Journal of Cancer Research and Clinical Oncology (2020) 146:2743–2752
https://doi.org/10.1007/s00432-020-03318-3

ORIGINAL ARTICLE – CANCER RESEARCH

# Identification of prognostic biomarkers for major subtypes of non-small-cell lung cancer using genomic and clinical data

**Anjali Lathwal[1] · Rajesh Kumar[2] · Chakit Arora[1] · Gajendra Pal Singh Raghava[1]**

Received: 11 March 2020 / Accepted: 8 July 2020 / Published online: 14 July 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

## Abstract

**Purpose** Intra-tumor heterogeneity and high mortality among patients with non-small-cell lung carcinoma (NSCLC) emphasize the need to identify reliable prognostic markers unique to each subtype.

**Methods** In this study, univariate cox regression and prognostic index (PI)-based approaches were used to develop models for predicting NSCLC patients' subtype-specific survival.

**Results** Prognostic analysis of TCGA dataset identified 1334 and 2129 survival-specific genes for LUSC (488 samples) and LUAD (497 samples), respectively. Individually, 32 and 271 prognostic genes were found and validated in GSE study exclusively for LUSC and LUAD. Nearly, 9–10% of the validated genes in each subtype were already reported in multiple studies thus highlighting their importance as prognostic biomarkers. Strong literature evidence against these prognostic genes like "*ELANE*" (LUSC) and "*AHSG*" (LUAD) instigates further investigation for their therapeutic and diagnostic roles in the corresponding cohorts. Prognostic models built on five and four genes were validated for LUSC [HR = 2.10, $p$ value = $1.86 \times 10^{-5}$] and LUAD [HR = 2.70, $p$ value = $3.31 \times 10^{-7}$], respectively. The model based on the combination of age and tumor stage performed well in both NSCLC subtypes, suggesting that despite having distinctive histological features and treatment paradigms, some clinical features can be good prognostic predictors in both.

**Conclusion** This study advocates that investigating the survival-specific biomarkers restricted to respective cohorts can advance subtype-specific prognosis, diagnosis, and treatment for NSCLC patients. Prognostic models and markers described for each subtype may provide insight into the heterogeneity of disease etiology and help in the development of new therapeutic approaches for the treatment of NSCLC patients.

**Keywords** NSCLC · Survival analysis · Prognostic biomarker · Cox univariate regression · Subtype-specific

## Introduction

Disruption in the signaling system that governs cell fate and development is the major initiation factor that contributes to tumorigenesis (Frost and Amos 2018). Lung cancer is the leading cause of cancer-related death worldwide (Siegel et al. 2018). It originates from transformed epithelial cells that form a heterogeneous undifferentiated malignant neoplasm (Molina et al. 2008). Its two major subtypes include small cell lung carcinoma (SCLC) and non-small cell lung carcinoma (NSCLC). NSCLC accounts for 85–90% of lung cancer cases. NSCLC's two major subtypes include squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD) (Chang et al. 2015).

The heterogeneous nature of the disease and lack of effective therapeutic options further contribute to poor prognosis and a low 5-year survival rate in NSCLC patients (Stewart

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s00432-020-03318-3) contains supplementary material, which is available to authorized users.

✉ Gajendra Pal Singh Raghava
raghava@iiitd.ac.in
http://webs.iiitd.edu.in/raghava/

[1] Department of Computational Biology, Indraprastha Institute of Information Technology-Delhi, Okhla Industrial Estate, Phase III (Near Govind Puri Metro Station), A-302 (R&D Block), New Delhi 110020, India

[2] Bioinformatics Centre, CSIR-Institute of Microbial Technology, Chandigarh, India

et al. 2004). However, the survival rate can be increased to 50% if patients diagnosed at an early stage (Matsuda et al. 2012). Currently, the available diagnosis method includes X-ray, CT scan, bronchoscopy, sputum cytology, and tissue histology. These traditional techniques have some limitations such as low sensitivity of X-rays, invasive nature of bronchoscopy, sputum cytology works only when cancer arises in the mid airway, and lack of availability of tissues for histological examination (Midthun 2011; Sheng et al. 2019).

The tumor and normal dataset generated by The Cancer Genome Atlas (TCGA) for both NSCLC subtypes has served the basis of numerous studies leading to the identification of diagnostic, prognostic, and novel therapeutic markers. Several genomic signatures have been identified for NSCLC subtypes, but these are still of limited use in terms of significance, biological relevance, and clinical application. The major factor contributing to this is the heterogeneity of the disease. Although several studies are in the literature which identify the gene signature from the meta-analysis of different cohorts available for lung cancer, but they are also suffering from non-reproducibility among cohorts (Tseng et al. 2012). One possible explanation could be that different gene signature is merely the separate aspect of the same molecular pathways/mechanism that leads to disease (Chen et al. 2010). Multiple pan-cancer studies suggest that the molecular mechanisms of carcinogenesis exhibit a high level of heterogeneity between LUAD and LUSC (Cancer Genome Atlas Research Network 2014; Chang et al. 2015; TCGA 2012), and hence, they should have separate therapeutic strategies. Drugs such as Bevacizumab (Avastin) can be used for LUSC patients, but have not tested so for LUAD (Sandler et al. 2006). Also, a few therapies proposed for LUAD often seem to be ineffective for LUSC patients (Rekhtman et al. 2012). Thus, it is important to contradistinguish between the two subtypes in a more specific way to design an effective therapeutic approach.

In this study, we have performed an integrative analysis by screening for survival-associated genes from TCGA dataset based on the median value of gene expression for lung cancer subtypes. Random forest variable hunting (RF-vh) was used for feature selection, and using these features as input, we have developed univariate cox regression-based prognostic models that can classify the patients in low- and high-risk groups within each subtype. Cancer-specific pathways and clinical information were further incorporated to identify a better risk assessment model for NSCLC patients. The significance of the study can be seen in terms that it not only captures the tumor heterogeneity by providing NSCLC subtype-specific survival-associated genes, but also prioritizes significant biological pathways when it comes to risk stratification among NSCLC patients. The identified candidate genes could serve as a potential discriminatory marker/target among the subtypes and, thus, may help clinicians and researchers in proposing better subtype-specific therapeutics for the treatment of NSCLC patients.

