Functional & Integrative Genomics (2024) 24:118
https://doi.org/10.1007/s10142-024-01388-x

RESEARCH

Integrated multi-omics analysis and machine learning to refine molecular subtypes, prognosis, and immunotherapy in lung adenocarcinoma

Tao Han[1,2] · Ying Bai[1,2] · Yafeng Liu[1,2] · Yunjia Dong[1,2] · Chao Liang[1,2] · Lu Gao[1,2] · Jiawei Zhou[1,2] · Jianqiang Guo[1,2] · Jing Wu[1,2,3,4] · Dong Hu[1,2,3,4]

Received: 23 December 2023 / Revised: 1 April 2024 / Accepted: 17 May 2024 / Published online: 27 June 2024

Abstract

Lung adenocarcinoma (LUAD) has a malignant characteristic that is highly aggressive and prone to metastasis. There is still a lack of suitable biomarkers to facilitate the refinement of precision-based therapeutic regimens. We used a combination of 10 known clustering algorithms and the omics data from 4 dimensions to identify high-resolution molecular subtypes of LUAD. Subsequently, consensus machine learning-related prognostic signature (CMRS) was developed based on subtypes related genes and an integrated program framework containing 10 machine learning algorithms. The efficiency of CMRS was analyzed from the perspectives of tumor microenvironment, genomic landscape, immunotherapy, drug sensitivity, and single-cell analysis. In terms of results, through multi-omics clustering, we identified 2 comprehensive omics subtypes (CSs) in which CS1 patients had worse survival outcomes, higher aggressiveness, mRNAsi and mutation frequency. Subsequently, we developed CMRS based on 13 key genes up-regulated in CS1. The prognostic predictive efficiency of CMRS was superior to most established LUAD prognostic signatures. CMRS demonstrated a strong correlation with tumor microenvironmental feature variants and genomic instability generation. Regarding clinical performance, patients in the high CMRS group were more likely to benefit from immunotherapy, whereas low CMRS were more likely to benefit from chemotherapy and targeted drug therapy. In addition, we evaluated that drugs such as neratinib, oligomycin A, and others may be candidates for patients in the high CMRS group. Single-cell analysis revealed that CMRS-related genes were mainly expressed in epithelial cells. The novel molecular subtypes identified in this study based on multi-omics data could provide new insights into the stratified treatment of LUAD, while the development of CMRS could serve as a candidate indicator of the degree of benefit of precision therapy and immunotherapy for LUAD.

Keywords Bioinformatics · Lung adenocarcinoma · Multi-omics · Machine learning · Immunotherapy · Prognosis · Drug screening

Tao Han and Ying Bai contributed equally to this work.

✉ Ying Bai
by0319_cpu@163.com

✉ Jing Wu
wujing8008@126.com

✉ Dong Hu
dhu@aust.edu.cn

[1] School of Medicine, Anhui University of Science and Technology, Huainan, Anhui, China

[2] Anhui Occupational Health and Safety Engineering Laboratory, Huainan, Anhui, China

[3] Key Laboratory of Industrial Dust Deep Reduction and Occupational Health and Safety of Anhui Higher Education Institute, Huainan, Anhui, China

[4] Key Laboratory of Industrial Dust Prevention and Control & Occupational Safety and Health of the Ministry of Education, Anhui University of Science and Technology, Huainan, Anhui, China

Introduction

Lung cancer is currently one of the leading causes of cancer-related deaths in the world, with approximately 2 million new cases and 1.76 million deaths annually (Thai et al. 2021). LUAD is one of the major histologic subtypes of lung cancer, accounting for 40% of lung cancers (Denisenko et al. 2018). Due to the highly invasive and metastatic characteristics of LUAD and the imperfections in early detection, patients often have metastatic cancer at the time of initial diagnosis, losing the opportunity for surgery and leading to unsatisfactory clinical outcomes (Lahiri et al. 2023). In recent years, cancer therapies including immune checkpoint blockers (Lentz et al. 2021), CAR-T cell therapy (Sterner and Sterner 2021), Tumor-specific vaccines (Sellars et al. 2022), and so on have made promising progress in cancer treatment, which prolonging survival time and improving the quality of life of cancer patients. However, the intratumor heterogeneity of LUAD, the susceptibility to drug resistance, and the side effects produced by the drugs make the benefit of treatment variable from patient to patient, and the 5-year survival rate of lung adenocarcinoma has remained at about 20% (Passaro et al. 2022; Siegel et al. 2021; Zhao et al. 2018; Jin et al. 2020). Therefore, nowadays, with the vigorous implementation of precision medicine, identifying unique molecular features of tumors and exploring novel biomarkers for effective stratified treatment and personalized management of LUAD patients are crucial for improving the survival of lung adenocarcinoma.

Anatomically developed TNM staging systems are currently key indicators for clinical decision-making and treatment regimen selection (Kay et al. 2017). Although the identification of TNM staging provides a relatively reliable basis for patient treatment, the high degree of intra- and inter-tumor heterogeneity that exists within LUAD itself makes the clinical outcomes of even patients at the same TNM stage variable (Caso et al. 2020; Yoo et al. 2022; Okayama et al. 2012). With the rapid development of genomics and high-throughput sequencing technologies, more and more biomarkers have begun to gradually enter clinical applications, including EGFR mutations (Castellanos et al. 2017; Sabbah et al. 2020), (tumor mutation burden) TMB (Addeo et al. 2021), PD1/PD-L1 (Yi et al. 2022; Ai et al. 2020), and so on, which have greatly promoted the process of precision tumor therapy (Wu and Qu 2015). However, there is still a number of LUAD patients who still fail to derive clinical benefit from these approaches, which may be due to the complex and variable tumor microenvironment within LUAD (Genova et al. 2021; Madeddu et al. 2022). A number of studies have shown that the identification of molecular subtypes based on comprehensive omics data can be a powerful means of stratifying treatment for patients. Bagaev et al. summarized tumor microenvironment-related features and performed unsupervised consensus clustering to classify cancer patients into four tumor microenvironmental subtypes facilitating the refinement of clinical immunotherapeutic regimens (Bagaev et al. 2021). Jiang et al. classified breast cancer patients into four stable molecular subtypes by combining genomics and transcriptomics data, providing new ideas for stratifying breast cancer patients (Jiang et al. 2019). In addition to this, prognostic biomarkers generated based on gene expression profiles have been widely used to assess the heterogeneity of LUAD. However, these models are often based on expression data of mRNA, microRNA (miRNA), or long-stranded noncoding RNA (lncRNA) with specific phenotypes (e.g., iron death (Shao et al. 2023), immunity (Han et al. 2023), and cellular pyroptosis (Hu et al. 2023) and so on), and the chosen modeling approach is relatively homogeneous, which allows for weak generalization of the established multigene prognostic signatures, thus limiting their wide range of applications. Here, we sought to identify novel biomarkers that can facilitate prognostic and therapeutic management of LUAD based on a combination of large-scale multi-omics data and advanced machine learning algorithms.

In this study, we combined mRNA expression profiles, lncRNA expression profiles, genomic mutations, and epigenomic DNA methylation data to identify novel LUAD consensus subtypes using 10 multi-omics integration programs. Subsequently, we developed consensus machine learning-related prognostic signature (CMRS) based on a combined framework of subtype related biomarkers and 10 machine learning algorithms. CMRS has strong prognostic predictive capabilities and demonstrated superior efficiency compared to other published prognostic signatures. CMRS can be used to assess tumor microenvironment, genome stability, immunotherapy, and drug therapies. Overall, our findings provide new insights for refining the molecular subtypes of LUAD and individualizing, precision therapy.

Materials and methods

Data download and preprocessing of LUAD

We first collected multi-omics data of lung adenocarcinoma (LUAD) from The Cancer Genome Atlas (TCGA) (https://portal.gdc.cancer.gov), including mRNA transcriptome expression (normalized by TPM), lncRNA transcriptome expression (normalized by TPM), DNA methylation and somatic mutation. The mRNA and lncRNA transcriptome profiles were obtained from the "TCGAbiolinks" R package; somatic mutation data and methylation profiles were downloaded from UCSC Xena (https://xenabrowser.net/); and clinical information on TCGA-LUAD patients was obtained from the previous study by Bagaev et al. (Bagaev et al. 2021). In addition, we obtained the corresponding data information from Gene Expression Omnibus (GEO) for four LUAD cohorts, GSE31210, GSE72094, GSE68465, GSE166722, two non-small-cell lung cancer (NSCLC) immunotherapy cohorts, GSE126044 and GSE135222, and one single-cell sequencing data, GSE148071. Furthermore, there is an immunotherapy clinical trial for metastatic urothelial cancer (mUC) from http://research-pub.gene.com/IMvigor210CoreBiologies. The portion of Fig. 1(flow chart) were drawn by using pictures from Servier Medical Art (http://smart.servier.com/). Servier Medical Art by Servier is licensed under a Creative Commons Attribution 3.0 Unported License (https://creativecommons.org/licenses/by/3.0/).

Multi-omics consensus integration analysis

To efficiently perform downstream clustering analysis, we matched 4 dimensions of omics information by sample ID and removed LUAD patients without survival information. Next, we performed multi-omics consensus typing of patients using the multi-omics and visualization of cancer subtypes integration R package ("MOVICS") (Lu et al. 2021a). The "getElites" function was first used to screen for genetic features. For continuous variables (mRNA, lncRNA, and methylation), we set the "method" parameter of the "getElites" function to "mad" to filter the genes with the highest degree of variation (the first 3000 for mRNA and lncRNA, and the first 5000 for methylation sites). We then set the "method" parameter to "cox" and combined it with the clinical data to identify genes with prognostic significance in each data dimension $P < 0.05$. For the mutation data, the number of mutations exceeding 60 in the

