International Journal of Molecular Sciences

MDPI

Article

Integrated Multi-Omics Analysis for Molecular Subtyping in NSCLC: A Cohort Study

Bin Yan [1], Zhangyu Wang [1], Yun Peng [2], Kefu Liu [1], Jian-Guo Zhou [3,*] and Shixiang Wang [1,*]

[1] Department of Biomedical Informatics, School of Life Sciences, Central South University, Changsha, China.
[2] Department of Urology, Peking University People's hospital, Beijing, China.
[3] Department of Oncology, The Second Affiliated Hospital of Zunyi Medical University, Zunyi, China.

* Correspondence: jianguo.zhou@zmu.edu.cn (J.-G.Z.); wangshx@csu.edu.cn (S.W.).

Abstract

Lung cancer remains the leading cause of cancer-related morbidity and mortality worldwide, with non-small cell lung cancer (NSCLC) accounting for approximately 85% of cases. Current histopathological classification and driver gene testing provide limited prognostic and therapeutic guidance due to intra-tumoral heterogeneity and incomplete characterization of the tumor microenvironment (TME). Here, we constructed single- and multi-omics molecular classification systems for NSCLC by integrating transcriptomic, genomic, epigenomic, proteomic, and TME data from TCGA LUAD and LUSC cohorts. Single-omics analyses revealed distinct molecular patterns, including DNA methylation subtypes associated with sex and histology. Multi-omics integration identified five consensus subtypes closely corresponding to histology, with three immune-activated (CS2–CS4) and two immunosuppressive (CS1, CS5) subtypes. While overall survival did not differ significantly, progression-free survival analysis highlighted CS4 as a high-risk subtype with *KRAS*-driven genomic alterations. These findings provide a framework for NSCLC molecular stratification and highlight molecular and immune features that could guide future research on targeted therapies and immunotherapy.

Keywords: non-small cell lung cancer; molecular subtyping; multi-omics integration; unsupervised clustering

Academic Editor: Firstname Lastname

Received: date
Revised: date
Accepted: date
Published: date

Citation: To be added by editorial staff during production.

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Lung cancer is one of the most prevalent malignancies worldwide. According to the GLOBOCAN 2022 data released by International Agency for Research on Cancer, in 2022, lung cancer had the highest incidence and estimated mortality rates among all cancers, making it once again the world's leading cancer[1]. Non-small cell lung cancer (NSCLC), which accounts for approximately 85% of lung cancer cases[2], has an overall 5-year survival rate of less than 20%. While early screening can reduce mortality, it is associated with a high false-positive rate. Moreover, the postoperative recurrence rate for stage I patients can reach 20%–40%, and the 5-year survival rate after recurrence and metastasis is only approximately 15%[3]. These phenomena highlight the limitations of current diagnostic and therapeutic approaches.

Currently, the clinical diagnosis and treatment of NSCLC rely primarily on histopathological classification and the Tumor-Node-Metastasis (TNM) staging system. However, this classification system has some limitations. For example, patients with

morphologically identical lung squamous cell carcinoma may have completely different prognoses due to distinct driver gene mutations (e.g., *NFE2L2*, *ASAH2*)[4]. Some early-stage patients may still experience recurrence due to molecular abnormalities, such as *EGFR* mutations, despite having a low tumor stage[5]. Conversely, some late-stage patients may benefit from targeted or immunotherapy due to specific molecular characteristics[6].

While driver gene testing has spurred progress in precision medicine for NSCLC, it is limited by several inherent constraints. These include: (1) a narrow testing scope, which predominantly focuses on a few well-established targets such as *EGFR*, *ALK*, and *ROS1*[7], potentially overlooking the broader molecular landscape of tumors; (2) the absence of identifiable driver mutations in a subset of patients, thereby excluding them from benefiting from targeted therapies; and (3) the near-inevitability of acquired resistance, exemplified by the development of resistance to osimertinib—a third-generation *EGFR* tyrosine kinase inhibitor (TKI)—within 2–3 years in most patients[8–10]. Furthermore, the information provided by driver gene testing alone is insufficient to reflect the complex interplay of factors such as the tumor microenvironment (TME), epigenetic modifications, and transcriptomic profiles, which collectively influence prognosis and treatment responses. Consequently, the traditional diagnostic and treatment model, which relies predominantly on histopathological classification, TNM staging, and driver gene testing, falls short of meeting the precise classification and treatment demands for NSCLC.

In recent years, multi-omics integration research has emerged as a new approach, covering genomics, transcriptomics, epigenomics, and proteomics, and offering a more comprehensive perspective for characterizing the molecular heterogeneity of NSCLC[11,12]. However, while some studies have classified tumors using TME as an independent omics layer[13–15], there is still a lack of systematic classification attempts that incorporate TME into a multi-omics integration system. Notably, key components of the TME, including immune cell infiltration and angiogenesis, are strongly associated with the survival outcomes of NSCLC patients and their responsiveness to immunotherapy, and may further modulate the efficacy of targeted therapies[16,17].

Based on these considerations, our study integrates multi-omics data from The Cancer Genome Atlas (TCGA) lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC) cohorts, including transcriptomics, somatic mutations, copy number variations, DNA methylation, and Reverse Phase Protein Array (RPPA) protein expression. In addition, we independently derived TME features from bulk transcriptome data. By performing consensus clustering at the single-omics level and applying ten multi-omics integration algorithms to construct consensus subtypes, we aim to establish a more comprehensive NSCLC classification system. This framework is intended to provide novel insights for precision diagnosis and treatment by systematically assessing the clinical prognosis and biological characteristics of different subtypes.

2. Results

2.1. Identification of NSCLC Molecular Subtypes by Single-Omics Consensus Clustering

To identify NSCLC molecular subtypes, we first examined six types of omics data, revealing pronounced differences in gene expression, DNA methylation, copy number variation, and protein expression, as generally expected for these histological subtypes (Supplementary Materials, Figure S1).

Unsupervised consensus clustering of gene expression data classified NSCLC samples into four transcriptomic subtypes (Figure 1). And the optimal number of clusters

was determined using the cumulative distribution function (CDF), Delta Area Plot, and Tracking Plot.. Clustering results for the other omics data indicated that, except for CNV data which favored five clusters, all other omics data were best represented by four clusters (Supplementary Materials, Figures S2–S6). These single-omics subtypes exhibited pronounced omics-specific features, providing a foundation for subsequent multi-omics integrative analysis.

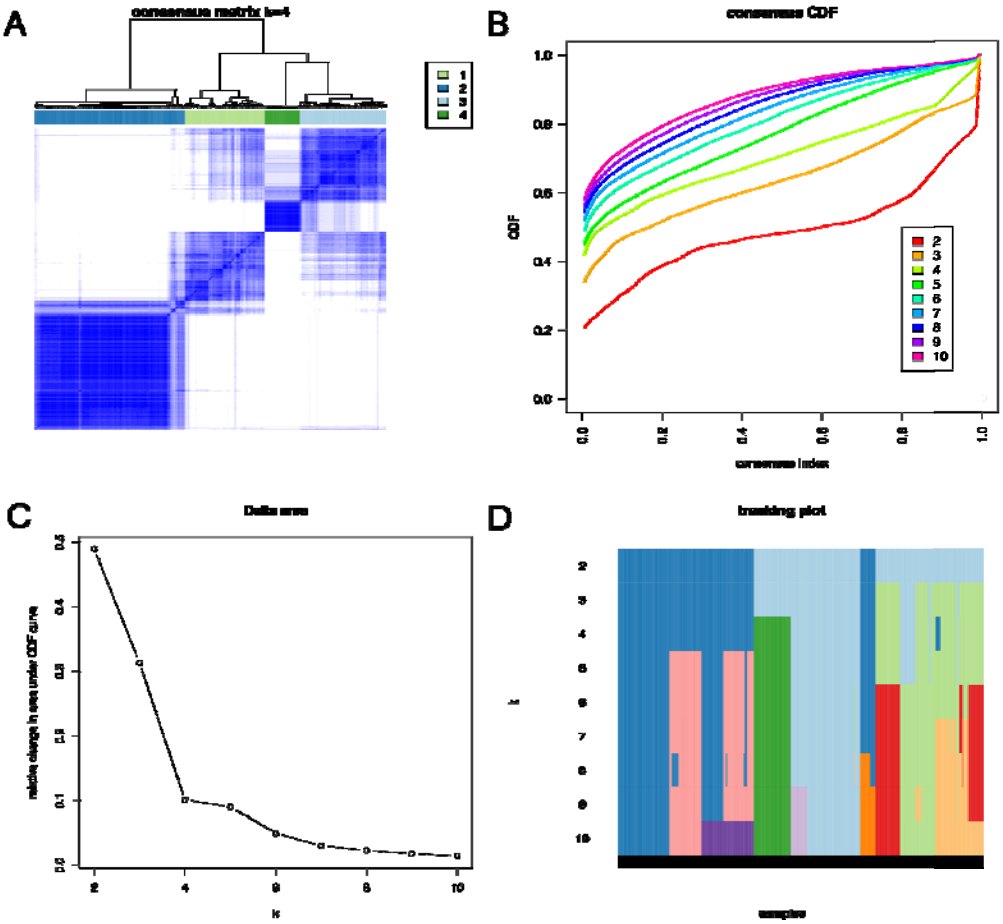


Figure 1. Consensus clustering results of gene expression data. (**A**) Consensus matrix heatmap at K = 4; (**B**) Cumulative distribution function plot; (**C**) Relative change in the area under the CDF curve (Delta Area Plot); (**D**) Sample classification tracking plot across different values of K.

2.2. Integrative Multi-Omics Clustering Reveals Five Robust NSCLC Subtypes

The optimal number of clusters for the multi-omics dataset was determined by calculating the clustering prediction index (CPI) and Gap statistic, both of which reached relatively high values at K = 5, indicating the best model performance at this resolution (Supplementary Materials, Figure S7).