## Methods

### Data acquisition and pre-processing

In our study, we have gathered RNA_seq expression data from TCGA for LUSC (488 samples) and LUAD (497 samples) using TCGA assembler (Zhu et al. 2014). Only samples for which clinical information was available were considered here. We started with a few clinical features and 20,530 genes for each subtype. As done in some previous studies (Wang et al. 2018), we first selected the genes having an expression in more than 50% of the samples which lead to 17,982, and 17,756 gene expression profiles for LUSC and LUAD cohorts, respectively. Quantile normalization was applied on these cohorts using R code to change the genomic expressions to a common scale and to remove the technical variation caused by noisy data (He et al. 2019; Mandelboum et al. 2019). We downloaded the GEO dataset using the accession number GSE42127 having 176 patients samples and 19,141 genes for the validation of our results. There were 133 and 43 patient samples of LUAD and LUSC in GEO dataset, respectively. There were total 16,295 common genes in GEO and TCGA LUSC datasets, whereas the number of common genes was 16,104 for LUAD. Following the standard protocol (Shi and Xu 2019), we processed and normalized the GEO dataset using the target matrix of the TCGA dataset. We mapped the Probe IDs to their corresponding gene symbols using the mapping scheme provided in the GEO database.

### Prognosis-associated genomic and clinical feature selection

To screen the genes associated with the overall survival (OS) in TCGA dataset, we performed cox proportional hazard regression on the two cohorts, i.e., LUSC and LUAD using the "Survival" package (V.2.42-6) in R (V.3.4.4, The R Foundation). Genes significantly related to the OS ($p$ value $< 0.05$) were selected. The identified prognostic genes for each subtype were then classified as bad prognostic markers (BPM), genes that were correlated with poor OS of the patient and good prognostic markers (GPM), genes that were correlated with better OS of the patient. For each dataset, the resulting genes were grouped into three sets as BPM, GPM, and combined (GPM + BPM) set. Similar to the previous studies done in the literature (Yunhe Liu et al. 2019a, b; Zhang et al. 2018), we have used the RF-vh with 100 iterations available in the "randomForestSRC" package of R to refine the gene pool. We iteratively scrutinized the

minimum possible gene set that can be used to build the predictive models so as to further enhance the scalability of the gene signature in the clinical setup. We also considered ten signaling pathways—cell cycle, HIPPO, MYC, NOTCH, NRF2, P53, PI3K-AKT, RAS, TGF-Beta, and WNT along with the apoptotic pathway which are likely to be cancer drivers (functional contributors) or therapeutic targets (Sanchez-Vega et al. 2018). The pathway genes that were significantly associated with OS were selected for further studies. Six clinical features in TCGA dataset such as age, gender, N staging, T staging, organ subdivision, and tumor stage were selected to evaluate their relevance with the OS of the patients. We have validated our final predictive models on the GEO dataset.

### Development of models

*Univariate cox regression* was implemented on the median cut-off value (Lathwal et al. 2019) of the identified prognostic gene expression using the 'survival' package (V.2.42-6) in R (V.3.4.4, The R Foundation). We have also evaluated the hazard ratio (HR) to indicate the survival probability for each group. We had the cox regression coefficient for each prognostic gene which represents the level change in the OS based on the unit change in the gene expression, keeping other factors constant. Several statistical tests such as log-rank test and Wald test were carried out for the assessment of the model. The Kaplan–Meier plots were also generated for the best models to compare the survival curves of high- and low-risk groups.

*Prognostic index (PI)*, similar to previous studies (Li et al. 2018; Wang et al. 2018), was formulated on $\boldsymbol{n}$ number of genes as follows:

$$\mathrm{PI} = \boldsymbol{\alpha}_1 Y_1 + \boldsymbol{\alpha}_2 Y_2 + \cdots + \boldsymbol{\alpha}_n Y_n,$$

where $\alpha$ is the regression coefficient for a gene $\boldsymbol{Y}$, obtained using a univariate cox-ph model.

For each set of subtypes, PI at median cut-offs was used to classify patients in high- and low-risk groups. PI model was built on the resultant gene sets that were obtained after the feature selection steps mentioned above. The patient samples with PI greater than median (PI) come under high-risk group and patients with PI less than equal to median (PI) were classified under the low-risk group. The corresponding statistical metrics such as HR, $p$ value, concordance, and standard error were also obtained.

### Evaluation metrics

Statistical analysis of the survival models was done using various metrics such as HR, log-rank, Wald test, and concordance index. HR predicts the death risk associated with the two groups. Log-rank test was used to explain the statistical significance of the survival curves of the two groups. The Wald test estimated the significance of explanatory variables used to calculate HR. The concordance index in terms of prediction accuracy (PA) measures the model's distinguishability between the high- and low-risk groups. Lower log-rank, $p$ value ($<0.05$), and a higher concordance value ($>0.5$) (Chaudhary et al. 2018; Dyrskjøt et al. 2017) imply a better prognostic model. Using the standard approach (Deng et al. 2020; Li et al. 2017), we validated our models (gene signature and clinical) and prognostic gene expressions onto the GEO dataset. We have given the complete workflow of the study in Fig. 1.

## Results

Survival analysis was performed on each TCGA dataset (LUSC and LUAD) to screen the prognostic genes. Selecting the median expression value of each gene as the cut-off for high- and low-risk groups, the cox univariate regression was used to extract 1334 and 2129 genes having some prognostic potential in LUSC and LUAD datasets, respectively. HR, p value, regression coefficients, concordance index, and standard error were calculated for these genes in each dataset and are provided in Supplementary S1 Tables 1 and 2. Genes were categorized among BPM and GPM sets for both subtypes using the criterion explained in the Methods section. RF-vh was used for selection of the gene panel to build a robust prognostic model for the prediction of the survival of patients in LUSC and LUAD. We obtained 26 genes from 637 BPM genes, 24 genes from 697 GPM genes, and 44 genes from combined set of 1334 genes for LUSC after applying RF-vh. Similarly, 41 genes from 1153 BPM genes, 34 genes from 976 GPM genes, and 89 genes from combined set of 2129 genes for LUAD were obtained. We have also provided the list of these genes in Supplementary S1 Table 3. We have generated the results using different machine learning (ML) models and PI-based models, but our PI models outperformed in terms of HR and PA, like similar studies done in past (Lathwal et al. 2019). However, results of our ML models are not shown here due to being less effective for this study.

### Prognostic index-based survival model

Prognostic index was used to build the different predictive models. PI-based models use the gene expressions and their corresponding regression coefficients obtained from univariate cox proportional hazard for input as mentioned in the Methods section.