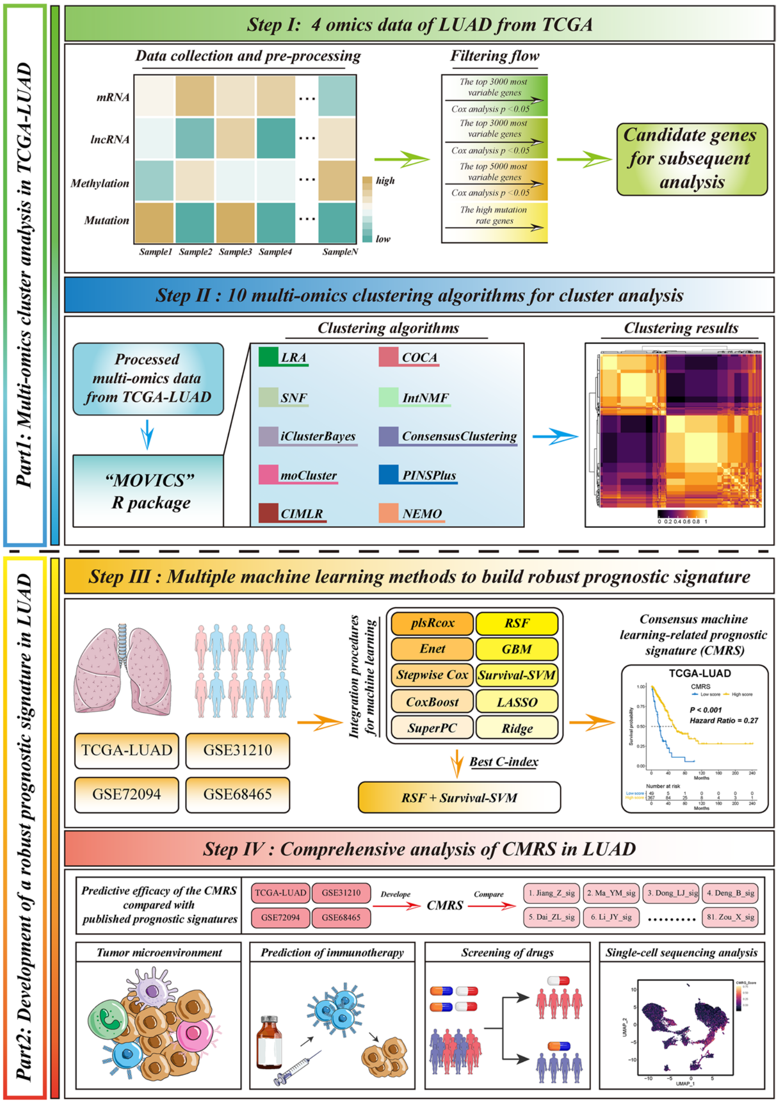
Part1: Multi-omics cluster analysis in TCGA-LUAD
Step I: 4 omics data of LUAD from TCGA
Data collection and pre-processing
mRNA
lncRNA
Methylation
Mutation
Sample1
Sample2
Sample3
Sample4
SampleN
high
low
Filtering flow
The top 3000 most variable genes
Cox analysis p <0.05
The top 3000 most variable genes
Cox analysis p <0.05
The top 5000 most variable genes
Cox analysis p <0.05
The high mutation rate genes
Candidate genes for subsequent analysis
Step II : 10 multi-omics clustering algorithms for cluster analysis
Processed multi-omics data from TCGA-LUAD
"MOVICS" R package
Clustering algorithms
LRA
COCA
SNF
IntNMF
iClusterBayes
ConsensusClustering
moCluster
PINSPlus
CIMLR
NEMO
Clustering results
0 0.2 0.4 0.6 0.8 1
Part2: Development of a robust prognostic signature in LUAD
Step III : Multiple machine learning methods to build robust prognostic signature
TCGA-LUAD
GSE31210
GSE72094
GSE68465
Integration procedures for machine learning
plsRcox
RSF
Enet
GBM
Stepwise Cox
Survival-SVM
CoxBoost
LASSO
SuperPC
Ridge
Best C-index
RSF + Survival-SVM
Consensus machine learning-related prognostic signature (CMRS)
TCGA-LUAD
CMRS
Low score
High score
P < 0.001
Hazard Ratio = 0.27
Survival probability
Months
Number at risk
Step IV : Comprehensive analysis of CMRS in LUAD
Predictive efficacy of the CMRS compared with published prognostic signatures
TCGA-LUAD
GSE31210
GSE72094
GSE68465
Develope
CMRS
Compare
1. Jiang_Z_sig
2. Ma_YM_sig
3. Dong_LJ_sig
4. Deng_B_sig
5. Dai_ZL_sig
6. Li_JY_sig
....
81. Zou_X_sig
Tumor microenvironment
Prediction of immunotherapy
Screening of drugs
Single-cell sequencing analysis
UMAP_1
UMAP_2
CMRG_Score

Fig. 1 The flowchart of LUAD subtype identification and CMRS construction. We synthesized 10 multi-omics clustering algorithms included in the "MOVICS" R package to identify new omics subtypes of LUAD. Afterward, the stable prognostic genes were further identified based on the subtypes. Using a combination of multiple machine learning algorithms, the prognostic signature with the highest C-index was screened. Finally, a CMRS was constructed based on the combination of RSF and Survival-SVM. Subsequently, we revealed the relationship between the CMRS and the prognosis, tumor microenvironment remodeling, immunotherapy benefit, and potential therapeutic agents of LUAD through comprehensive analysis

TCGA-LUAD patients in this study was used as a threshold to filter the genes. Afterward, the outcome data of these four dimensions were summarized and included in our study for further analysis.

After the initial signature selection, we further determined the optimal number of clusters for our study using the “getClustNum” function in the “MOVICS” package. Subsequently, we applied the “getMOIC” function for cluster analysis. Afterward, we integrated the results of the different algorithms using the “getConsensusMOIC” function as well as the default parameters provided by the “MOVICS” package to improve the robustness of the clustering and obtained the final clustering results through this integration process.

Finally, to test the stability of molecular subtyping, we compared the Nearest Template Prediction (NTP) (The top 100 genes upregulated in CS1 and CS2, respectively, were used to perform the NTP) and Partitioning Around Medoids (PAM) classifiers to the CSs. Further, the clustering results were validated in a validation cohort based on NTP, and the survival differences between subtypes were analyzed using Kaplan-Meier (K-M) curves analysis.

Analysis of molecular features between LUAD subtypes

To investigate the unique molecular feature profiles in the CSs, we first used the PROGENy (Schubert et al. 2018) method to evaluate the oncogenic-related pathway activity in LUAD. Afterward, we obtained the genes that were highly expressed in each subtype, setting a threshold of $|\log_2 FC| > 1$ with a *P-value* < 0.05, and analyzed the heterogeneity of invasive features among patients with different subtypes using the method of Yoo et al. (Yoo et al. 2022); the gene sets were mainly obtained from the MSigDB database (https://www.gsea-msigdb.org/gsea/msigdb). Additionally, we collected and compared the invasiveness score (IVS) of TCGA-LUAD patients in CSs (Yoo et al. 2022).

Transcriptional regulatory networks were constructed by “RTN” R package, including candidate regulators associated with cancerous chromatin remodeling collected from Lu et al. (Lu et al. 2021b). The 26 stemness gene sets scores of LUAD were calculated using single-sample gene set enrichment analysis (ssGSEA), which were obtained from the study of Zheng et al. (Zheng et al. 2022); meanwhile, we utilized the mRNAsi distributed by Malta et al. calculation method to calculate the degree of stemness activity of LUAD (Malta et al. 2018).

Concerning the mutational landscape, **1).** We calculated the TMB for each sample and visualized the genes with mutation frequencies above 20%. **2).** As described by Lu et al, we applied the “deconstructSigs” package to extract mutation signatures for each LUAD patient and selected mutation signature 1 (age-related), mutation signature 2 (APOBEC activity-related), mutation signature 4 (smoking-related), and mutation signature 5 (unknown) for visualization, these mutations occur more frequently in LUAD (Lu et al. 2021b; Alexandrov et al. 2013); **3).** Information on the types of mutations produced in LUAD patients was obtained from Thorsson et al. to explore the differences in mutation types between subtypes (Thorsson et al. 2018).

Comprehensive omics subtype-derived prognostic signature

Next, we constructed a consensus prognostic model for LUAD. First, we used the “limma” method in “MOVICS” R package to analyze the differences between the subtypes and obtained the genes that were up-regulated in CS1 and met $\log_2 FC > 1$, *P-value* < 0.05. Next, we performed univariate Cox analysis on the above genes simultaneously in TCGA-LUAD, GSE68465, GSE72094, and GSE31210. The shared genes that had prognostic significance ($P < 0.05$) in all cohorts were used as candidate genes for the construction of prognostic models.

Next, we selected the TCGA cohort with relatively complete clinical information as the training set and the other cohorts as the validation set. We submitted the above prognostic biomarkers to a comprehensive machine learning analysis program that integrated 10 classical algorithms: CoxBoost, stepwise Cox, Lasso, Ridge, elastic net (Enet), survival support vector machine (survival-SVM), generalized boosted regression model (GBM), supervised principal components (SuperPC), partial least Cox (plsRcox) and Random Survival Forest (RSF). After that, we selected the best consensus prognostic model (consensus machine learning-related prognostic signature, CMRS) for LUAD based on the average C-index of the 3 GEO-LUAD cohorts.

Prognostic value and potential clinical value of CMRS

We first calculated CMRS based on the generated model for each LUAD sample in the training and validation sets, after which the prognostic significance of CMRS was assessed by K-M survival curves. Further, we systematically retrieved 81 prognostic signatures associated with lung adenocarcinoma (Supplementary table.1) and calculated scores for each sample based on published coefficients (some of these prognostic signatures were excluded due to gene deletions in some LUAD cohorts). The ability of all signatures in each cohort to predict prognosis was assessed by the C-index. Further, to enhance the clinical utility of the CMRS, we constructed a nomogram based on CMRS. Calibration curve was plotted to characterize accuracy, and decision curves were used to calculate the clinical benefit to patients.

Immune features and mutational heterogeneity in different CMRS groups

First, based on the “IOBR”(Zeng et al. 2021) package, we collected previously published TME-associated features and used the ssGSEA method to calculate enrichment scores for each sample and comprehensively analyzed tumor microenvironmental differences between samples. Similarly, we obtained the 28 immune cell gene sets from the study by Charoentong et al. and used the ssGSEA method for estimation and comparison (Charoentong et al. 2017). Subsequently, we used the method of Thorsson et al. (Thorsson et al. 2018) to visualize mRNA expression of immunomodulators in patients with high and low CMRS (median normalized expression levels); expression versus methylation (correlation of gene expression with DNA methylation beta values); amplification frequency, and deletions frequency. Spearman correlation analyses were performed to assess the relationship between TMB and CMRS. Finally, differences in the type of mutations present in patients with high and low CMRS were assessed using the Wilcoxon test.

Comprehensive analysis of CMRS-based immunotherapy prediction

For immunotherapy response, we first compared the correlation between immune checkpoints and CMRS, and compared TIP score (http://biocc.hrbmu.edu.cn/TIP/) in the high CMRS and low CMRS groups. Subsequently, TIDE scores as well as immunotherapy response of LUAD patients with different CMRS groups were calculated and compared by the Tumor Immune Dysfunction and Exclusion (TIDE) algorithm (http://tide.dfci.harvard.edu/). Next, we applied CMRS to two NSCLC immunotherapy cohorts to evaluate its predictive value for immunotherapy. The IMvigor210 cohort is a large immunotherapy cohort for mUC and has more complete clinical treatment information (Mariathasan et al. 2018). In the IMvigor210 cohort, we evaluated the predictive efficiency of CMRS in terms of patient survival in delayed response to immunotherapy.

CMRS-based screening of potential therapeutic agents

Volcano plots showed differential genes in patients in the high and low CMRS groups; subsequent GSEA analysis revealed the activation status of oncogenic pathways among low CMRS patients. Subsequently, the “oncoPredict” R package (Maeser et al. 2021) analysis was used to show the differences in drug sensitivity between the different CMRS groups. Next, potential therapeutic agents were developed for the high CMRS group. 1). we obtained drug sensitivity data from CTRP (https://portals.broadinstitute.org/ctrp) and PRISM Repurposing datasets (https://depmap.org/portal/prism/), and obtained expression data of cell lines from the Cancer Cell Line Encyclopedia (CCLE, https://sites.broadinstitute.org/ccle/) database; 2). The CTRP and PRISM datasets possess AUC values, with lower AUC values indicating increased sensitivity to the compound. Moreover, as the common chemotherapeutic agent for LUAD, we further selected cisplatin to validate the scientific validity and rigor of this approach; 3). We analyzed the differences in drug responses between the high CMRS (top 10%) and low CMRS (bottom 10%) groups based on the Wilcoxon rank-sum test and set the threshold $\log_2$FC > 0.1 to identify compound; 4). Next, we further screen compounds with negative correlation coefficients between AUC values and CMRS (set threshold R < -0.3, Spearman test); 5). Finally, we screened potential drugs for use in patients in the high CMRS group by the overlap of compounds obtained in 3) and 4).