Following this, consensus clustering was performed using ten machine learning algorithms. A total of 551 NSCLC samples were stratified into five robust multi-omics consensus subtypes (CS1–CS5), with the following sample sizes and proportions: CS1, 116 cases (21.1%); CS2, 113 cases (20.5%); CS3, 140 cases (25.4%); CS4, 125 cases (22.7%); and CS5, 57 cases (10.3%) (Figure 2). Clustering quality assessed by silhouette coefficients (Figure 2C) showed that CS2 and CS4 exhibited higher average widths, indicating that these clusters represent well-defined molecular subgroups suitable for downstream

analyses. In contrast, CS3 displayed a relatively low silhouette width (0.25), suggesting lower within-cluster similarity and potential overlap with neighboring subtypes.

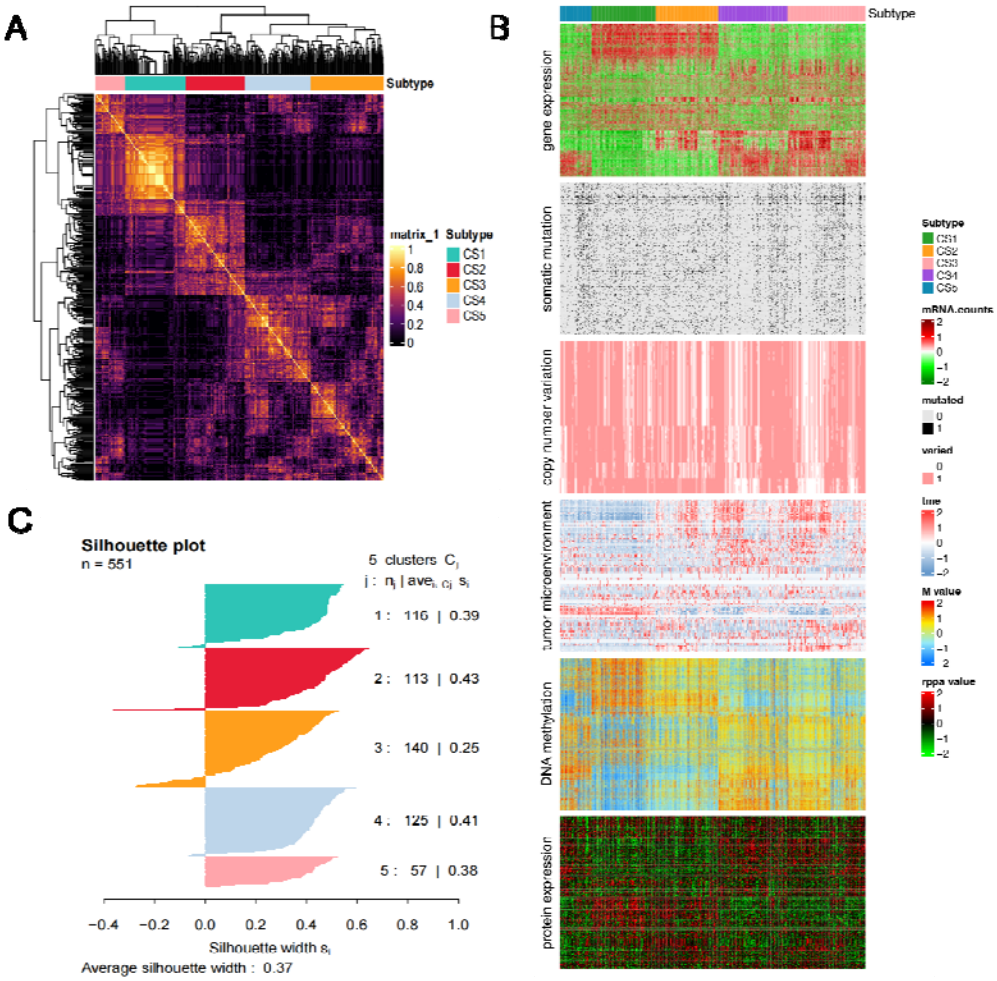


Figure 2. Multi-omics consensus subtypes of NSCLC. (**A**) Consensus matrix heatmap of the five subtypes; (**B**) Molecular landscapes of the five subtypes across six omics layers; (**C**) Clustering quality assessment based on silhouette coefficients.

2.3. Cross-Omics Consistency and Clinical Relevance of Single- and Multi-Omics Subtypes in NSCLC

To explore the associations between multi-omics consensus subtypes, single-omics subtypes, and clinical features (including histological type, clinical stage, and gender), we performed heatmap visualization (Figure 3A). The results revealed that certain subtype classifications (e.g., MOVICS, CNV, and MET) were strongly associated with histological type. Notably, the MET subtypes also showed strong correlations with patient gender, whereas proteomic and tumor microenvironment subtypes exhibited weaker associations.

We further examined the composition of histological types across different subtype results using stacked bar plots (Figure 3B, 3C, and Supplementary Figure S8). The multi-omics classification divided NSCLC into five consensus subtypes, which were highly correlated with histological type (Figure 3B). LUAD samples predominantly clustered in MOVICS clusters 3, 4, and 5, comprising 108 (87.9%), 114 (91.2%), and 57 (100%) patients

respectively. LUSC samples were mainly distributed in clusters 1 and 2, with 109 (94.0%) and 109 (96.5%) patients, respectively. These five subtypes were subsequently designated CS1–CS5 for downstream analyses.

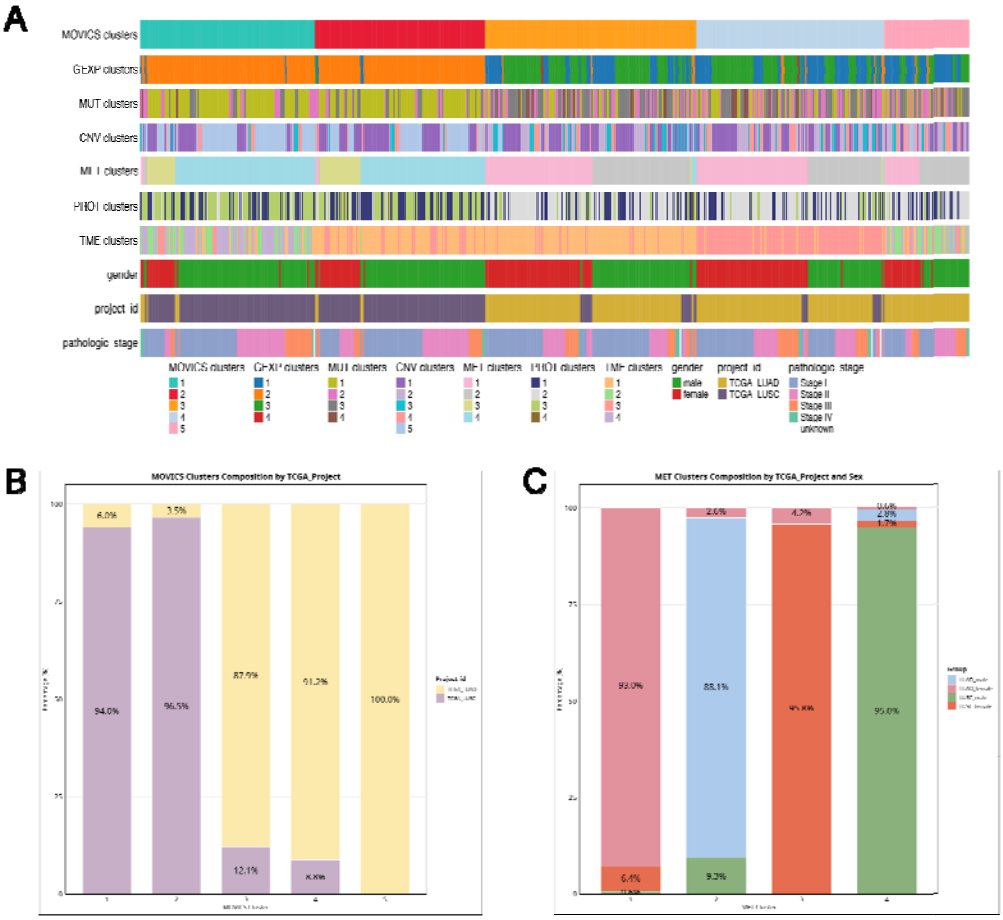


Figure 3. Associations of MOVICS and single-omics subtypes with clinical features. (**A**) Heatmap showing the concordance between MOVICS consensus subtypes, single-omics clustering results, and clinical features (gender, project, and stage); (**B**) Distribution of histological subtypes (LUAD vs. LUSC) across the MOVICS consensus subtypes. (**C**) Association of DNA methylation (MET) subtypes with histological type and sex.

For single-omics subtypes (excluding DNA methylation), GEXP, CNV, MUT, PROT, and TME clusters exhibited diverse enrichment patterns: some clusters were LUAD-dominant (e.g., GEXP 1,3; CNV 2–4), others LUSC-dominant (e.g., GEXP 2; CNV 5; MUT 1), and several displayed a relatively balanced distribution between the two histological types (e.g., PROT 1; TME 1,2) (Supplementary Materials, Figure S8).

Figure 3C further shows the distribution of the four methylation subtypes stratified by histological type and gender. MET clusters 1 and 2 were enriched in LUAD, while clusters 3 and 4 were enriched in LUSC. Methylation subtypes displayed striking associations with both histology and gender: MET cluster 1 comprised mostly female LUAD (93.0%), cluster 2 was predominantly male LUAD (88.1%), cluster 3 was mostly female LUSC (95.8%), and cluster 4 was predominantly male LUSC (95.0%).

2.4. Prognostic Impact of Multi-Omics Subtypes on Survival Outcomes

In the overall survival (OS) analysis, the Kaplan–Meier curves showed no signifi-

(Log-rank p = 0.53) (Figure 4A). This lack of separation may reflect the limited sample size or insufficient biological heterogeneity across subtypes, resulting in reduced statistical power. In the multivariable Cox regression analysis, after adjusting for clinical stage, clinical stage emerged as an independent prognostic factor for OS. In contrast, the hazard ratios of the consensus subtypes were modest, and the corresponding confidence intervals encompassed 1, suggesting no independent prognostic value (Figure 4C).

