



## OPEN A three-subtype prognostic classification based on base excision repair and oxidative stress genes in lung adenocarcinoma and its relationship with tumor microenvironment

Wen Rao<sup>1,2,5</sup>, Qin Zhang<sup>3,5</sup>, Xiaoyan Dai<sup>1</sup>, Yuxin Yang<sup>1</sup>, Zhang Lei<sup>1</sup>, Xunjie Kuang<sup>1</sup>, He Xiao<sup>1</sup>, Jianwu Zhu<sup>1</sup>, Yanli Xiong<sup>1</sup>, Dong Wang<sup>4</sup>✉ & Lujie Yang<sup>1</sup>✉

Unrepaired DNA damage is the initiation of mutation and tumor-specific biological characteristics. Oxidative stress and base excision repair (BER) are the two main pathways to cope with oxidative DNA damage, which is closely related to the heterogeneity of Lung adenocarcinoma (LUAD), but their relationship with tumor biological characteristics is unclear, and a molecular subtyping based on comprehensive BER and oxidative stress gene expression is lacking. 501 samples from The Cancer Genome Atlas (TCGA) were classified into three subtypes based on genes related to BER and oxidative stress through hierarchical agglomerative cluster analysis. By integrating the nearest template prediction (NTP), four GEO datasets and 52 samples from our institution were analyzed for validation. Bioinformatic analysis was performed to define the diverse molecular characteristics, mutation background, tumor microenvironment, and prognosis. Three subtypes with distinct gene signatures were identified: relatively high BER and low oxidative stress gene expression (C1), low BER gene and high oxidative stress gene expression (C2), and high expression of both BER and oxidative stress genes (C3). C2 was characterized by a low mutation frequency in TP53 (29%) and a high mutation frequency in EGFR (20%), whereas a high frequency of mutation was seen in C3 in STK11 and KEAP1 genes. Additionally, differentially expressed genes among the three subtypes were particularly enriched in immune-related pathways, and the abundance of immune cells and Immunophenoscore were significantly higher in C2, while the Tumor Immune Dysfunction and Exclusion (TIDE) score was lower in C2, indicating a better response to immunotherapy. C2 was also associated with an improved survival outcome compared with C1 and C3, and this finding was validated in 978 samples from four independent GEO datasets and 52 samples at our institution by the NTP algorithm. The three-subtype classifications based on BER and oxidative stress gene expression offers potential for predicting the survival and response to immunotherapy of LUAD patients.

**Keywords** Lung adenocarcinoma, Base excision repair, Oxidative stress, Prognosis, Tumor microenvironment

Lung cancer is the second most common morbidity of malignance and the leading cause of cancer-related death worldwide. Lung adenocarcinoma (LUAD) is the most prevalent histological subtype, making up approximately

<sup>1</sup>Cancer Center, Daping Hospital and Army Medical Center of PLA, Army Medical University, No.10 Changjiangzhi Rd, Yuzhong District, Chongqing, People's Republic of China. <sup>2</sup>The 75th Group Army Hospital, Dali, Yunnan, People's Republic of China. <sup>3</sup>Department of Radiation Oncology, Sichuan Cancer Hospital and Institute, Sichuan Cancer Center, Cancer Hospital Affiliate to School of Medicine, University of Electronic Science and Technology of China, Sichuan, People's Republic of China. <sup>4</sup>Chongqing University Qianjiang Hospital, Chongqing, People's Republic of China. <sup>5</sup>Wen Rao and Qin Zhang shared first authorship. ✉email: dongwang64@hotmail.com; yanglujie@tmmu.edu.cn

40% of lung cancer<sup>1</sup>. The development of molecular targeted agents and immune checkpoint inhibitors (ICIs) has brought about revolutionary progress to advanced-stage LUAD, and many biomarkers have been uncovered to guide the treatment, such as epidermal growth factor receptor (EGFR) and programmed cell death-1 (PD-1)/programmed cell death ligand-1 (PD-L1)<sup>2</sup>. However, due to tumor heterogeneity, these biomarkers exhibit varying predictive abilities regarding treatment response and prognosis across different patients. Thus, it is an urgent priority to further investigate the heterogeneous biological characteristics of LUAD and establish reliable molecular classifications that distinguish the prognosis and treatment of LUAD patients.

The development of cancer encompasses a multistage, multistep process involving cellular and molecular alterations. Lung tissues are consistently exposed to intracellular metabolic oxidants and environmental toxicants<sup>3</sup>, such as cigarette smoke and polluted air exposure. Antioxidant defense system play a prominent role in counteracting oxidative stress. An imbalance in oxidative agents and antioxidant defense mechanisms leads to excessive reactive chemical species [e.g. reactive oxygen (ROS) and reactive nitrogen (RNS)], which can damage macromolecules including lipids, proteins and nucleic acids<sup>4</sup>. Previous experimental evidence have indicated that exposure to ROS can lead to genomic damage and subsequent cellular transformation in mouse fibroblasts<sup>5</sup>, and a deficiency in the antioxidant superoxide dismutase enzyme can result in dramatically increased cancer phenotypes in mice<sup>6</sup>. Additionally, studies have shown that oxidative stress is heightened and antioxidant molecule levels are diminished in the advanced stages of lung cancer<sup>7,8</sup>.

Multiple researches have demonstrated that nuclear factor erythroid 2-like factor 2 (NRF2), a master regulator of the antioxidant enzymes, plays a central role in redox homeostasis and protecting against tumorigenesis<sup>9,10</sup>. Intriguingly, recent evidence also shows that tumors can increase the transcription of antioxidant genes to maintain oxidative homeostasis, a process that relates to cancer development and metastases. For instance, 30% of lung cancers increase the transcription of antioxidant genes by acquiring stabilizing mutations in NRF2. Constitutive activation of NRF2 involves in promoting cancer development and metastasis<sup>11</sup>. Furthermore, compared with normal tissue and tumor-adjacent tissue, the expression of SOD2 is significantly upregulated in metastases, and it is thought to be a marker of progression from growth to metastasis<sup>12</sup>. Combined, the balance between the oxidative stress and antioxidants in tumorigenesis and progression of lung cancer remains mostly unknown.

Unrepaired DNA damage induced by oxidative stress is a major cause of mutations and carcinogenesis. Of all types of DNA damage, oxidative damage is the most common type, and unrepaired oxidative damage can convert to fatal DNA double-strand break (DSB). Hence, base excision repair (BER) pathway, the major contributor in repairing DNA lesions induced by oxidative stress, is essential for the instability of genome<sup>13</sup>. Many studies have revealed the defects in the BER pathway are associated with the initiation and progression of tumors. For instance, MYH/OGG1 double mutant mice have elevated susceptibility to tumor initiation, likely due to the accumulation of 8-oxoguanine base damage in the genome<sup>14</sup>. Moreover, inherited mutations in DNA glycosylases MUTYH and NTHL1, which resolve oxidative base lesions, are linked to an increased risk of colorectal cancer<sup>15</sup>. In non-small cell lung cancer (NSCLC), evidence indicates that gene polymorphisms in BER are associated with the risk of tumorigenesis and the survival of patients<sup>16,17</sup>. Furthermore, enhanced BER and antioxidant activity can promote therapeutic resistance to certain genotoxic agents in tumor cells<sup>18,19</sup>. Collectively, experimental and clinical studies demonstrate that genes involved in oxidative damage repair pathways are pivotal in initiation or progression of cancer, including LUAD.

Gene expression based subtyping is widely accepted as a relevant means of disease stratification<sup>20,21</sup>. However, no established molecular subtypes based on comprehensive oxidative stress and BER pathway gene expression patterns have been proposed in LUAD. In this study, we constructed a three-subtype molecular classification for LUAD based on BER and oxidative stress related genes. Subsequent analyses demonstrated that these three molecular subtypes are associated with distinct clinical characteristics, mutational backgrounds, tumor microenvironments and prognosis. The classification can help identify new prognostic subgroups and could provide a basis for selecting specific targets for future drug development.

