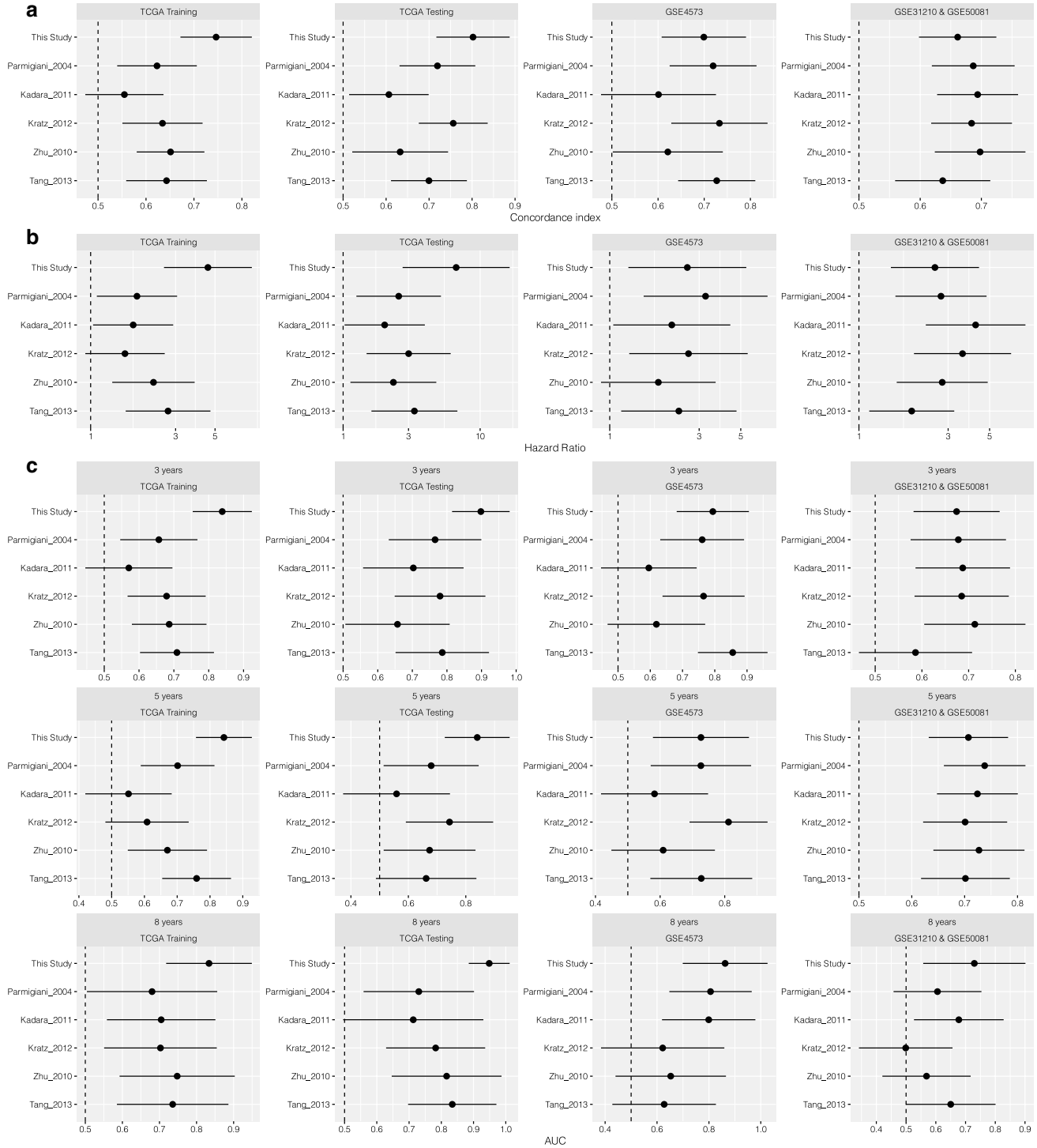

Fig. 5 Benchmarking performance in four different datasets using **a** *C*-index where our model outperforms the rest in TCGA datasets but not the external validations sets; **b** Hazard ratio where the same pattern can be seen; and **c** AUC at four different time points where in addition to superior performance in TCGA datasets, our model also has the best performance at predicting 8-year survival in both external validation sets