PI was estimated using the same set of genes as mentioned above and the results corresponding to each subtype

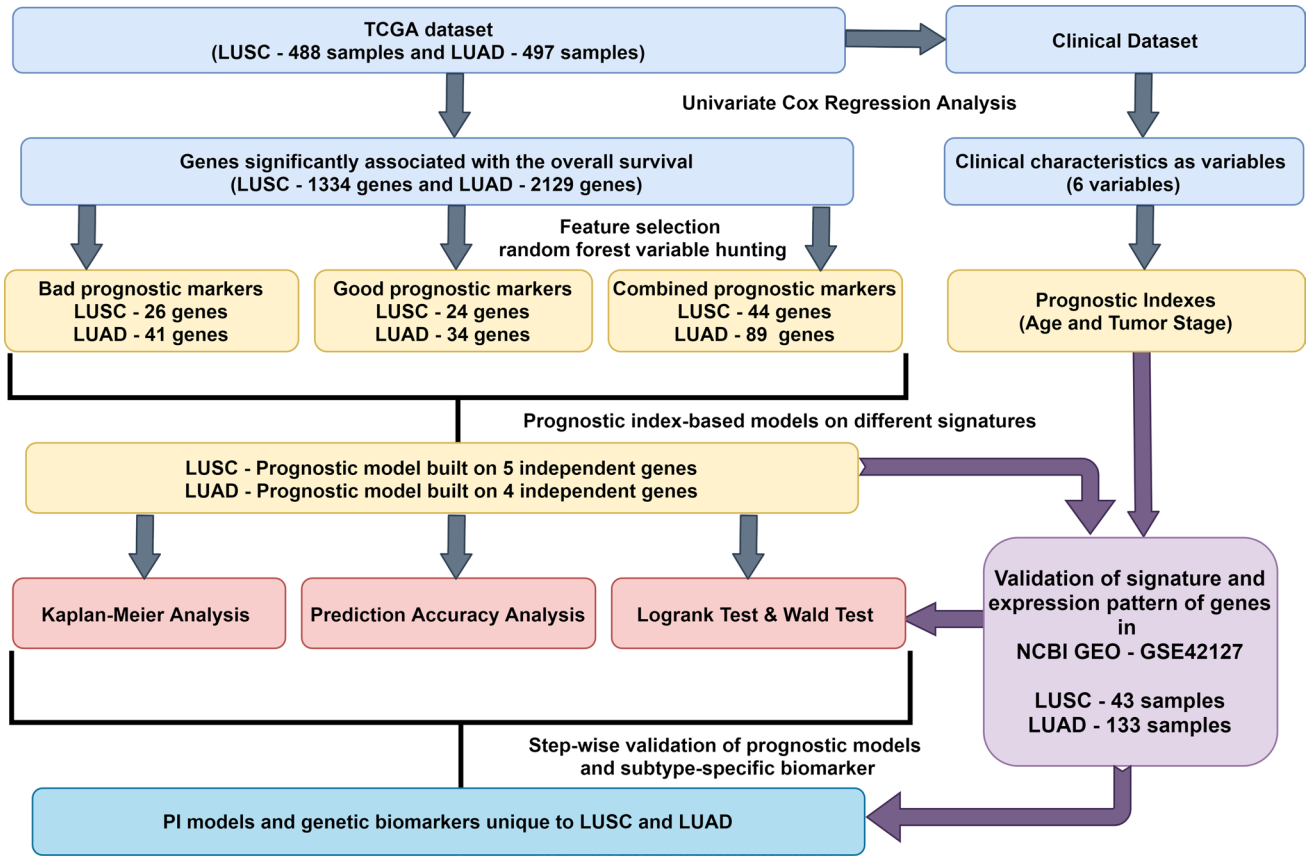


**Fig. 1** Workflow of the study

**Table 1** Statistics obtained for each dataset based on prognostic index model

| | HR | *p* value | PA (%) | Std. error |
|---|---|---|---|---|
| LUSC | | | | |
| BPM ($n=26$) | 2.21 | $1.83\times10^{-6}$ | 61 | 0.0225 |
| GPM ($n=24$) | 2.20 | $2.23\times10^{-6}$ | 59 | 0.0237 |
| Combined ($n=44$) | 2.04 | $1.50\times10^{-5}$ | 60 | 0.0226 |
| LUAD | | | | |
| BPM ($n=41$) | 1.85 | $1.08\times10^{-3}$ | 60 | 0.0273 |
| GPM ($n=34$) | 1.70 | $3.78\times10^{-3}$ | 58 | 0.0270 |
| Combined ($n=89$) | 2.10 | $9.58\times10^{-5}$ | 63 | 0.0260 |

*n* denotes the number of genes

are shown in Table 1. The patients with PI less than equal to median (PI) were grouped in low risk and patients with PI greater than median (PI) were grouped in high risk. Using the gene sets of best models for LUSC (BPM) and LUAD (Combined) cohorts, we iteratively searched for a minimal subset of genes that can discriminate the patients among high- and low-risk groups in the respective cohorts. We started with the set of two genes and stopped the incremental iteration whenever the gene set performed comparable to the existing model. This way, we came up with the gene signature based on five (*KIF16B, KLK7, LONRF3, OPLAH,* and *RIPK3*) and four genes (*AHSG, DKK1, MGAT5B,* and *NEMP2*) for LUSC [HR = 2.10 and $p$ value = $1.86\times10^{-5}$] and LUAD [HR = 2.70 and $p$ value = $3.31\times10^{-7}$], respectively. We have given the Kaplan–Meier plots for the same in the Fig. 2. We have also tested the performance of the models in the counter cohorts and found that model for LUSC does not work for LUAD and vice-versa. These results confirmed that prognostic models will work dedicatedly in their respective cohorts only.