Single-cell RNA-seq analysis

We used the “Seurat” R package (Hao et al. 2021) to process single-cell RNA-seq data. We selected 18 lung adenocarcinoma samples from GSE148071 with the following selection criteria: **a)** each cell should express more than 200 genes and less than 5000 genes; **b)** each gene should be expressed in at least 3 cells; and **c)** the content of mitochondrial RNA in each cell should be less than 30%. Afterward, we normalized the expression values using the “NormalizeData” function. Subsequently, 2000 highly variable genes were identified using the “FindVariableFeatures” function. These genes were then centered using the “ScaleData” function. Next, the “RunPCA” function is used to downscale the obtained highly variable genes. Finally, we use the “harmony” package to integrate the data.

Uniform Manifold Approximation and Projection (UMAP) analysis was performed to visualize cell populations. Subsequently, cell subpopulations were manually annotated and marker genes of the cells were visualized using violin maps.

The “AddModuleScore” function is a function used to calculate and add scores for specific gene sets to individual cells or clusters of cells. The “AddModuleScore” function was used to evaluate the scores of the 13 genes related to CMRS.

Statistical analysis

For the comparison of the two groups, the Student's t-test or Wilcoxon rank sum test was used for testing. For comparison of more than two groups, the one-way ANOVA test or the Kruskal-Wallis test was used for testing. Fisher's exact test or

Chi-square test was performed for the categorical variables. Patients in each cohort were categorized into high and low subgroups based on the median value of the CMRS score, and differential expression analysis was performed by the "limma" R package. Pearson or Spearman correlation analysis was performed to analyze the correlation between two continuous variables. Multi-omics clustering was accomplished by the "MOVICS" R package. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$, $****p < 0.0001$. All statistical analyses were performed in R v.4.1.2.

Result

Identification and validation of multi-omics consensus molecular subtypes in LUAD

The workflow of this study is shown in Fig. 1. We first obtained four dimensions of omics data from TCGA-LUAD, including mRNA expression, lncRNA expression, methylation, and mutation status. The Cluster Prediction Index (CPI) and Gaps-statistics algorithms included in the "MOVICS" R package were then utilized to derive the optimal number of clusters as 2 (Fig. 2A). Next, we comprehensively identified 2 comprehensive omics subtypes (CSs), including CS1 and CS2, using 10 multi-omics integrated clustering algorithms (Fig. 2B). Subsequent silhouette scores also indicated that the clustering worked well (Fig. 2C). We found that among the different subtypes, LUAD have distinct molecular expression patterns across the transcriptome (mRNA and lncRNA), epigenetic methylation and somatic mutations that are significantly different (Fig. 2D). Notably, in terms of the defined immune subtypes, we found that the poor prognostic C1 and C2 immune subtypes were more enriched in the CS1; in contrast, the favorable prognostic C3 immune subtype was more enriched in the CS2, which suggests the existence of distinctly different TME attributes among the CSs. Subsequently, survival curves revealed a strong correlation between CSs and survival, including overall survival (OS) ($P = 0.001$; Fig. 2E) and progression-free survival (PFS) ($P = 0.017$; Fig. 2F);

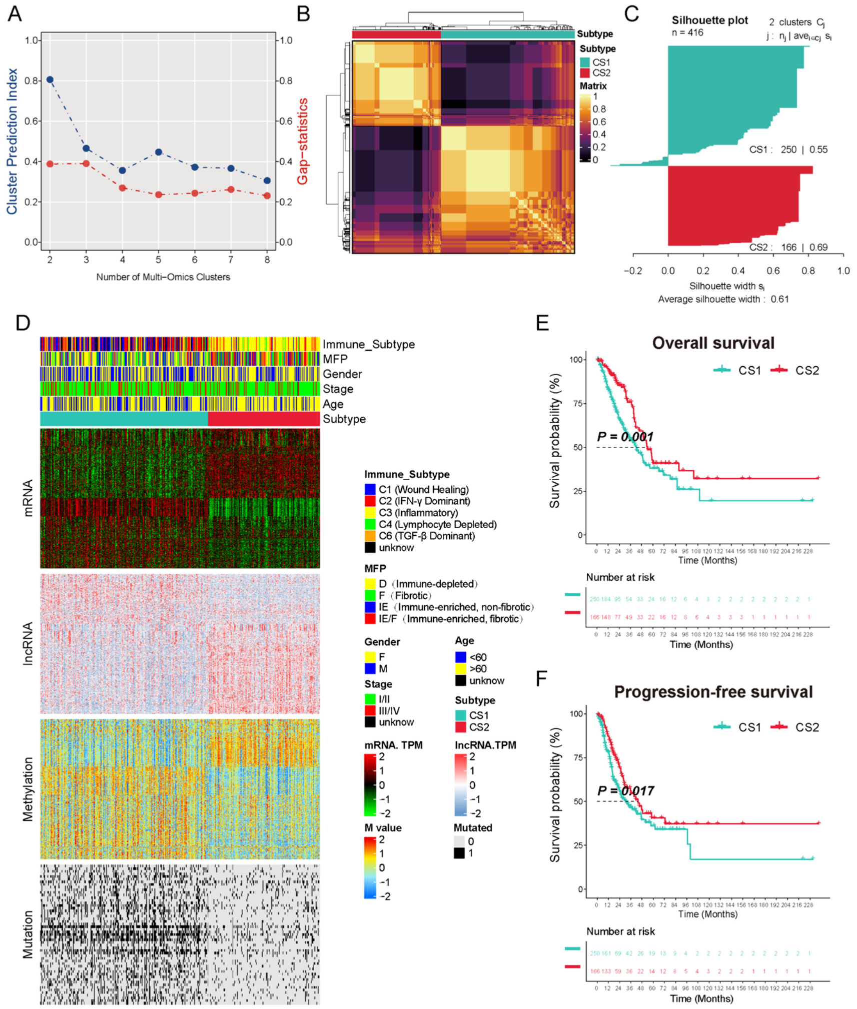

Fig. 2 Consensus subtypes for multi-omics integration of LUAD. (**A**) The Cluster Prediction Index and Gap Statistical Analysis of the multi-omics clusters. (**B**) Consensus clustering matrix of novel prognostic subtypes based on 10 algorithms. (**C**) The sample similarity of each subgroup was assessed by calculating the Silhouette score. (**D**) Comprehensive heatmap of shared subtypes including mRNA, lncRNA, DNA methylation sites, and mutated genes. K-M curve analysis of different survival outcomes between CSs, including OS (**E**) and PFS (**F**)

among them, CS1 patients exhibited a significantly poorer prognosis.

Next, we also evaluated the consistency of CSs with the nearest template prediction (NTP) and partition around medoids (PAM) algorithms in the TCGA cohort (*P < 0.001*; Supplementary Fig. 1 A-B). Further, we simultaneously expanded the molecular subtypes into multiple external cohorts using both NTP and PAM classifiers and compared the results of the two classifiers. The results show that the populations obtained by both classifiers are in high agreement, indicating good stability of the molecular subtypes (*P < 0.001*; Supplementary Fig. 1 C-E). In addition, we visualized the NTP classification results and analyzed the survival differences. The results showed that in all cohorts, LUAD patients with CS1 consistently exhibited poorer survival outcomes (*P < 0.001*; Supplementary Fig. 1 F-G).

Distinctive biological features in multi-omics molecular subtypes

Next, we also tried to explore the differences in molecular characteristics between the CSs. The results showed that there existed differences in the activity of common oncogenic pathways among different CSs. Among them, EGFR, Hypoxia, and PI3K pathways were significantly more active in CS1 subtype; p53 pathway was more actively expressed in CS2 subtype (Fig. 3A). Consistent with the subtype-related prognostic results, K-M curves showed that EGFR (*P < 0.001*), PI3K (*P = 0.007*), and Hypoxia (*P < 0.001*) oncogenic pathways enriched in CS1 were significantly associated with poor prognosis (Fig. 3B). Subsequently, we analyzed the differences in invasiveness characteristics between CS1 and CS2; the results showed that CS1 was significantly enriched in cell cycle, angiogenesis, EMT, and metastasis upregulation-related pathways; CS2 was enriched in tumor suppressor gene (TSG) (Zhao et al. 2013), and metastatic activity downregulation-related pathways (Fig. 3C). Invasiveness score (IVS) is a score developed by Yoo et al. (Yoo et al. 2022) to estimate invasiveness activity in patients. Our study demonstrated a significantly higher level of IVS in CS1 than in CS2 (*P < 0.001*) (Fig. 3D). Further, we used the TCGA comic subtypes as a training set to cluster the patients in the GSE166722 cohort into CS1 and CS2 subtypes using the NTP method, and we similarly found that patients in the CS1 subtype had predominantly invasion attributes, whereas those in the CS2 subtype had predominantly indolent attributes (Fig. 3E). These results suggest that the characteristics of CS1 patients mainly include active tumor proliferation and invasive features.

We then further investigated the differences in transcriptomics within the CSs, and the differential activity of cancerous chromatin remodeling regulators highlighted other potential patterns of differential regulation among CSs (Fig. 3F). It is suggested that the formation of epigenetically driven transcriptional networks may be causative for the differentiation of these subtypes. Cancer stem cells (CSCs) have an important role in inducing tumor metastasis and recurrence (Ayob and Ramasamy 2018). Therefore, we analyzed the enrichment of 26 stemness gene sets in patients with different CSs. The results showed that most of the stemness gene sets including Hs_ESC_Assou, Hs_EC_Skotheim, and Plurinet were enriched in CS1. The mRNAsi was similarly significantly up-regulated in CS1 patients (*P < 0.001*) (Fig. 3G).

We next compared the heterogeneity of mutation patterns between CSs, and we found that the TMB was higher in the CS1 group, where the classical oncogene TP53 and KRAS were more frequently mutated in CS1, and the prevalent mutational signature in LUAD showed that there were differences in the distribution of mutational signatures between different subtypes. (Fig. 3H). Consistently, in terms of known mutation types, patients in the CS1 group all had more frequent mutations compared to those in CS2 (*P < 0.001*) (Fig. 3I).