(CDH2), suggesting a potential role of ADGRF1 in the EMT process [6]. NDUFB1, a subunit of mitochondrial oxidative phosphorylation complex I, is related to bioenergetic pathways, including the electron transport chain and mitochondrial ATP synthesis coupled to electron transport [38]. FZD8, which codes for a Wnt pathway receptor, mediates

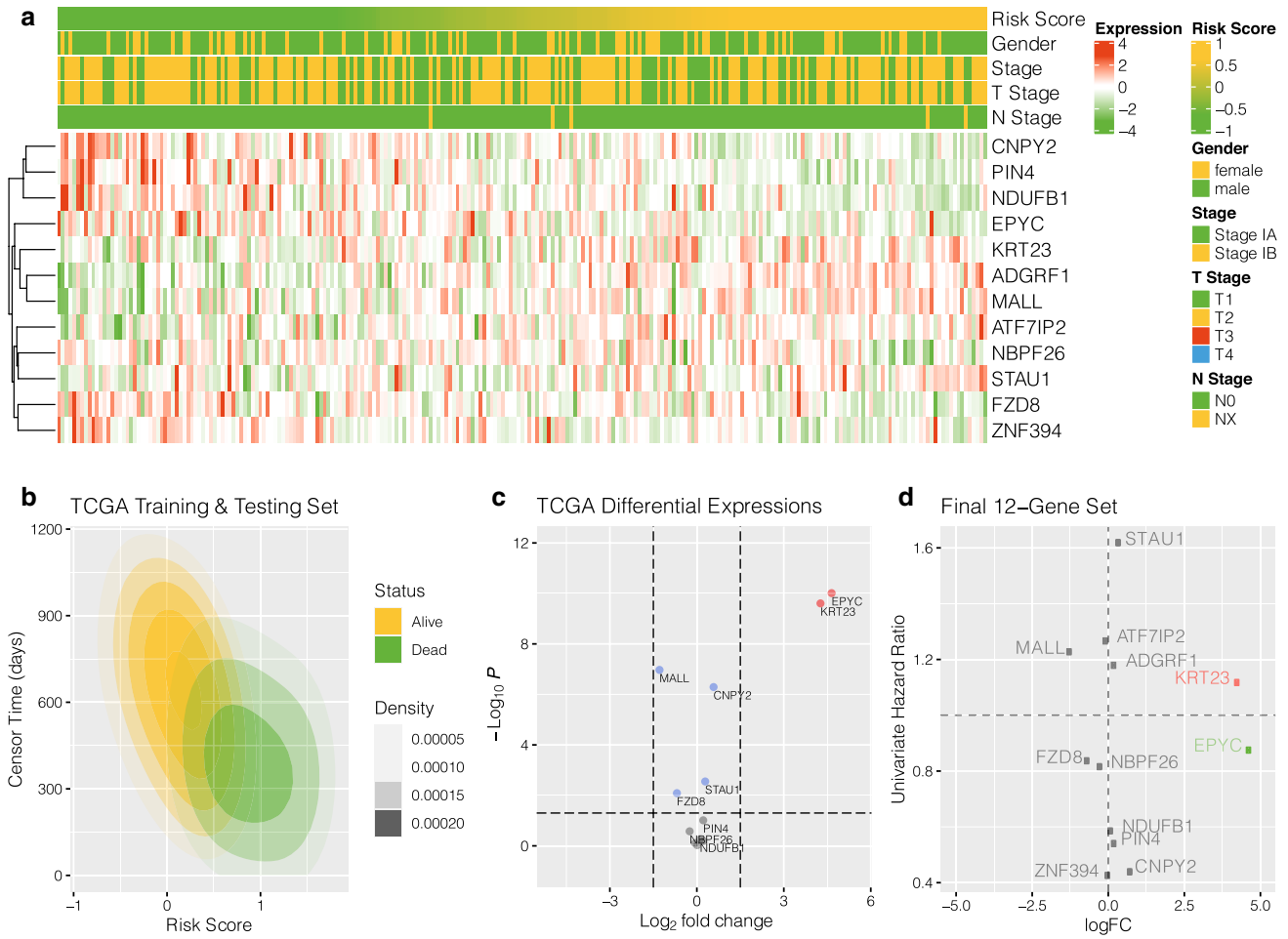

Fig. 6 Final model evaluation. **a** Heatmap of the 12 genes in the full TCGA dataset of stage I samples, with samples sorted from low to high-risk scores horizontally. **b** Density plot showing the cluster of deceased patients towards the right side of the graph with higher risk scores. **c** Volcano plot in the full TCGA dataset of stage I samples, with fold change denoting the differential expression in tumour samples compared to normal solid tissue. **d** Prognostic properties of the 12 genes in relation to whether they are differentially expressed in tumours. Colours of the gene names: grey = not significantly differentially expressed in tumours, red = negative prognostic factor up-regulated in tumours; green = positive prognostic factor up-regulated in tumour

the interaction of c-Met and Wnt/β-catenin signalling in head and neck squamous carcinomas [32] and is reported to mediate resistance to chemotherapy [40] and promote bone metastasis in prostate cancer [19]. Through targeting FZD8, miR-99b-5p was shown to inhibit NSCLC cell proliferation, migration and invasion [21]. Several lines of pharmaceutical inquiries leveraged such roles of FZD8. For example, a phase I clinical trial involved OMP-54F28, a recombinant fusion protein of the extracellular domain of the FZD8 receptor and a human IgG1 Fc fragment. OMP-54F28 binds Wnt ligands, blocking the signalling of the Wnt/FZD signalling pathway, a key oncologic pathway implicated in tumour cell de-differentiation and cancer stem cell (CSC) function [29]. Canopy homolog 2 (CNPY2) is an endoplasmic reticulum (ER) luminal protein. It contributes to tumour invasion and metastasis by activating the AKT/GSK3β pathway and modulating the EMT pathway [9]. In lung adenocarcinoma, MiR-30a-3p was