## Universal NSCLC model poorly discriminates among subtypes

This section focuses on investigating whether a universal prognostic model exists for the NSCLC cohort. We found that there were no common prognostic genes among each set of subtypes obtained after applying RF-vh. Therefore, to build a prognostic model that can be used universally across the NSCLC patients, we extracted 1381 prognostic genes for the NSCLC complete cohort using univariate cox regression,

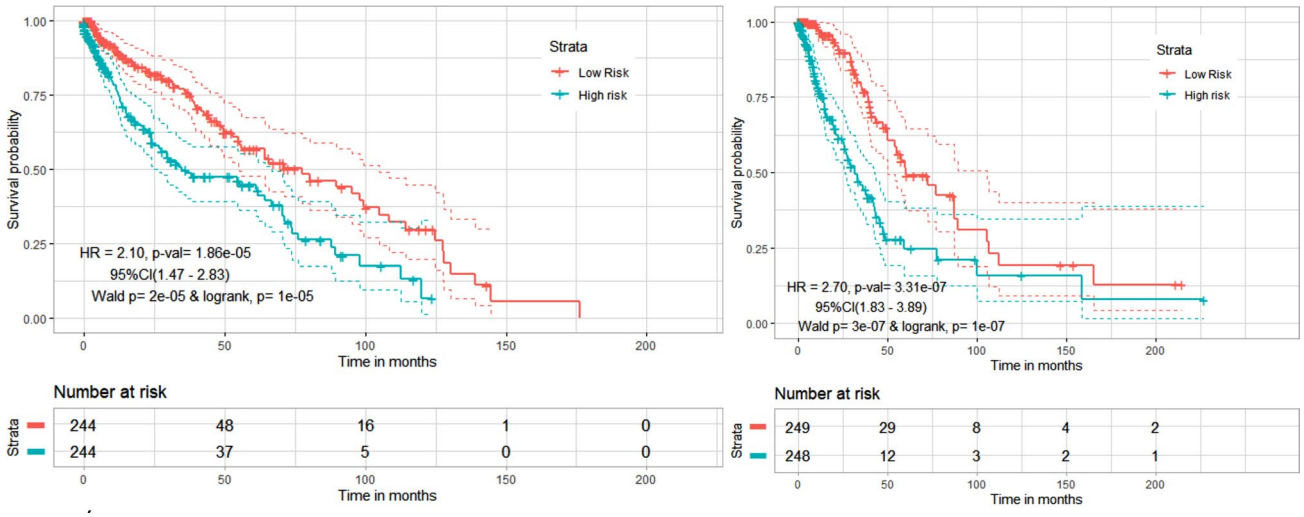


**Fig. 2** Kaplan–Meier plots for the top models used in prognostication of NSCLC patients in TCGA dataset. **a** In LUSC subtype, patients having PI > median (PI) are at high risk as compared to the patients with PI ≤ median (PI) with HR = 2.10 ($p$ value = $1.86 \times 10^{-5}$) and PA = 61%. **b** In LUAD subtype, patients having PI > median (PI) are at high risk than the patients with PI ≤ median (PI) with HR = 2.70 ($p$ value = $3.31 \times 10^{-7}$) and PA = 68%

**Table 2** Statistics obtained for each dataset based on PI- and ML-based model for NSCLC

| Cohorts | HR | $p$ value | PA (%) | Std. error |
|---|---|---|---|---|
| LUSC | 1.73 | $7.46 \times 10^{-4}$ | 58 | 0.0240 |
| LUAD | 1.82 | $1.28 \times 10^{-3}$ | 58 | 0.0270 |
| NSCLC Cohort | 1.79 | $1.87 \times 10^{-6}$ | 58 | 0.0180 |

out of which 682 were BPM and 699 were GPM genes. We found 68 overlapping prognosis-related genes across subtypes and the NSCLC complete cohort which formed the basis of our universal models.

All the information on these gene sets can be found in Supplementary S1 Table 4. We have summarized the results along with all the statistical details of the model such as HR, $p$ value, PA, and the standard error in Table 2. When compared to previous subtype-specific prognostic models, we found that this model poorly discriminates among the subtypes in terms of HR and PA.

## Survival analysis of cancer-specific pathway genes

To further comprehend the differences between the LUSC and LUAD, we extracted the gene set of 11 cancer-specific pathways as explained in Methods section. Prognostic potential of the pathway genes was calculated by applying cox univariate regression and all the potential genes found in each pathway are listed in Supplementary S1 Table 5. These genes were then used in developing the different prognostic models. We built models using a similar approach as explained in the above sections. The goal was to identify the important biological markers for risk stratification in NSCLC subtypes which can improve the PA with better HR and provide us with few insights on heterogeneity at the pathway level. All the details of the models corresponding to each pathway can be found in Supplementary S1 Table 6. We observed that none of the models in each subtype outperformed the previous models in terms of HR and PA. Also, the models based on the apoptotic pathway genes have shown the good performance with HR greater than 2.0, in both LUSC and LUAD. These results further strengthened our hypothesis that NSCLC patients should be considered in a more subtype-specific way while building reliable diagnostic and prognostic strategies.

## Risk estimation using clinical features

To investigate the correlation between different clinical features and the survival of NSCLC patients, cox univariate regression model was implemented. The data were transformed to binary using the criteria mentioned in the strata column of Table 3. We have found that none of the clinical features are of much importance in the case of LUSC patients except tumor stage.

However, in LUAD samples, tumor and N stage give the performance comparable to the best prognostic model so far. The result statistics of the models for LUSC and LUAD using all relevant clinicopathological features is shown in Table 3.

**Table 3** Statistics of clinical features-based risk stratification model for NSCLC subtypes

| | Strata | HR | *p* value | PA (%) | Std error |
|---|---|---|---|---|---|
| LUSC (samples) | | | | | |
| Age (488) | < 65 vs ≥ 65 | 1.28 | $1.70\times10^{-1}$ | 52 | 0.0220 |
| Gender (488) | Female vs male | 1.06 | $7.70\times10^{-1}$ | 51 | 0.0220 |
| N staging (482) | N0 vs N1, N2, N3 | 1.32 | $1.00\times10^{-1}$ | 51 | 0.0230 |
| T staging (488) | T1, T2 vs T3, T4 | 1.48 | $5.04\times10^{-2}$ | 53 | 0.0200 |
| Tumor stage (484) | I, II vs III, IV | 1.48 | $3.50\times10^{-2}$ | 53 | 0.0200 |
| Organ subdivision (459) | Left vs right | 1.51 | $6.14\times10^{-1}$ | 50 | 0.0250 |
| LUAD (samples) | | | | | |
| Age (497) | < 65 vs ≥ 65 | 1.32 | $1.38\times10^{-1}$ | 54 | 0.0260 |
| Gender (497) | Female vs male | 0.94 | $7.29\times10^{-1}$ | 49 | 0.0280 |
| N staging (485) | N0 vs N1, N2, N3 | 2.62 | $1.97\times10^{-7}$ | 63 | 0.0270 |
| T staging (494) | T1, T2 vs T3, T4 | 2.39 | $2.69\times10^{-4}$ | 57 | 0.0240 |
| Tumor stage (489) | I, II vs III, IV | 2.80 | $4.84\times10^{-8}$ | 64 | 0.0270 |
| Organ subdivision (483) | Left vs right | 1.08 | $6.79\times10^{-1}$ | 51 | 0.0280 |