Development of the CMRS based on a comprehensive machine learning program

We obtained 234 genes up-regulated in CS1 (Supplementary table.2) in TCGA-LUAD, followed by univariate Cox analysis in the TCGA, GSE72094, GSE31210, and GSE68465 cohorts (Supplementary table.3). Then, screened 87 prognostic genes shared among all four LUAD cohorts (Supplementary Fig. 2; Supplementary table.4). Next, we submitted these genes to a comprehensive machine learning program to construct the consensus machine learning-related prognostic signature (CMRS). The TCGA cohort was used as the training set for model construction, and after that, based on the average C-index of the test cohort as the ranking of model efficiency, we found that the prognostic model developed by the RSF + Survival-SVM method has the highest average C-index of 0.668, which can be used as the best prognostic CMRS (Fig. 4A). CMRS is mainly composed of 13 key genes (Supplementary table.5), and these genes were significantly associated with poor prognosis in all 4 cohorts (Fig. 4B). We then calculated the CMRS scores for each sample in all cohorts and analyzed the correlation between CMRS and survival. The results showed that patients with low CMRS have poorer clinical outcomes in all cohorts, including TCGA-LUAD (Hazard Ratio = 0.27, *P < 0.001*), GSE31210 (Hazard Ratio = 0.15, *P < 0.001*), GSE72094 (Hazard Ratio = 0.37, *P < 0.001*), GSE68465 (Hazard Ratio = 0.43, *P < 0.001*) (Fig. 4C).

Next, we systematically examined the multi-omics phenotype of CMRS-related genes in TCGA pan-cancer by the “GSCALite” public server (http://bioinfo.life.hust.edu.cn/web/GSCALite/). The results showed that these genes

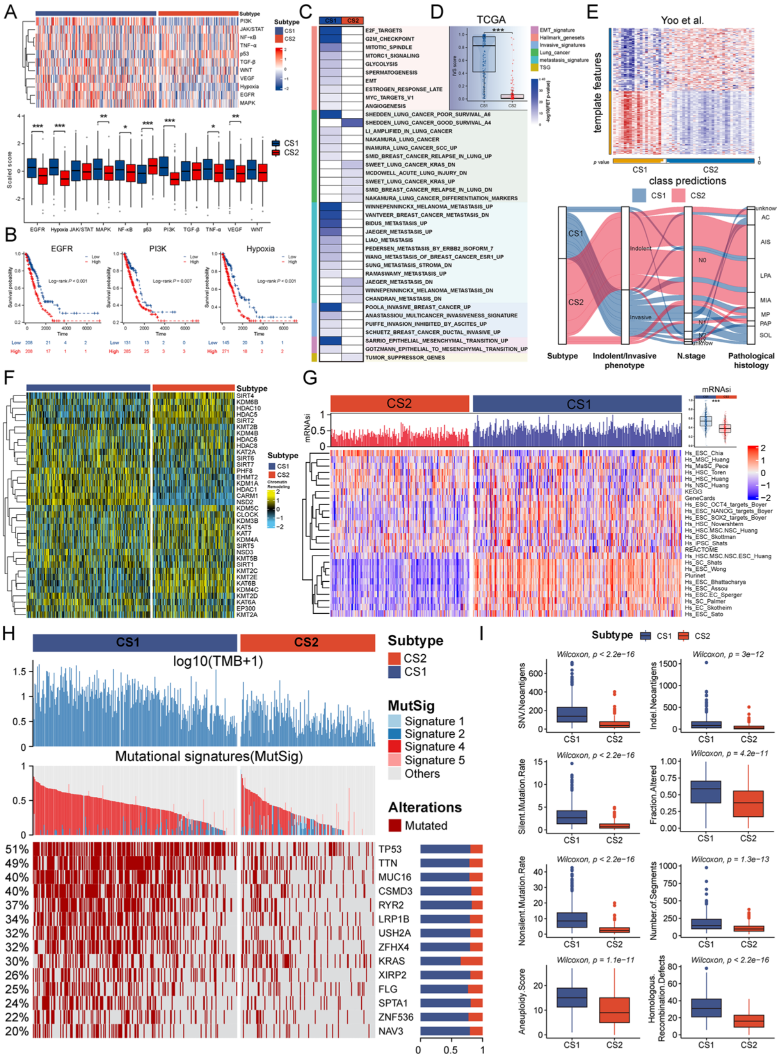

Fig. 3 Multi-omics landscape between different subtypes. (**A**) Heatmap and subgroup comparisons of LUAD cancer pathway activity calculated according to the PROGENy method. (**B**) K-M survival curve analysis of oncogenic pathways. (**C**) Association of CSs with MSigDB gene sets, including Hallmark, lung cancer-related genes, and tumor invasion- or metastasis-related gene sets, using Fisher's exact test. (**D**) Distribution of invasiveness score (IVS) of TCGA-LUAD within different CSs. (**E**) Estimation of subtypes of the GSE166722 cohort using the NTP method (where acinar (AC), micropapillary (MP), papillary (PAP), and solid (SOL) are pathologically aggressive histologic subtypes; MIA, AIS, and lepidic predominant (LPA) for noninvasive or minimally invasive subtypes). (**F**) Heatmap of potential regulatory factor activity associated with cancerous chromatin remodeling. (**G**) Heatmap of the distribution of the 26 stemness scores in different subtypes and comparative mRNAsi grouping. (**H**) Landscape of genomic alterations in CSs. Distribution of TMB, the relative contribution of the mutation signatures, and mutated genes with mutation frequencies greater than 20% are shown in the top-to-bottom panels. (**I**) Comparison of the eight mutation types across subtypes

were commonly highly expressed in multiple cancer tissues across most cancer types (Supplementary Fig. 3A) and were significantly associated with poor prognosis (Supplementary Fig. 3B). We also found a positive correlation between mRNA expression levels and copy number variation (CNV) of CMRS genes in most cancer types, especially ECT2 (Supplementary Fig. 3C). Analysis of CNV frequency changes showed that CNV of CMRS-related genes have significant differences across cancer types (Supplementary Fig. 3D-F). Furthermore, we similarly found that the methylation levels of CMRS genes in most cancer types were different between tumor and normal samples (Supplementary Fig. 4A). Meanwhile, at the pan-cancer level, the methylation levels of these genes were negatively correlated with the mRNA expression levels of these genes (Supplementary Fig. 4B). It is worth mentioning that the methylation levels of different CMRS-related genes were also correlated with the prognosis of patients to some extent (Supplementary Fig. 4C). These results suggest that CMRS genes may cause differences in survival outcomes by influencing epigenetic changes. In

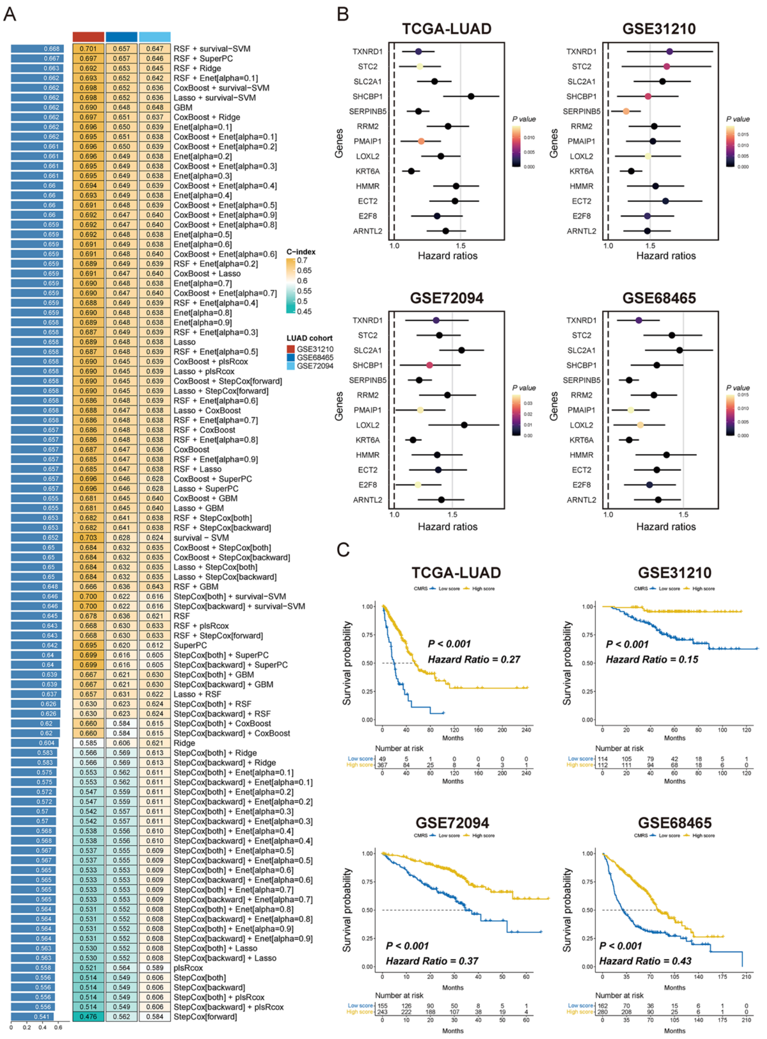

Fig. 4 Prognostic value of CMRS. (**A**) Multiple machine-learning models were built by synthesizing machine-learning frameworks. The C-index of each model was calculated from the TCGA-LUAD, GSE31210, GSE72094, and GSE68465 cohorts and sorted by the average C-index of the validation set (GSE31210, GSE72094, and GSE68465). (**B**) Forest plot of univariate Cox regression analysis of key genes in the training and validation cohorts. (**C**) Survival analysis plot of CMRS in TCGA-LUAD, GSE31210, GSE72094 and GSE68465 cohorts

addition, we found that CMRS genes could generally activate the Cell Cycle pathway and had a significant inhibitory effect on the hormone AR pathway (Supplementary Fig. 4D-E). In terms of mutation patterns, we found that these genes are more commonly mutated mainly in cancers such as UCEC, SKCM, COAD, and STAD (Supplementary Fig. 5A-B).

Comparison of CMRS with published prognostic signatures in LUAD

In recent years, with the rapid development of high-throughput sequencing, it has made possible stratified treatment strategies for tumors and precise and personalized treatment for patients (Reel et al. 2021). A large number of studies have been conducted to develop prognostic signatures based on machine learning methods to predict cancer prognosis (Greener et al. 2022). Therefore, here, we collected a total of 81 prognostic signatures for LUAD (Supplementary table.1) to compare the predictive accuracy of CMRS and these models. Among them, some signatures in the LUAD cohort were excluded due to the lack of constituent genes for prognostic signatures. We compared the average C-index of CMRS and 81 prognostic signatures in the TCGA-LUAD, GSE31210, GSE72094, and GSE68465 cohorts, and we found that CMRS is the 6th in the TCGA cohort, the 2nd in GSE31210, and the 1st in GSE72094 and GSE68465. This revealed the robustness of CMRS (Fig. 5A). Notably, some of the prognostic signatures had high C-index in the TCGA-LUAD cohort, but performed poorly in the other cohorts, which may be caused by the overfitting that these prognostic signatures are produced by training in TCGA. We then explored the association between CMRS and common clinicopathologic features, and we found that men and advanced lung adenocarcinoma patients were predominant in low CMRS (Fig. 5B).