demonstrated to down-regulate CNPY2, suppressing tumour cell proliferation and migration [35]. In human colorectal cancer, up-regulation of CNPY2 inhibited the activity of p53 and thus enhanced tumour growth and angiogenesis, and inhibited cell apoptosis [39]. Serum CNPY2 isoform 2 was also shown to be an effective biomarker for the early detection of CRCs [24]. We noted that the protective roles of CNPY2 and FZD8 revealed in this study seem to contradict the cited literature. We hypothesise that the inconsistencies observed could be attributed to the different underlying biology in the tumour models investigated in the literature. Further interrogation is warranted into the roles of these genes in LUSC.

KRT23 codes for a member protein of the keratin family. In one study, the inhibition of KRT23 by a KRT23-specific

siRNA repressed the endogenous hTERT protein and cell telomerase activity, and significantly inhibited tumour cell growth in vitro and in vivo. In another study of colorectal cancers, KRT23 promoted cancer stem cell properties and increased the expression of CD133 and CD44 [41]. Furthermore, KRT23 participates in EMT progression and interacts with p21 to mediate PI3K/AKT/GSK3β pathway in hepatocellular carcinoma (HCC) development [11]. Results from this study support the tumorigenic role of KRT23 in the LUSC model. MALL encodes for a protein of the MAL proteolipid family. Overexpression of MALL could suppress HCT116 and SW480 cell proliferation and inhibit HCT116 migration, and is demonstrated to reduce vessel invasion, disease recurrence and metastasis and death ($p=0.027$) [36]. STAU1 encodes for Stauden Double-Stranded RNA Binding Protein 1. A study showed that depletion of another protein coded by SNHG5 induces cell cycle arrest and apoptosis in vitro and limits tumour outgrowth in vivo, while depletion of STAU1 rescues the apoptosis induced after SNHG5 knockdown [7], demonstrating STAU1's oncogenic role. In summary, it is evident that the present algorithm succeeded in selecting plausible prognostically relevant genes that can be corroborated by previous research on tumourigenesis mechanisms.