## Materials and methods

### Patients and datasets

The RNA-seq and mutational data of LUAD patients and corresponding clinical data were downloaded from TCGA (<https://tcga-data.nci.nih.gov/tcga/>). Gene expression profile based on microarray data and paired clinical information were retrieved from GEO database (<http://www.ncbi.nlm.nih.gov/geo/>). After removing patients with non-LUAD and incomplete follow-up data, we collected 501 samples from TCGA as the training set, and 978 samples from four GEO datasets as validation cohorts (226 samples from GSE31210<sup>22</sup>, 128 samples from GSE50081<sup>23</sup>, 442 samples from GSE68465<sup>24</sup>, 182 samples from GSE41271<sup>25</sup>). GSE31210 and GSE50081 were based on GPL570 platform, while GSE68465 was based on GPL96 platform and GSE41271 was based on GPL6884 platform. In addition, 52 LUAD patients who were pathologically diagnosed and surgically resected in Daping hospital from 2011 to 2013 were enrolled in this study. The TNM staging of all patients was T<sub>1-2</sub>N<sub>1-2</sub>M<sub>0</sub>, which was evaluated based on the International Union Against Cancer (UICC) 8th Edition TNM lung cancer staging criteria. RNA sequencing of the 52 patients was performed with Illumina Hiseq 2500. The genes were retained that had at least 10 counts for 50% samples and converted into FPKM by using annotation file “Biomart.annotations.hg38” and function “countToFPKM” from R package “fpkm”. The format of log2 scale in expression matrix was used for the subsequent statistical analysis. The detailed clinical data of the patients was shown in Supplementary Table 1.

### Unsupervised clustering based on BER and oxidative stress related genes

To identify different patterns mediated by BER and oxidative stress genes, a total of 37 BER genes and 30 oxidative stress genes were obtained from GeneCards database (<https://www.genecards.org/>) (Supplementary Table 2). Hierarchical agglomerative clustering was used for sample clustering using the R package “ConsensusClusterPlus”.

### Analysis of differentially expressed genes (DEGs) and mutational profile

R packages “limma” was used to identify DEGs between each cluster and the filter criteria was  $|\log_2 \text{Foldchange (FC)}| > 1$  and false discovery rate (FDR)  $< 0.01$ . Venn diagram was used to display the overlapping genes of DEGs. R package “maftools” was used to perform mutational profile of each cluster.

### Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) enrichment analysis

KEGG and GO enrichment analysis were conducted by The Database for Annotation, Visualization and Integrated Discovery (DAVID, <https://david.ncicrf.gov/>) to reveal the key pathway and unique biological significance of overlapping DEGs. Fisher’s exact test was used to identify the significant KEGG pathway and GO terms (FDR  $< 0.05$ ). R package “ggplot2” was used for visualization of KEGG and GO enrichment.

### Analysis of the protein-protein (PPI) network

The Retrieval of Interacting Genes (STRING) (<http://string-db.org/>) was used to analyze the interactive relationships among overlapping DEGs to construct protein-protein interaction (PPI) network and a combined score  $> 0.4$  were selected as statistically significant. Cytoscape was used to construct the PPI network map. The plug-in Molecular Complex Detection (MCODE) was used to select the prime module from the PPI network. The plug-in cytoHubba of Cytoscape was used to analyze and identify the top 7 crucial genes in the PPI network.

### Evaluation of tumor microenvironment and immunotherapy efficacy

xCell (<https://xcell.ucsf.edu/>) and CIBERSORT (<https://cibersortx.stanford.edu/>) algorithms were used to evaluate the levels of infiltration immune cells of each LUAD sample, while xCell and ESTIMATE (<https://bioinformatics.mdanderson.org/estimate/>) algorithm was applied to perform the calculation of the immune score, stromal score and estimate score based on the proportion of immune cells and stromal cells<sup>26,27</sup>. The Tumor Immune Dysfunction and Exclusion (TIDE) algorithm was utilized to model the primary mechanisms of tumor immune evasion and had been applied to predict the immunotherapy efficacy in many types of solid tumor, especially in LUAD and melanoma<sup>28</sup>. And the Immunophenoscore (IPS) from The Cancer Immunome Atlas (TCIA) database (<https://tcia.at/>) was used to predict the immunotherapy response (anti-PD-1 and anti-CTLA 4)<sup>29</sup>.

### Nearest template prediction validation

Nearest template prediction (NTP) algorithm was utilized to make cluster prediction confidence for each single patient, which provided a method to flexibly apply to cross-platform, cross-species, and multiclass predictions without any optimization of analysis parameters<sup>30,31</sup>. In the validation cohort of LUAD in GEO datasets and our institution, NTP algorithm was used. We use R package CMScaller (<https://github.com/Lothelab/CMScaller>) to predict subtypes of any independent cohorts through NTP. This package is originally designed for classification of CMS. The original R scripts for CMS classification in the package CMScaller contain all functions that are essential to carry out NTP algorithm. Therefore, these main functions were directly adopted for our classification without further modification in R scripts. Briefly, “subDEG” was firstly utilized to identify DEGs of each originally assigned subtype compared against all rest subtypes in order to generate template markers for each subtype. The template for each subtype was generated by using function “ntpMakeTemplates” with top 100 DEGs to average the expression values of template markers across original samples within one consensus subtype. Finally, function “ntp” was applied to each sample in an independent normalized expression matrix. The distances between expression values of components within three templates and corresponding expression profile of any given single sample was measured with cosine distance. Each sample was assigned with subtype that had the nearest cosine distance with that subtype template. Permutation of 1000 times was used to obtain the estimation of false discovery rate for each sample. Only samples with FDR  $< 0.05$  were included into the subsequent statistical analysis.

### Tumor mutation burden (TMB) calculation and comparison

The TMB data of all 501 LUAD samples were downloaded from <http://www.cbioportal.org/>. According to the record of samples in TCGA database, only non-synonymous including missense mutation, nonsense mutation, splice site, frame shift insertion, frame shift indel, in frame deletion, in frame insertion and multi-hit mutation were included in our calculation of mutation counts, and the differences of TMB among the three groups were compared using Kruskal-Wallis test.

### Statistic

All statistical analyses were performed using R or GraphPad Prism version 7.0 (GraphPad, La Jolla, CA, USA). Comparisons of proportions were used Fisher’s exact test or Chi-square test, as appropriate. Kruskal-Wallis test was performed to compare the difference among three groups. The Kaplan-Meier analysis was conducted by the “survminer” package. The log-rank test was applied to compare the survival statistics of categorical variables. All statistical tests were two-sided.  $P < 0.05$  was considered statistically significant.

## Results

### Identification of the three LUAD molecular subtypes based on BER and oxidative stress gene expression

Accumulating evidence indicates dysregulation of BER and oxidative stress capacity in LUAD<sup>32–34</sup>. The workflow diagram was shown in Supplementary Fig. 1. To confirm these findings, we firstly assessed the expression and mutation profile of BER and oxidative stress genes in LUAD tumor samples and adjacent tissues in the TCGA database (Supplementary Fig. 2A, B). Compared to adjacent normal tissues, 78.4% of genes (29/37) in the BER pathway (e.g., APEX1, APTX and XRCC1) and 30% of genes (9/30) in the oxidative stress pathway (e.g., GCLC, GSR and MGST1) were significantly overexpressed in tumor samples. In contrast, half of the oxidative stress genes (15/30) and only three BER genes were significantly downregulated in tumor samples (Supplementary Fig. 2A, B). However, the mutation profiles of BER and oxidative stress genes were similar between tumor samples and adjacent tissues, with all genes having a mutation frequency of less than 5% (Supplementary Fig. 2C, D). These results indicated that BER and oxidative stress genes are relatively preserved at the nucleotide level, whereas their expression showed considerable heterogeneity in LUAD.

Next, we performed an unsupervised clustering analysis to distinguish 501 tumor samples from TCGA database based on the expression of BER and oxidative stress genes. The results showed that all samples can be distinctly classified into three subtypes (Fig. 1A). Cumulative distribution function (CDF) was utilized to identify the optimal number of clusters, which was defined by the smallest difference of the area under consensus cluster. This analysis indicated the optimal number of clusters was 3 (Fig. 1B). Principal component analysis confirmed the significant differences in gene expression among the three subtypes (Fig. 1C). As presented in the unsupervised clustering heatmap of Fig. 1D, the three LUAD subtypes were labeled as follows: C1 has relatively high expression of BER genes and low expression of oxidative stress genes; C2 has low expression of BER genes and high expression of oxidative stress genes; C3 has high expression of both BER and oxidative stress genes. Specifically, 81.1% of BER genes (30/37) and 86.6% of oxidative stress genes (26/30) exhibited significant differential expression among the three subtypes. For the 30 BER differentially expressed genes (DEGs), the expression pattern were similar between C1 and C3, with only 2 genes overexpressed and 6 genes underexpressed in C1 compared to C3, respectively. Conversely, compared to C1 and C3, we observed that 70% BER DEGs (21/30) had significantly lower expression (including APEX1, APTX, FEN1, etc.), while only 13.3% BER-related DEGs (4/30) had higher expression in C2 (including APF and NEIL1). Combined, we characterized C1 and C3 as having relatively high expression of BER DEGs, while C2 by relatively low expression of BER DEGs among the three subtypes. As for oxidative stress DEGs, the relatively high expression genes accounted to 11.5% (3/26), 46.2% (12/26) and 46.2% (12/26) in C1, C2, C3, respectively. Combined, C1 had a relatively low expression in oxidative stress DEGs compared with C2 and C3. The expression of single gene among the three subtypes was shown in Supplementary Table 3.