Notably, the progression-free survival (PFS) analysis revealed significant differences among the five subtypes (Log-rank p = 0.016) (Figure 4B). Multivariable Cox regression revealed that the CS4 subtype had the highest risk (HR = 1.605, 95% CI 1.075–2.40, p = 0.0213), although its confidence interval was relatively wide. Compared with CS1, the hazard ratios for CS2, CS3, and CS5 were not statistically significant (Figure 4D).

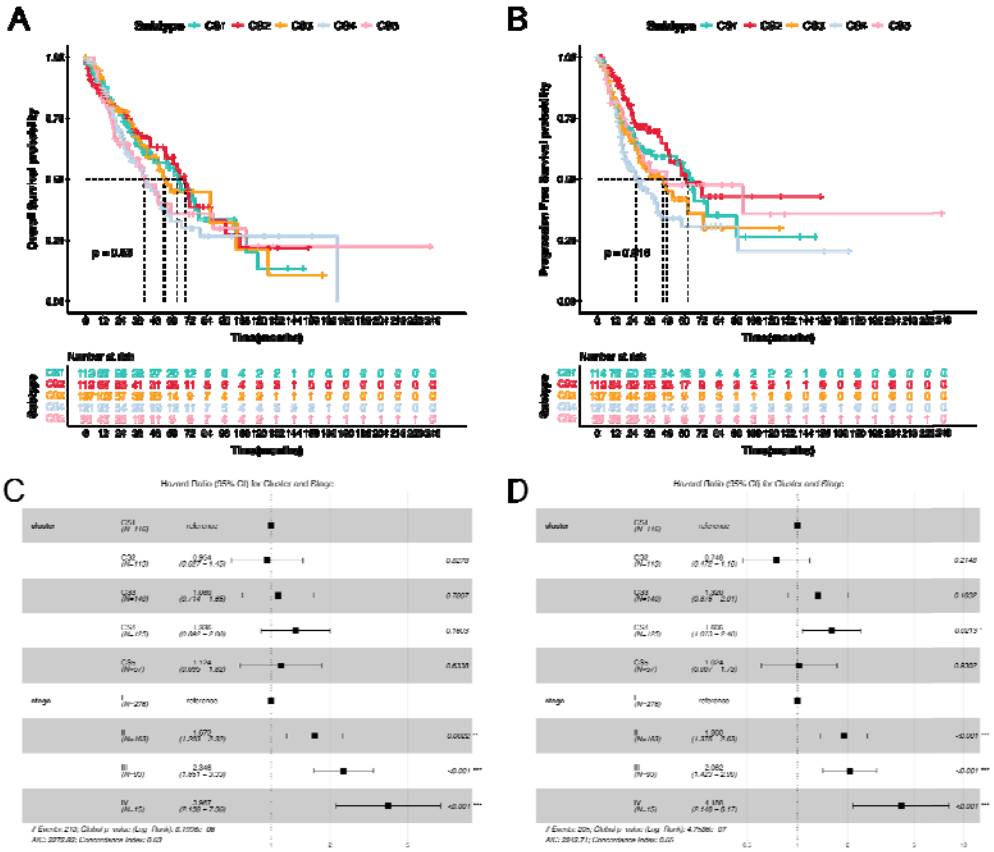


Figure 4. Survival analysis of multi-omics consensus subtypes: Kaplan-Meier curves and Cox regression results for OS and PFS. (**A**) Kaplan–Meier curves of OS stratified by multi-omics consensus subtypes; (**B**) Kaplan-Meier curves of PFS stratified by multi-omics consensus subtypes; (**C**) Multivariable Cox regression analysis for OS (consensus subtypes and clinical stage); (**D**) Multivariable Cox regression analysis for PFS (consensus subtypes and clinical stage).

2.5. Subtype-Specific Molecular Characteristics and Functional Pathway Landscapes.

2.5.1. Mutational Landscape Reveals Subtype-specific Alterations

Genetic mutations play a pivotal role in tumor initiation and progression[18]. In this study, we compared the mutational landscapes of the five multi-omics consensus subtypes. As shown in Figure 5A, a total of 18 significantly mutated genes were identified across the five consensus subtypes. *TP53*, *TTN*, *SYNE1*, *FAM135B*, *KMT2D*, and

composed of LUSC, whereas *KRAS*, *ASXL3*, *STK11*, *EGFR*, *LPPR4*, and *CLCN1* were more commonly mutated in CS3–CS5, which were enriched for LUAD. Supplementary Table 1 further highlights subtype-specific mutational profiles: CS1 was enriched for *TTN*, *FAM135B*, *NPAP1*, and *CDKN2A* mutations; CS2 for *TP53*, *SYNE1*, *KMT2D*, *LRRK2*, and *NFE2L2*; CS4 for *KRAS* and *EGFR*; and CS5 for *KEAP1*, *ASXL3*, *STK11*, and *CLCN1*.

Tumor mutational burden (TMB), a surrogate for neoantigen load, has been linked to improved response to immune checkpoint inhibitors[19,20]. The transition-to-transversion (Ti/Tv) ratio also carries clinical significance in NSCLC; for instance, smoking-related C>A transversions are associated with poor prognosis, whereas *APOBEC*-driven C>T transitions may predict immunotherapy sensitivity[21]. In our analysis, TMB differed significantly across subtypes ($p < 0.05$), with CS1 highest and CS3 lowest. LUSC-dominated subtypes (CS1, CS2) had higher TMB than LUAD-dominated ones (CS3–CS5), suggesting greater genomic instability. Ti/Tv analysis further showed a higher proportion of T>C transitions in CS1 and CS2.

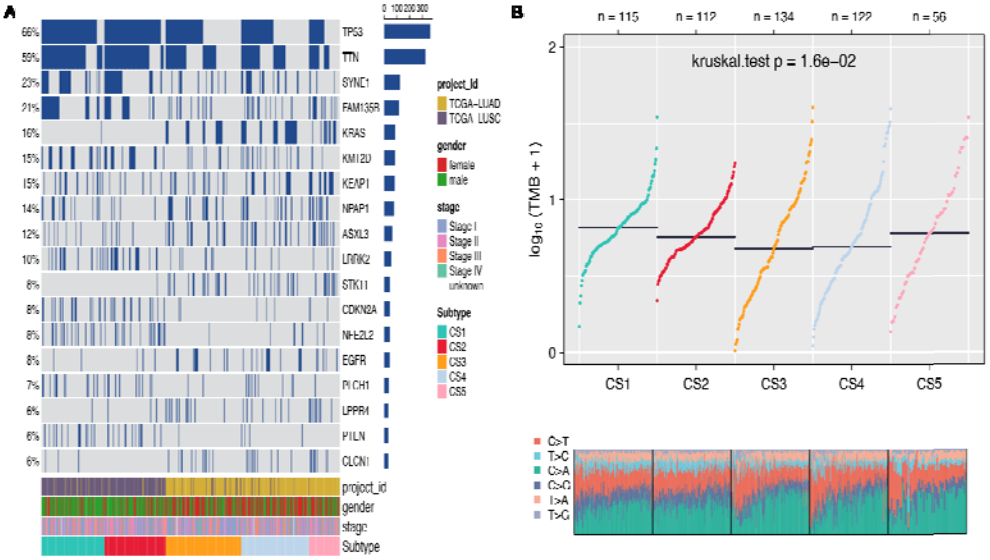


Figure 5. Mutation landscape of the five multi-omics consensus subtypes. (A) Waterfall plots of significantly mutated genes in each subtype; (B) Comparison of TMB and Ti/Tv ratios across subtypes.

2.5.2. Distinct Molecular Signatures of Multi-Omics Subtypes

Through differential expression analysis, we identified subtype-specific genes across the five multi-omics consensus subtypes (Supplementary Materials, Figure S9A). Members of the keratin gene family (e.g., *KRT5*, *KRT6A*, *KRT6B*) were markedly upregulated in the LUSC-dominant subtypes (CS1 and CS2) but downregulated in the LUAD-dominant subtypes (CS3−CS5). From these analyses, we selected 200 subtype-specific signature genes for each subtype (100 upregulated and 100 downregulated). Their distinct expression profiles across subtypes supported the robustness of our classification and provided a molecular basis for external validation (Supplementary Materials, Figure S9B, S9C).

2.5.3. Subtype-Specific Functional Pathways

We performed subtype-specific gene set enrichment analyses (GSEA) for each of the five multi-omics subtypes and then compared the enrichment patterns across subtypes. KEGG pathway analysis (Figure 6) revealed that CS2, CS3, and CS4 consistently exhib-

naling pathway," "Cytokine–cytokine receptor interaction," and "Phagosome," whereas CS1 and CS5 showed consistent downregulation, suggesting the presence of two distinct immune-related subtype groups. GO term enrichment comparison (Supplementary Materials, Figure S10) further demonstrated that gene sets associated with "leukocyte-mediated immunity," "immunoglobulin production," "lymphocyte-mediated immunity," and "B cell-mediated immunity" were consistently upregulated in CS2 and CS3 and downregulated in CS1 and CS5, largely corroborating the KEGG enrichment patterns.