The most important limitation of the present study is the lack of PCR validation on formalin-fixed and paraffin-embedded (FFPE) tissues to assess the clinical applicability of the 12-gene model. The current set of gene signatures were derived from an RNA-Seq platform using snap-frozen tissue samples which are not readily available in routine clinical practices. In addition, several other prognostic factors considered for early-stage LUSC in practice including surgical margin, vascular invasion and histological grade were unfortunately not available in the datasets. It was thus difficult to verify if the present model offers additional prognostic power over those factors. In future studies, the model would benefit from blinded, multi-centre validations to further assess its clinical utility. Further investigation into the individual prognostic genes identified in this study is also encouraged to explore their potential clinical utilities.

Conclusion

In summary, the paper constructed an independent 12-gene prognostic model using a novel K-Cox gene selection algorithm with comparable or even better performance than the top gene expression-based prognostic models for LUSC/NSCLC in the literature. In view of the current lack of clinical risk-stratification tools for stage I and especially stage IA lung squamous cell carcinomas, upon further validation, this genetic model could prove useful as a complementary tool on top of existing clinical prognostic factors to guide clinical decision making and recommend individualised treatment plans.

Author contributions (1) Conception and design: KW, YL and JL; (2) administrative support: JL and RC; (3) provision of study materials or patients: JL and RC; (4) collection and assembly of data: KW, YL and JW; (5) data analysis and interpretation: KW, YL and JW; (6) manuscript writing: all authors; (7) final approval of manuscript: all authors.

Funding This study was supported by the National Natural Science Foundation of China (81573178). The study was also supported by Jiangsu Key Laboratory of Preventive and Translational Medicine for Geriatric Diseases and the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

Data availability All data used in the study were downloaded from publicly available sources, see the methods section for their corresponding index numbers.

Declarations

Conflict of interest All authors have completed the ICMJE uniform disclosure form. The authors have no conflicts of interest to declare.

Ethical statement The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. There was no need for ethical approval as all data in this study were downloaded from public databases (TCGA), and the data processing met the TCGA publication guidelines (https://cancergenome.nih.gov/publications/guidelines).

Reporting checklist The TRIPOD reporting checklist was used for this study.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Aran D, Camarda R, Odegaard J, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun. 2017;8:1–14. https://doi.org/10.1038/s41467-017-01027-z.
2. Arriagada R, Bergman B, Dunant A, et al. Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small-cell lung cancer. N Engl J Med. 2004;350:351–60. https://doi.org/10.1056/NEJMoa031644.
3. Barta JA, Powell CA, Wisnivesky JP. Global Epidemiology of Lung Cancer. Annals of Global Health. 2019;85(1):8, 1–16. https://doi.org/10.5334/aogh.2419
4. Bhat RR, Yadav P, Sahay D, et al. GPCRs profiling and identification of GPR110 as a potential new target in HER2+ breast cancer. Breast Cancer Res Treat. 2018;170:279–92. https://doi.org/10.1007/s10549-018-4751-9.
5. Butts CA, Ding K, Seymour L, et al. Randomized phase III trial of vinorelbine plus cisplatin compared with observation in completely resected stage IB and II non-small-cell lung cancer: updated survival analysis of JBR-10. J Clin Oncol. 2010;28:29–34. https://doi.org/10.1200/JCO.2009.24.0333.