**Table 4** Statistics obtained for PI-based model in TCGA and GEO datasets

| | HR | *p* value | PA (%) | Std. error |
|---|---|---|---|---|
| LUSC ($n=5$) | | | | |
| TCGA | 2.10 | $1.86\times10^{-5}$ | 61 | 0.0215 |
| GEO | 2.53 | $4.00\times10^{-2}$ | 60 | 0.0589 |
| LUAD ($n=4$) | | | | |
| TCGA | 2.70 | $3.31\times10^{-7}$ | 68 | 0.0262 |
| GEO | 2.50 | $4.00\times10^{-3}$ | 63 | 0.0339 |

*n* Denotes the number of genes

## External validation of the prognostic models

To evaluate the robustness of the final prognostic gene signatures and clinical feature-based prognostic models, we downloaded the GEO dataset with accession number GSE42127. The dataset has been transformed and normalized using the target matrix of the TCGA dataset for both LUSC and LUAD. We performed the cox univariate regression analysis to authenticate the predictive power of our gene signatures and the results of which can be found in the Table 4. Clearly, the final models for LUSC and LUAD also performed well in our validation GEO cohorts.

We have generated the Kaplan–Meier plots for the visual representation of comparison between the high- and low-risk groups among the LUSC and LUAD cohorts of the validation dataset in Fig. 3. We can see that our proposed gene signature also performed well in the validation sets with fairly good predictive power.

We have also tried validating the robustness of our clinical feature-based prognostic models for NSCLC subtypes. However, there were only three clinical features (gender, age, and tumor Stage) available for the GEO patient's samples. We built and validated the clinical model using these limited features in the similar manner as done in Table 4. In validation set, none of the independent clinical features were found to be significantly associated with the overall survival of the patients when univariate cox regression analysis was performed. However, in both TCGA and GEO studies, age and tumor stage are the two features that showed some prognostic potential.

## Combination of age and tumor stage works best for risk stratification in both NSCLC subtypes

We tried to build a prognostic model using the combination of clinical features to come up with a better and simplistic model for risk assessment among NSCLC patients. Entries corresponding to every clinical feature under consideration are categorized as 1, −1, and 0 for high risk, low risk, and unavailable, respectively. This categorization was done to handle the missing clinical data and, thus, ensured fixed-length vectors of patient samples. Different linear combinations of two or more features were computed to obtain the best results.

The sum of different linear combinations was termed as "SUM" which was used as the cut-off to stratify patients in high- and low-risk groups. Patients with SUM ≤ median (SUM) were categorized as high risk and patients with SUM > median (SUM) were listed as low risk groups. We evaluated the significance of age and tumor stage in predicting the risk for LUSC and LUAD patients in both testing and validation sets. The results corresponding to each subtype using a combination of clinical features for TCGA and GEO datasets can be found in Table 5. The results signifies that age and tumor stage can be good prognostic predictors when used in combination for both LUSC and LUAD subtypes,

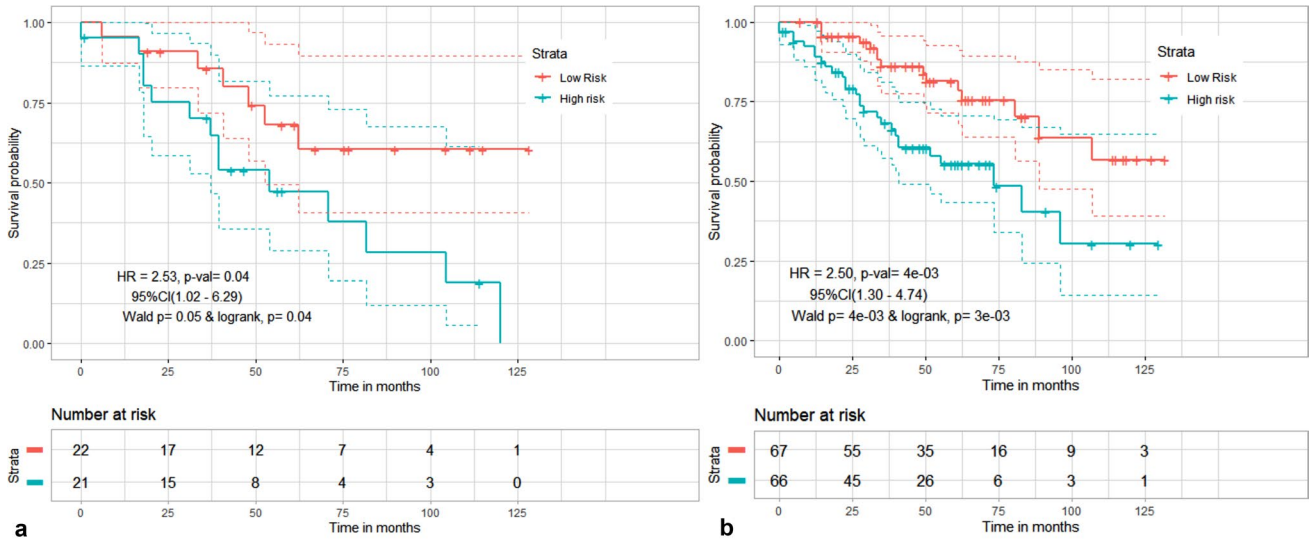


**Fig. 3** Kaplan–Meier plots for the top models used in prognostication of NSCLC patients in GEO dataset. **a** In LUSC subtype, patients having PI > median (PI) are at high risk as compared to the patients with PI ≤ median (PI) with HR = 2.53 ($p$ value = $4.00 \times 10^{-2}$) and PA = 60%. **b** In LUAD subtype, patients having PI > median (PI) are at high risk than the patients with PI ≤ median (PI) with HR = 2.50 ($p$ value = $4.00 \times 10^{-3}$) and PA = 63%

**Table 5** Statistics of combinatorial clinical feature-based prognostic model for NSCLC subtypes

| | HR | $p$ value | PA (%) | Std error |
|---|---|---|---|---|
| LUSC (clinical features—age, tumor stage) | | | | |
| TCGA | 1.74 | $1.06 \times 10^{-2}$ | 53 | 0.0173 |
| GEO | 3.90 | $6.29 \times 10^{-3}$ | 60 | 0.0476 |
| LUAD (clinical features—age, tumor stage) | | | | |
| TCGA | 2.72 | $2.85 \times 10^{-6}$ | 61 | 0.0265 |
| GEO | 2.40 | $1.53 \times 10^{-2}$ | 56 | 0.0335 |

despite having very distinctive histological features, oncogenic origins, and treatment paradigm.