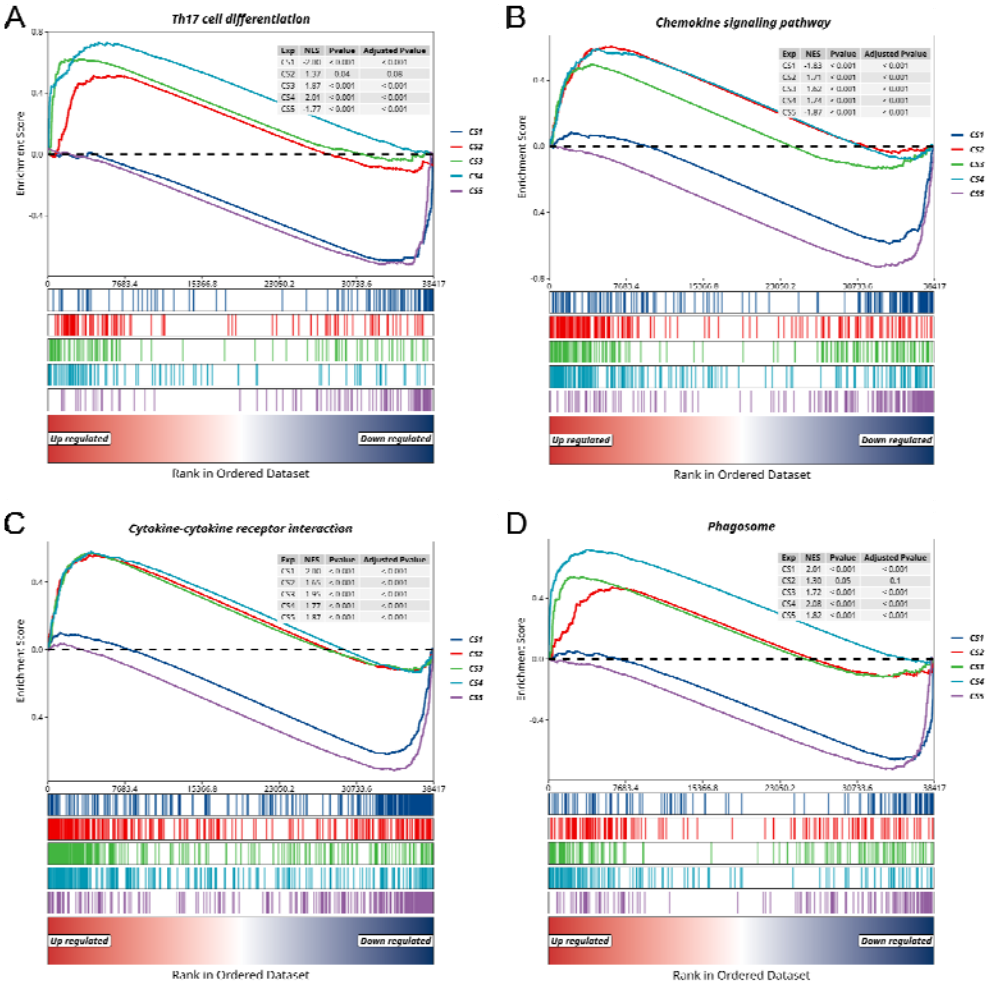


Figure 6. Comparison of selected immune-related KEGG pathway enrichment among five multi-omics consensus subtypes. (**A**) KEGG pathway enrichment for "Th17 cell differentiation"; (**B**) KEGG pathway enrichment for "Chemokine signaling pathway"; (**C**) KEGG pathway enrichment for "Cytokine–cytokine receptor interaction" ;(**D**) KEGG pathway enrichment for "Phagosome".

3. Discussion

In this study, we constructed single- and multi-omics molecular classification systems for NSCLC by integrating transcriptomic, genomic, epigenomic, proteomic, and tumor microenvironment data. Analyses at the single-omics level revealed distinct patterns: DNA methylation subtypes were strongly associated with sex and histology, likely reflecting smoking-induced methylation alterations[22], whereas proteomic subtypes were largely independent of histopathological classification. This suggests that tumors

of different origins may converge on similar functional modules, including proliferation, immune evasion, and metabolic adaptation.

Integration of these multi-omics layers yielded consensus subtypes that corresponded closely with histology: CS1 and CS2 were predominantly LUSC, whereas CS3–CS5 were mainly LUAD. Beyond tissue-of-origin characteristics, the subtypes exhibited consistent immune profiles across histological subtypes. Specifically, CS2, CS3, and CS4 showed immune-activated features, whereas CS1 and CS5 displayed immunosuppressive traits, suggesting the presence of two distinct immune subtypes that may differ in their responses to immunotherapy.

Despite these pronounced molecular distinctions, OS did not differ significantly among subtypes. This lack of association may reflect several factors, including the high proportion of early-stage patients (I–II, 79.5%), unaccounted treatment heterogeneity, and the limited sample size of CS5 (n = 57). Interestingly, PFS analysis revealed that CS4 was associated with higher risk, consistent with its *KRAS*-driven genomic alterations. These findings suggest that patients in this subtype could potentially benefit from MEK inhibitor-based combination immunotherapy[23].

Nevertheless, several limitations warrant consideration: (1) The TCGA cohort includes only LUAD and LUSC, excluding rare subtypes such as large-cell neuroendocrine carcinoma and sarcomatoid carcinoma, which harbor distinctive molecular alterations (e.g., TP53 and *STK11*/*KEAP1* co-mutations or *RB1*/*TP53* co-deletion)[24]. Their absence may limit the biological generalizability of the classification. (2) The lack of microbiome and metabolome data may constrain the completeness of subtype characterization[25]. (3) Key fusion events (*ALK*, *ROS1*, *RET*) were not systematically incorporated; for example, *CD74-ROS1* rearrangements are common in *ROS1*-positive LUAD and predict sensitivity to crizotinib[26], and their omission may result in misclassification of fusion-driven subtypes.(4) TCGA lacks targeted therapy and immune checkpoint inhibitor response data, preventing evaluation of subtype-treatment associations and limiting clinical translational potential.

Importantly, some planned analyses were not completed in the present study. Future work will aim to: (1) optimize single-omics clustering to identify more robust and biologically meaningful subtypes; (2) systematically compare multi-omics subtypes with established TCGA classifications; (3) validate the classification in independent external cohorts; and (4) further characterize the genomic and tumor microenvironment features of the multi-omics consensus subtypes. These efforts are expected to refine NSCLC molecular stratification and provide a more reliable framework for precision oncology.

4. Materials and Methods

4.1. Data Source and Preprocessing

We downloaded TCGA non-small cell lung cancer (NSCLC) data from the UCSC Xena website (https://xena.ucsc.edu/), including LUAD and LUSC cohorts: RNA-seq gene expression data (counts format, n = 1129, normalized and log2-transformed across pan-cancer samples); DNA methylation data (Illumina Human Methylation 450K, 503 LUAD samples and 412 LUSC samples); copy number variation (CNV) data (Gene Level Copy Number, ASCAT3, 503 LUAD samples and 490 LUSC samples); protein expression data (RPPA, 365 LUAD samples and 328 LUSC samples); and somatic mutation data (Pan-Cancer Atlas, 513 LUAD samples and 480 LUSC samples, binary encoding with 1 indicating nonsynonymous mutations and 0 indicating wild type).

To reduce data dimensionality and facilitate clustering analysis, the preprocessing and feature selection of the other five omics datasets were performed as follows: (1) So-

matic mutation data: retain genes mutated in at least 5% of samples; (2) Protein expression data: remove proteins with missing values; (3) RNA-seq data: Z-score normalization for each gene across samples, followed by selection of highly variable genes based on median absolute deviation (MAD); (4) CNV data: rank genes based on the variation frequency, defined as the proportion of samples with copy number not equal to 2 (derived from ASCAT3); (5) DNA methylation data: rank probes based on standard deviation to select highly variable loci.

In addition, clinical phenotypes, survival data, and TPM expression data were downloaded from UCSC Xena for downstream analyses, while MAF-formatted mutation data were obtained separately using the TCGAblinks R package[27]. For simplicity in downstream analyses and figure annotations, the six omics layers are referred to by the following variable names: GEXP (gene expression), MUT (somatic mutation), CNV (copy number variation), MET (DNA methylation), PROT (protein expression), and TME (tumor microenvironment features).

4.2. TME Feature Calculation

TME features were computed based on RNA-seq data downloaded from the GDC portal (https://portal.gdc.cancer.gov/) for TCGA LUAD and LUSC cohorts. Seven built-in immune infiltration algorithms implemented in the IOBR R package (CIBERSORT, ESTIMATE, quanTIseq, TIMER, IPS, MCPCounter, and EPIC) were applied to quantify the infiltration levels of various immune cell types and other TME characteristics for each sample[28]. These features were subsequently incorporated into the multi-omics integration analysis and used for subtype construction.

4.3. Identification of Single-Omics Molecular Subtypes

Unsupervised clustering of the six omics datasets was performed using the ConsensusClusterPlus R package with hierarchical clustering[29]. The number of features used for clustering in each omics dataset was as follows: the top 20% highly variable genes for RNA-seq (4,106), the top 1% most variable probes for DNA methylation (8,064), the top 10% genes ranked by variation frequency for CNV (5,794), 651 somatic mutation features, 206 protein expression features, and 96 TME features.

Clustering parameters were set as follows: maxK = 10, reps = 1000, pItem = 0.8, pFeature = 1, clusterAlg = "hc", innerLinkage = "ward.D2", and finalLinkage = "ward.D2". The distance parameter was adjusted according to the omics type: “binary” for mutation data, “euclidean” for CNV data, and “pearson” for all other omics types.

4.4. Multi-Omics Consensus Clustering

A total of 551 samples with available data across six omics layers were stratified using the MOVICS package for multi-omics consensus clustering[30]. Missing values were imputed and high-variance features were selected using the built-in “getElites” function. The parameters of “getElites” were configured to balance the number of features across the different omics layers: the top 3,000 features with the highest MAD were retained for RNA-seq and DNA methylation data, 50% of proteins with the highest MAD (243 proteins) were retained for RPPA data, all 96 TME features were retained, and the top 3,000 CNV features with the highest standard deviation were selected. CNV values were binarized such that samples with copy number not equal to 2 were considered as having a CNV, and those equal to 2 as no CNV. For somatic mutation data, 619 genes mutated in at least 5% of samples were retained.