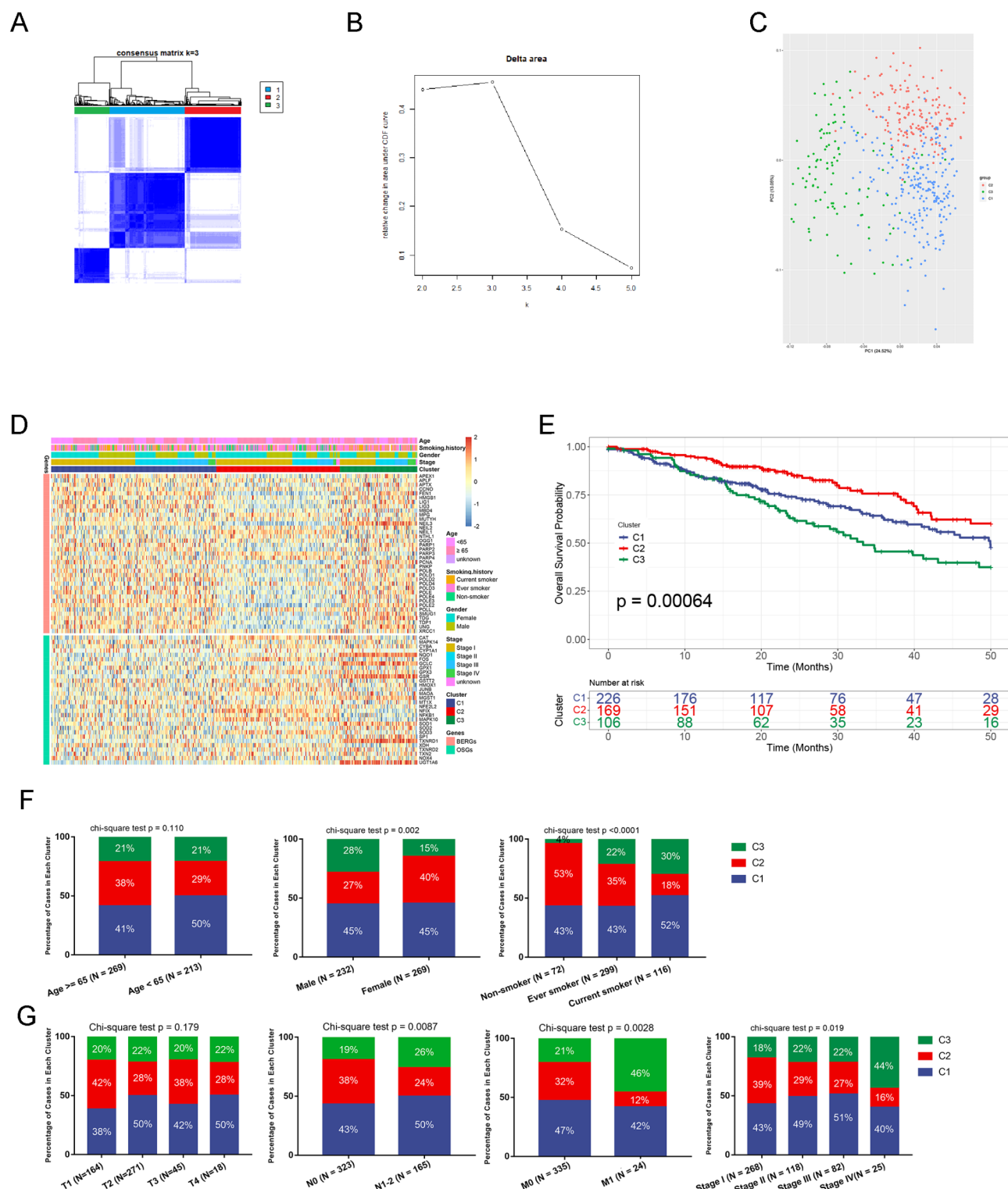
Finally, we investigated the biological and clinical ramifications of the three subtypes. We first analyzed whether the LUAD subtypes would predict the clinic outcome of patients. The log-rank test comparisons among the three molecular subtypes revealed statistically significant prognostic stratification, with the C2 subgroup demonstrating the most favorable survival outcome (C1 VS C2  $P=0.0265$ , C1 vs. C3  $P=0.0615$ , C2 vs. C3  $P=0.0002$ ) (Fig. 1E). Second, we analyzed the association between the LUAD subtypes and specific clinical characteristics. No significant difference was observed in age among the three subtypes ( $P>0.05$ ). However, when considering gender and smoking, a significantly higher proportion of females (44%) and non-smokers (53%) was observed in the C2 subtype (Fig. 1F). In contrast, C1 had a higher proportion of males and smokers (all  $P<0.05$ ). Third, we investigated the association of LUAD subtypes and pathological characteristics. As shown in Fig. 1G, no significant difference was observed among the three subtypes with respect to the T stage. Conversely, the N stage, M stage and clinical stage of the C2 classification were generally earlier than those of the other two subtypes, with 44% of C3 patients being staged as stage IV (all  $P<0.05$ ). Taken together, the results suggested that the subtypes based on BER and oxidative stress gene expression were generally distinct, a feature that might help identify new prognostic subgroups in LUAD.

### Mutation profiles of the three subtypes

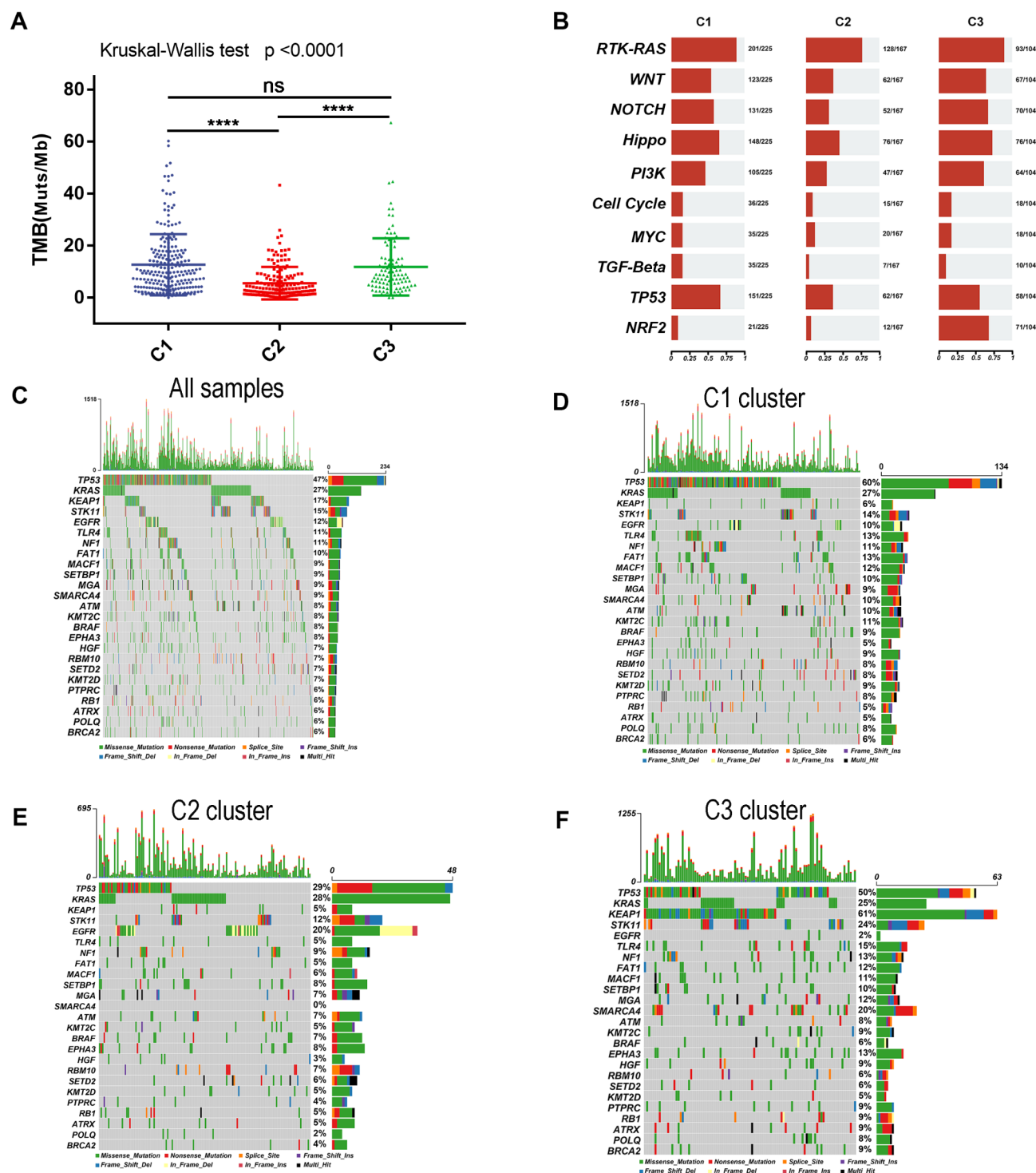
Genetic alterations are closely associated with distinct molecular characteristics across several tumor types<sup>35,36</sup>. To explore the molecular features of the three LUAD subtypes, the mutational landscape of all 501 samples from TCGA database were charted. We first evaluated the tumor mutation burden (TMB) of each sample. The results showed that C2 has a significantly lower TMB compared with C1 and C3 ( $P<0.001$ ) (Fig. 2A). Second, we analyzed whether cancer-related signaling pathways were enriched in the different subtypes by analyzing genetic mutation profiles. Consistent with the results of TMB, we revealed that C2 was lowly enriched in all signaling pathways. Interestingly, the C3 subtype was enriched in the NRF2 pathway, whereas neither C1 nor C2 was enriched in this response (Fig. 2B).