Next, to better predict the clinical outcomes of LUAD, we developed a nomogram based on CMRS, clinical stage, age, gender, and smoking status (Supplementary Fig. 6A). The calibration plot showed high agreement between the

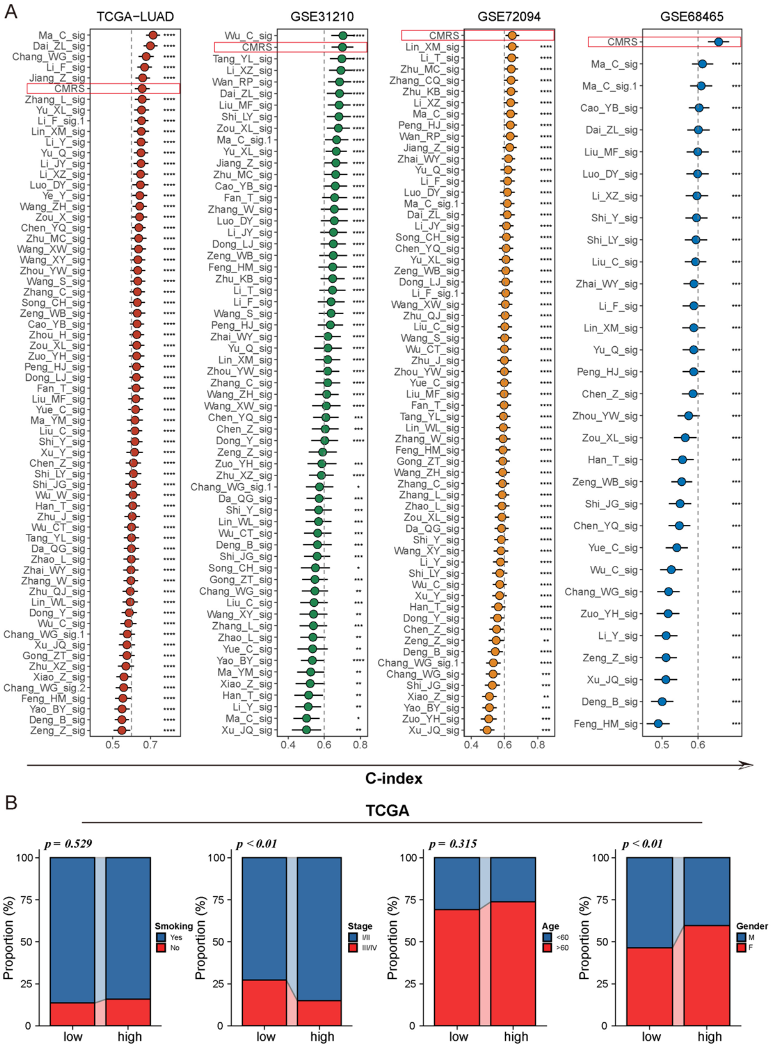

Fig. 5 Comparison between consensus machine learning-related prognostic signature (CMRS) and 81 prognostic signatures. (**A**) Comparison of C-index between CMRS and 81 published LUAD signatures (some prognostic signatures were excluded due to genetic incompleteness within the cohort). (**B**) The stacked plot of the distribution of CMRS versus clinicopathologic features (Chi-square test)

predicted and ideal curves (Supplementary Fig. 6B). The DCA curves also show that the nomogram has a good predictive benefit (Supplementary Fig. 6C).

CMRS is associated with the reshaping of the tumor microenvironment

The heterogeneity of the tumor microenvironment is considered to be one of the important factors affecting cancer patients and anticancer therapy (Xiao and Yu 2021). Here, we evaluated the tumor microenvironment of LUAD using the “IOBR” R package (Zeng et al. 2021). As expected, we found significant differences in TME signatures between the high and low CMRS groups, where signatures associated with tumor proliferation and metastasis including DDR, CellCycle_Reg, CellCycle, and EMT2 were more enriched in the low CMRS group, which partly explains why the prognosis of the low CMRS group was worse than that of the high CMRS group (Fig. 6A; Supplementary Fig. 7A-C). In terms of immune cell infiltration, we found that immune cells such as Type 17 helper cells, Plasmacytoid dendritic cells, Immature B cells, and Activated B cells were more active in the high CMRS group, whereas Activated CD4 T cell and CD56bright natural killer cell were significantly enriched in the low CMRS group. (Fig. 6B).

In the field of cancer therapy, targeted therapies against immunomodulators (IM) have been widely put into clinical trials, demonstrating the importance of IM for anticancer therapy. We evaluated the correlation between IM and mRNA expression, mRNA expression and methylation, amplification frequency and deletion frequency under different CMRS groups based on the protocol of Thorsson et al. The results showed that IM differed significantly in different CMRS groups. mRNA expression of immune checkpoints including CD276, CD274, and PDCD1LG2 were highly expressed in the low CMRS group; the expression of most of the antigen-presenting molecules including the HLA family was up-regulated in the high CMRS group; which implies that immune cells in the high CMRS group tended to be more immunoreactive. In addition, the methylation levels

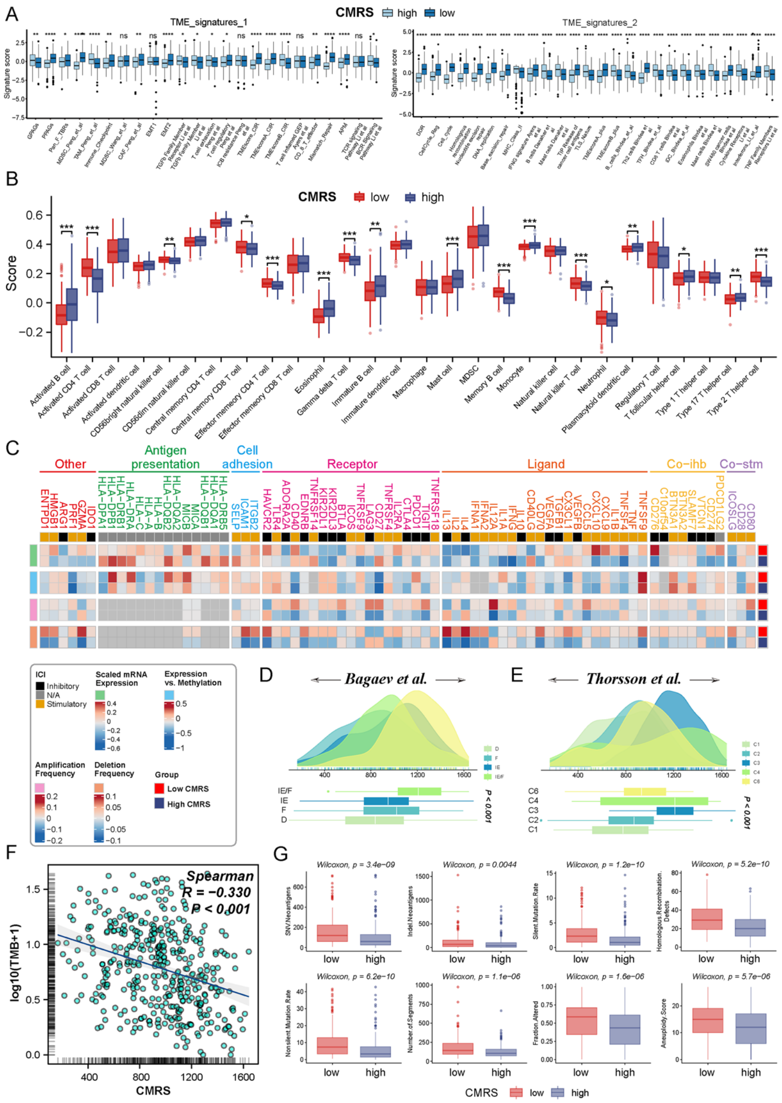

Fig. 6 Association of CMRS with molecular features of the tumor microenvironment. (**A**) Differences in the distribution of TME signatures between patients with high and low CMRS. (**B**) Differences in immune cell infiltration between patients with high and low CMRS. (**C**) From top to bottom: mRNA expression (median expression level after normalization using scale); correlation between mRNA expression and methylation; amplification frequency (difference between the proportion of samples in a given CMRS group in which amplification of IM occurs versus the proportion of all samples in which amplification occurs); deletion frequency (the difference between the proportion of samples in a given CMRS group in which IM undergoes deletion and the proportion of deletions in all samples). (**D**-**E**) Distribution of CMRS expression in known TME subtypes: Bagaev et al: IE/F (Immune-enriched, fibrotic), IE (Immune-enriched, non-fibrotic), F (Fibrotic), and D (Immune-depleted); Thorsson et al: C1 (Wound Healing), C2 (IFN-γ Dominant), C3 (Inflammatory), C4 (Lymphocyte Depleted) and C6 (TGF-β Dominant). (**F**) Scatter plot of the correlation between CMRS and TMB. (**G**) Boxplot of mutation type distribution between different CMRS groups

of CD40, CD27, and CTLA4 showed a negative correlation with mRNA expression levels, suggesting epigenetic silencing of these immune genes. The occurrence of CNV also affects the function of IM. We found that the low CMRS group exhibited higher frequent amplification and deletion of IM genes compared to the high CMRS group, suggesting that more pronounced genomic instability exists in the low CMRS group: IM genes including TIGIT, PDCD1, and LAG3 had higher amplification frequencies in the low CMRS group; meanwhile, IM genes including IL13, IL2, IL4, IFNA1 and IFNA2, which can activate T cell functions, showed frequent deletions in the low CMRS group (Fig. 6C).

Next, to gain a deeper understanding of the relationship between CMRS and the defined classification of TME subtypes. We found that CMRS scores were higher in the IE/F subtype and lower in the D subtype (Fig. 6D); among the immune subtypes defined by Thorsson et al. (Thorsson et al. 2018), CMRS scores were higher in the C3 subtype and, relatively, lower in the C1 and C2 subtypes (Fig. 6E). Notably, the C3 subtype mainly showed low to moderate proliferation of tumor cells and low mutational variability. In contrast, C1 and C2 subtypes of tumors are characterized by high proliferation, high mutation rates, and susceptibility to immune escape. Thus, we further analyzed the correlation between CMRS and genomic instability. As expected, CMRS showed a negative correlation with TMB (R= -0.33, *P < 0.001*) (Fig. 6F). And all common mutation types appeared more frequently in the low CMRS group (*P < 0.05*) (Fig. 6G). Overall, CMRS can be a highly efficient indicator for assessing the reshaping of the tumor microenvironment, immune cell infiltration, and mutation frequency in LUAD, and the abnormalities of immune function and frequent mutations may also be one of the reasons why patients with low CMRS have a shorter survival time.

CMRS can efficiently predict immunotherapy response

Considering the relevance of CMRS to the immune microenvironment, we next explored and evaluated the role of CMRS in immunotherapy. We analyzed the correlation between CMRS and common immune checkpoints in four LUAD cohorts. The results showed that CMRS showed a negative correlation with immune checkpoints in all cohorts (Supplementary Fig. 8A). In addition, Tracking Tumor Immunophenotype (TIP) analysis also revealed differences in anti-tumor immune processes between high and low CMRS groups (Supplementary Fig. 8B), suggesting the potential immunotherapy predictive ability of CMRS. Further, we evaluated the TIDE scores and immunotherapy response rates of these four LUAD cohorts using the TIDE algorithm analysis. The results showed that CMRS showed a negative correlation with TIDE in all cohorts (Fig. 7A), and high CMRS exhibited a better therapeutic effect to immunotherapy (*P [Chi-square test] < 0.001*). (Fig. 7B). Subsequently, we evaluated the predictive efficacy of CMRS in two NSCLC cohorts receiving immunotherapy, and results showed that patients in the high CMRS group tended to exhibit longer progression-free survival time after receiving immunotherapy, (GSE126044, *P = 0.0267*; GSE135222, *P = 0.0301*) (Fig. 7C). Furthermore, we followed another large mUC immunotherapy cohort (IMvigor210) to assess the predicted generalization ability of immunotherapy for CMRS. We found that in the IMvigor210 cohort, CMRS still showed better immunotherapy efficacy prediction, which means patients in the high CMRS group had better survival outcomes, although the *P-value* was not significant (Fig. 7D). Considering the possible delayed clinical effects of immunotherapy. We further compared the difference in long-term survival (LTS) between patients in the high and low CMRS groups after 3 months of immunotherapy (Fig. 7E), and also assessed the restricted mean survival (RMS) of patients at 6 and 12 months (Fig. 7F); the results consistently showed that the patient group with higher CMRS showed better prognostic outcomes, with the benefit of immunotherapy obtained greater.