The optimal number of clusters (K) was determined based on the CPI and gap statistics. Subsequently, the ten multi-omics integration algorithms implemented in the MOVICS package (intNMF, moCluster, SNF, PINSPlus, NEMO, CIMLR, COCA,

iClusterBayes, ConsensusClustering, and LRAcluster) were applied to derive clustering solutions, followed by consensus integration across these methods to obtain robust molecular subtypes. Subtype stability was evaluated using silhouette scores.

4.5. Evaluation of Genetic Alterations among Different Subtypes

The "compMut" function from the MOVICS package was applied to analyze mutation features among subtypes, identifying significantly different and frequently mutated genes (mutation frequency >5%, $p < 0.05$). TMB was calculated as the number of nonsynonymous mutations per million bases.

4.6. Subtype-Specific Transcriptomic Profiling

Differential expression analysis among subtypes was performed using the "runDEA" function from the MOVICS package, with the limma method selected for statistical testing. Multi-omics volcano plots were generated using the jjVolcano function from the scRNAtoolVis package[31]. Subtype-specific biomarkers were further identified by selecting genes with the most significant log2 fold changes (adjusted $p < 0.05$) while ensuring that the selected biomarkers were not shared with other subtypes.

4.7. Gene Set Enrichment Analysis

Differentially expressed genes were ranked according to log2 fold change (log2FC). GSEA was conducted with the R package clusterProfiler for GO and KEGG pathways[32,33]. Enrichment results were visualized using enrichPlot and GseaVis to characterize the biological functions and signaling pathways associated with each subtype[34].

4.8. Statistical Analysis

All data preprocessing, statistical analyses, and visualizations were performed in the R environment (version 4.4.1). Kaplan–Meier survival curves and Cox proportional hazards regression were performed using the survival package, and forest plots were generated with survminer. All tests were two-sided, with $p < 0.05$ considered statistically significant.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/doi/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Author Contributions: Conceptualization, S.W. and J.-G.Z.; methodology, B.Y., Z.W.; investigation, B.Y.; formal analysis, B.Y.; visualization, B.Y., Z.W. and Y.P.; writing—original draft preparation, B.Y.; writing—review and editing, B.Y., Z.W., Y.P., K.L., S.W. and J.-G.Z.; supervision, S.W. and J.-G.Z.; project administration, S.W. and K.L.; funding acquisition, S.W. and J.-G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Natural Science Foundation of China (Grant No. 82303953), Hunan Provincial Natural Science Foundation of China (Grant No. 2025JJ40079), Central South University Startup Funding, Noncommunicable Chronic Diseases-National Science and Technology Major Project (Grant No. 2023ZD0502105), Ministry of Education in China Liberal arts and Social Sciences Foundation (Grant No. 24YJCZH462), Youth Science and Technology Elite Talent Project of Guizhou Provincial Department of Education (Grant No.QJJ-2024-333), Excellent Young Talent Cultivation Project of Zunyi City (Zunshi Kehe HZ (2023) 142), Future Science and Technology Elite Talent Cultivation Project of Zunyi Medical University (ZYSE 2023-02), and the Key Program of the Education Sciences Planning of Guizhou Province (Grant No.7).

Institutional Review Board Statement: Not applicable. Ethical review and approval were waived for this study, as all data were obtained from the publicly available TCGA database, for which patient consent and ethical approval were already secured by the original investigators.

Informed Consent Statement: Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. TCGA data are available from The Cancer Genome Atlas (https://portal.gdc.cancer.gov/), and processed TCGA datasets were accessed via the UCSC Xena platform (https://xena.ucsc.edu/). No new data were generated in this study.

Acknowledgments: We are grateful for resources from the Bioinformatics Platform, Furong Laboratory and Bioinformatics Center, Xiangya Hospital, Central South University.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CNV	Copy number variation
CPI	Clustering prediction index
CS	Consensus subtype
GSEA	Gene set enrichment analysis
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
MAD	Median absolute deviation
NSCLC	Non-small cell lung cancer
OS	Overall survival
PFS	Progression-free survival
RPPA	Reverse phase protein array
TCGA	The Cancer Genome Atlas
Ti/Tv	The transition-to-transversion
TKI	Tyrosine kinase inhibitor
TMB	Tumor mutational burden
TME	Tumor microenvironment
TNM	Tumor-Node-Metastasis

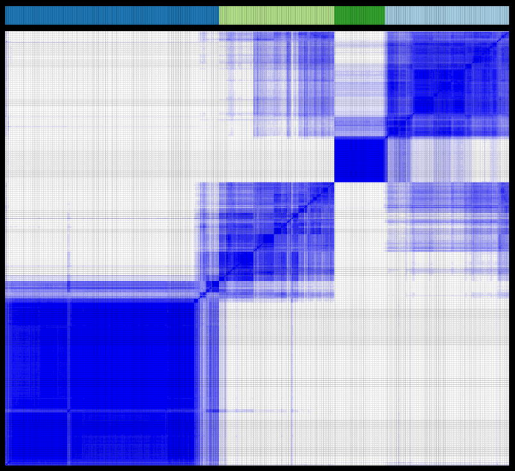
References

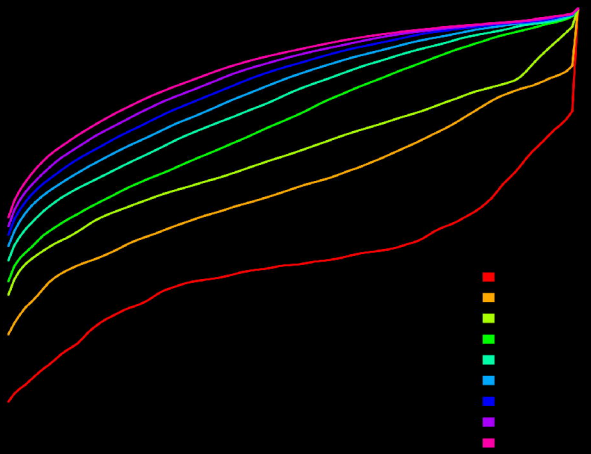
1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global Cancer Statistics 2022: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA. Cancer J. Clin.* **2024**, *74*, 229–263, doi:10.3322/caac.21834.
2. Herbst, R.S.; Morgensztern, D.; Boshoff, C. The Biology and Management of Non-Small Cell Lung Cancer. *Nature* **2018**, *553*, 446–454, doi:10.1038/nature25183.
3. Wang, C.; Li, J.; Chen, J.; Wang, Z.; Zhu, G.; Song, L.; Wu, J.; Li, C.; Qiu, R.; Chen, X.; et al. Multi-Omics Analyses Reveal Biological and Clinical Insights in Recurrent Stage I Non-Small Cell Lung Cancer. *Nat Commun* **2025**, *16*, 1477, doi:10.1038/s41467-024-55068-2.
4. Xu, Y.; She, Y.; Li, Y.; Li, H.; Jia, Z.; Jiang, G.; Liang, L.; Duan, L. Multi-Omics Analysis at Epigenomics and Transcriptomics Levels Reveals Prognostic Subtypes of Lung Squamous Cell Carcinoma. *Biomed. Pharmacother.* **2020**, *125*, 109859, doi:10.1016/j.biopha.2020.109859.
5. Yang, F.; Sun, K.; Li, F.; Li, X.; Shi, J.; Sun, X.; Hong, Y.; Jiang, G.; Zhu, Y.; Song, X. The Prognostic Impact of Epidermal Growth Factor Receptor Mutation in Clinical Stage I Lung Adenocarcinoma. *Ann. Thorac. Surg.* **2024**, *117*, 1111–1119, doi:10.1016/j.athoracsur.2023.05.031.
6. Hung, Y.P.; Chirieac, L.R. How Should Molecular Findings Be Integrated in the Classification for Lung Cancer? *Transl. Lung Cancer Res.* **2020**, *9*, doi:10.21037/tlcr-20-153.