Additional mutational analyses revealed that TP53 (47%), KRAS (27%), KEAP1 (17%), STK11 (15%) and EGFR (12%) had the highest mutant frequency for all samples, and that single nucleotide variation was the most common mutation type (Fig. 2C). Nevertheless, the mutant frequency of these 5 genes varied among the different subtypes. TP53 mutations were significantly more frequent in the C1 (60%) and C3 (50%) subtypes but lower in C2 (29%). KRAS mutations were equally distributed among the three subtypes (C1 27%; C2 28% and C3 25%). The highest mutation frequency involving STK11 was seen in C3 (24%), compared to C1 (14%) and C2 (12%). C3 had the highest mutation frequency of KEAP1 (61%), a finding generally consistent with prior work demonstrating that co-occurring mutant patterns in STK11 and KEAP1 are common in tumors. Notably, EGFR mutations were significantly higher only in C2 (20%) (all  $P<0.05$ ) (Fig. 2D, F and Supplementary Table 4). Collectively, the results indicate that the three subtypes have markedly different mutation patterns, which could contribute to the subtype heterogeneity.





**Fig. 1.** Construction of the three-subtype classification based on BER and oxidative stress genes and their characteristics of LUADs in TCGA database. **(A)** Consensus matrix plot of all 501 LUAD samples (k=3). **(B)** Consensus cluster CDF curve of three to five classification. **(C)** Principal component analysis of all 501 LUAD samples. **(D)** Unsupervised heatmap clustering of all samples based on BER and oxidative stress genes. **(E)** Kaplan–Meier plot of the three subtypes (up) and their follow-up time (down). **(F)** Clinical characteristics of the three subtypes including age, gender and smoking. **(G)** Pathological characteristics of the three subtypes including T stage, N stage, M stage and total stage.



**Fig. 2.** Mutation profile analysis of the different three subtypes in TCGA database. **(A)** Tumor mutation burden of the three subtypes. **(B)** Cancer-related signal pathways enrichment based on mutation profile. **(C)** Mutation frequency and mutation type of common mutated genes in all samples. **(D–F)** Mutation frequency and mutation type of common mutated genes in C1(D), C2(E) and C3(F) subtypes. The T test was used to determine differences between tumor and normal tissues. A  $p$  value of  $< 0.05$  /  $< 0.01$  /  $< 0.001$  /  $< 0.0001$  denoted by an asterisk \* / \*\* / \*\*\* / \*\*\*\* in figures.

### Gene expression characteristics of the three subtypes

It stands to reason that variation in the tumor-associated mutation patterns could lead to distinct expression signatures. To explore this, we investigated the DEGs among the three subtypes to define the gene expression characteristics of each. A total of 271 genes were identified as DEGs (Supplementary Fig. 3), and the Kyoto

Encyclopedia of Genes (KEGG) and Genomes (GO) were used to uncover enriched signaling pathways associated with those DEGs. KEGG enrichment analysis showed that the DEGs were predominantly enriched in immune-related pathways and diseases, such as asthma, graft versus host disease, and antigen processing and presentation (Fig. 3A). Meanwhile, GO enrichment analysis, categorized into Biological Process (BP), Cellular Component (CC) and Molecular Function (MF) also highlighted pathways involved in immunity. For example, antigen processing and presentation, interferon-gamma-mediated signaling and the global immune response were found to be significantly enriched by BP analysis. MHC class II protein complex, plasma membrane and transport vesicle membrane were enriched as defined by CC. MHC class II receptor activity, peptide antigen binding and lipopeptide binding were uncovered by MF enrichment (Fig. 3B).

To further define the central DEGs among the three subtypes, we performed a Protein-Protein Interaction (PPI) network analysis (Fig. 3C). The results identified seven of the 271 genes, which were revealed as the DEGs with the strongest interaction (i.e., CD33, CD74, CD69, CD1C, TLR2, TLR7, and MRC1), and also tightly associate with tumor immune regulation (Fig. 3D). It should be noted that the highest expression of all seven genes was observed in C2, while the lowest expression was found in C3 (Fig. 3E). These results indicated that there is significant variability in tumor immune regulation across the three subtypes, where C2 might be an immune-active subtype, and C1 and C3 may represent immune-suppressive subtypes with low expression of immune-related gene clusters.

### Tumor microenvironment analysis of the three subtypes

We next investigated whether the LUAD subtypes were associated with distinct tumor immune microenvironments. xCell was used to evaluate the infiltration of immune cells. As shown in Fig. 4A, the abundance of majority immune cells [e.g., B cell, dendritic cell (DC), naïve CD4 T cell] was significant higher in C2, but lower in C3 among the three subtypes. Conversely, infiltration of Treg, Th1 and Th2 cells was lower in C2 (all  $P < 0.05$ ) (Supplementary Table 5). The highest Immune Score, Stroma Score and Microenvironment Score was observed in C2, whereas C3 had the lowest scores among the different subtypes (Fig. 4B). TIDE is an algorithm to evaluate the possibility of tumor immune escape based on the gene expression profile of tumor samples. Our results showed that the C2 subtype had the lowest TIDE score, indicating the least likelihood of tumor immune evasion (Fig. 4C). IPS scores were used as a general indicator of immunogenetic and immunotherapeutic response across samples. In line with the TIDE result, C2 had the highest IPS score, suggesting an optimal tumor immune microenvironment in C2 (Fig. 4D). CIBERSORT was used to further evaluate the fraction of stromal and immune cells in tumor samples. Our result demonstrated that memory B cell and dendritic cells were significantly enriched in C2 (Supplementary Fig. 4A). By the ESTIMATE algorithm, we calculated that C2 had the highest Immune Score, Stroma Score and ESTIMATE score, whereas C3 was determined to have the poorest infiltration of major immune cells among the three subtypes (Supplementary Fig. 4B). In summary, our results revealed that LUAD subtypes are highly associated with the tumor immune microenvironment, which are likely to influence the immunotherapeutic response.