Screening for potential therapeutic drugs based on CMRS

Next, we further analyzed the differences in biological functions present between patients in the high CMRS and low CMRS groups. (Fig. 8A) showed the differential gene distribution present between the different CMRS groups. GSEA analysis revealed the presence of significant activation of oncogenic pathways including cell cycle, hypoxia, EMT, and angiogenesis-related oncogenic pathways in low CMRS (Fig. 8B). Subsequent drug sensitivity analysis showed that patients with low CMRS demonstrated significantly higher sensitivity to commonly used anticancer drugs compared to patients with high CMRS (Fig. 8C). Considering that chemotherapeutic agents are primarily therapeutic mechanisms that target the ability of cells to divide and proliferate (Dasari and Tchounwou 2014), we believe that this result (Fig. 8C) may be related to the highly proliferative malignant nature of tumors in patients with low CMRS. Next, in order to better treat LUAD in the high CMRS group. We screened potential therapeutic agents for the high CMRS group. To ensure the robustness of our approach, we used cisplatin, a chemotherapeutic agent widely used in lung adenocarcinoma, as a mean of verifying that the sensitivities derived from the algorithm were consistent with established clinical approaches. A previous study showed that ERCC1 is a prognostic biomarker for patients with advanced non-small cell lung cancer receiving cisplatin chemotherapy (Ceppi et al. 2006), and our algorithm gave similar results (Fig. 8D), which means it was shown that patients with lower expression of ERCC1 demonstrated better therapeutic outcomes to cisplatin treatment. This result demonstrates the accuracy of our algorithm. (Fig. 8E) demonstrates our workflow to explore potential drugs for patients with high CMRS. After systematic screening, we finalized 2 CTRP-derived drugs (neratinib, oligomycin A; Fig. 8F) and 4 PRISM-derived drugs (prednisolone, BIBU-1361, aspirin, and RITA; Fig. 8G). The estimated AUC values of these drugs were not only significantly negatively correlated with CMRS, but also significantly lower in the high CMRS group, and may be able to serve as potential therapeutic agents for LUAD patients in the high CMRS group.

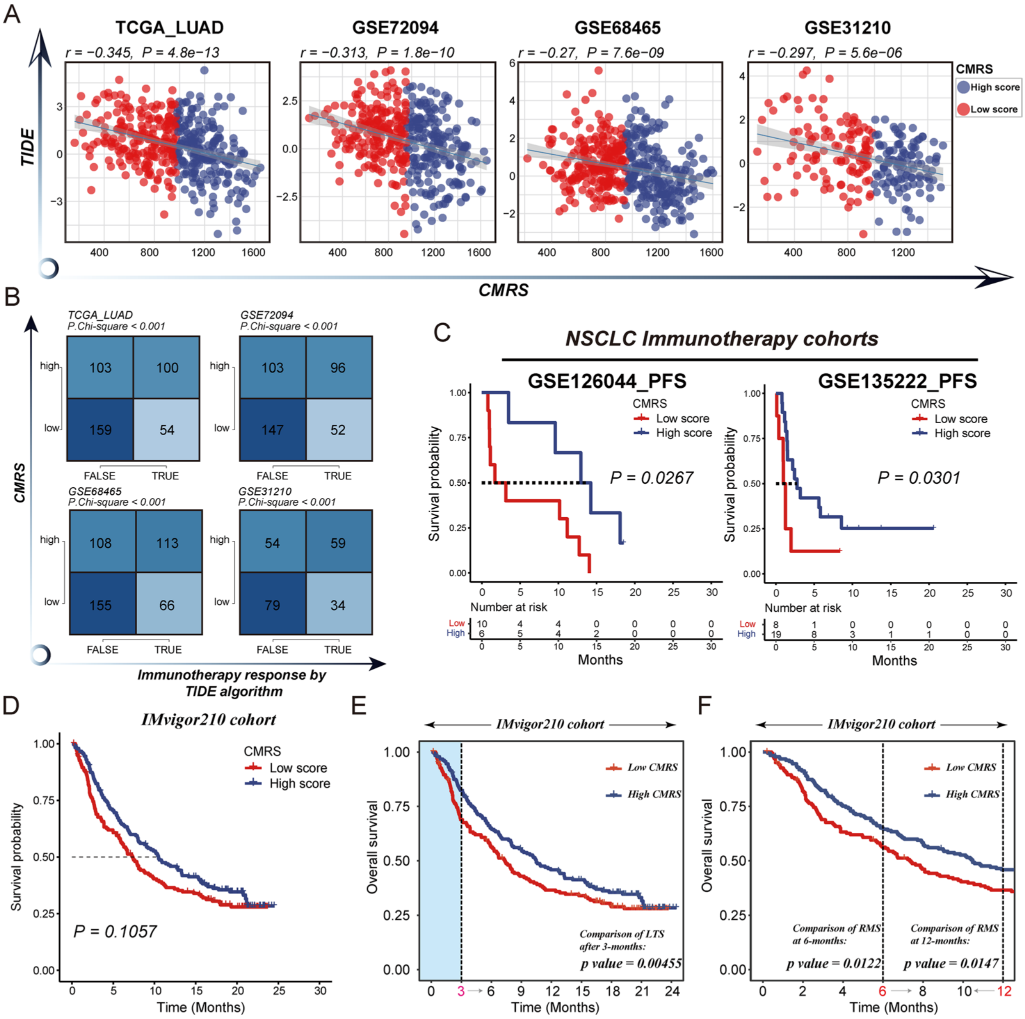

Fig. 7 Predictive value of CMRS in immunotherapy. (**A**) Scatter-plot of correlation between CMRS and TIDE score in the 4 LUAD cohorts. (**B**) Relationship between high and low CMRS and response to immunotherapy in the 4 LUAD cohorts, Chi-square test was used. (**C**) K-M plot of CMRS in predicting patient progress-free survival (PFS) in two NSCLC immunotherapy cohorts (GSE126044 and GSE135222). (**D**) Survival plot of CMRS in an immunotherapy cohort (IMvigor210) of mUC. (**E**) The difference in the long-term survival (LTS) between the high CMRS group and the low CMRS group after 3 months of treatment. (**F**) Difference in the restricted mean survival (RMS) at 6 and 12 months of treatment in the high and low CMRS groups

Single-cell sequencing analysis of CMRS-related genes

Next, we further analyzed the cellular localization of CMRS-associated genes in a single-cell dimension. We first acquired 18 lung adenocarcinoma samples within GSE148071. We clustered and annotated the cells and visualized them in a dimensionality reduction using the UMAP method. The results showed that the cells were categorized into 21 cell clusters (Fig. 9A) and annotated into 7 major classes (Fig. 9B), including Epithelial cells, Myeloid cells, B-cells, Fibroblasts, T-cells, Endothelial cells as well as other cells. (Fig. 9C) shows the biomarkers for each cell type. After that, we considered the expression distribution of the 13 genes that make up the CMRS, and we found that most of the genes were predominantly distributed in epithelial cells, and some of the genes, including SLC2A1 and TXNRD1, were also activated in Myeloid cells (Fig. 9D). Subsequently, we calculated these genes as CMRG_score using the function "AddModuleScore" and used UMAP plots for visualization (Fig. 9E). In addition, we found that CMRG_score was more actively expressed mainly in epithelial cells (Fig. 9F). (Fig. 9G) shows the distribution of CMRG_score expression in each sample.

Discussion

Cancer occurs as a result of the interaction of multiple factors, including gene mutations, epigenetic variants, changes in cellular identity, chronic inflammation and tumor

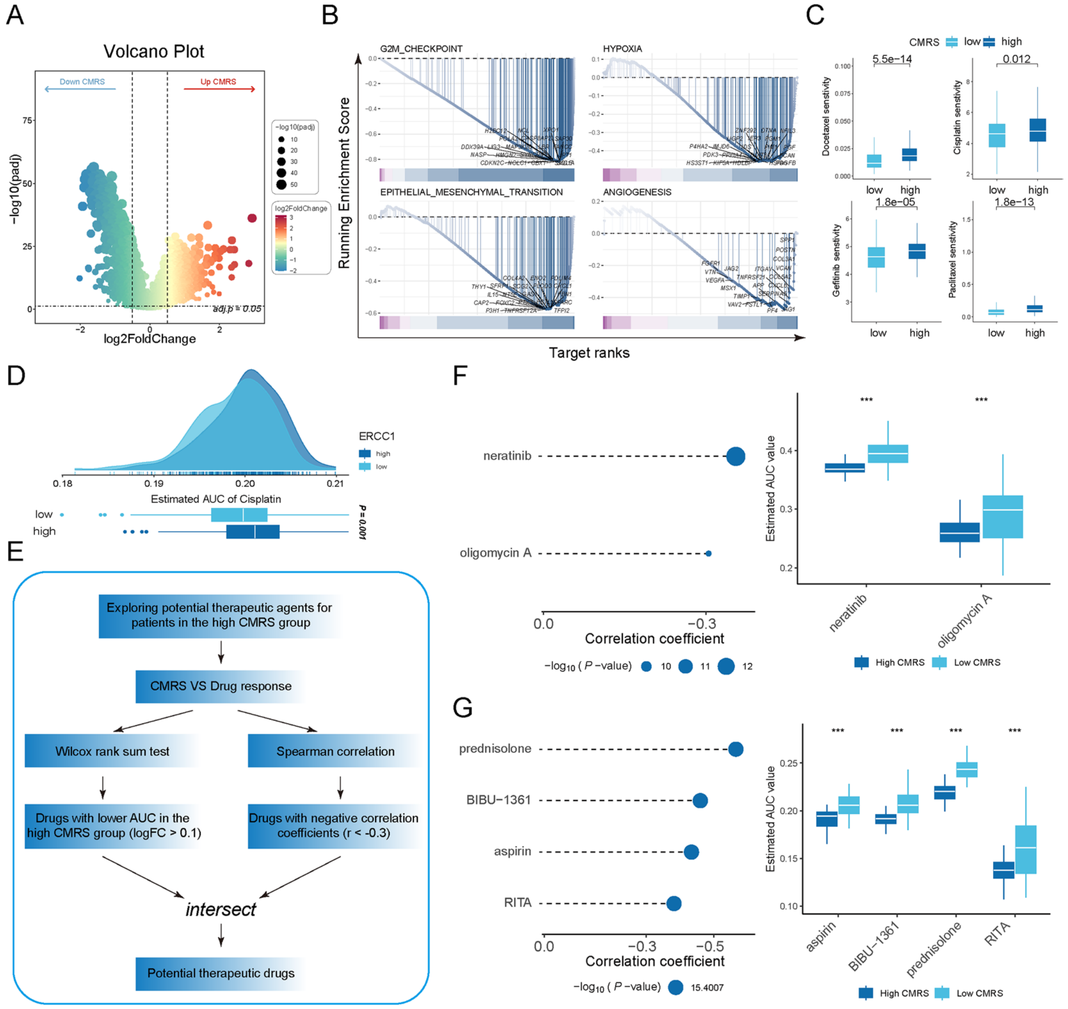

Fig. 8 Potential drug screening for patients. (**A**) Volcano plot showing differential gene distribution between patients in different CMRS groups. (**B**) GSEA analysis shows significantly activated oncogenic pathways in patients with low CMRS group. (**C**) Distribution of drug IC_{50} values in different CMRS groups (analyzed using "oncoPredict" R package). (**D**) To validate the feasibility of the computational algorithm by predicting the relationship between ERCC1 expression and cisplatin sensitivity. (**E**) The computational pipeline is used to screen potential drugs for patients in the high CMRS group. The results of correlation analysis and differential drug response analysis of (**F**) CTRP-derived compounds and (**G**) PRISM-derived compounds, and the lower values on the estimated AUC in boxplots imply greater drug sensitivity