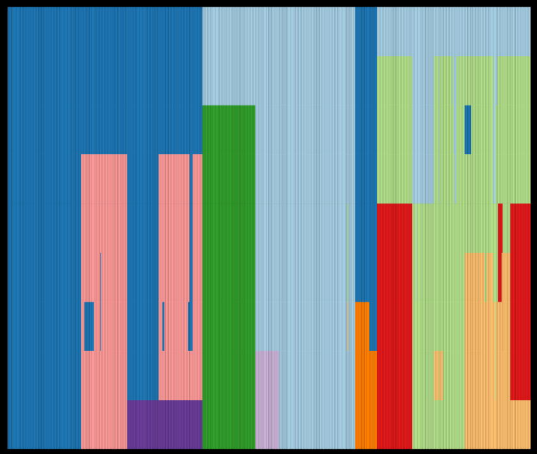
7. Yang, S.-R.; Schultheis, A.M.; Yu, H.; Mandelker, D.; Ladanyi, M.; Büttner, R. Precision Medicine in Non-Small Cell Lung Cancer: Current Applications and Future Directions. *Semin. Cancer Biol.* **2022**, *84*, 184 – 198, doi:10.1016/j.semcancer.2020.07.009.
8. Ramalingam, S.S.; Vansteenkiste, J.; Planchard, D.; Cho, B.C.; Gray, J.E.; Ohe, Y.; Zhou, C.; Reungwetwattana, T.; Cheng, Y.; Chewaskulyong, B.; et al. Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC. *N Engl J Med* **2020**, *382*, 41–50, doi:10.1056/NEJMoa1913662.
9. Cheng, Y.; He, Y.; Li, W.; Zhang, H.-L.; Zhou, Q.; Wang, B.; Liu, C.; Walding, A.; Saggese, M.; Huang, X.; et al. Osimertinib versus Comparator EGFR TKI as First-Line Treatment for EGFR-Mutated Advanced NSCLC: FLAURA China, a Randomized Study. *Target Oncol* **2021**, *16*, 165–176, doi:10.1007/s11523-021-00794-6.
10. Zhou, X.; Zeng, L.; Huang, Z.; Ruan, Z.; Yan, H.; Zou, C.; Xu, S.; Zhang, Y. Strategies beyond 3rd EGFR-TKI Acquired Resistance: Opportunities and Challenges. *Cancer Med.* **2025**, *14*, e70921, doi:10.1002/cam4.70921.
11. Parra, E.R.; Zhang, J.; Duose, D.Y.; Gonzalez-Kozlova, E.; Redman, M.W.; Chen, H.; Manyam, G.C.; Kumar, G.; Zhang, J.; Song, X.; et al. Multi-Omics Analysis Reveals Immune Features Associated with Immunotherapy Benefit in Patients with Squamous Cell Lung Cancer from Phase III Lung-MAP S1400I Trial. *Clin. Cancer Res.* **2024**, *30*, 1655, doi:10.1158/1078-0432.CCR-23-0251.
12. Zheng, Y.; Tang, L.; Liu, Z. Multi-Omics Analysis of an Immune-Based Prognostic Predictor in Non-Small Cell Lung Cancer. *BMC Cancer* **2021**, *21*, 1322, doi:10.1186/s12885-021-09044-4.
13. Huang, J.; Li, J.; Zheng, S.; Lu, Z.; Che, Y.; Mao, S.; Lei, Y.; Zang, R.; Liu, C.; Wang, X.; et al. Tumor Microenvironment Characterization Identifies Two Lung Adenocarcinoma Subtypes with Specific Immune and Metabolic State. *Cancer Science* **2020**, *111*, 1876–1886, doi:10.1111/cas.14390.
14. Zeng, D.; Yu, Y.; Qiu, W.; Ou, Q.; Mao, Q.; Jiang, L.; Wu, J.; Wu, J.; Luo, H.; Luo, P.; et al. Immunotyping the Tumor Microenvironment Reveals Molecular Heterogeneity for Personalized Immunotherapy in Cancer. *Adv. Sci.* **2025**, *12*, 2417593, doi:10.1002/advs.202417593.
15. Chen, H.; Zhang, T.; Zhang, Y.; Wu, H.; Fang, Z.; Liu, Y.; Chen, Y.; Wang, Z.; Jia, S.; Ji, X.; et al. Deciphering the Tumor Microenvironment Cell – Infiltrating Landscape Reveals Microenvironment Subtypes and Therapeutic Potentials for Nonsquamous NSCLC. *JCI Insight* **2022**, *7*, doi:10.1172/jci.insight.152815.
16. Xue, H.; Fan, Y.; Li, Y.; Zhao, Q.; Zhang, X.; Zhao, P.; Liu, Z. Tumor-Infiltrating Lymphocytes in NSCLC: From Immune Surveillance to Immunotherapy. *Front. Immunol.* **2025**, *16*, doi:10.3389/fimmu.2025.1610998.
17. Lim, J.U.; Lee, E.; Lee, S.-Y.; Cho, H.J.; Ahn, D.H.; Hwang, Y.; Choi, J.Y.; Yeo, C.D.; Park, C.K.; Kim, S.J. Current Literature Review on the Tumor Immune Micro-Environment, Its Heterogeneity and Future Perspectives in Treatment of Advanced Non-Small Cell Lung Cancer. *Transl. Lung Cancer Res.* **2023**, *12*, doi:10.21037/tlcr-22-633.
18. Iranzo, J.; Martincorena, I.; Koonin, E.V. Cancer-Mutation Network and the Number and Specificity of Driver Mutations. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E6010–E6019, doi:10.1073/pnas.1803155115.
19. Jardim, D.L.; Goodman, A.; Gagliato, D. de M.; Kurzrock, R. The Challenges of Tumor Mutational Burden as an Immunotherapy Biomarker. *Cancer Cell* **2021**, *39*, 154–173, doi:10.1016/j.ccell.2020.10.001.
20. Meng, G.; Liu, X.; Ma, T.; Lv, D.; Sun, G. Predictive Value of Tumor Mutational Burden for Immunotherapy in Non-Small Cell Lung Cancer: A Systematic Review and Meta-Analysis. *PLOS One* **2022**, *17*, e0263629, doi:10.1371/journal.pone.0263629.
21. de Alencar, V.T.L.; Figueiredo, A.B.; Corassa, M.; Gollob, K.J.; Cordeiro de Lima, V.C. Lung Cancer in Never Smokers: Tumor Immunology and Challenges for Immunotherapy. *Front. Immunol.* **2022**, *13*, doi:10.3389/fimmu.2022.984349.
22. Huang, T.; Chen, X.; Hong, Q.; Deng, Z.; Ma, H.; Xin, Y.; Fang, Y.; Ye, H.; Wang, R.; Zhang, C.; et al. Meta-Analyses of Gene Methylation and Smoking Behavior in Non-Small Cell Lung Cancer Patients. *Sci. Rep.* **2015**, *5*, 8897, doi:10.1038/srep08897.
23. Han, J.; Liu, Y.; Yang, S.; Wu, X.; Li, H.; Wang, Q. MEK Inhibitors for the Treatment of Non-Small Cell Lung Cancer. *J. Hematol. Oncol.* **2021**, *14*, 1, doi:10.1186/s13045-020-01025-7.
24. George, J.; Walter, V.; Peifer, M.; Alexandrov, L.B.; Seidel, D.; Leenders, F.; Maas, L.; Müller, C.; Dahmen, I.; Delhomme, T.M.; et al. Integrative Genomic Profiling of Large-Cell Neuroendocrine Carcinomas Reveals Distinct Subtypes of High-Grade Neuroendocrine Lung Tumors. *Nat. Commun.* **2018**, *9*, 1048, doi:10.1038/s41467-018-03099-x.
25. Cheng, J.; Wang, H. Lung Microbiome Alterations Correlate with Immune Imbalance in Non-Small Cell Lung Cancer. *Front. Immunol.* **2025**, *16*, doi:10.3389/fimmu.2025.1589843.
26. Li, W.; Fei, K.; Guo, L.; Wang, Y.; Shu, C.; Wang, J.; Ying, J. *CD74/SLC34A2-ROS1* Fusion Variants Involving the Transmembrane Region Predict Poor Response to Crizotinib in NSCLC Independent of *TP53* Mutations. *J. Thorac. Oncol.* **2024**, *19*, 613–625, doi:10.1016/j.jtho.2023.12.009.
27. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; et al. TCGAbiolinks: An R/Bioconductor Package for Integrative Analysis of TCGA Data. *Nucleic Acids Res.* **2016**, *44*, e71, doi:10.1093/nar/gkv1507.

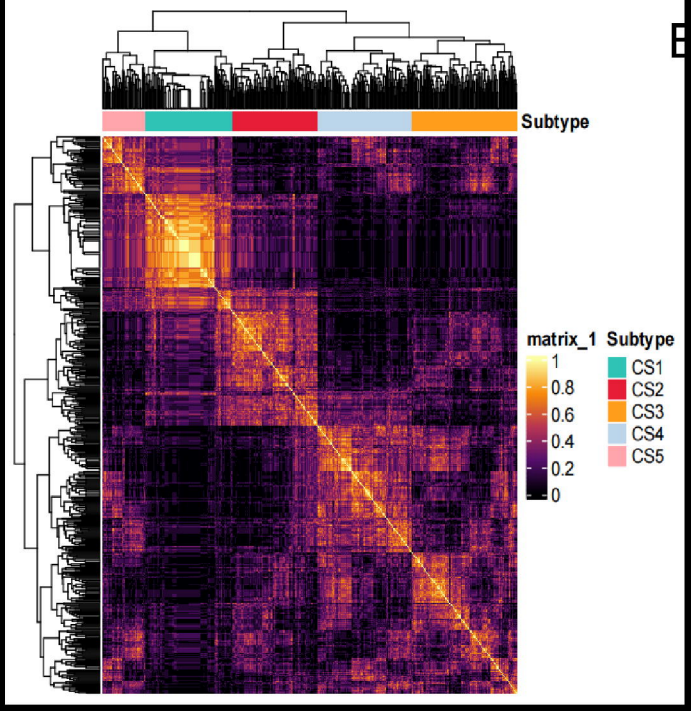
28. Zeng, D.; Fang, Y.; Qiu, W.; Luo, P.; Wang, S.; Shen, R.; Gu, W.; Huang, X.; Mao, Q.; Wang, G.; et al. Enhancing Immuno-Oncology Investigations through Multidimensional Decoding of Tumor Microenvironment with IOBR 2.0. *Cell Rep. Methods* **2024**, *4*, doi:10.1016/j.crmeth.2024.100910.
29. Wilkerson, M.D.; Hayes, D.N. ConsensusClusterPlus: A Class Discovery Tool with Confidence Assessments and Item Tracking. *Bioinformatics* **2010**, *26*, 1572–1573, doi:10.1093/bioinformatics/btq170.
30. Lu, X.; Meng, J.; Zhou, Y.; Jiang, L.; Yan, F. MOVICS: An R Package for Multi-Omics Integration and Visualization in Cancer Subtyping. *Bioinformatics* **2021**, *36*, 5539–5541, doi:10.1093/bioinformatics/btaa1018.
31. Zhang, J. scRNAtoolVis: Useful Functions to Make Your scRNA-Seq Plot More Cool. Available online: https://github.com/junjunlab/scRNAtoolVis (accessed on 28 August 2025).
32. Yu, G. Thirteen Years of clusterProfiler. *Innovation* **2024**, *5*, doi:10.1016/j.xinn.2024.100722.
33. Yu, G.; Wang, L.-G.; Han, Y.; He, Q.-Y. clusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters. *OMICS: A Journal of Integrative Biology* **2012**, *16*, 284–287, doi:10.1089/omi.2011.0118.
34. Zhang, J.; Li, H.; Tao, W.; Zhou, J. GseaVis: An R Package for Enhanced Visualization of Gene Set Enrichment Analysis in Biomedicine. *Med Research* **2025**, *1*, 131–135, doi:10.1002/mdr2.70000.