### Validation of the prognosis and central DEGs in four GEO datasets

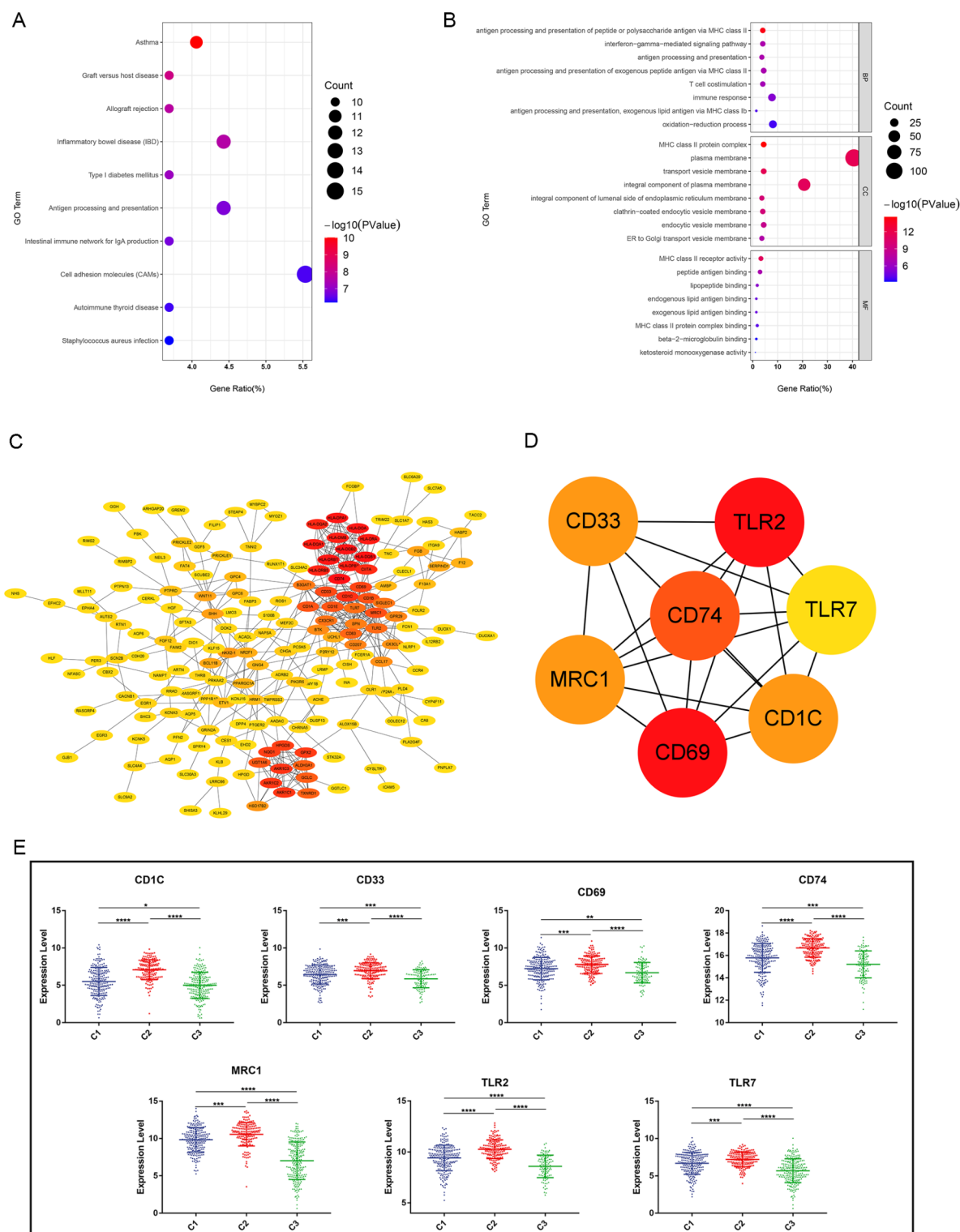
Four GEO LUAD datasets, i.e., GSE31210, GSE50081, GSE68465 and GSE41271, spanning a total of 978 samples, were included in the validation cohort to verify the earlier prognosis and DEG observations in the three subtypes (Fig. 5). Since the sample number of the validation cohort was relatively small, in order to yield reliable clustering harboring similar biological features with the TCGA training set, the NTP algorithm was utilized to mapped the genomic signatures of the different subtypes onto the four independent LUAD cohort datasets. In our study, template genes were derived from DEGs among the three subtypes, 130 template genes were optimal to reach maximal number of successful assignments via NTP algorithm. 30 genes in C1 and 50 genes in C2 and C3 were identified as template gene sets to define the subtype of the samples in the GEO database (Supplementary Table 6). The distribution of each molecular subtype in the TCGA and validation cohorts is shown in Supplementary Table 7.

The unsupervised heatmap indicated that the template gene sets we employed performed well in classifying GEO samples (Supplementary Fig. 5), supporting that the 130 gene set can robustly classify LUAD from the GEO database into three subtypes. As such, the classifications based on NTP were used for subsequent analyses.

LUAD subtypes were highly associated with disease outcome as measured by OS. Kaplan-Meier plots showed that the C2 subtype had significantly longer OS compared with the other two subtypes in all four GEO cohorts; either the C3 or the C1 exhibited the shorter survival outcome in the GSE31210 (log-rank  $P < 0.05$ ) and GSE68465 cohorts (log-rank  $P < 0.05$ ) or GSE50081 (log-rank  $P < 0.05$ ) and GSE41271 cohorts (log-rank  $P = 0.063$ ), respectively (Fig. 5). In line with the results of the TCGA analysis, the expression of the central DEGs in all four GEO datasets was significantly higher in the C2 subtype compared with the other two subtypes (Supplementary Fig. 6). These studies robustly validated our subtype classification scheme and verify that C2 is the subtype with the best prognosis, whereas the C1 and C3 subtypes are linked to poorer prognosis.

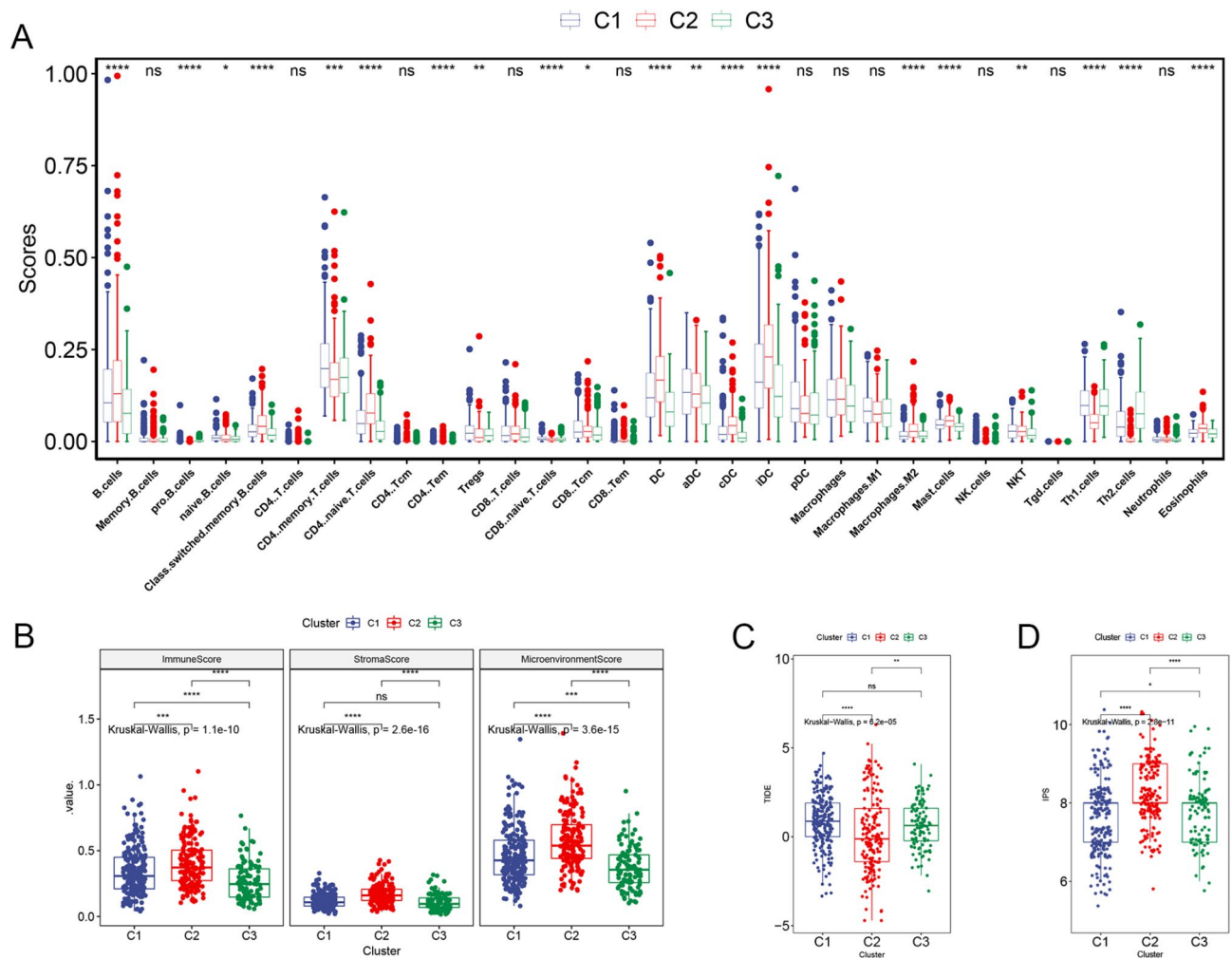
### Validation using a LUAD cohort from our institution

To further validate the results of the TCGA and GEO database analyses, we constructed a separate validation cohort within our institution, consisting of 52 LUAD patients who underwent radical resection. Clinical characteristics of the patients are listed in Supplementary Tables 1, and relapse-free survival (RFS) was evaluated as the outcome for each subject. Amongst the 52 samples, 16 were classified into C1 (30.8%), 19 into C2 (36.5%), and 11 into C3 (21.2%); 6 were unable to be classified (11.5%), since no significant P value was obtained for any of the subtypes, and they were therefore excluded from further analysis. We then determined the clinical relevance of our LUAD classifications via a prognostic analysis of RFS. Patients classified into C2 had a better outcome



**Fig. 3.** Different gene expression character analysis of the three subtypes in TCGA database. **(A)** KEGG enrichment analysis of DEGs among the three subtypes. **(B)** GO enrichment analysis of DEGs among the three subtypes in biological process, cellular component and molecular function. **(C)** Protein-protein interaction network analysis using all 271 DEGs. The color represents the number of interactive proteins. From yellow to red, the darker the color, the greater the number of interacting proteins. **(D)** Seven central DEGs among the three subtypes with the strongest interaction. The color represents the number of interactive proteins. From yellow to red, the darker the color, the greater the number of interacting proteins. **(E)** The expression of the seven central DEGs in C1, C2 and C3, respectively. The T test was used to determine differences between tumor and normal tissues. A p value of  $<0.05$  /  $<0.01$  /  $<0.001$  /  $<0.0001$  denoted by an asterisk \*/\*\*/\*\*\*/\*\*\*\* in figures.





