Original Article

CD8$^+$ T-Cell Signatures as Prognostic and Immunotherapy Response Predictors in Non-Small Cell Lung Cancer

(NSCLC / single-cell / bulk RNA sequencing / CD8$^+$ T-cell marker genes / prognostic signature / immunotherapy response)

TIENAN ZHAO, SARINDER KAUR DHILLON

Data Science and Bioinformatics Laboratory, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia

Abstract. Non-small cell lung carcinoma (NSCLC) represents the majority of lung cancer cases, comprising approximately 85 % of the total. The five-year survival rate for NSCLC patients remains discouragingly low. Recently, immunotherapy has emerged as a promising approach. Nevertheless, only a minority of patients experience considerable benefits from these treatments. This highlights the critical need for effective biomarkers that can predict both patient prognosis and response to immunotherapy. CD8$^+$ T cells play a crucial role in cancer immunotherapy. Their presence within tumours is generally indicative of a favourable prognosis and increased efficacy of immunotherapy. This study was undertaken to identify and authenticate a novel biomarker signature based on CD8$^+$ T-cell marker genes, to prognosticate therapeutic responses in individuals afflicted with NSCLC. This in-depth study was based on a total of 1,200 samples, which included four NSCLC specimens analysed through single-cell RNA sequencing (scRNA-seq), 1,000 NSCLC samples obtained from The Cancer Genome Atlas (TCGA) and 196 NSCLC specimens collected from the GSE37745 cohort. In patients with NSCLC, those presenting a favourable risk profile demonstrated notable elevations in specific immune cells while concurrently exhibiting reductions in other types. CD8$^+$ T cells, with their established role in inducing apoptosis in cancer cells, have emerged as crucial predictors and modulators of treatment strategies for NSCLC patients. The combination of single-cell and bulk RNA sequencing has produced a biomarker signature, emphasizing the CD8$^+$ T cells' crucial role in NSCLC prognosis and treatment.

Received April 2, 2024. Accepted October 9, 2024.

Corresponding author: Sarinder Kaur Dhillon, Data Science and Bioinformatics Laboratory, Institute of Biological Sciences, Faculty of Science, Universiti Malaya, Wilayah Persekutuan, 50603 Kuala Lumpur, Malaysia; Phone and Fax: (+603) 796 767 41; E-mail: sarinder@um.edu.my

Abbreviations: APCs – antigen-presenting cells, ANXA2 – annexin A2, AUC – area under the receiver operating characteristic curve, CCR – chemokine receptor, CTSL – cathepsin L, DEGs – differentially expressed genes, ECM – extracellular matrix, EREG – epiregulin, EGFR – epidermal growth factor receptor, GEO – Gene Expression Omnibus, GO – Gene Ontology, IDCs – immature dendritic cells, KEGG – the Kyoto Encyclopedia of Genes and Genomes, LASSO – least absolute shrinkage and selection operator, LUAD – lung adenocarcinoma, NSCLC – non-small cell lung carcinoma, OS – overall survival, PCA – principal component analysis, PDCs – plasmacytoid dendritic cells, RNA-seq – RNA sequencing, ROC – receiver operating characteristic curve, scRNA-seq – single-cell RNA sequencing, TCGA – The Cancer Genome Atlas, TCMGs – CD8$^+$ T-cell marker genes, TFH – T follicular helper cells, TIL – tumour-infiltrating lymphocytes, TPM – transcripts per kilobase million, Y-IFN – type II interferon.

Introduction

Non-small cell lung carcinoma (NSCLC) represents the predominant type of lung cancer worldwide, accounting for about 85 % of all cases (Chhikara and Parang, 2023). Despite continuous efforts over the years leading to significant progress in diagnosing and managing NSCLC, the five-year survival rate remains dismally low (Spigel et al., 2022). Immunotherapy has gained recognition within this framework as a ground-breaking approach, especially in advanced NSCLC cohorts. Notably, antibody therapies targeting cellular immune checkpoints have become essential treatment strategies (Lurienne et al., 2020). However, only a subset of patients achieves substantial benefits from these treatments, highlighting the urgent need for effective biomarkers to predict patient outcomes and response to immunotherapy.

CD8$^+$ T cells, central to immune defence, are pivotal in determining the success of immunotherapeutic approaches in oncology (Tanoue et al., 2019). Recent research has established a significant association between tumour-infiltrating CD8$^+$ T cells and enhanced patient outcomes, as well as improved efficacy of immunotherapies (Krishna et al., 2020; Raskov et al., 2021; Philip

and Schietinger, 2022). Although CD8+ T-cell marker genes (TCMGs) have been extensively studied, a comprehensive and reliable biomarker signature has yet to be found. Such a signature could significantly improve prognostic precision and predict the immunotherapy response in NSCLC patients, potentially marking a turning point in the battle against this challenging disease. The scientific community continues its relentless pursuit of ground-breaking discoveries that may alter the therapeutic landscape for NSCLC.

The advent of single-cell sequencing technology provides an in-depth understanding of the mechanisms of action of tumour-infiltrating CD8+ T cells in NSCLC (Liu et al., 2021). Utilizing single-cell RNA sequencing (scRNA-seq), scientists are focused on unravelling the complex interactions within the tumour microenvironment, identifying key players in the immune response to NSCLC (Chen et al., 2019; Cortellini et al., 2019; Qu et al., 2020). This cutting-edge technique heralds a new era, possibly revealing a ground-breaking biomarker signature that could transform prognostic and therapeutic approaches for NSCLC patients.

In this study, the fusion of scRNA-seq with bulk RNA sequencing (bulk RNA-seq) seeks to identify and validate an innovative signature of biomarkers grounded in TCMGs, aiming to assess both the prognosis and the immunotherapeutic response in NSCLC patients. The study begins with an extensive analysis of numerous publicly available scRNA-seq and bulk RNA-seq datasets, focusing on genes closely related to CD8+ T-cell behaviour. After obtaining a curated gene set, we moved to the next phase, which is applying advanced statistical and bioinformatic techniques to develop an innovative predictive model for prognosis and immunotherapy trajectories. This approach offers valuable insights into the NSCLC's complexity, guiding the field toward personalized immunotherapies and improved patient prognoses.

Material and Methods

Data collection

This investigation incorporated a comprehensive dataset comprising 1,200 specimens, which included four NSCLC samples with scRNA-seq data, 1,000 NSCLC specimens sourced from The Cancer Genome Atlas (TCGA) and an additional 196 NSCLC samples derived from the GSE37745 cohort (https://www.ncbi.nlm.nih.gov/geo/). To elucidate TCMGs specific to NSCLC, scRNA-seq data of four NSCLC patients from GSE117570 were analysed from Tumour Immune Single-cell Hub 2 (Han et al., 2023). Additionally, we collected extensive tumour transcriptomic data for 1,080 NSCLC samples from TCGA-LUAD and TCGA-LUSC via the UCSC Xena platform (available at: https://xenabrowser.net/). After removing samples without corresponding clinical information, 1,000 samples were analysed. This extensive dataset facilitated identification of survival-related genes and formulation of prognostic signatures with significant implications. To validate our findings externally, we employed the GSE37745 dataset from the GEO database (Barrett et al., 2013). For harmonization and enhanced comparability between TCGA samples and the GEO dataset, the TCGA RNA-seq data were transformed into Transcripts Per Kilobase Million (TPM) values, a unit conducive to thorough analysis and evaluation (Wagner et al., 2012). In order to confirm the predictive capacity of TCMGs in immunotherapy response, we collected transcriptomic data along with matching clinical details from individuals treated with anti-PD-L1 in the IMvigor210 cohort. These resources were accessed from Mariathasan et al. (2018), thus providing a comprehensive understanding of the nuances of immunotherapy response.

Unveiling CD8+ T-cell profiles by diving into scRNA-seq data

We initiated the analysis by utilizing the scRNA-seq data sourced from the GSE117570 dataset, accessible via the TISCH2 database (Han et al., 2023). By comparing the distinctive profiles of CD8+ T cells with other cellular counterparts, we successfully found 220 differentially expressed genes, based on a fold change greater than 1.5 and an adjusted P value lower than 0.05. In parallel, we downloaded a comprehensive collection of 2,533 immune-related genes from ImmPort (Bhattacharya et al., 2014) and InnateDB (Breuer et al., 2013) databases. We effectively compared the DEGs with the immune-related gene pool by employing a Venn diagram. This comprehensive analysis revealed a notable intersection, highlighting 108 genes demonstrating a significant overlap between the DEGs and the immune-related genes.