6. Byers LA, Diao L, Wang J, et al. An epithelial-mesenchymal transition gene signature predicts resistance to EGFR and PI3K inhibitors and identifies Axl as a therapeutic target for overcoming EGFR inhibitor resistance. Clin Cancer Res. 2013. https://doi.org/10.1158/1078-0432.CCR-12-1558.
7. Damas ND, Marcatti M, Côme C, et al. SNHG5 promotes colorectal cancer cell survival by counteracting STAU1-mediated mRNA destabilization. Nat Commun. 2016. https://doi.org/10.1038/ncomms13875.
8. Der SD, Sykes J, Pintilie M, et al. Validation of a histology-independent prognostic gene signature for early-stage, non-small-cell lung cancer including stage IA patients. J Thorac Oncol. 2014;9:59–64. https://doi.org/10.1097/JTO.0000000000000042.
9. Dou Y, Lei J-Q, Guo S-L, et al. The CNPY2 enhances epithelial-mesenchymal transition via activating the AKT/GSK3β pathway in non-small cell lung cancer. Cell Biol Int. 2018;42:959–64. https://doi.org/10.1002/cbin.10961.
10. Douillard JY, Rosell R, De Lena M, et al. Adjuvant vinorelbine plus cisplatin versus observation in patients with completely resected stage IB-IIIA non-small-cell lung cancer (Adjuvant Navelbine International Trialist Association [ANITA]): a randomised controlled trial. Lancet Oncol. 2006;7:719–27. https://doi.org/10.1016/S1470-2045(06)70804-X.
11. Guo D, Ma W, Wang R, et al. KRT23 acts as an oncogene in hepatocellular carcinoma by regulating P21 via PI3K/AKT/GSK3β pathway. Res Sq. 2020. https://doi.org/10.21203/rs.3.rs-34680/v1.
12. Heon S, Johnson BE. Adjuvant chemotherapy for surgically resected non-small cell lung cancer. J Thorac Cardiovasc Surg. 2012;144:S39–42.
13. Hirsch FR, Scagliotti GV, Mulshine JL, et al. Lung cancer: current therapies and new targeted treatments. Lancet. 2017;389:299–311.
14. Brierley JD, Gospodarowicz MK, Wittekind C, editors. TNM classification of malignant tumours. 8th ed. Wiley; 2016.
15. Jemal A, Torre L, Soerjomataram I, Bray F. Lung cancer. In: Jemal A, Torre L, Soerjomataram I, Bray F, editors. The cancer atals. 3rd ed. The American Cancer Society; 2019. p. 38.
16. Kinoshita T. Ideal prognostic model in lung squamous cell carcinoma. Transl Lung Cancer Res. 2020;9(3):430–431.
17. Kratz JR, He J, Van Den Eeden SK, et al. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. Lancet. 2012;379:823–32. https://doi.org/10.1016/S0140-6736(11)61941-7.
18. Kratz JR, Van Den Eeden SK, He J, et al. A prognostic assay to identify patients at high risk of mortality despite small, node-negative lung tumors. J Am Med Assoc. 2012;308:1629–31.
19. Li Q, Ye L, Zhang X, et al. FZD8, a target of p53, promotes bone metastasis in prostate cancer by activating canonical Wnt/β-catenin signaling. Cancer Lett. 2017;402:166–76. https://doi.org/10.1016/j.canlet.2017.05.029.
20. Bin LS, Tan SJ, Lim WT, Lim CT. A merged lung cancer transcriptome dataset for clinical predictive modeling. Sci Data. 2018;5:1–8. https://doi.org/10.1038/sdata.2018.136.
21. Liu R, Chen Y, Shou T, et al. miRNA-99b-5p targets FZD8 to inhibit non-small cell lung cancer proliferation, migration and invasion. Onco Targets Ther. 2019;12:2615–21. https://doi.org/10.2147/OTT.S199196.
22. Okayama H, Kohno T, Ishii Y, et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res. 2012. https://doi.org/10.1158/0008-5472.CAN-11-1403.
23. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clin Cancer Res. 2004;10:2922–7. https://doi.org/10.1158/1078-0432.CCR-03-0490.
24. Peng J, Ou Q, Pan Z, et al. Serum CNPY2 isoform 2 represents a novel biomarker for early detection of colorectal cancer. Aging (Albany NY). 2018;10:1921–31. https://doi.org/10.18632/aging.101512.