**Fig. 4.** Tumor microenvironment analysis and immunotherapeutic response prediction of the three subtypes in TCGA database. (A–C) *xCell* showed the tumor immune infiltration of the three subtypes. (A) Immune cells infiltration of the three subtypes. (B) Total Immune Score, Stromal Score and Microenvironment Score of the three subtypes. (C) Tumor Immune Dysfunction and Exclusion algorithm showed the TIDE score of the three subtypes. (D) Immunophenoscore algorithm showed the IPS score of the three subtypes. The T test was used to determine differences between tumor and normal tissues. *A* *p* value of <0.05 / <0.01 / <0.001 / <0.0001 denoted by an asterisk \* / \*\* / \*\*\* / \*\*\*\* in figures.

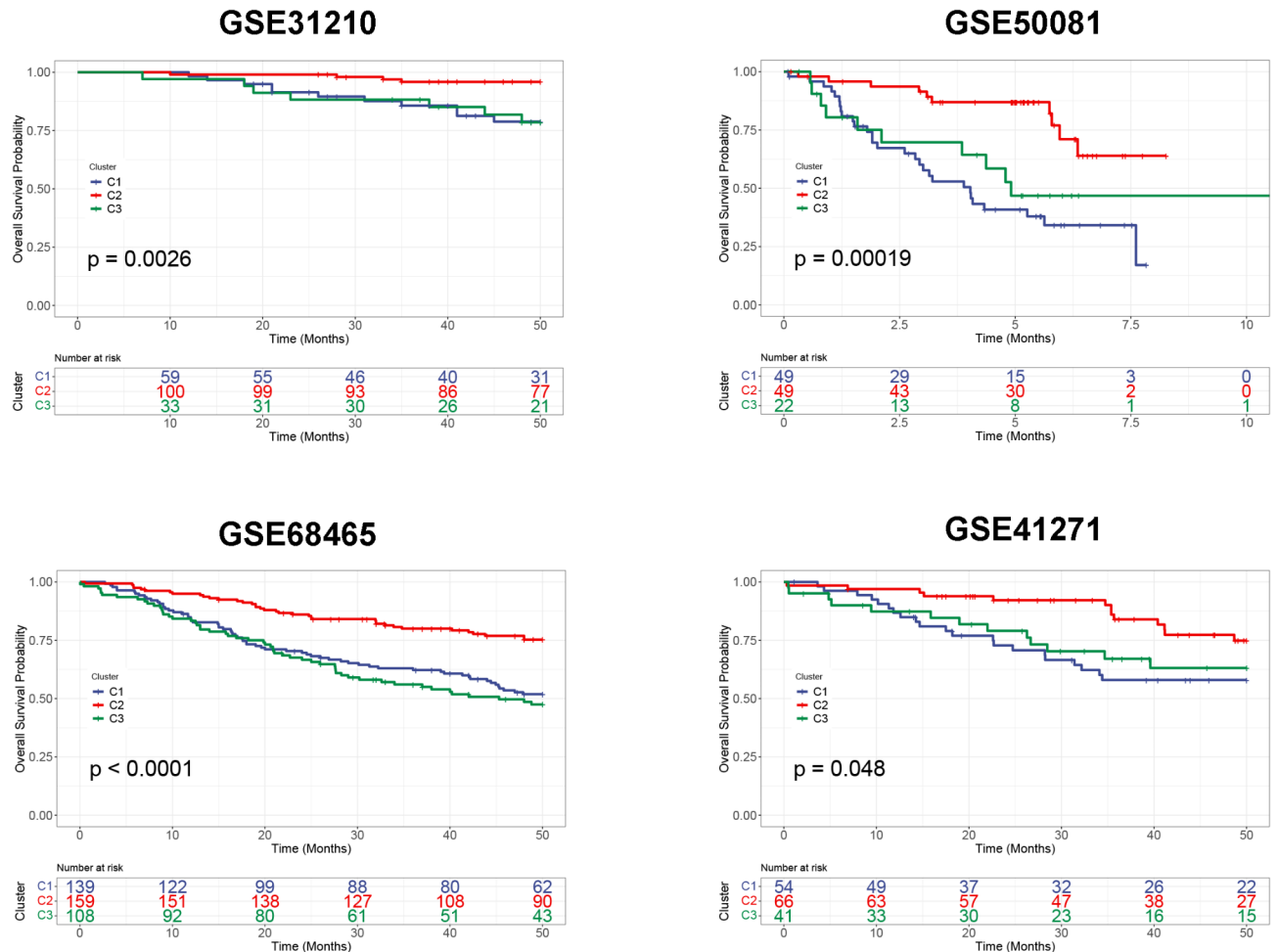
(mean RFS of 40.48 months) compared to C1 and C3 (mean RFS of 22.69 and 22.27 months, respectively,  $P = 0.0036$ ) (Fig. 6).

Taken together, the results of our validation cohort corroborate those from TCGA and GEO database studies, supporting the conclusion that our classification strategy can effectively define distinct LUAD subtypes and might help identify new prognostic subgroups in LUAD.

## Discussion

LUAD is a heterogeneous disease with distinct biological and molecular characteristics as well as clinical outcomes. Abnormal BER and oxidative stress capacity is strongly associated with the progression and therapeutic resistance of lung cancer cells. Although specific genes related to BER and oxidative stress, such as NRF2 and cytoplasmic APEX1, have been reported to be tightly linked to the development and progression of lung cancer<sup>37,38</sup>, the biological signatures based on comprehensive oxidative stress and BER gene expression patterns have been incompletely investigated in LUAD.

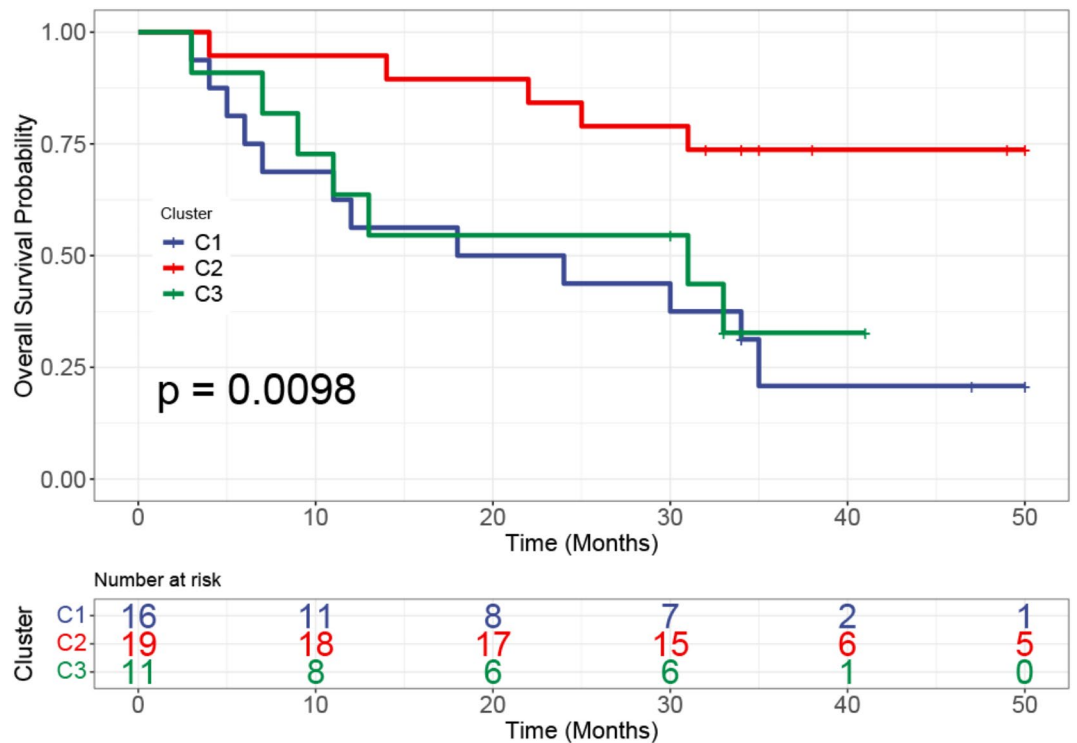
In this study, we found that LUAD could be separated into three subtypes based on BER and oxidative stress gene expression profiles through unsupervised hierarchical clustering. Each subtype exhibited distinct gene expression patterns related to BER and oxidative stress (Fig. 1D). Our subtypes had stable and reproducible classification performance in TCGA training cohort and GEO validation cohorts, demonstrating that our subtypes are statistically robust. Moreover, the three subtypes were associated with prognosis in LUAD. Analysis of four independent GEO datasets and one dataset within our institution confirmed that patients classified to the C2 subtype had the best prognosis, while C1 and C3 subtypes were linked to poorer outcomes.