Construction and validation of the prognostic potential of TCMGs

To assess the prognostic significance of TCMGs for overall survival (OS) in the TCGA NSCLC cohort, a univariate Cox regression analysis was carried out. This analysis identified prognostic genes based on a P value less than 0.05, revealing genes with notable prognostic relevance. To reduce the likelihood of overfitting, these prognostic genes were subject to an exhaustive evaluation using the advanced least absolute shrinkage and selection operator (LASSO) Cox proportional hazards regression method, facilitated by the "glmnet" package (Tibshirani, 1997). Using the "cv.glmnet" function, we conducted a detailed 10-fold cross-validation to identify the best model setup. The tuning parameter, labelled as "lambda", was thoughtfully determined using the 1 – SE (standard error) criterion, ensuring high accuracy in model selection. As a result, a complete catalogue of genes possessing non-zero beta coefficients were identified, representing the essence of prognostic value. Subsequently, a stepwise multivariate Cox regression analy-

sis was applied, incorporating the gene signatures identified through the LASSO Cox regression. This detailed analysis allowed the pinpointing of specific gene signatures with significant prognostic implications. Our risk model was based on a meticulously formulated linear combination, integrating the mRNA expression levels of these genes with their corresponding risk coefficients. By adopting the median cut-off value as a pivotal criterion, we systematically categorized patients into low-risk and high-risk groups. This approach significantly enhanced the precision of our prognostic assessments, enabling more detailed and insightful stratification that potentially informs personalized treatment decisions. For validation of the prognostic efficacy and discriminative ability of these TCMGs, we utilized the "survivalROC" package (Heagerty and Zheng, 2005) to compute the area under the curve (AUC). The Kaplan-Meier method was used in the survival analysis, while the log-rank test, conducted employing the "survminer" R package, assessed the significance in statistics of deviations in survival outcomes. Additionally, we validated the predictive power of our gene signature through comprehensive survival analysis and AUC calculations, using data from the GEO dataset. To assess the statistical significance of the differences between the high-risk group and low-risk group for each clinicopathological factor, the χ^2 test was adopted.

Pathway and function enrichment analysis

We analysed our genetic data using Gene Ontology (GO) and The Kyoto Encyclopedia of Genes and Genomes (KEGG) by utilizing the "clusterProfiler" R package. This approach allowed us to categorize genes into functionally related groups and to identify relevant biological pathways, enhancing our understanding of the molecular mechanisms involved in our study (Yu et al., 2012). The Gene Ontology (GO) analysis was done using the "enrichGO" function within the "clusterProfiler" R package, leveraging the extensive genome-wide annotation packages, specifically org.Hs.eg.db, which the Bioconductor project has carefully curated. This approach allowed for detailed exploration of gene functions, biological processes and cellular components associated with our dataset (Gentleman et al., 2004). For deeper exploration of the biological pathways, we employed the "enrichKEGG" function of "clusterProfiler". By offering direct access to the most recent version of the KEGG database via a web API, the tool made it simpler to acquire crucial pathway information. This feature significantly facilitated the comprehensive functional analysis by allowing for rapid and efficient retrieval of the data related to biological pathways, thereby enhancing the depth and scope of our study insights into gene functions and interactions. Throughout our analysis, we set a stringent significance threshold of $P < 0.05$ to ensure identification of notably enriched GO terms and KEGG pathways. This threshold was crucial in discerning our dataset's most significant functional enrichments, guiding us toward meaningful biological insights.

Immune cell infiltration analysis and immunotherapy response prediction

To acquire an in-depth understanding of the complex dynamics of immune cell infiltration and associated variations in gene sets, we performed a detailed analysis using immune cell infiltration analysis. The CIBERSORT algorithm, which is renowned for precisely identifying infiltration patterns of 22 immune cell types by their gene expression profiles, was utilized for analysing the complex nature of immune cell infiltration in depth (Newman et al., 2019). This method provided a crucial framework for examining the proportionate distribution of immune cell infiltration between high-risk and low-risk groups, thus revealing the complexities of immune cell behaviour within the tumour microenvironment. Finally, PD-L1 mRNA expression data were extracted from the RNA-sequencing datasets of the TCGA NSCLC cohort. PD-L1 is a key biomarker applied for forecasting responses to immune checkpoint blockade intervention.

Statistical analysis

We conducted a detailed comparison of categorized variables among distinct risk groups, employing the Wilcoxon rank-sum test for its proven reliability and precision in non-parametric statistical analysis. To enhance the depth of our prognostic assessment, we carried out both univariate and multivariate Cox regression analyses. These analyses were pivotal in determining the significance of TCMGs in conjunction with various clinicopathological factors. We applied a P value threshold of < 0.05, enabling identification of statistically significant associations. To mitigate the effects of multiple testing, the Benjamini-Hochberg procedure was strategically employed to adjust P values, thus maintaining the validity of our analysis. Our data analysis and visualization were performed using R (version 4.1.0), which is accessible at: http://www.R-project.org.

Results

Decoding $CD8^+$ T-cell signatures through scRNA-seq analysis of GSE117570

In our in-depth examination of the scRNA-seq data from GSE117570, we applied Principal Component Analysis (PCA) for dimensionality reduction. Focusing on variable genes, this method allowed us to distinguish 16 unique cell clusters (Fig. 1A). Our next step was to elucidate the identities of these clusters. To achieve that, we referred to the Human Primary Cell Atlas (https://rdrr.io/), which provided valuable insights into the cellular composition of our samples. This investigation revealed that cluster 2 predominantly consisted of $CD8^+$ T cells (Fig. 1B). Notably, the gene expression profile of this particular cluster revealed a significant 220 differentially expressed genes among the 16 clusters (https://github.com/sarinderkaur/CD8-T-Cell-Signatures-as-