## Prognostic biomarker unique to LUSC and LUAD

In the above sections, we have built and validated the various prognostic models based on gene signature and clinical features for NSCLC subtypes. We have identified 32 and 271 prognostic biomarkers specific to LUSC and LUAD subtype, respectively. These biomarkers were also found to be significantly associated with survival of the patients in the validation dataset and list of all such genes can be found in Supplementary S1 Table 7. We have investigated the relevance of these biomarkers in the other published studies and found that around 9–10% of these prognostic biomarkers were already reported in the literature (Cheng et al. 2020; Liu et al. 2019a, b; Lv and Lei 2020; Ma et al. 2020a, b; Relli et al. 2018). This indicates that our marker genes are of high importance having in more than one study and needs further investigation for their therapeutic potential in a more subtype-specific manner. Interestingly, literature evidence suggests that one of our identified prognostic gene, *UIMC1* is involved in DNA damage repair mechanism and is in clinical trial (NCT00883480) exclusively for LUSC treatment (Yan et al. 2007). However, we could not verify this gene from our GEO study. Another study showed the relevance of *ELANE* gene specific to advancement of LUSC, not LUAD (Yang et al. 2017). Also, *AHSG* gene, i.e., used in prognostic model, has already been specifically targeted in a clinical trial having patent number (CA2847188A1) for LUAD patients. Using DisGeNET (Piñero et al. 2020), we were able to select 12 prognostic genes (*ALOX5, HELLS, KPNA2, NRAS, RRM2, SFTPC, TK1, TOP2A, TPI1, TYMS, BZW2,* and *SFXN1*) that are particularly playing role in the progression and development of LUAD. All these genes are experimentally validated in several studies and have also shown great prognostic potential in both TCGA as well as GEO datasets for LUAD patients. These evidences are strongly supporting the credibility of our results. We suggest that these genes may be further exploited for their therapeutic, prognostic, and diagnostic potential, specifically in LUAD.

## Therapeutic potential of identified prognostic genes specific to subtype

There are several prospective strategies such as gene insertion and silencing, immune modulation, and targeting gene expression using inhibitors/activators for the treatment of cancer. In our study, each marker gene stratifies patients in high- and low-risk groups with significant difference

among the gene expression levels of two cohorts. For example, *APH1A* gene is greatly up regulated in the high-risk cohort as compared to low-risk cohort. Thus, strong inhibitors against *APH1A* gene can be suggested to poor prognosis patients for better outcome. We have found 32 and 271 genes with greater degree of up regulation in poor prognostic patients as compared to good prognostic ones in LUSC and LUAD, respectively. We have used DGIdb resource (https://www.dgidb.org/search_interactions) to investigate the therapeutic potential of the identified prognostic genes specific to each subtype. This database maintains latest information of drug–gene interactions identified from the experimental studies. We found that 6 out of 32 (LUSC) and 58 out of 271 (LUAD) identified prognostic genes have inhibitory drugs designed against them. We have summarized the details of the corresponding drug–gene interactions along with the source and PMIDS of the study in the Supplementary S1 Tables 8 and 9. Data analysis further highlights that the identified prognostic genes may be attractive drug targets for subtype-specific lung cancer treatment as some of the interacting drugs are already in clinical trial against some other genes for lung cancer. This emphasizes on the fact that already approved drugs may be repurposed for our identified marker genes. For example, paclitaxel drug which shows interaction with the identified prognostic *AURKB* gene is well-known treatment for lung cancer, Alisertib drug (also interacting to *AURKB* gene) in combination with osimertinib is currently in clinical trial for the treatment of *EGFR* mutant lung cancer (NCT04085315). Thus, it is clear that the identified prognostic genes may hold a great therapeutic potential, and can be used for designing strategies based on inhibitors, agonist, and antagonists.

## Discussion

Genomic and epigenomic alterations within the genome favor the tumorigenesis (Kumar et al. 2019). Several genomic alterations occur in cell signaling pathways that control cell death, cell division, and cell fate. Despite the advancement in lung cancer treatment, the survival rate among the patients is very poor (Song et al. 2018). The possible reason could be the inability of the drug to relieve all patients and the lack of an effective biomarker for the identification of lung cancer. Although some statistical methods have been developed for the prognosis prediction that are based on a gene-centric approach (Yuan et al. 2014; Zhao et al. 2015), but they are still of limited use. It has also been observed that patients respond differently to treatment because of heterogeneity among non-small-cell lung cancer molecular subtypes, which makes it necessary to capture the heterogeneity for the better management of patients. Previous studies exploited the differences in genomic alterations among LUSC and LUAD to explain the variation in the OS rates of the NSCLC patients. They identified *SNTG1* and *LRRK2* genes to be significantly associated with the OS in LUSC and LUAD, respectively (Meng et al. 2019). However, we observed that these two genes were not associated with OS in both LUSC and LUAD at mRNA expression level. These findings provoke the need of investigation to understand why a gene with some prognostic importance at one biological level fails to depict the same at other levels. To fill the gap between the different biological molecular mechanisms of the disease and to increase the predictive power of a multi-gene signature that operates in different pathways, we performed an integrative study that takes into consideration the disease-related pathway and gene expression information.