microenvironment (Jassim et al. 2023). Therefore, comprehensive analysis of multidimensional comprehensive omics data from patients will contribute to our further understanding of disease-specific regulatory mechanisms (Oh et al. 2021). Currently, most of the comprehensive omics studies on LUAD have focused on individual dimensions of comprehensive omics data, such as mRNA expression, mutation status of genes, etc. (Feng et al. 2023); meanwhile, the selection of clustering methods based on comprehensive omics data is also different, which may make it limited to a certain extent when applied on a large scale (Ma et al. 2022).

Here, we successfully identified two LUAD subtypes with different molecular features by integrating multidimensional comprehensive omics data and combining 10 advanced clustering algorithms, and the stability and prognostic variability of the subtypes were validated in different cohorts. Among them, patients with CS1 have higher tumor invasive, more mutation frequency, and generalized activation of oncogenic pathways, which ultimately constituted the poor clinical outcome of CS1. Notably, we found that patients with the CS1 molecular subtype were predominantly enriched for the C1 (Wound Healing, characterized by elevated expression of angiogenic genes and a high rate of tumor proliferation) and C2 (IFN-γ Dominant, characterized by a high degree of lymphocytic infiltration, M1 macrophage infiltration, and a high rate of tumor proliferation) immune subtypes. This suggests that although there is a strong immune response within the TME in CS1 patients,

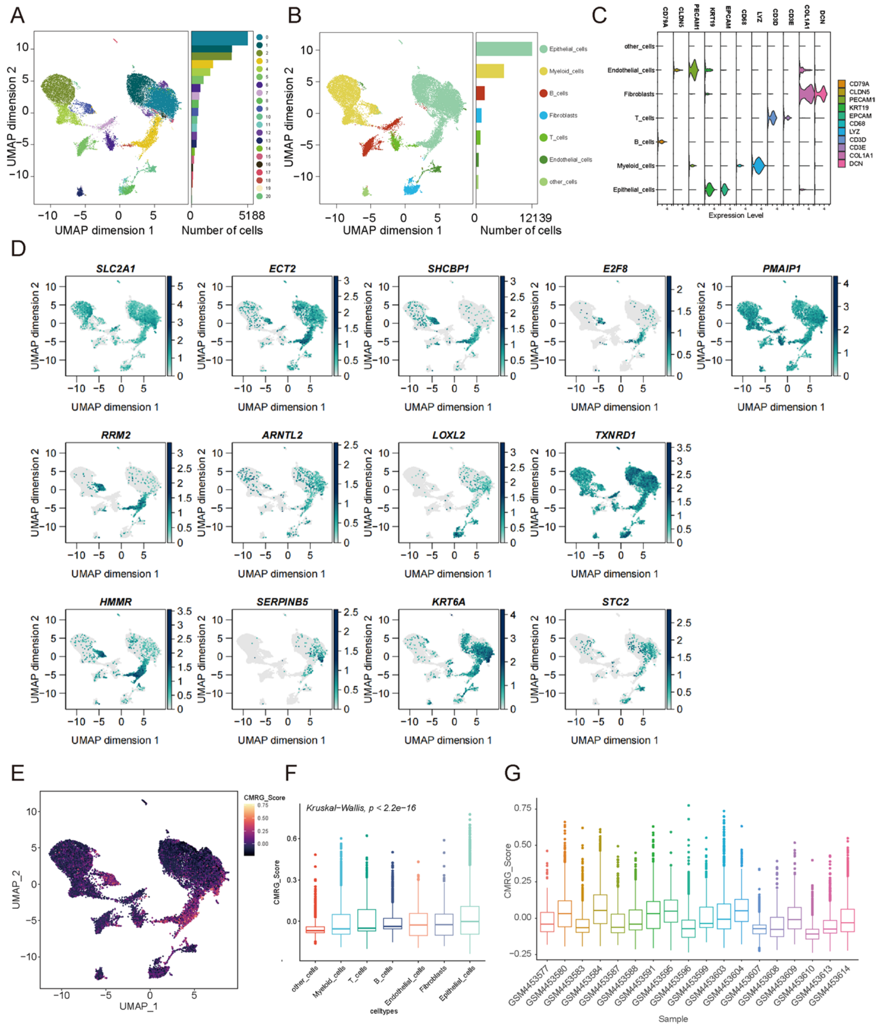

Fig. 9 ScRNA-Seq reveals CMRS-related gene expression patterns. UMAP plots of 21 cell clusters (**A**) and 7 major cell types (**B**). (**C**) Violin plots of cellular markers. (**D**) UMAP plot of expression distribution of CMRS constituent genes. (**E**) Calculation of the expression of (comprehensive omics subtypes related genes score) CMRG_score in major cell types based on "AddModuleScore" function. Expression distribution of CMRG_score in cell types (**F**) and samples (**G**)

the high proliferation rate and high mutation frequency of tumor cells result in the inability to control tumor progression even with an active immune response; in contrast, CS2 patients predominantly enriched the C3 (Inflammatory, mainly characterized by elevated Th17 and Th1, and a lower proliferation rate) immune subtype; this implies that CS1 and CS2 patients may have different degrees of benefit when treated with targeted immune responses. Overall, the comprehensive omics subtypes identified in this study may be of possible value for precise stratification of LUAD.

Machine learning algorithms are currently the dominant means of building efficient predictive markers (Lee et al. 2021). We constructed an efficient prognostic signature (CMRS) by an integrated program containing 10 machine-learning algorithms (Wang et al. 2022). For the 13 genes constituting the CMRS, we detailed them in (Supplementary table.5). Compared to most existing prognostic signatures for LUAD, CMRS demonstrate robust and stable prognostic prediction efficiency. Overfitting is an important issue that cannot be ignored by machine learning algorithms when constructing models, i.e., they show good performance in the model training set, yet fail to achieve satisfactory predictions in the external cohort (Deo 2015). Considering that cohorts originating from TCGA , which are often used as training sets, this partly explains why the performance of some of the prognostic signatures is superior to CMRS in the TCGA cohort but does not show good efficiency in other LUAD cohorts. In addition, the nomogram constructed based on CMRS and clinicopathologic features also showed good efficiency in survival prediction, which is expected to be useful as a prognostic indication for LUAD patients.

In terms of tumor microenvironment expression patterns, we identified significant TME heterogeneity between the high and low CMRS groups. Among them, not only cell proliferation and EMT-related signatures were significantly enriched in the low-CMRS group, but also immune checkpoints including CD274 (PD-L1), CD276, PDCD1, and TIGIT were actively expressed in the low-CMRS group; these results imply that the low-CMRS group had higher proliferation rates, more malignant characteristics of the tumor cells and the immune microenvironment was biased towards immunosuppression. Furthermore, the amplification and deletion of IM genes, which are prevalent in the low-CMRS group, largely lead to the abnormal function of the immune system and thus the inability to control the malignant progression of the tumors (Gubin and Vesely 2022; O'Donnell et al. 2019). This phenomenon implies that patients in the low-CMRS group may not be able to benefit well from immunotherapy. We verified this speculation under the subsequent TIDE algorithm. Compared with the low CMRS group, the high CMRS group showed a better immunotherapy response rate. Similar results were validated in an independent NSCLC immunotherapy cohort and a large mUC immunotherapy cohort. These results consistently suggest that CMRS may serve as strong evidence for identifying immunotherapy-sensitive populations.

GSEA analysis showed that there was a generalized activation of pathways related to EMT, cell cycle and angiogenesis in the low CMRS group. This is consistent with the phenotype of high proliferation rate in low CMRS. Furthermore, in terms of clinical treatment, the low-CMRS group demonstrated sensitivity to common chemotherapeutic agents such as Docetaxel and Cisplatin. It is well known that the mechanism of action of chemotherapeutic agents is mainly to inhibit tumor development by hindering cell division and proliferation (Dasari and Tchounwou 2014; Jin et al. 2023). Therefore, we believe that it may be the high proliferative characteristics of tumor cells in the low-CMRS group that make patients in the low-CMRS group more likely to benefit from chemotherapeutic agents. Afterward, in order to expand the clinical treatment scope of the high CMRS group. We screened neratinib, oligomycin A, and other drugs as potential agents targeting the high CMRS group based on the protocol summarized in previous studies (Yang et al. 2021). Among them, Neratinib, as a tyrosine kinase inhibitor, has been mainly used in the treatment of breast cancer, and it can achieve anticancer effects by blocking the transduction of HER1, HER2, and HER4 signaling pathways (Guo et al. 2023). A study on oligomycin A has shown that Oligomycin A can be effective by inducing apoptosis in Hela cells mediated by TRAIL (He et al. 2013). More clinical trials are needed in the future to confirm the potential role of these drugs in the therapeutic field of LUAD in the high CMRS group.

Overall, in this study, the novel molecular subtypes and robust prognostic signature identified, which are promising in terms of clinical implications for LUAD. However, the limitations of this study should also be recognized. First, all samples in this study were retrospective, and future validation for CMRS should be performed in prospective multicenter cohorts. Second, further experimental exploration is still needed for the specific mechanisms between CMRS-related genes and tumor progression. Third, the effects of potential drugs screened based on CMRS in LUAD still need to be confirmed by further clinical trials.

Conclusion

In summary, this study identified two molecular subtypes with different molecular feature heterogeneity in LUAD through multi-omics consensus clustering, which may be able to drive the refinement of stratified and precise treatment of LUAD. The CMRS developed by the framework of machine learning algorithms has a strong prognostic prediction capability and demonstrated the good ability to identify immunotherapy-sensitive populations. This study may helpful for the early diagnosis and the personalized treatment in LUAD.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10142-024-01388-x.

Author contributions DH and JW: conception and design, and study supervision. TH, YB, YD, CL, LG, YL, JZ, JG: development of methodology, performance of assays, analysis and interpretation of data, and writing of the manuscript. DH, JW and YB: review of the manuscript.