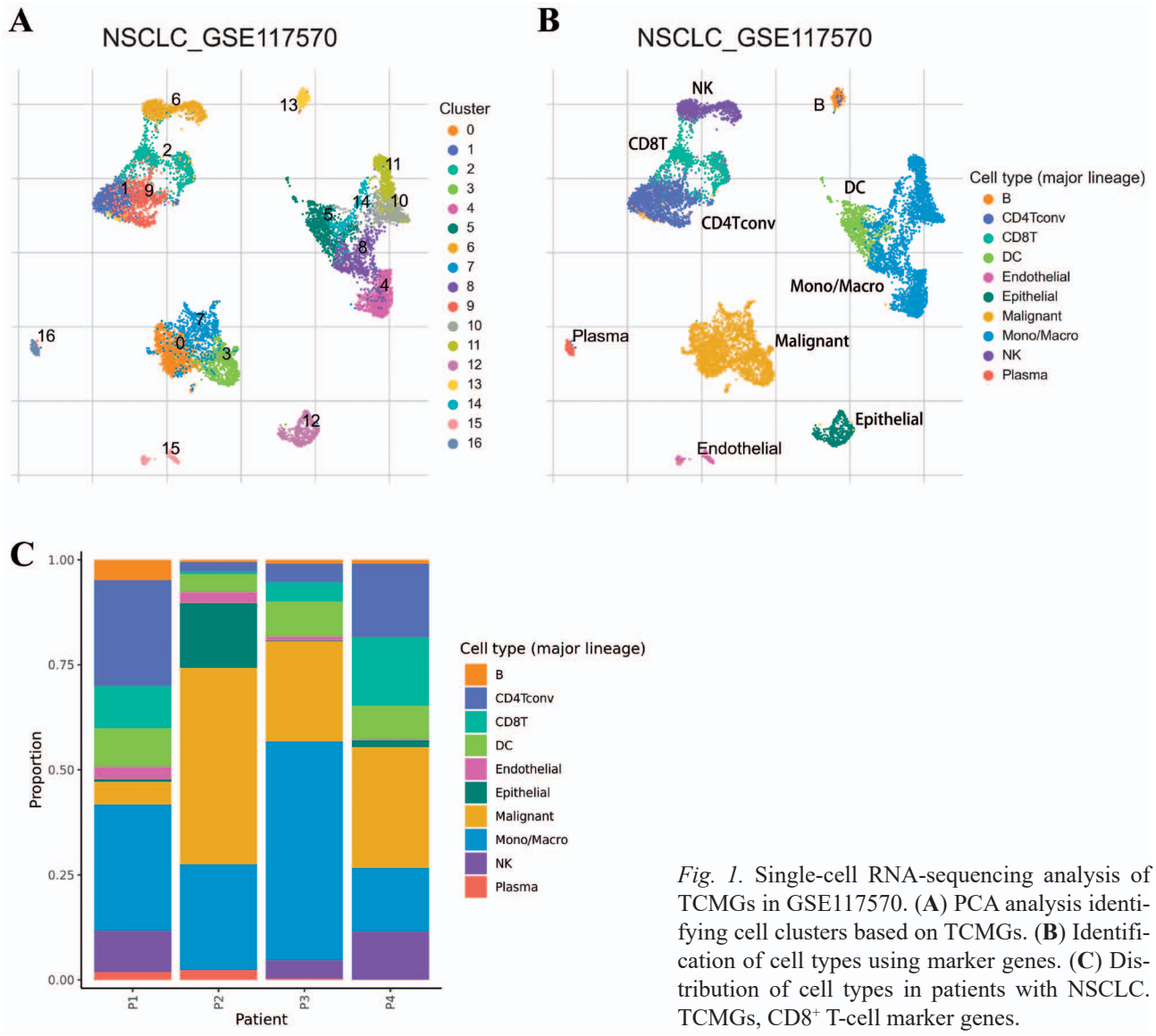


Fig. 1. Single-cell RNA-sequencing analysis of TCMGs in GSE117570. (**A**) PCA analysis identifying cell clusters based on TCMGs. (**B**) Identification of cell types using marker genes. (**C**) Distribution of cell types in patients with NSCLC. TCMGs, CD8+ T-cell marker genes.

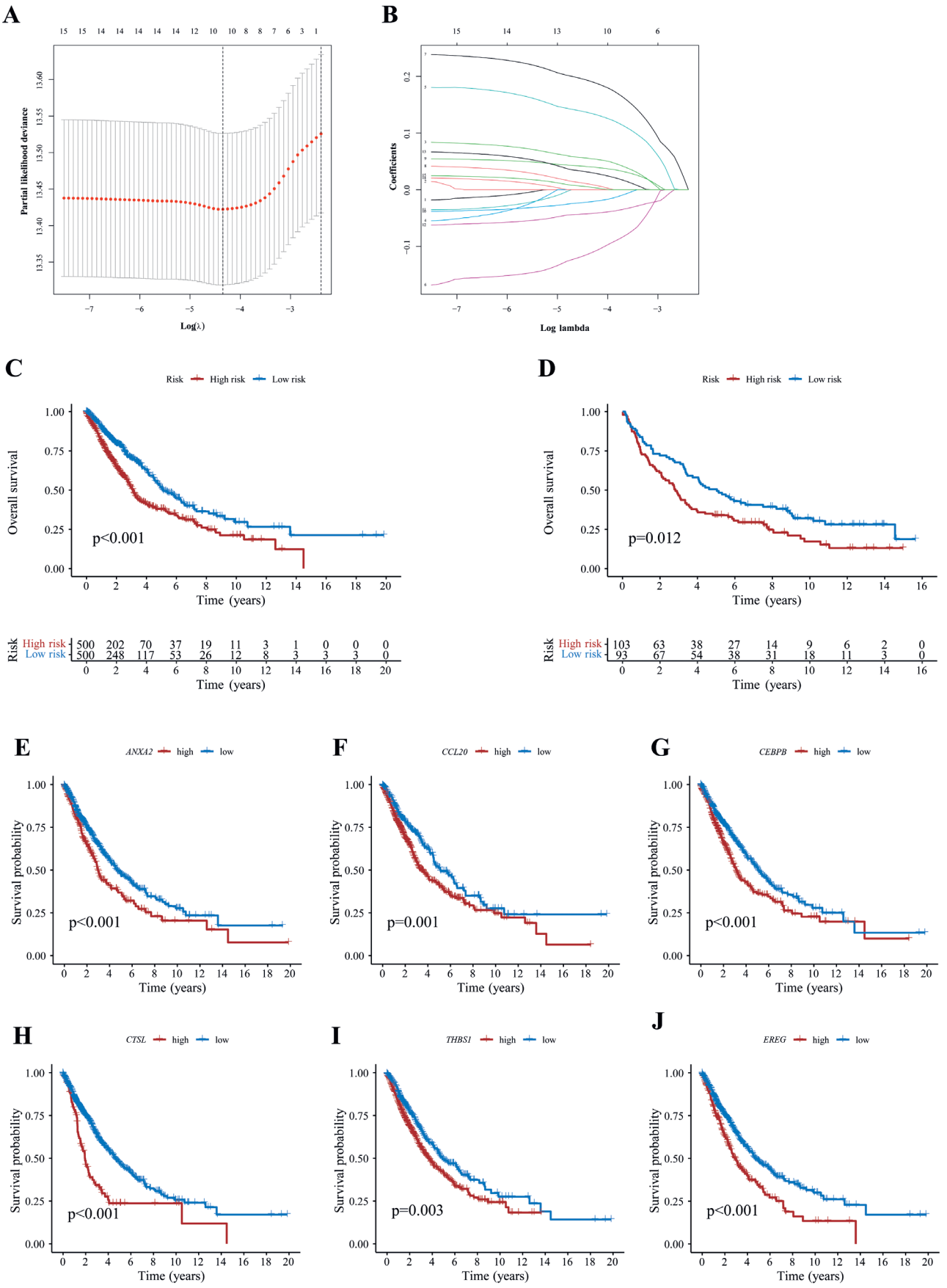
Prognostic-and-Immunotherapy-Response-Predictors-in-NSCLC). To delve deeper into the cellular landscape of NSCLC, we carefully selected the top 10 cell clusters from four primary NSCLC samples, to explore their hidden complexities (Fig. 1C).

Identification and analysis of NSCLC-associated TCMGs

From our investigation using ImmPort (https://www.immport.org) and InnateDB (http://www.innatedb.com/), we identified 2,533 immune-related genes directly involved in the molecular landscape of NSCLC. Venn diagram analysis elucidated those 108 genes intersected between CD8+ T differentially expressed genes and those from the aforementioned databases (Supplementary Fig. S1), henceforth referred to as NSCLC-associated TCMGs. For the development of a predictive signature based on TCMGs, we employed the NSCLC patient cohort from TCGA as our primary dataset. Through univariate Cox regression, we identified 15 TCMGs significantly connected to OS, detailed in Supplementary Fig. S2. Progressing to LASSO Cox regression analysis, we carefully evaluated these 15 TCMGs (Fig. 2A and B), ultimately narrowing down to 10 key genes (*ANXA2, CCL20, CEBPB, CTSL, THBS1, EREG, SPP1, HLA-DMB, CXCL17, TRBC1*) for further analysis. To determine the TCMG risk score, we used the following equation (1):

$$0.055 * \text{ANXA2} + 0.030 * \text{CCL20} + 0.135 * \text{CEBPB} + 0.192 * \text{CTSL} + 0.010 * \text{THBS1} + 0.040 * \text{EREG} + 0.004 * \text{SPP1} - 0.112 * \text{HLA-DMB} - 0.042 * \text{CXCL17} - 0.021 * \text{TRBC1} \quad \text{(Equation 1)}$$

We then segmented the patient cohort into high-risk and low-risk groups based on the median risk score. Remarkably, survival analysis revealed a noteworthy


A
Partial likelihood deviance
Log(λ)
B
Coefficients
Log lambda
C
Risk High risk Low risk
Overall survival
p<0.001
Time (years)
High risk 500 202 70 37 19 11 3 1 0 0 0
Low risk 500 248 117 53 26 12 8 3 3 3 0
D
Risk High risk Low risk
Overall survival
p=0.012
Time (years)
High risk 103 63 38 27 14 9 6 2 0
Low risk 93 67 54 38 31 18 11 3 0
E
ANXA2 high low
Survival probability
p<0.001
Time (years)
F
CCL20 high low
Survival probability
p=0.001
Time (years)
G
CEBPB high low
Survival probability
p<0.001
Time (years)
H
CTSL high low
Survival probability
p<0.001
Time (years)
I
THBS1 high low
Survival probability
p=0.003
Time (years)
J
EREG high low
Survival probability
p<0.001
Time (years)