In this study, we have used the RNA_seq expression and clinical data of LUSC and LUAD from TCGA and GEO for training and validation, respectively. We aim to identify survival-associated genes for each subtype by applying cox regression and other statistical measures. We developed different prognostic index-based models to stratify the patients in high- and low-risk groups for NSCLC subtypes. The obtained results were further supported by the experimental evidences. We have found 32 and 271 prognostic marker genes, some of which were also validated in other published studies for LUSC and LUAD, respectively. These genes can be further investigated for therapeutic and diagnostic potentials in more subtype-specific interventions. Also, 90% of the identified prognostic biomarkers are novel and need further investigation. We showed that for the risk stratification of patients in LUAD and LUSC, different gene set and pathways come out to be important. For the LUSC, apoptosis and p53 pathway genes are more important and for LUAD genes involved in apoptosis, PI3K-AKT, and WNT pathways are more important (Supplementary S1 Tables 5 and 6). We have also shown that despite having different oncogenic origins, age and tumor stage (in combination) are good predictors for both NSCLC subtypes. Our best models outperformed the existing prognostic models for NSCLC subtypes in terms of HR estimation and significance of prediction (Meng et al. 2019). The limitation of the present study is that we have considered only the gene expression data and clinicopathological features. In the future, we hope to apply this approach to the other levels of genomic data such as methylation, copy number change, and miRNA data.

The significance of the present study can be seen in terms that it can effectively explain the basis of heterogeneity among NSCLC subtypes at different levels. Our study revealed the subtype-specific prognostic genes, thus providing little insight into the biology of the disease etiology. The identified genes have been poorly investigated and thus deserve the attention of clinicians and researchers to

propose reliable prognostic as well as therapeutic strategies for NSCLC patients in a subtype-specific manner.

**Author contributions** Conceptualization, AL, RK, and GPSR; methodology, AL, RK, and GPSR; formal analysis, AL, RK, and GPSR.; investigation, AL, RK, and GPSR; code development, AL; visualization and figures, AL, RK, and GPSR; interpretation of data and results, AL, RK, CA and GPSR; supervision, GPSR; project administration, GPSR; funding acquisition, GPSR; writing and editing, AL, RK, CA, and GPSR. All authors have read and agreed to the published version of the manuscript.

**Funding** There is no funding available for this paper.

**Data availability** The data that support the findings of this study are openly available on TCGA website as well as in the supplementary files.

## Compliance with ethical standards

**Conflict of interest** The authors declare no conflict of interest.

## References

Cancer Genome Atlas Research Network (2014) Comprehensive molecular profiling of lung adenocarcinoma. Nature 511(7511):543–550. https://doi.org/10.1038/nature13385

Chang JT-H, Lee YM, Huang RS (2015) The impact of the Cancer Genome Atlas on lung cancer. Transl Res J Lab Clin Med 166(6):568–585. https://doi.org/10.1016/j.trsl.2015.08.001

Chaudhary K, Poirion OB, Lu L, Garmire LX (2018) Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res Off J Am Assoc Cancer Res 24(6):1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853

Chen J, Sam L, Huang Y, Lee Y, Li J, Liu Y, Xing HR, Lussier YA (2010) Protein interaction network underpins concordant prognosis among heterogeneous breast cancer signatures. J Biomed Inform 43(3):385–396. https://doi.org/10.1016/j.jbi.2010.03.009

Cheng YQB, Shao N, Jia Y, Song Q, Tan B, Wang J (2020) Prognostic value of immune-related genes in the tumor microenvironment of lung adenocarcinoma and lung squamous cell carcinoma. Aging 12(6):4757–4777. https://doi.org/10.18632/aging.102871

Deng F, Chen D, Wei X, Lu S, Luo X, He J, Liu J, Meng T, Yang A, Chen H (2020) Development and validation of a prognostic classifier based on HIF-1 signaling for hepatocellular carcinoma. Aging 12(4):3431–3450. https://doi.org/10.18632/aging.102820

Dyrskjøt L, Reinert T, Algaba F, Christensen E, Nieboer D, Hermann GG, Mogensen K, Beukers W, Marquez M, Segersten U, Høyer S, Ulhøi BP, Hartmann A, Stöhr R, Wach S, Nawroth R, Schwamborn K, Tulic C, Simic T et al (2017) Prognostic Impact of a 12-gene progression score in non-muscle-invasive bladder cancer: a prospective multicentre validation study. Eur Urol 72(3):461–469. https://doi.org/10.1016/j.eururo.2017.05.040

Frost HR, Amos CI (2018) A multi-omics approach for identifying important pathways and genes in human cancer. BMC Bioinform 19(1):479. https://doi.org/10.1186/s12859-018-2476-8

He X, Xu H, Zhao W, Zhan M, Li Y, Liu H, Tan L, Lu L (2019) POPDC3 is a potential biomarker for prognosis and radioresistance in patients with head and neck squamous cell carcinoma. Oncol Lett 18(5):5468–5480. https://doi.org/10.3892/ol.2019.10888

Kumar R, Patiyal S, Kumar V, Nagpal G, Raghava GPS (2019) In silico analysis of gene expression change associated with copy number of enhancers in pancreatic adenocarcinoma. Int J Mol Sci. https://doi.org/10.3390/ijms20143582

Lathwal A, Arora C, Raghava GPS (2019) Prediction of risk scores for colorectal cancer patients from the concentration of proteins involved in mitochondrial apoptotic pathway. PLoS ONE 14(9):e0217527. https://doi.org/10.1371/journal.pone.0217527

Li B, Feng W, Luo O, Xu T, Cao Y, Wu H, Yu D, Ding Y (2017) Development and validation of a three-gene prognostic signature for patients with hepatocellular carcinoma. Sci Rep. https://doi.org/10.1038/s41598-017-04811-5

Li P, Ren H, Zhang Y, Zhou Z (2018) Fifteen-gene expression based model predicts the survival of clear cell renal cell carcinoma. Medicine (United States) 97(33):e11839. https://doi.org/10.1097/MD.0000000000011839

Liu Y, Wu L, Ao H, Zhao M, Leng X, Liu M, Ma J, Zhu J (2019a) Prognostic implications of autophagy-associated gene signatures in non-small cell lung cancer. Aging Impact J LLC 11(23):11440–11462. https://doi.org/10.18632/aging.102544

Liu Y, Yang N, Peng X, Liu G, Zhong H, Liu L (2019b) One-lincRNA and five-mRNA based signature for prognosis of multiple myeloma patients undergoing proteasome inhibitors therapy. Biomed Pharmacother Biomed Pharmacother 118:109254. https://doi.org/10.1016/j.biopha.2019.109254

Lv Z, Lei T (2020) Systematical identifications of prognostic meaningful lung adenocarcinoma subtypes and the underlying mutational and expressional characters. BMC Cancer. https://doi.org/10.1186/s12885-019-6462-y