**Fig. 5.** Validation of the diverse prognosis in four GEO datasets. K–M plots showed the OS of four GEO datasets including GSE31210, GSE68465, GSE50081 and GSE41271.

We further investigated the differences in mutant frequency and profiles among the three subtypes. Notably, C3 exhibited the highest mutation frequency in KEAP1, STK11 and SMARCA4 genes, a finding that is in accordance with previous observations showing frequent co-mutations between SMARCA4, STK11 and KEAP1<sup>39</sup>. In contrast, patients belonging to the C1 and C2 subtype harbored lower STK11 mutation frequency. Loss-of-function mutations in KEAP1 are one of the most frequent mutational events in NSCLC with a prevalence of about 20%. Previous research has found that KEAP1 is a major negative regulator of NRF2 by increasing NRF2's ubiquitination<sup>9</sup>. NRF2 is a critical regulator of cellular oxidative stress in many cancers<sup>40</sup>, often regarded as a driver gene in cancer initiation or metastasis by regulating proliferation, metabolic reprograms, and redox balance<sup>41</sup>. Our finding regarding NRF2 is also consistent with our gene expression signatures, where the C3 subtype harbored higher oxidative stress gene expression than C1 and C2. Additionally, EGFR mutations were frequently observed in the C2 subtype. Prior clinical studies found that EGFR mutations are more frequently detected in females and exclusively present with STK11, KEAP1 and KRAS mutations<sup>42</sup>. Given the higher EGFR mutations frequency in C2, the EGFR pathway might be a potential drug target for these patients. Our data also suggest the BER and oxidative stress related genes may involve in EGFR mutation, though the underlying mechanism warrant further investigation. Collectively, the results highlight that the three subtypes have markedly different mutational backgrounds and genomic instability, features that likely play an important role in giving rise to the observed subtype heterogeneity.

As mentioned above, patients belonging to the C2 subtype display a higher TMB and a lower STK11 mutation frequency, implying a link between such genomic alterations and an optimal tumor immune microenvironment. DEG analysis of the three subtypes further supports these findings. Seven central DEGs associated with tumor immune responses were identified among the three subtypes (Fig. 3D). In our analysis, all the seven genes (i.e., CD69, CD74, CD33, TLR2, TLR7, CD1C, and MRC1) were significantly more highly expressed in C2 but lower in C3. CD69 is a biomarker of T cell activation and proliferation<sup>43</sup>. CD74 is a transmembrane protein functioning as a protein trafficker and immune regulator for T cell and B cell development, as well as dendritic cells (DC) motility<sup>44</sup>. CD33, a differentiation marker of myeloid cells, is a therapeutic target for acute myeloid leukemia. However, CD33 is also expressed in mature cells, including NK cells, macrophage and DC, participating in innate immune responses<sup>45</sup>. In particular, CD33 contributes to leukocyte functionality by inhibiting proliferation



**Fig. 6.** Validation of the diverse prognosis in our original validated dataset. K–M plots showed the overall survival of validation cohort in our institution.

and inducing apoptosis, potentially influencing anti-tumor immunity<sup>46</sup>. TLR2 and TLR7, both members of the Toll-like receptor (TLR) family, are classical pattern recognition receptors<sup>47</sup>. Growing efforts have put forth in exploring the potential application of TLR agonists in arousing an anti-tumor response. CD1c+ was identified as a biomarker of a major subset of DC<sup>48</sup>, which are closely related to the activation of naïve T cells<sup>49</sup>. Mannose receptor C type 1 (MRC1), also known as CD206, is a biomarker for M2-like macrophage polarization and associated with a poor survival outcome<sup>50</sup>. A meta-analysis has reported that high expression of MRC1 or an increased MRC1/CD68 ratio are related to a higher hazard ratio in various types of malignancies, including hepatocellular carcinoma, colon cancer and gastric cancer<sup>51</sup>.

Notably, we found that the C2 subtype was associated with an activated-immune microenvironment, which was further supported upon multiple immune scores. Combined, our data suggest significant heterogeneity in the tumor microenvironment among the three LUAD subtypes, and tumors classified as C2 exhibit higher immune response gene expression and better prognosis, in contrast to C3 tumors that express lower levels of immune response genes and display a poor prognosis.

Our investigations have uncovered three classification clusters in LUAD based on BER and oxidative stress gene expression, which can be used to distinguish the prognosis of LUAD patients and predict their treatment efficacy. Nevertheless, there are important limitations of the current work to mention. For example, only bulk gene expression data were utilized for subtype construction, which may lead to the bias of our study without inclusion of mutational profile and other epigenetic information. As a result, our study did not extensively characterize the main immune cell variability or functional phenotypes among the three subtypes. Furthermore, due to lack of the mutation profile in GEO and Daping hospital cohorts, we failed to compare the genetic background in three subtypes in those datasets, we would further investigate this in our future work. Thus, the underlying mechanisms of the heterogeneous tumor microenvironment remains unresolved. Moreover, the tumor sample numbers in the validation cohort of our institution are relatively small. To sum up, the translation of our classification into clinical implication still requires many efforts.

As such, more clinical studies are warranted to validate our subtype classification in terms of the prognosis of LUAD patients, and more importantly, in predicting the response to immunotherapy.

## Conclusion

Pathological staging, often used as the main predictor of prognosis and treatment options, still fails to accurately predict clinical outcome or guide the treatment of LUAD disease due to the enormous heterogeneity among patients. Our classification reported herein reveals genomic complexity of LUAD based on BER and oxidative stress gene expression that might serve as promising subtyping to predict patient prognosis and immunotherapeutic responses.

## Data availability

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Received: 23 October 2024; Accepted: 9 April 2025