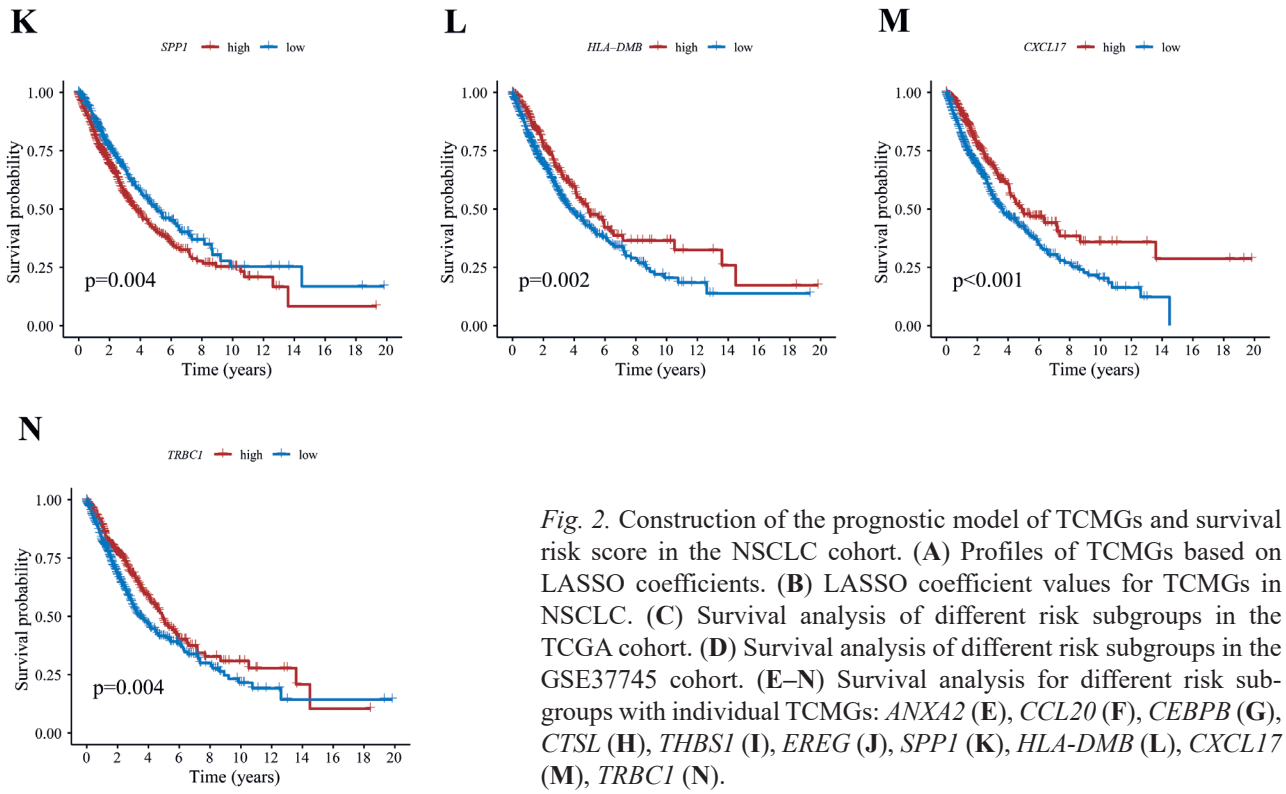


Fig. 2. Construction of the prognostic model of TCMGs and survival risk score in the NSCLC cohort. (**A**) Profiles of TCMGs based on LASSO coefficients. (**B**) LASSO coefficient values for TCMGs in NSCLC. (**C**) Survival analysis of different risk subgroups in the TCGA cohort. (**D**) Survival analysis of different risk subgroups in the GSE37745 cohort. (**E–N**) Survival analysis for different risk subgroups with individual TCMGs: *ANXA2* (**E**), *CCL20* (**F**), *CEBPB* (**G**), *CTSL* (**H**), *THBS1* (**I**), *EREG* (**J**), *SPP1* (**K**), *HLA-DMB* (**L**), *CXCL17* (**M**), *TRBC1* (**N**).

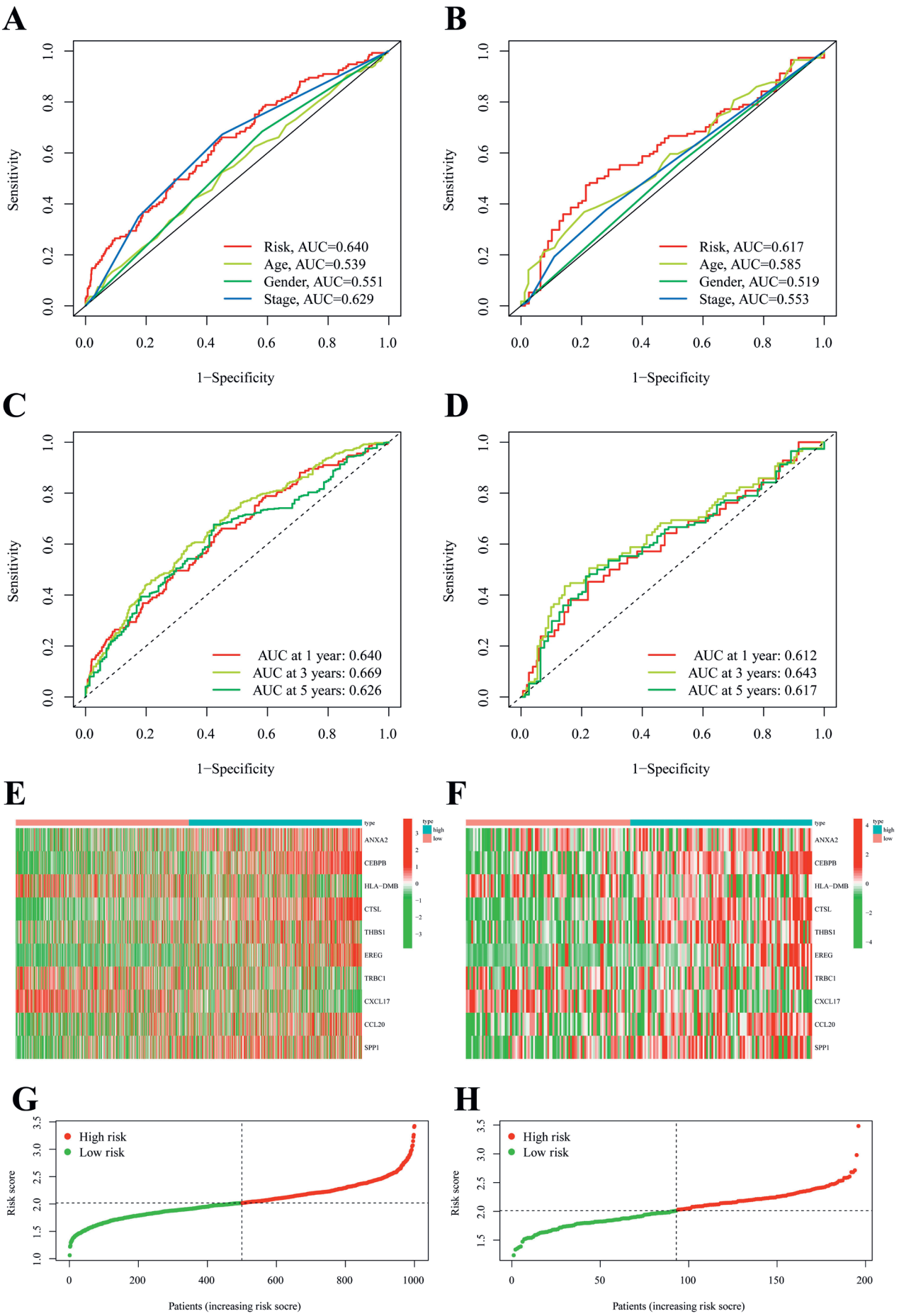
variance in outcomes between these groups. Figures 2C and D depict a notable trend in both the TCGA cohort and GSE37745, where the OS is markedly lower in patients within the low-risk group than in those categorized as high-risk. We observed a notable correlation in detailed examination of TCMG mRNA expression within the TCGA NSCLC cohort. Higher expression levels of *ANXA2, CCL20, CEBPB, CTSL, THBS1, EREG* and *SPP1* were associated with decreased OS (Fig. 2 E–K), while lower expression levels of *HLA-DMB*, *CXCL17* and *TRBC1* showed similar trends (Fig. 2L–N).