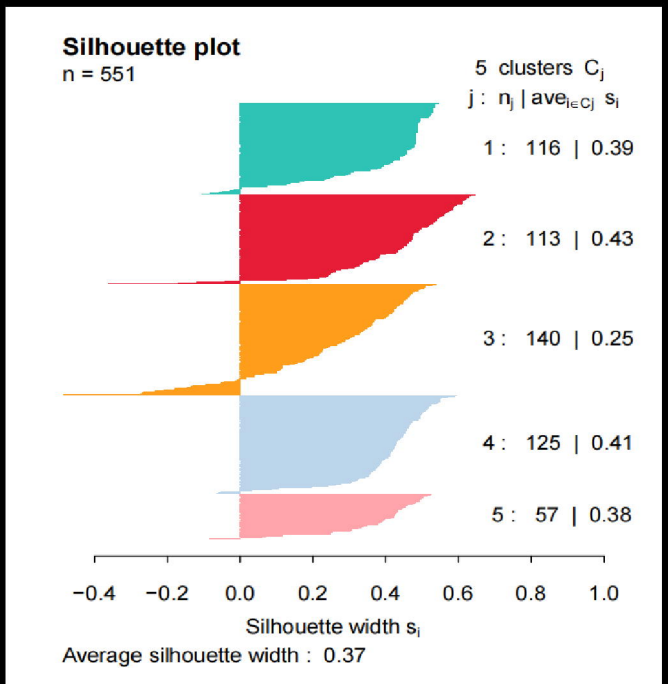
Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