Ma B, Geng Y, Meng F, Yan G, Song F (2020a) Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a machine learning method. J Cancer 11(5):1288–1298. https://doi.org/10.7150/jca.34585

Ma X, Ren H, Peng R, Li Y, Ming L (2020b) Identification of key genes associated with progression and prognosis for lung squamous cell carcinoma. PeerJ 8:e9086. https://doi.org/10.7717/peerj.9086

Mandelboum S, Manber Z, Elroy-Stein O, Elkon R (2019) Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias. PLoS Biol 17(11):e3000481. https://doi.org/10.1371/journal.pbio.3000481

Matsuda A, Yamaoka K, Tango T (2012) Quality of life in advanced non-small cell lung cancer patients receiving palliative chemotherapy: a meta-analysis of randomized controlled trials. Exp Ther Med 3(1):134–140. https://doi.org/10.3892/etm.2011.368

Meng F, Zhang L, Ren Y, Ma Q (2019) The genomic alterations of lung adenocarcinoma and lung squamous cell carcinoma can explain the differences of their overall survival rates. J Cell Physiol 234(7):10918–10925. https://doi.org/10.1002/jcp.27917

Midthun DE (2011) Screening for lung cancer. Clin Chest Med 32(4):659–668. https://doi.org/10.1016/j.ccm.2011.08.014

Molina JR, Yang P, Cassivi SD, Schild SE, Adjei AA (2008) Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 83(5):584–594. https://doi.org/10.4065/83.5.584

Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong LI (2020) The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 48(D1):D845–D855. https://doi.org/10.1093/nar/gkz1021

Rekhtman N, Paik PK, Arcila ME, Tafe LJ, Oxnard GR, Moreira AL, Travis WD, Zakowski MF, Kris MG, Ladanyi M (2012) Clarifying the spectrum of driver oncogene mutations in biomarker-verified squamous carcinoma of lung: lack of EGFR/KRAS and presence of PIK3CA/AKT1 mutations. Clin Cancer Res Off J Am Assoc Cancer Res 18(4):1167–1176. https://doi.org/10.1158/1078-0432.CCR-11-2109

Relli V, Trerotola M, Guerra E, Alberti S (2018) Distinct lung cancer subtypes associate to distinct drivers of tumor progression. Oncotarget 9(85):35528–35540. https://doi.org/10.18632/oncotarget.26217

Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, Chakravarty D, Daian F, Gao Q, Bailey MH, Liang WW, Foltz SM, Shmulevich I, Ding L, Heins Z et al (2018) Oncogenic signaling pathways in the cancer genome atlas. Cell 173(2):321–337.e10. https://doi.org/10.1016/j.cell.2018.03.035

Sandler A, Gray R, Perry MC, Brahmer J, Schiller JH, Dowlati A, Lilenbaum R, Johnson DH (2006) Avastin in metastatic NSCLC-Sandler data (Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. N Engl J Med 355(24):2542–2550. https://doi.org/10.1056/NEJMoa061884

Sheng M, Xie X, Wang J, Gu W (2019) A pathway-based strategy to identify biomarkers for lung cancer diagnosis and prognosis. Evolut Bioinf Online 15:1176934319838494. https://doi.org/10.1177/1176934319838494

Shi M, Xu G (2019) Development and validation of GMI signature based random survival forest prognosis model to predict clinical outcome in acute myeloid leukemia. BMC Med Genom. https://doi.org/10.1186/s12920-019-0540-5

Siegel RL, Miller KD, Jemal A (2018) Cancer statistics, 2018. CA Cancer J Clin 68(1):7–30. https://doi.org/10.3322/caac.21442

Song YJ, Tan J, Gao XH, Wang LX (2018) Integrated analysis reveals key genes with prognostic value in lung adenocarcinoma. Cancer Manag Res 10:6097–6108. https://doi.org/10.2147/CMAR.S168636

Stewart DJ, Tomiak E, Shamji FM, Maziak DE, MacLeod P (2004) Phase II study of alternating chemotherapy regimens for advanced non-small cell lung cancer. Lung Cancer (Amsterdam, Netherlands) 44(2):241–249. https://doi.org/10.1016/j.lungcan.2003.10.009

TCGA (2012) Comprehensive genomic characterization of squamous cell lung cancers: pathways and integrated analysis (Supplementary material). Nature 489(7417):519–525. https://doi.org/10.1038/nature11404

Tseng GC, Ghosh D, Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res 40(9):3785–3799. https://doi.org/10.1093/nar/gkr1265

Wang Y, Ren F, Chen P, Liu S, Song Z, Ma X (2018) Identification of a six-gene signature with prognostic value for patients with endometrial carcinoma. Cancer Med 7(11):5632–5642. https://doi.org/10.1002/cam4.1806

Yan J, Kim Y-S, Yang X-P, Li L-P, Liao G, Xia F, Jetten AM (2007) The ubiquitin-interacting motif containing protein RAP80 interacts with BRCA1 and functions in DNA damage repair response. Can Res 67(14):6647–6656. https://doi.org/10.1158/0008-5472.CAN-07-0924

Yang S, Chen L, Chan DW, Li QK, Zhang H (2017) Protein signatures of molecular pathways in non-small cell lung carcinoma (NSCLC): comparison of glycoproteomics and global proteomics. Clin Proteom. https://doi.org/10.1186/s12014-017-9166-9

Yuan Y, Van Allen EM, Omberg L, Wagle N, Amin-Mansour A, Sokolov A, Byers LA, Xu Y, Hess KR, Diao L, Han L, Huang X, Lawrence MS, Weinstein JN, Stuart JM, Mills GB, Garraway LA, Margolin AA, Getz G, Liang H (2014) Assessing the clinical utility of cancer genomic and proteomic data across tumor types. Nat Biotechnol 32(7):644–652. https://doi.org/10.1038/nbt.2940

Zhang J, Xu M, Gao H, Guo J-C, Guo Y-L, Zou M, Wu X-F (2018) Two protein-coding genes act as a novel clinical signature to predict prognosis in patients with ovarian serous cystadenocarcinoma. Oncol Lett 15(3):3669–3675. https://doi.org/10.3892/ol.2018.7778

Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S (2015) Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform 16(2):291–303. https://doi.org/10.1093/bib/bbu003

Zhu Y, Qiu P, Ji Y (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods 11(6):599–600. https://doi.org/10.1038/nmeth.2956

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