Prognostic efficacy of TCMG risk score in NSCLC cohorts

To confirm the prognostic utility of TCMGs, we established a training set from NSCLC patients in the TCGA database, with GSE37745 patients forming the validation cohort. We thoroughly analysed the clinical characteristics across different risk categories in the TCGA dataset. This analysis elucidated a significant correlation between the calculated risk score and principal variables, including age, gender and stage of disease, as depicted in Supplementary Fig. S3. The ROC curves for the training group showed that the AUC for the risk score, age, gender and stage were 0.640, 0.539, 0.551 and 0.629, respectively (Fig. 3A). Variations in AUC values were also observed in the GSE37745 cohort, with the risk score at 0.617, age at 0.585, gender at 0.519 and disease stage at 0.553 (Fig. 3B). Furthermore, the survival rates over 1-, 3- and 5-year intervals in both cohorts yielded AUC values exceeding the 0.6 threshold (Fig. 3C and D). The study further identified that the seven genes deemed high-risk (*ANXA2, CCL20, CEBPB, CTSL, THBS1, EREG* and *SPP1*) showed increased expression in the high-risk group, indicating their potential as markers or contributors to heightened risk. Concurrently, heatmaps for the TCGA and GSE37745 datasets provided a clear depiction of these findings, as shown in Figure 3E and F. Additionally, a higher risk score consistently correlated with increased mortality and shorter survival, evident in both TCGA (Fig. 3G and I) and GSE37745 cohorts (Fig. 3H and J). The distribution plots of patient survival status, alongside survival analysis and ROC curves, displayed similar patterns across the respective risk groups in both cohorts. These findings collectively highlight the robustness and effectiveness of the TCMG risk score in predicting prognostic results in individuals diagnosed with NSCLC.

Influence of TCMGs on pathway enrichment in NSCLC risk

We employed gene sets from the GO and KEGG pathways for a comprehensive enrichment analysis of the DEGs linked with TCMG risk groups. Our GO analysis revealed distinct molecular involvement for each risk group. Genes associated with the high-risk group


A
Sensitivity
1−Specificity
Risk, AUC=0.640
Age, AUC=0.539
Gender, AUC=0.551
Stage, AUC=0.629
B
Sensitivity
1−Specificity
Risk, AUC=0.617
Age, AUC=0.585
Gender, AUC=0.519
Stage, AUC=0.553
C
Sensitivity
1−Specificity
AUC at 1 year: 0.640
AUC at 3 years: 0.669
AUC at 5 years: 0.626
D
Sensitivity
1−Specificity
AUC at 1 year: 0.612
AUC at 3 years: 0.643
AUC at 5 years: 0.617
E
type
high
low
ANXA2
CEBPB
HLA−DMB
CTSL
THBS1
EREG
TRBC1
CXCL17
CCL20
SPP1
F
type
high
low
ANXA2
CEBPB
HLA−DMB
CTSL
THBS1
EREG
TRBC1
CXCL17
CCL20
SPP1
G
High risk
Low risk
Risk score
Patients (increasing risk socre)
H
High risk
Low risk
Risk score
Patients (increasing risk socre)

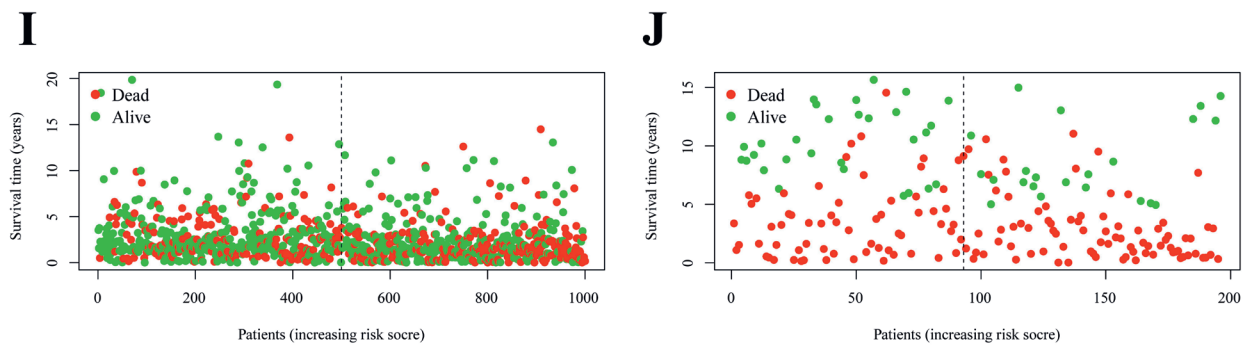


Fig. 3. Risk score analysis of the TCMG signature in the NSCLC cohort. (**A–B**) ROC curves predicting overall survival sensitivity and specificity based on different risk scores, age, gender and stage in TCGA (**A**) and GSE37745 (**B**); (**C–D**) OC curves predicting 1-, 3- and 5-year survival sensitivity and specificity based on different risk scores in TCGA (**C**) and GSE37745 (**D**); (**E–F**) expression characteristics of the 10 identified TCMGs in TCGA (**E**) and GSE37745; **(G–H)** distribution of risk scores in TCGA (**G**) and GSE37745 (**H**); (**I–J**) survival status of different risk subgroups in TCGA (**I**) and GSE37745 (**J**); (**K**) clinical characteristics of different risk subgroups in TCGA.

predominantly interacted with external protective matrices, such as the extracellular matrix and cellular membranes. Conversely, genes in the low-risk group were notably associated with immunoglobulin-related cellular components, indicating distinct biological roles, as shown in Figure 4A and B. KEGG pathway analysis provided additional insights, revealing a significant correlation between the high-risk gene group and pathways regulating cell cycle regulation, cytokine receptor interactions, extracellular matrix (ECM) receptor engagement, focal adhesion and NOD-like receptor signalling. These pathways suggest a complex interplay of cell proliferation, adhesion and immune signalling. Nevertheless, the low-risk gene group evidenced a remarkable propensity for metabolic pathways, highlighting differences in cellular metabolic activities (Fig. 4C and D).

Immune cell dynamics and TCMGs in NSCLC risk stratification

To elucidate the complex interplay between TCMGs and immune cell infiltration in NSCLC, we initially focused on the allocation of immune cells among different risk categories in the TCGA cohort (Fig. 5A). Utilizing CIBERSORT (http://cibersort.stanford.edu/), our analysis indicated that NSCLC patients with low-risk scores showed a significant increase in B memory cells, plasma cells, CD8+ T cells, monocytes and other specific immune cells, suggesting a distinct immunological profile. Conversely, these patients showed a reduced presence of resting NK cells, M0 macrophages, activated mast cells, eosinophils and neutrophils (Fig. 5B). Further extending our analysis, we compared immune function scores between the high-risk and low-risk groups. The data we obtained indicated that patients with low-risk scores had enhanced scores in several immune function areas, including antigen-presenting cells (APCs), B cells, CD8+ T cells, HLA, immature dendritic cells (IDCs), mast cells, plasmacytoid dendritic cells (PDCs), T-cell co-stimulation, T helper cells, T follicular helper cells (TFH), tumour-infiltrating lymphocytes (TIL) and type II interferon (Y-IFN) response. However, these patients showed decreased scores in chemokine receptor (CCR) expression, macrophages, para inflammation and Treg sectors (Fig. 5C). Additionally, our observations indicated markedly reduced expression of CD274, an essential immunoregulatory molecule, in the low-risk group (Fig. 5D). Intriguingly, our analysis revealed a distinct positive correlation between CD274 expression and the risk score, suggesting a potential involvement of CD274 in determining the prognostic landscape of NSCLC (Fig. 5E).