25. Pignon JP, Tribodet H, Scagliotti GV, et al. Lung adjuvant cisplatin evaluation: a pooled analysis by the LACE collaborative group. J Clin Oncol. 2008;26:3552–9. https://doi.org/10.1200/JCO.2007.13.9030.
26. Postmus PE, Kerr KM, Oudkerk M, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO clinical practice guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2017;28:iv1–21. https://doi.org/10.1093/annonc/mdx222.
27. Raponi M, Zhang Y, Yu J, et al. Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. Cancer Res. 2006;66:7466–72. https://doi.org/10.1158/0008-5472.CAN-06-1191.
28. Slaughter DP, Southwick HW, Smejkal W. "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. Cancer. 1953;6:963–8. https://doi.org/10.1002/1097-0142(195309)6:5%3c963::AID-CNCR2820060515%3e3.0.CO;2-Q.
29. Smith DC, Gordon M, Messersmith W, et al. Abstract B79: a first-in-human Phase 1 study of anti-cancer stem cell (CSC) agent OMP-54F28 (FZD8-Fc) targeting the WNT pathway in patients with advanced solid tumors. Mol Cancer Ther. 2013;12:B79. https://doi.org/10.1158/1535-7163.TARG-13-B79.
30. Strauss GM, Herndon JE, Maddaus MA, et al. Adjuvant paclitaxel plus carboplatin compared with observation in stage IB non-small-cell lung cancer: CALGB 9633 with the cancer and leukemia group B, radiation therapy oncology group, and North Central cancer treatment group study groups. J Clin Oncol. 2008;26:5043–51. https://doi.org/10.1200/JCO.2008.16.4855.
31. Subramanian J, Simon R. Gene expression-based prognostic signatures in lung cancer: ready for clinical use? J Natl Cancer Inst. 2010;102:464–74.
32. Sun S, Liu S, Duan SZ, et al. Targeting the c-Met/FZD8 signaling axis eliminates patient-derived cancer stem-like cells in head and neck squamous carcinomas. Cancer Res. 2014;74:7546–59. https://doi.org/10.1158/0008-5472.CAN-14-0826.
33. Tang H, Wang S, Xiao G, et al. Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. Ann Oncol. 2017;28:733–40.
34. Tang H, Xiao G, Behrens C, et al. A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. Clin Cancer Res. 2013;19:1577–86. https://doi.org/10.1158/1078-0432.CCR-12-2321.
35. Wang H, Kanmangne D, Li R, et al. miR-30a-3p suppresses the proliferation and migration of lung adenocarcinoma cells by downregulating CNPY2. Oncol Rep. 2019;43:646–54. https://doi.org/10.3892/or.2019.7423.
36. Wang X, Fan J, Yu F, et al. Decreased MALL expression negatively impacts colorectal cancer patient survival. Oncotarget. 2016;7:22911–27. https://doi.org/10.18632/oncotarget.8094.
37. Woodard GA, Jones KD, Jablons DM. Lung cancer staging and prognosis. In: Cancer treatment and research. Kluwer Academic Publishers; 2016. p. 47–75.
38. Xu H, Ma J, Wu J, et al. Gene expression profiling analysis of lung adenocarcinoma. Braz J Med Biol Res. 2016;49:1–11. https://doi.org/10.1590/1414-431X20154861.
39. Yan P, Gong H, Zhai X, et al. Decreasing CNPY2 expression diminishes colorectal tumor growth and development through activation of p53 pathway. Am J Pathol. 2016;186:1015–24. https://doi.org/10.1016/j.ajpath.2015.11.012.
40. Yin S, Xu L, Bonfil RD, et al. Tumor-initiating cells and FZD8 play a major role in drug resistance in triple-negative breast

cancer. Mol Cancer Ther. 2013;12:491–8. https://doi.org/10.1158/1535-7163.MCT-12-1090.
41. Zhang N, Zhang R, Zou K, et al. Keratin 23 promotes telomerase reverse transcriptase expression and human colorectal cancer growth. Cell Death Dis. 2017;8:e2961–e2961. https://doi.org/10.1038/cddis.2017.339.
42. Zhu C-Q, Ding K, Strumpf D, et al. Prognostic and predictive gene signature for adjuvant chemotherapy in resected non-small-cell lung cancer. J Clin Oncol. 2010;28:4417–24. https://doi.org/10.1200/JCO.2009.26.4325.
43. Zhu CQ, Tsao MS. Prognostic markers in lung cancer: is it ready for prime time? Transl Lung Cancer Res. 2014;3:149–58.