Funding This study was sponsored by Scientific Research Foundation for High-level Talents of Anhui University of Science and Technology (2022y jrc14), Youth Fund of Anhui University of Science and Technology (QNZD202203), Open Research Fund of Anhui Province Engineering Laboratory of Occupational Health and Safety (AYZJSGCLK202201002), Collaborative Innovation Project of Colleges and Universities of Anhui Province (GXXT-2020-058), Anhui Province Engineering Laboratory of Occupational Health and Safety (AYZJSGCLK202201001, AYZJSGCLK 202202001), Key Laboratory of Industrial Dust Deep Reduction and Occupational Health and Safety of Anhui Higher Education Institutes (AYZJSGXLK202202002). The funders had no role in the study design, data collection, analysis, and decision to publish or preparation of the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflicts of Interest The authors declare no competing interests.

References

Addeo A, Friedlaender A, Banna GL, Weiss GJ (2021) TMB or not TMB as a biomarker: that is the question. Crit Rev Oncol Hematol 163:103374

Ai L, Xu A, Xu J (2020) Roles of PD-1/PD-L1 pathway: signaling, cancer, and beyond. Adv Exp Med Biol 1248:33–59

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV et al (2013) Signatures of mutational processes in human cancer. Nature. 500(7463):415–421

Ayob AZ, Ramasamy TS (2018) Cancer stem cells as key drivers of tumour progression. J Biomed Sci 25(1):20

Bagaev A, Kotlov N, Nomie K, Svekolkin V, Gafurov A, Isaeva O et al (2021) Conserved pan-cancer microenvironment subtypes predict response to immunotherapy. Cancer Cell 39(6):845–865.e7

Caso R, Sanchez-Vega F, Tan KS, Mastrogiacomo B, Zhou J, Jones GD et al (2020) The underlying tumor genomics of predominant histologic subtypes in lung adenocarcinoma. J Thorac Oncol 15(12):1844–1856

Castellanos E, Feld E, Horn L (2017) Driven by mutations: the predictive value of mutation subtype in EGFR-mutated non-small cell lung cancer. J Thorac Oncol 12(4):612–623

Ceppi P, Volante M, Novello S, Rapa I, Danenberg KD, Danenberg PV et al (2006) ERCC1 and RRM1 gene expressions but not EGFR are predictive of shorter survival in advanced non-small-cell lung cancer treated with cisplatin and gemcitabine. Ann Oncol 17(12):1818–1825

Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D et al (2017) Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. Cell Rep 18(1):248–262

Dasari S, Tchounwou PB (2014) Cisplatin in cancer therapy: molecular mechanisms of action. Eur J Pharmacol 740:364–378

Denisenko TV, Budkevich IN, Zhivotovsky B (2018) Cell death-based treatment of lung adenocarcinoma. Cell Death Dis 9(2):117

Deo RC (2015) Machine learning in medicine. Circulation. 132(20):1920–1930

Feng H-M, Zhao Y, Yan W-J, Li B (2023) Genomic and immunogenomic analysis of three prognostic signature genes in LUAD. BMC Bioinform 24(1):19

Genova C, Dellepiane C, Carrega P, Sommariva S, Ferlazzo G, Pronzato P et al (2021) Therapeutic implications of tumor microenvironment in lung cancer: focus on immune checkpoint blockade. Front Immunol 12:799455

Greener JG, Kandathil SM, Moffat L, Jones DT (2022) A guide to machine learning for biologists. Nat Rev Mol Cell Biol 23(1):40–55

Gubin MM, Vesely MD (2022) Cancer immunoediting in the era of immuno-oncology. Clin Cancer Res 28(18):3917–3928

Guo L, Shao W, Zhou C, Yang H, Yang L, Cai Q et al (2023) Neratinib for HER2-positive breast cancer with an overlooked option. Mol Med 29(1):134

Han T, Liu Y, Wu J, Bai Y, Zhou J, Hu C et al (2023) An immune indicator based on BTK and DPEP2 identifies hot and cold tumors and clinical treatment outcomes in lung adenocarcinoma. Sci Rep 13(1):5153

Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A et al (2021) Integrated analysis of multimodal single-cell data. Cell 184(13):3573–3587.e29

He L, Jang JH, Choi HG, Lee SM, Nan MH, Jeong SJ et al (2013) Oligomycin a enhances apoptotic effect of TRAIL through CHOP-mediated death receptor 5 expression. Mol Carcinog 52(2):85–93

Hu J, Song Y, Cai X, Halina H, Qiao K, Lu J et al (2023) A novel pyroptosis gene expression-based risk score for survival in gastric cancer. Front Endocrinol (Lausanne) 14:1120216

Jassim A, Rahrmann EP, Simons BD, Gilbertson RJ (2023) Cancers make their own luck: theories of cancer origins. Nat Rev Cancer 23(10):710–724

Jiang Y-Z, Ma D, Suo C, Shi J, Xue M, Hu X et al (2019) Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies. Cancer Cell 35(3):428–440.e5

Jin H, Wang L, Bernards R (2023) Rational combinations of targeted cancer therapies: background, advances and challenges. Nat Rev Drug Discov 22(3):213–234

Jin R, Wang X, Zang R, Liu C, Zheng S, Li H et al (2020) Desmoglein-2 modulates tumor progression and osimertinib drug resistance through the EGFR/Src/PAK1 pathway in lung adenocarcinoma. Cancer Lett 483:46–58

Kay FU, Kandathil A, Batra K, Saboo SS, Abbara S, Rajiah P (2017) Revisions to the tumor, node, metastasis staging of lung cancer (8th edition): rationale, radiologic findings and clinical implications. World J Radiol 9(6):269–279

Lahiri A, Maji A, Potdar PD, Singh N, Parikh P, Bisht B et al (2023) Lung cancer immunotherapy: progress, pitfalls, and promises. Mol Cancer 22(1):40

Lee D, Park Y, Kim S (2021) Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches. Brief Bioinform 22(3):bbaa188

Lentz RW, Colton MD, Mitra SS, Messersmith WA (2021) Innate immune checkpoint inhibitors: the next breakthrough in medical oncology? Mol Cancer Ther 20(6):961–974

Lu X, Meng J, Su L, Jiang L, Wang H, Zhu J et al (2021b) Multi-omics consensus ensemble refines the classification of muscle-invasive bladder cancer with stratified prognosis, tumour microenvironment and distinct sensitivity to frontline therapies. Clin Transl Med 11(12):e601

Lu X, Meng J, Zhou Y, Jiang L, Yan F (2021a) MOVICS: an R package for multi-omics integration and visualization in cancer subtyping. Bioinformatics. 36(22-23):5539–5541

Ma C, Wu M, Ma S (2022) Analysis of cancer omics data: a selective review of statistical techniques. Brief Bioinform 23(2)

Madeddu C, Donisi C, Liscia N, Lai E, Scartozzi M, Macciò A (2022) EGFR-mutated non-small cell lung cancer and resistance to immunotherapy: role of the tumor microenvironment. Int J Mol Sci 23(12):6489

Maeser D, Gruener RF, Huang RS (2021) oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. Brief Bioinform 22(6):bbab260

Malta TM, Sokolov A, Gentles AJ, Burzykowski T, Poisson L, Weinstein JN et al (2018) Machine learning identifies stemness features associated with oncogenic dedifferentiation. Cell 173(2):338–354.e15

Mariathasan S, Turley SJ, Nickles D, Castiglioni A, Yuen K, Wang Y et al (2018) TGFβ attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. Nature. 554(7693):544–548

O'Donnell JS, Teng MWL, Smyth MJ (2019) Cancer immunoediting and resistance to T cell-based immunotherapy. Nat Rev Clin Oncol 16(3):151–167

Oh M, Park S, Kim S, Chae H (2021) Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations. Brief Bioinform 22(1):66–76

Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R et al (2012) Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res 72(1):100–111

Passaro A, Brahmer J, Antonia S, Mok T, Peters S (2022) Managing resistance to immune checkpoint inhibitors in lung cancer: treatment and novel strategies. J Clin Oncol Off J Am Soc Clin Oncol 40(6):598–610

Reel PS, Reel S, Pearson E, Trucco E, Jefferson E (2021) Using machine learning approaches for multi-omics data analysis: a review. Biotechnol Adv 49:107739

Sabbah DA, Hajjo R, Sweidan K (2020) Review on Epidermal Growth Factor Receptor (EGFR) structure, signaling pathways, interactions, and recent updates of EGFR inhibitors. Curr Top Med Chem 20(10):815–834

Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S et al (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun 9(1):20

Sellars MC, Wu CJ, Fritsch EF (2022) Cancer vaccines: building a bridge over troubled waters. Cell. 185(15):2770–2788

Shao J, Zhang Y, Chang Z, Du S, Li W, Bai Y et al (2023) A novel ferroptosis-related microRNA signature with prognostic value in osteosarcoma. Acta Biochim Biophys Sin Shanghai 55(11):1758–1769

Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer statistics, 2021. CA Cancer J Clin 71(1):7–33

Sterner RC, Sterner RM (2021) CAR-T cell therapy: current limitations and potential strategies. Blood Cancer J 11(4):69

Thai AA, Solomon BJ, Sequist LV, Gainor JF, Heist RS (2021) Lung cancer. Lancet 398(10299):535–554

Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang T-H et al (2018) The immune landscape of cancer. Immunity 48(4):812–830.e14

Wang L, Liu Z, Liang R, Wang W, Zhu R, Li J et al (2022) Comprehensive machine-learning survival framework develops a consensus model in large-scale multicenter cohorts for pancreatic cancer. Elife 11:e80150

Wu L, Qu X (2015) Cancer biomarker detection: recent achievements and challenges. Chem Soc Rev 44(10):2963–2997

Xiao Y, Yu D (2021) Tumor microenvironment as a therapeutic target in cancer. Pharmacol Ther 221:107753

Yang C, Huang X, Li Y, Chen J, Lv Y, Dai S (2021) Prognosis and personalized treatment prediction in TP53-mutant hepatocellular carcinoma: an in silico strategy towards precision oncology. Brief Bioinform 22(3):bbaa164

Yi M, Zheng X, Niu M, Zhu S, Ge H, Wu K (2022) Combination strategies with PD-1/PD-L1 blockade: current advances and future directions. Mol Cancer 21(1):28

Yoo S, Sinha A, Yang D, Altorki NK, Tandon R, Wang W et al (2022) Integrative network analysis of early-stage lung adenocarcinoma identifies aurora kinase inhibition as interceptor of invasion and progression. Nat Commun 13(1):1592

Zeng D, Ye Z, Shen R, Yu G, Wu J, Xiong Y et al (2021) IOBR: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. Front Immunol 12:687975

Zhao M, Sun J, Zhao Z (2013) TSGene: a web resource for tumor suppressor genes. Nucleic Acids Res 41(Database issue):D970–D9D6

Zhao X, Li X, Zhou L, Ni J, Yan W, Ma R et al (2018) LncRNA HOXA11-AS drives cisplatin resistance of human LUAD cells via modulating miR-454-3p/Stat3. Cancer Sci 109(10):3068–3079

Zheng H, Liu H, Li H, Dou W, Wang J, Zhang J et al (2022) Characterization of stem cell landscape and identification of stemness-relevant prognostic gene signature to aid immunotherapy in colorectal cancer. Stem Cell Res Ther 13(1):244

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