Discussion

In this study, we discovered a biomarker signature centred on the genetic markers of CD8+ T cells. These cells are crucial cytotoxic lymphocytes, instrumental in targeting and eradicating malignant cells primarily through induction of apoptosis. The combination of single-cell and bulk RNA sequencing enabled identification of this unique signature, highlighting the power of integrating genomic technologies. This development significantly advances our understanding of the essential role of CD8+ T cells in prognosis and personalized therapy for NSCLC patients.

In our detailed analysis of the scRNA-seq data from GSE117570, we utilized PCA to navigate the complexities of the dataset, identifying 16 distinct cell clusters. Significantly, one cluster was identified as particularly noteworthy, distinguished by a predominant composition of CD8+ T cells. The differential gene expression patterns within this cluster indicated the involvement of specific genes in modulating CD8+ T-cell activity. These

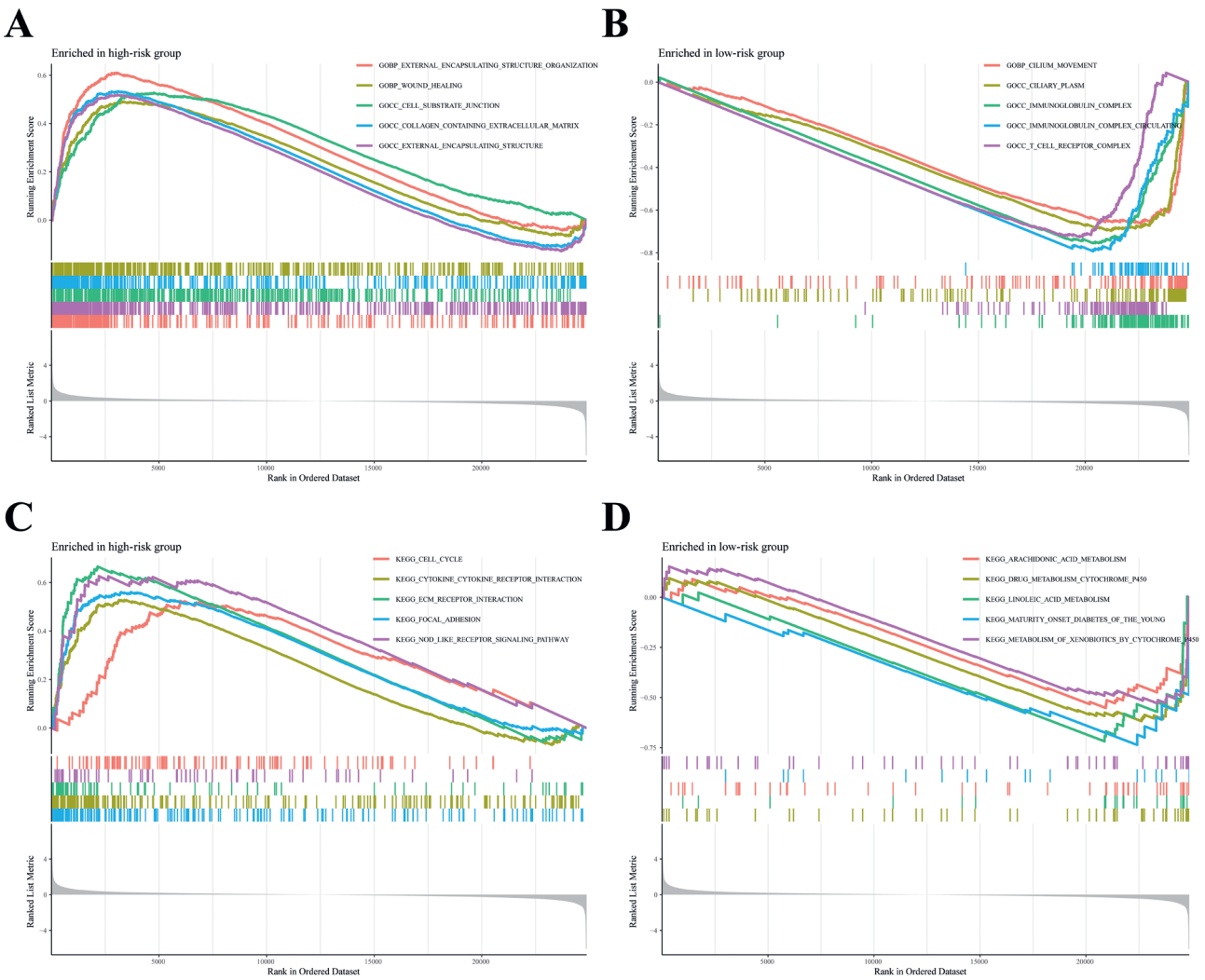


Fig. 4. Functional enrichment analyses were conducted on the transcriptome of TCMR risk groups. (**A–B**) The top five pathways with the highest normalized enrichment scores in GO gene sets were identified for both high-risk (**A**) and low-risk (**B**) groups; (**C–D**) the top five pathways with the highest normalized enrichment scores within the KEGG gene sets were determined for both high-risk (**C**) and low-risk (**D**) groups.

findings align with existing research, which has consistently highlighted the critical role of tumour-infiltrating $CD8^+$ T cells in NSCLC. Such cells are integral in determining the patient prognosis and in influencing the efficacy of immunotherapeutic responses (Zhuang et al., 2010; Donnem et al., 2015).

Our data-driven methodology successfully identified a powerful 10-gene prognostic signature for NSCLC, consisting of *ANXA2, CCL20, CEBPB, CTSL, THBS1, EREG, SPP1, HLA-DMB, CXCL17* and *TRBC1*. These genes distinctly stratify patients into high-risk and low-risk categories. Delving deeper into this signature, we observe a complex interplay among these genes, each uniquely associated with various cellular functions and pathways. Notably, annexin A2 *(ANXA2)* plays a multifaceted role in cellular processes such as proliferation, differentiation and migration, emerging as a key player in this signature (Wang et al., 2012; Wang and Lin, 2014; Feng et al., 2017). Epiregulin (EREG), an important epidermal growth factor receptor (EGFR) ligand, is involved in various cancers (Yun et al., 2012; Jing et al., 2016; He et al., 2019), including lung cancer (Sunaga and Kaira, 2015). Another significant gene, cathepsin L (*CTSL*), is known for its role in tumour invasion and metastasis (Olson and Joyce, 2015; Sudhan and Siemann, 2015). Utilizing LASSO Cox regression analysis, these 10 genes collectively form a robust biomarker signature. This signature not only correlates with survival rates but also provides insights into the molecular mechanisms underpinning NSCLC prognosis and response to immunotherapy. This discovery marks a significant advancement in personalized treatment strategies, offering a new perspective on the management of NSCLC.