Published online: 13 May 2025

## References

- Sung, H. et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Cancer J. Clin.* **71**(3), 209–249 (2021).
- Bodor, J. N., Boumber, Y. & Borghaei, H. Biomarkers for immune checkpoint inhibition in non-small cell lung cancer (NSCLC). *Cancer* **126**(2), 260–270 (2020).
- Ma, Q. Role of nrf2 in oxidative stress and toxicity. *Annu. Rev. Pharmacol. Toxicol.* **53**, 401–426 (2013).
- Basu, A. K. DNA damage, mutagenesis and cancer. *Int. J. Mol. Sci.* **19**(4) (2018).
- Du, C. et al. Mitochondrial ROS and radiation induced transformation in mouse embryonic fibroblasts. *Cancer Biol. Ther.* **8**(20), 1962–1971 (2009).
- Palma, F. R. et al. Mitochondrial superoxide dismutase: what the established, the intriguing, and the novel reveal about a key cellular redox switch. *Antioxid. Redox. Signal.* **32**(10), 701–714 (2020).
- Puig-Vilanova, E. et al. Oxidative stress, redox signaling pathways, and autophagy in cachectic muscles of male patients with advanced COPD and lung cancer. *Free Radic. Biol. Med.* **79**, 91–108 (2015).
- Hayes, J. D., Dinkova-Kostova, A. T. & Tew, K. D. Oxidative stress in cancer. *Cancer Cell.* **38**(2), 167–197 (2020).
- Yamamoto, M., Kensler, T. W. & Motohashi, H. The KEAP1-NRF2 system: a thiol-based sensor-effector apparatus for maintaining redox homeostasis. *Physiol. Rev.* **98**(3), 1169–1203 (2018).
- Wang, J. et al. Nestin regulates cellular redox homeostasis in lung cancer through the Keap1–Nrf2 feedback loop. *Nat. Commun.* **10**(1), 5043 (2019).
- Lignitto, L. et al. Nrf2 activation promotes lung cancer metastasis by inhibiting the degradation of Bach1. *Cell* **178**(2), 316–329 e18 (2019).
- Miar, A. et al. Manganese superoxide dismutase (SOD2/MnSOD)/catalase and SOD2/GPx1 ratios as biomarkers for tumor progression and metastasis in prostate, colon, and lung cancer. *Free Radic. Biol. Med.* **85**, 45–55 (2015).
- Zhao, S., Tadesse, S. & Kidane, D. Significance of base excision repair to human health. *Int. Rev. Cell. Mol. Biol.* **364**, 163–193 (2021).
- Xie, Y. et al. Deficiencies in mouse Myh and Ogg1 result in tumor predisposition and G to T mutations in codon 12 of the K-ras oncogene in lung tumors. *Cancer Res.* **64**(9), 3096–3102 (2004).
- Weren, R. D. et al. NTHL1 and MUTYH polyposis syndromes: Two sides of the same coin? *J. Pathol.* **244**(2), 135–142 (2018).
- Su, Y. et al. DNA repair gene polymorphisms in relation to Non-Small cell lung cancer survival. *Cell. Physiol. Biochem.* **36**(4), 1419–1429 (2015).
- Karahalil, B., Bohr, V. A. & Wilson, D. M. Impact of DNA polymorphisms in key DNA base excision repair proteins on cancer risk. *Hum. Exp. Toxicol.* **31**(10), 981–1005 (2012).
- Kara, A., Ozgur, A., Nalbantoglu, S. & Karadag, A. DNA repair pathways and their roles in drug resistance for lung adenocarcinoma. *Mol. Biol. Rep.* **48**(4), 3813–3825 (2021).
- Vens, C. & Begg, A. C. Targeting base excision repair as a sensitization strategy in radiotherapy. *Semin Radiat. Oncol.* **20**(4), 241–249 (2010).
- Marisa, L. et al. Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. *PLoS Med.* **10**(5), e1001453 (2013).
- Li, Y. et al. A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC Genom.* **18**(1), 1–13 (2017).
- Okayama, H. et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res.* **72**(1), 100–111 (2012).
- Der, S. D. 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.* **9**(1), 59–64 (2014).
- Director's Challenge Consortium for the Molecular Classification of et al. Gene expression-based survival prediction in lung adenocarcinoma: A multi-site, blinded validation study. *Nat. Med.* **14**(8), 822–827 (2008).
- Sato, M. et al. Human lung epithelial cells progressed to malignancy through specific oncogenic manipulations. *Mol. Cancer Res.* **11**(6), 638–650 (2013).
- Aran, D., Hu, Z. & Butte, A. J. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* **18**(1), 220 (2017).
- Chen, B., Khodadoust, M. S., Liu, C. L., Newman, A. M. & Alizadeh, A. A. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol. Biol.* **1711**, 243–259 (2018).
- Jiang, P. et al. Signatures of T cell dysfunction and exclusion predict cancer immunotherapy response. *Nat. Med.* **24**(10), 1550–1558 (2018).
- Charoentong, P. et al. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell. Rep.* **18**(1), 248–262 (2017).
- Zhang, G. et al. Smooth muscle cell fate decisions Decipher a high-resolution heterogeneity within atherosclerosis molecular subtypes. *J. Transl. Med.* **20**(1), 568 (2022).
- Hoshida, Y. Nearest template prediction: a single-sample-based flexible class prediction with confidence assessment. *PLoS One* **5**(11), e15543 (2010).
- Zhang, L. et al. OGG1 co-inhibition antagonizes the tumor-inhibitory effects of targeting MTH1. *Redox Biol.* **40**, 101848 (2021).
- Moon, E. J. & Giaccia, A. Dual roles of NRF2 in tumor prevention and progression: Possible implications in cancer treatment. *Free Radic. Biol. Med.* **79**, 292–299 (2015).
- Di Maso, V. et al. Transcriptional Up-Regulation of APE1/Ref-1 in hepatic tumor: role in hepatocytes resistance to oxidative stress and apoptosis. *PLoS One* **10**(12), e0143289 (2015).
- Guo, X., Tang, Y. & Zhu, W. Distinct esophageal adenocarcinoma molecular subtype has subtype-specific gene expression and mutation patterns. *BMC Genom.* **19**(1), 769 (2018).
- Lin, J. et al. Multimodule characterization of immune subgroups in intrahepatic cholangiocarcinoma reveals distinct therapeutic vulnerabilities. *J. Immunother. Cancer* **10**(7). (2022).
- Wu, H. et al. Subcellular localization of apurinic endonuclease 1 promotes lung tumor aggressiveness via NF- $\kappa$ B activation. *Oncogene* **29**(30), 4330–4340 (2010).
- Zimta, A.-A. et al. The role of Nrf2 activity in cancer development and progression. *Cancers* **11**(11), 1755 (2019).
- Di Federico, A., De Giglio, A., Parisi, C. & Gelsomino, F. STK11/LKB1 and KEAP1 mutations in non-small cell lung cancer: Prognostic rather than predictive? *Eur. J. Cancer* **157**, 108–113 (2021).



40. Satoh, H., Moriguchi, T., Takai, J., Ebina, M. & Yamamoto, M. Nrf2 prevents initiation but accelerates progression through the Kras signaling pathway during lung carcinogenesis. *Cancer Res.* **73**(13), 4158–4168 (2013).
41. de la Rojo, M., Chapman, E. & Zhang, D. D. NRF2 and the hallmarks of cancer. *Cancer Cell.* **34**(1), 21–43 (2018).
42. Skoulidis, F. & Heymach, J. V. Co-occurring genomic alterations in non-small-cell lung cancer biology and therapy. *Nat. Rev. Cancer* **19**(9), 495–509 (2019).
43. Cibrian, D. & Sanchez-Madrid, F. CD69: From activation marker to metabolic gatekeeper. *Eur. J. Immunol.* **47**(6), 946–953 (2017).
44. Su, H., Na, N., Zhang, X. & Zhao, Y. The biological function and significance of CD74 in immune diseases. *Inflamm. Res.* **66**(3), 209–216 (2017).
45. Walter, R. B. Investigational CD33-targeted therapeutics for acute myeloid leukemia. *Expert Opin. Investig. Drugs.* **27**(4), 339–348 (2018).
46. Crocker, P. R., Paulson, J. C. & Varki, A. Siglecs and their roles in the immune system. *Nat. Rev. Immunol.* **7**(4), 255–266 (2007).
47. Hennessy, E. J., Parker, A. E. & O'Neill, L. A. Targeting toll-like receptors: Emerging therapeutics? *Nat. Rev. Drug Discov.* **9**(4), 293–307 (2010).
48. Stankovic, B. et al. Immune cell composition in human non-small cell lung cancer. *Front. Immunol.* **9**, 3101 (2018).
49. Bourdely, P. et al. Transcriptional and functional analysis of CD1c(+) human dendritic cells identifies a CD163(+) subset priming CD8(+)CD103(+) T cells. *Immunity* **53**(2), 335–352 (2020).
50. Larionova, I. et al. Tumor-associated macrophages in human breast, colorectal, lung, ovarian and prostate cancers. *Front. Oncol.* **10**, 566511 (2020).
51. Debacker, J. M., Gondry, O., Lahoutte, T., Keyaerts, M. & Huvenne, W. The prognostic value of CD206 in solid malignancies: A systematic review and meta-analysis. *Cancers (Basel)* **13**(14) (2021).

## Acknowledgements

We thank National Natural Science Foundation of China for funding supporting.

## Author contributions

Conceptualization, L.Y. and D.W.; methodology, W.R. and H.X.; software, X.D. and Q.Z.; validation, J.Z. and Y.X.; formal analysis, Y.Y.; writing—original draft preparation, L.Y.; writing—review and editing, W.R.; visualization, Y.Y. and X.K.; supervision, Q.Z. and L.Z.; funding acquisition, Q.Z., J.Z. and D.W.. All authors have read and agreed to the published version of the manuscript.

## Funding

This study is funded by National Natural Science Foundation of China (Granted Number: 82173203); Natural Science Foundation of Sichuan Province (Granted Number: 2024NSFSC1915); Excellent Youth Foundation of Sichuan Cancer Hospital & Institute (Granted Number: YB2024015); Qianjiang Central Hospital Cancer National Clinical Key Specialized Project (Granted Number: QJCH2004-01).

## Declarations

## Competing interests

The authors declare no competing interests.

## Ethical approval and consent to participate

This study was conducted in accordance with the guideline of the Helsinki Declaration and was approved by the Ethics Committee of Army Medical Center of PLA I (2017 number 30). Informed consent was obtained from all individual participants included in the study.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-98088-8>.

**Correspondence** and requests for materials should be addressed to D.W. or L.Y.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025