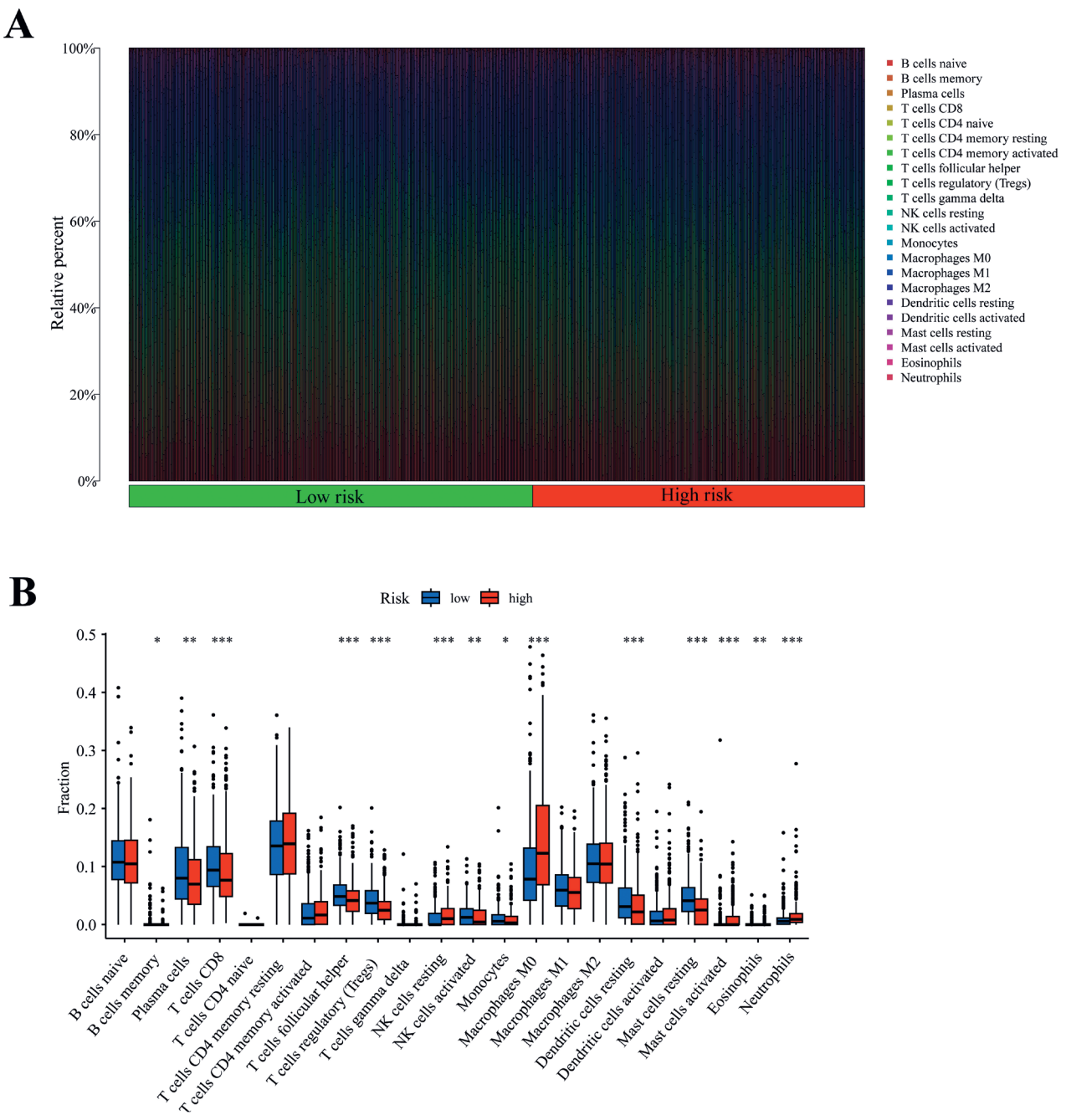


Fig. 5. Association between the TCMGs and the immune cell infiltration in NSCLC. (**A**) Distribution of immune cells in different risk subgroups in TCGA. (**B**) Differences in immune cell infiltration between the two risk subgroups.

Our validation process, utilizing the extensive TCGA and GEO datasets, has provided significant insights into the predictive accuracy of our novel gene signature. The consistency observed between the training (TCGA) and testing (GEO) cohorts underscores the reliability and potential clinical relevance of our findings. Additionally, the strong correlation between an elevated risk score, as determined by our signature, and increased mortality, along with decreased survival time, further emphasizes the clinical utility of this signature in prognosticating outcomes for NSCLC patients.

Our extensive enrichment analysis of DEGs within the high- and low-risk groups, based on our identified TCMGs, has provided a detailed understanding of their functional implications. This analysis has revealed a distinct division of cellular roles and pathways between these groups. These insights are crucial in elucidating the complex influence of TCMGs on the pathogenesis and progression of NSCLC. However, it is essential to emphasize the need for further in-depth research. Such research would aim to understand the biological significance of these findings fully and to explore their poten-

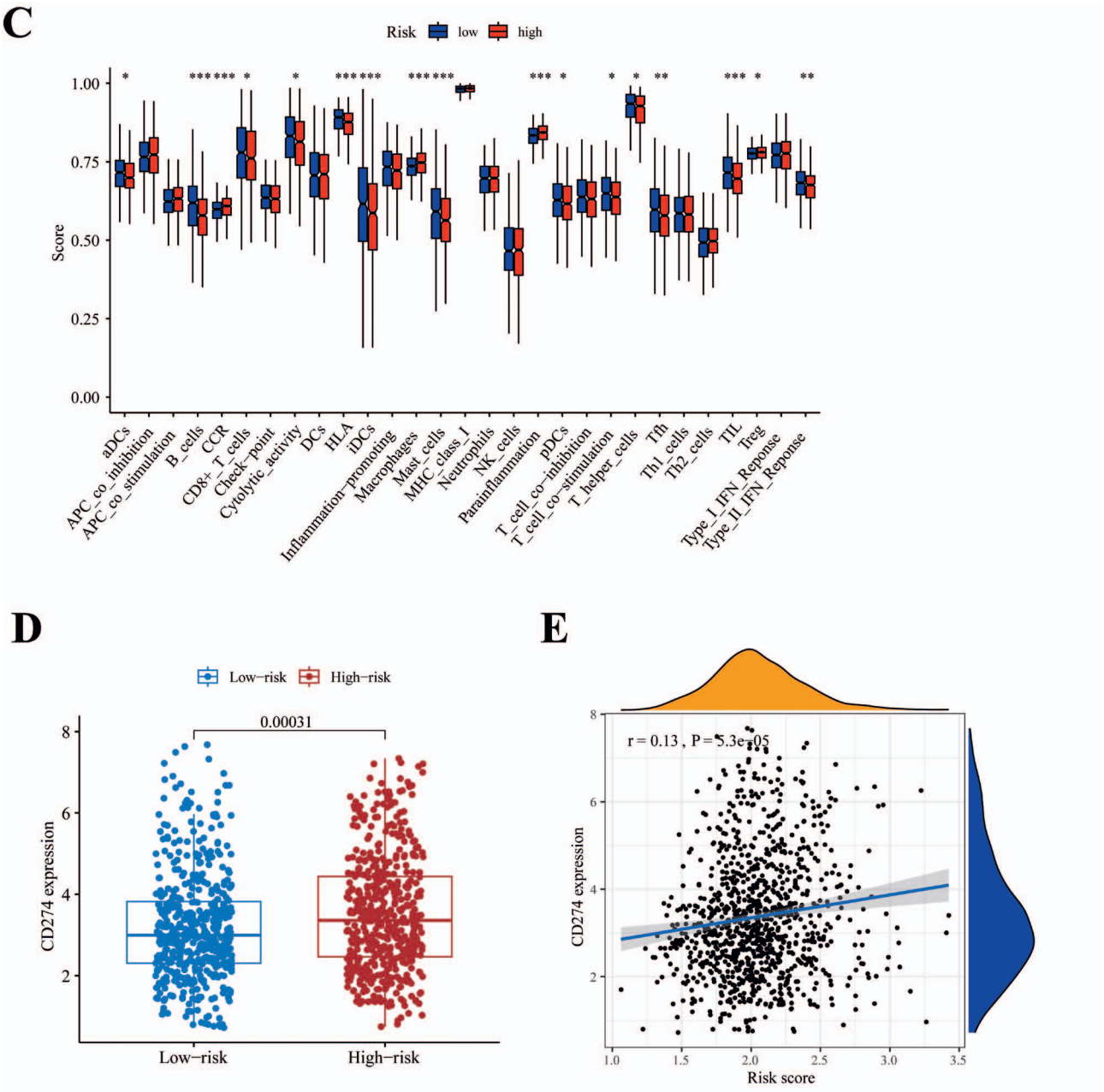


Fig. 5. Association between the TCMGs and the immune cell infiltration in NSCLC. (**C**) Differential immune function scores between the two risk subgroups. (**D**) CD274(PD-L1) expression in the two risk subgroups. (**E**) Correlation between CD274 expression and risk score.

tial in developing targeted therapeutic strategies for NSCLC.

Our study illuminates the complex interplay between TCMGs and the dynamic immune landscape in NSCLC. Notably, we observed distinct patterns of immune cell infiltration, including B memory cells, plasma cells and $CD8^+$ T cells, among others, which varied significantly between high-risk and low-risk patient groups. These variations highlight the significant influence of TCMGs on shaping the tumour microenvironment. Furthermore, the correlation between these variations in immune cell composition and risk stratification likely plays a critical role in the differing outcomes observed in NSCLC patients. This finding underscores the importance of our TCMGs in determining prognosis and influencing the treatment responses in NSCLC.

A particularly interesting finding of our study is the positive correlation between *CD274 (PDL1)* expression and the derived risk score. This observation gains significance in light of the extensive research on PD-L1 as a predictive biomarker for immunotherapy response in NSCLC. Given the PD-L1 established role as a key tar-

get in NSCLC immunotherapy (D'incecco et al., 2015; Schmidt et al., 2015), our TCMG signature potentially offers valuable insights for clinical decision-making, specifically in the use of PD-L1 inhibitors for NSCLC patients. However, it is crucial to acknowledge the need for further detailed research. This is particularly significant in light of the intricate and multifaceted nature of responses observed in immune checkpoint blockade therapy.

We must recognize the limitations of our study, as they highlight avenues for future research. The primary limitation is its retrospective nature, which introduces the possibility of selection bias. Therefore, the validation of our findings through well-designed, prospective studies involving diverse populations is crucial to confirm their validity and robustness. Additionally, there is a clear need for functional experiments. These experiments are essential to determine the specific roles of the identified TCMGs in the pathogenesis of NSCLC and their influence on the response to immunotherapeutic treatments.

In conclusion, our study significantly advances NSCLC research by identifying a novel 10-gene signature based on TCMGs. This signature represents a significant step forward in the prognostication of NSCLC and potentially guides immunotherapeutic strategies for this complex disease. Our findings contribute to a paradigm shift, enhancing our understanding of the dynamic relationship between $CD8^+$ T cells and NSCLC. Moving forward, these insights offer a basis for developing personalized and effective treatments, presenting promising avenues for patients confronting this formidable medical condition.

References

Barrett, T., Wilhite, S. E., Ledoux, P. et al. (2013) NCBI GEO: archive for functional genomics data sets – update. https://doi.org/10.1093/nar/gks1193

Bhattacharya, S., Andorf, S., Gomes, L. et al. (2014) ImmPort: disseminating data to the public for the future of immunology. *Immunol. Res.* **58**, 234-239.

Breuer, K., Foroushani, A. K., Laird, M. R. et al. (2013) InnateDB: systems biology of innate immunity and beyond – recent updates and continuing curation. *Nucleic Acids Res.* **41**, D1228-D1233.

Chen, H., Chong, W., Teng, C. et al. (2019) The immune response-related mutational signatures and driver genes in non-small-cell lung cancer. *Cancer Sci.* **110**, 2348-2356.

Chhikara, B. S., Parang, K. (2023) Global Cancer Statistics 2022: the trends projection analysis. *Chem. Biol. Lett.* **10**, 451.

Cortellini, A., Chiari, R., Ricciuti, B. et al. (2019) Correlations between the immune-related adverse events spectrum and efficacy of anti-PD1 immunotherapy in NSCLC patients. *Clin. Lung Cancer* **20**, 237-247.e1.

D'incecco, A., Andreozzi, M., Ludovini, V. et al. (2015) PD-1 and PD-L1 expression in molecularly selected non-small-cell lung cancer patients. *Br. J. Cancer* **112**, 95-102.

Donnem, T., Hald, S. M., Paulsen, E. E. et al. (2015) Stromal $CD8^+$ T-cell density – a promising supplement to TNM staging in non-small cell lung cancer. *Clin. Cancer Res.* **21**, 2635-2643.

Feng, X., Liu, H., Zhang, Z. et al. (2017) Annexin A2 contributes to cisplatin resistance by activation of JNK-p53 pathway in non-small cell lung cancer cells. *J. Exp. Clin. Cancer Res.* **36**, 123.

Gentleman, R. C., Carey, V. J., Bates, D. M. et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.

Han, Y., Wang, Y., Dong, X. et al. (2023) TISCH2: expanded datasets and new tools for single-cell transcriptome analyses of the tumour microenvironment. *Nucleic Acids Res.* **51**, D1425-D1431.

He, M., Jin, Q., Chen, C. et al. (2019) The miR-186-3p/EREG axis orchestrates tamoxifen resistance and aerobic glycolysis in breast cancer cells. *Oncogene* **38**, 5551-5565.

Heagerty, P. J., Zheng, Y. (2005) Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92-105.

Jing, C., Jin, Y. H., You, Z. et al. (2016) Prognostic value of amphiregulin and epiregulin mRNA expression in metastatic colorectal cancer patients. *Oncotarget* **7**, 55890-55899.

Krishna, S., Lowery, F. J., Copeland, A. R. et al. (2020) Stem-like CD8 T cells mediate response of adoptive cell immunotherapy against human cancer. *Science* **370**, 1328-1334.

Liu, J., Chang, H. W., Huang, Z. M. et al. (2021) Single-cell RNA sequencing of psoriatic skin identifies pathogenic Tc17 cell subsets and reveals distinctions between $CD8^+$ T cells in autoimmunity and cancer. *J. Allergy Clin. Immunol.* **147**, 2370-2380.

Lurienne, L., Cervesi, J., Duhalde, L. et al. (2020) NSCLC immunotherapy efficacy and antibiotic use: a systematic review and meta-analysis. *J. Thorac. Oncol.* **15**, 1147-1159.

Mariathasan, S., Turley, S. J., Nickles, D. et al. (2018) TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544-548.

Newman, A. M., Steen, C. B., Liu, C. L. et al. (2019) Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773-782.

Olson, O. C., Joyce, J. A. (2015) Cysteine cathepsin proteases: regulators of cancer progression and therapeutic response. *Nat. Rev. Cancer* **15**, 712-729.

Philip, M., Schietinger, A. (2022) $CD8^+$ T cell differentiation and dysfunction in cancer. *Nat. Rev. Immunol.* **22**, 209-223.

Qu, J., Jiang, M., Wang, L. et al. (2020) Mechanism and potential predictive biomarkers of immune checkpoint inhibitors in NSCLC. *Biomed. Pharmacother.* **127**, 109996.

Raskov, H., Orhan, A., Christensen, J. P. et al. (2021) Cytotoxic $CD8^+$ T cells in cancer and cancer immunotherapy. *Br. J. Cancer* **124**, 359-367.

Schmidt, L. H., Kümmel, A., Görlich, D. et al. (2015) PD-1 and PD-L1 expression in NSCLC indicate a favorable prognosis in defined subgroups. *PLoS One* **10**, e0136023.

Spigel, D. R., Faivre-Finn, C., Gray, J. E. et al. (2022) Five-year survival outcomes from the PACIFIC trial: durvalumab after chemoradiotherapy in stage III non-small-cell lung cancer. *J. Clin. Oncol.* **40**, 1301-1311.

Sudhan, D. R., Siemann, D. W. (2015) Cathepsin L targeting in cancer treatment. *Pharmacol. Ther.* **155**, 105-116.

Sunaga, N., Kaira, K. (2015) Epiregulin as a therapeutic target in non-small-cell lung cancer. *Lung Cancer (Auckl.)* **6**, 91-98.

Tanoue, T., Morita, S., Plichta, D. R. et al. (2019) A defined commensal consortium elicits CD8 T cells and anti-cancer immunity. *Nature* **565**, 600-605.

Tibshirani, R. (1997) The lasso method for variable selection in the Cox model. *Stat. Med.* **16**, 385-395.

Wagner, G. P., Kin, K., Lynch, V. J. (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281-285.

Wang, C. Y., Chen, C. L., Tseng, Y. L. et al. (2012) Annexin A2 silencing induces G2 arrest of non-small cell lung cancer cells through p53-dependent and -independent mechanisms. *J. Biol. Chem.* **287**, 32512-32524.

Wang, C. Y., Lin, C. F. (2014) Annexin A2: its molecular regulation and cellular expression in cancer development. *Dis. Markers* **2014**, https://doi.org/10.1155/2014/308976

Yu, G., Wang, L. G., Han, Y. et al. (2012) clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287.

Yun, J., Song, S. H., Park, J. et al. (2012) Gene silencing of EREG mediated by DNA methylation and histone modification in human gastric cancers. *Lab. Invest.* **92**, 1033-1044.

Zhuang, X., Xia, X., Wang, C. et al. (2010) A high number of $CD8^+$ T cells infiltrated in NSCLC tissues is associated with a favorable prognosis. *Appl. Immunohistochem. Mol. Morphol.* **18**, 24-28.