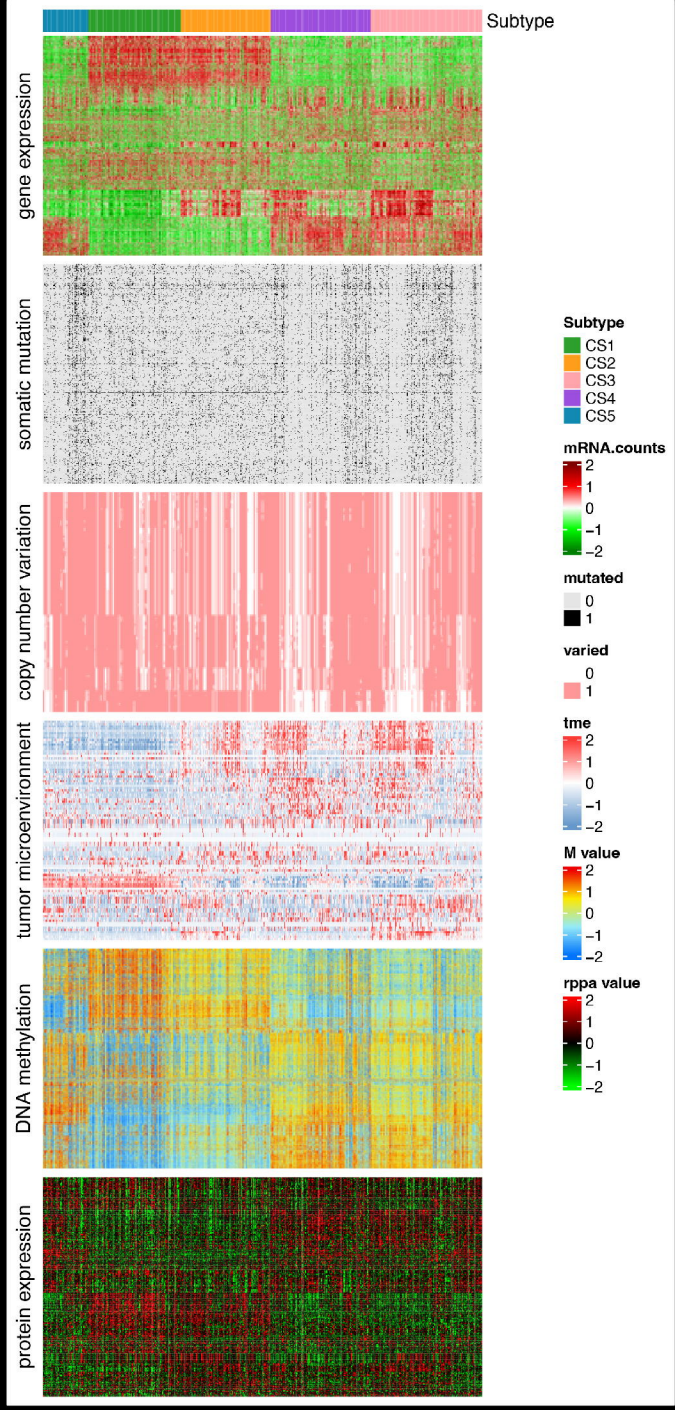


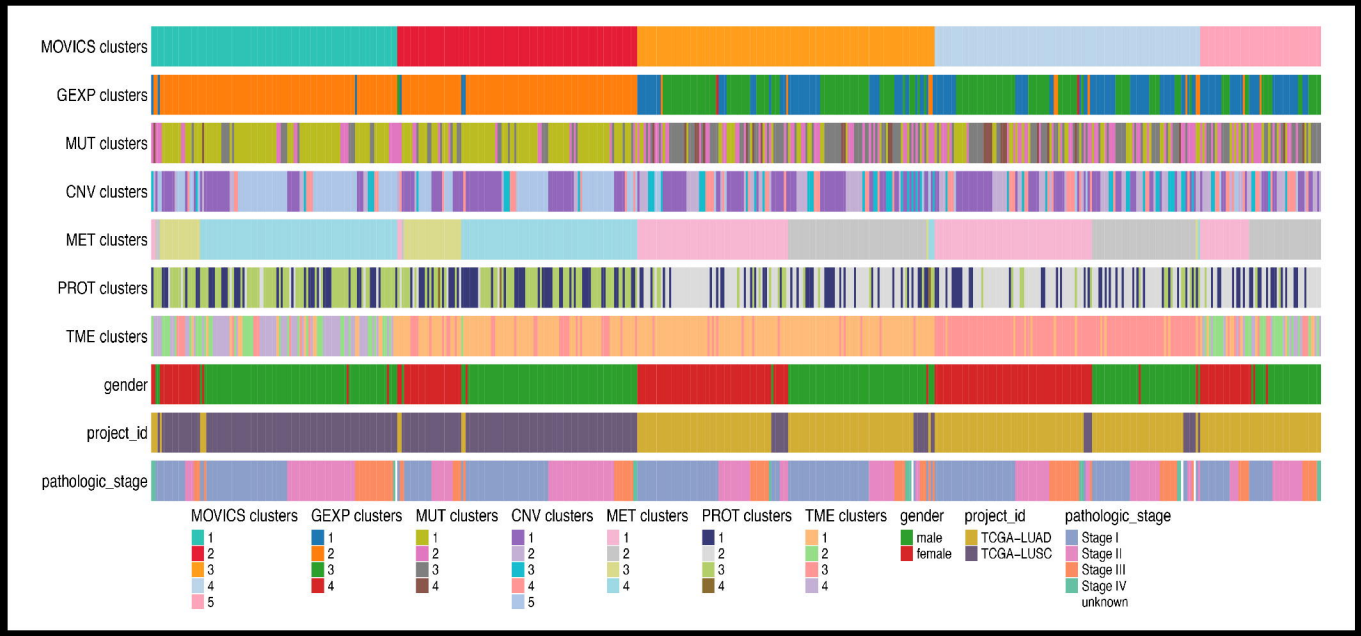


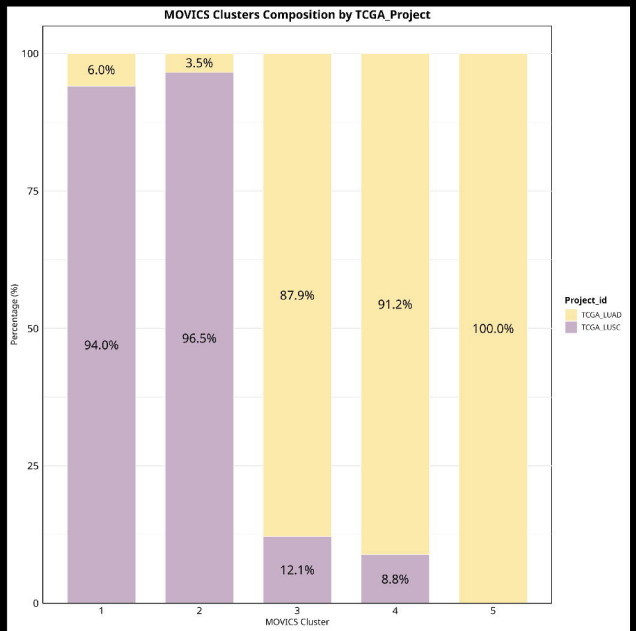


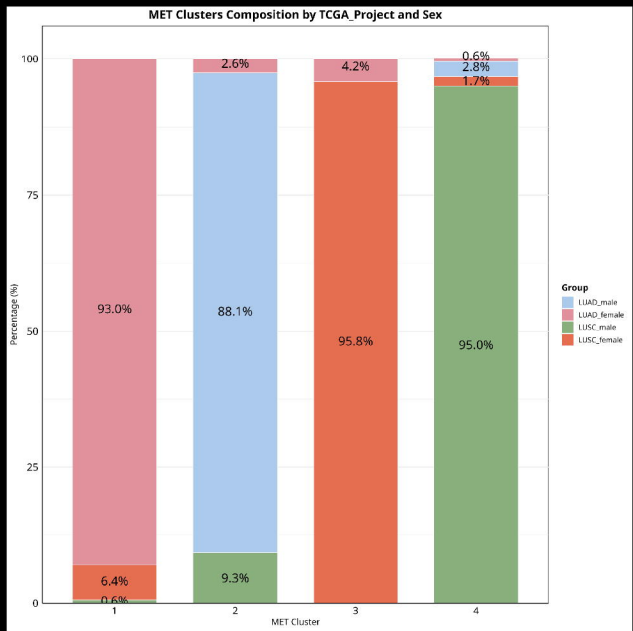

Subtype
matrix_1
1
0.8
0.6
0.4
0.2
0
Subtype
CS1
CS2
CS3
CS4
CS5

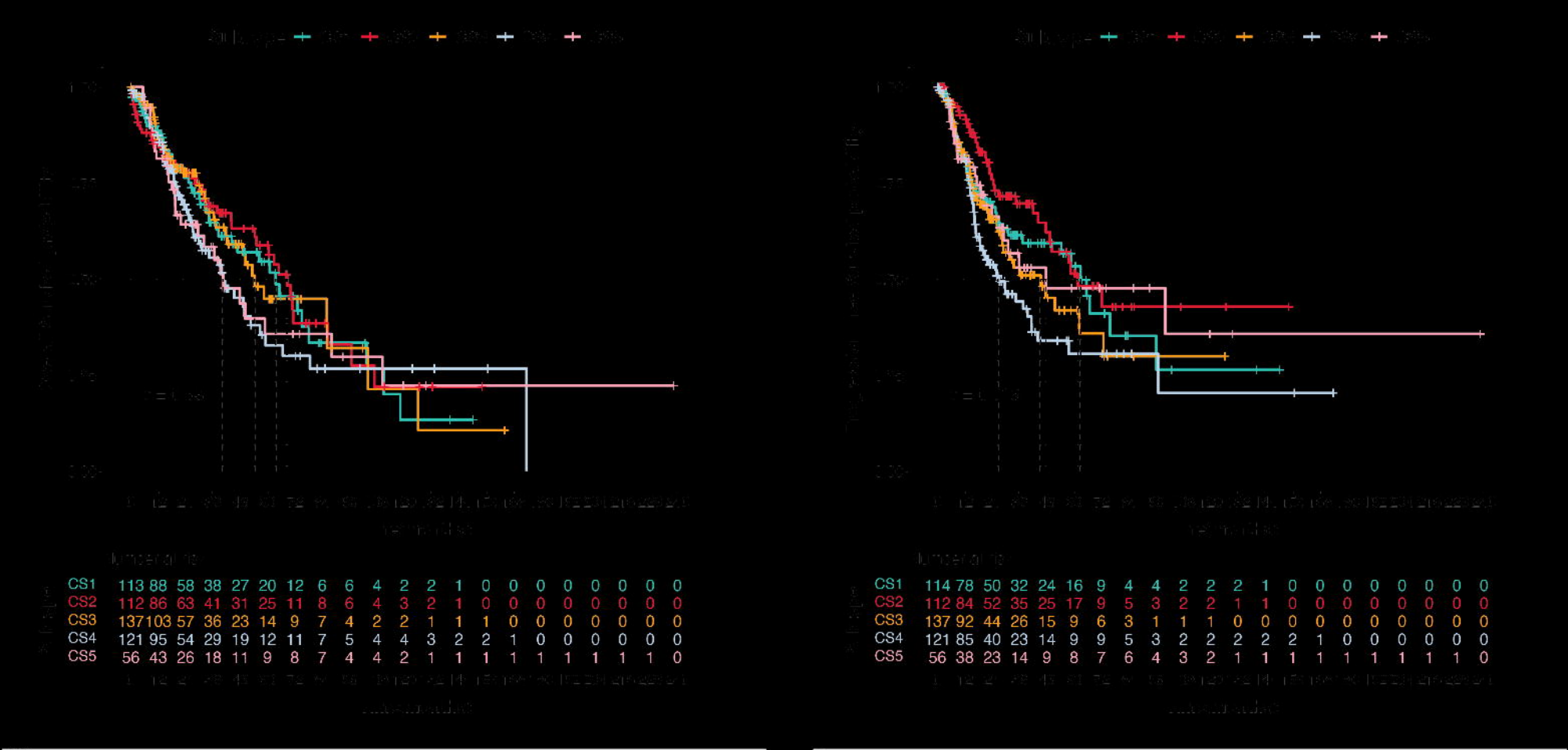

Silhouette plot
n = 551
5 clusters C_j
j : n_j | $ave_{i \in Cj}$ s_i
1 : 116 | 0.39
2 : 113 | 0.43
3 : 140 | 0.25
4 : 125 | 0.41
5 : 57 | 0.38
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width s_i
Average silhouette width : 0.37

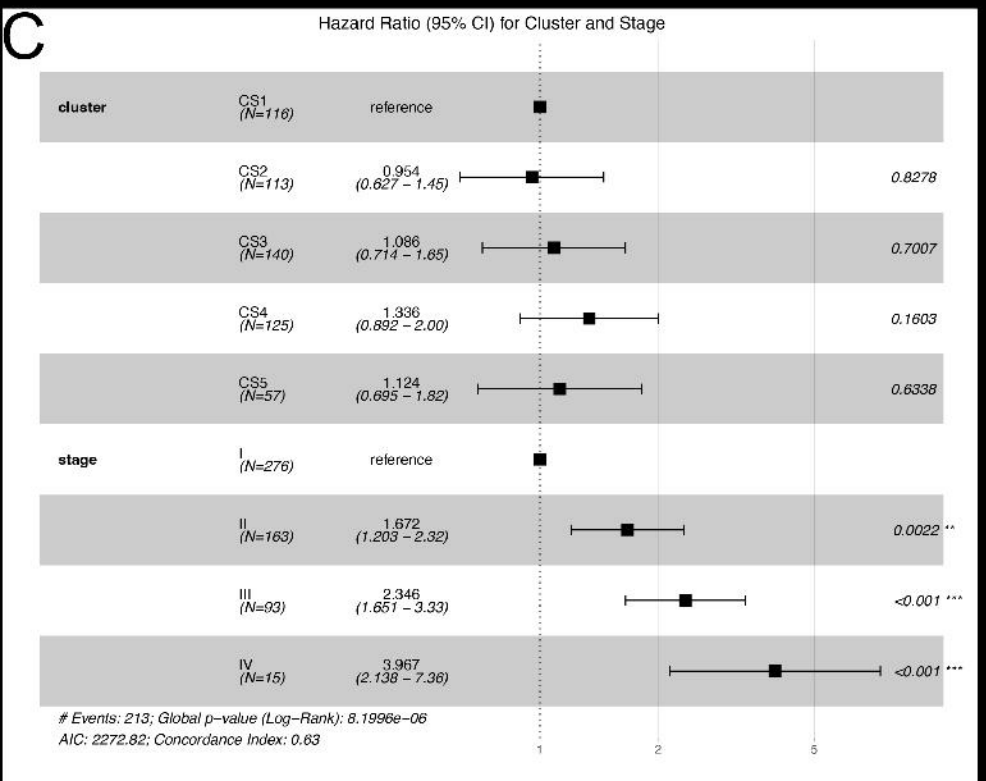

Subtype
gene expression
somatic mutation
copy number variation
tumor microenvironment
DNA methylation
protein expression
Subtype
CS1
CS2
CS3
CS4
CS5
mRNA.counts
2
1
0
−1
−2
mutated
0
1
varied
0
1
tme
2
1
0
−1
−2
M value
2
1
0
−1
−2
rppa value
2
1
0
−1
−2

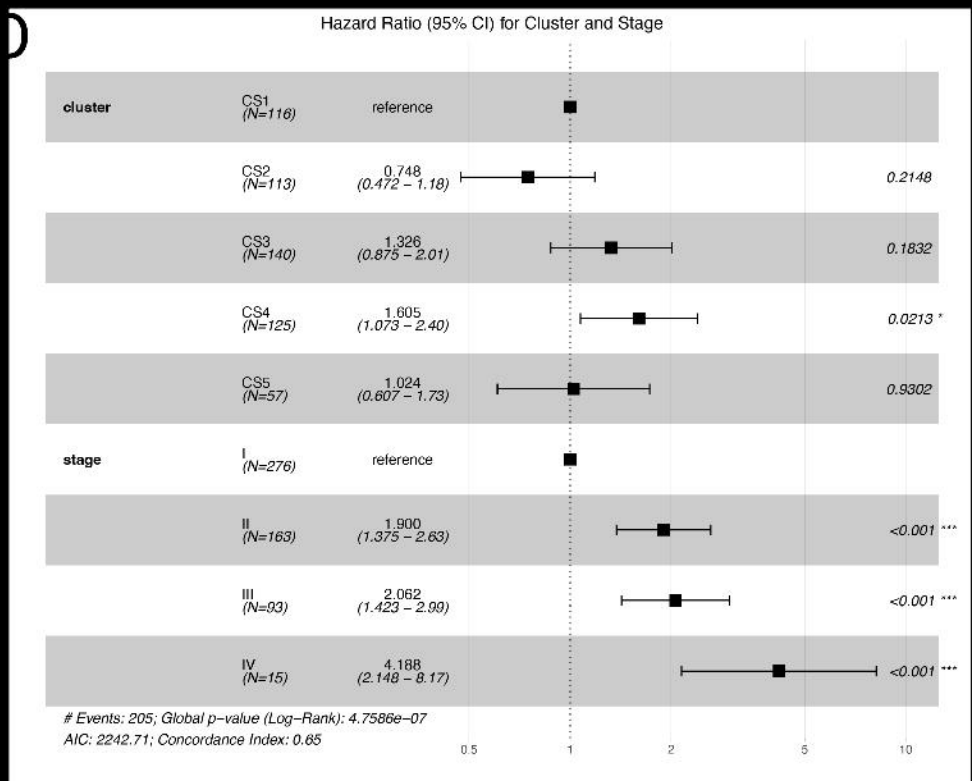

MOVICS clusters
GEXP clusters
MUT clusters
CNV clusters
MET clusters
PROT clusters
TME clusters
gender
project_id
pathologic_stage
MOVICS clusters
1
2
3
4
5
GEXP clusters
1
2
3
4
MUT clusters
1
2
3
4
CNV clusters
1
2
3
4
5
MET clusters
1
2
3
4
PROT clusters
1
2
3
4
TME clusters
1
2
3
4
gender
male
female
project_id
TCGA-LUAD
TCGA-LUSC
pathologic_stage
Stage I
Stage II
Stage III
Stage IV
unknown

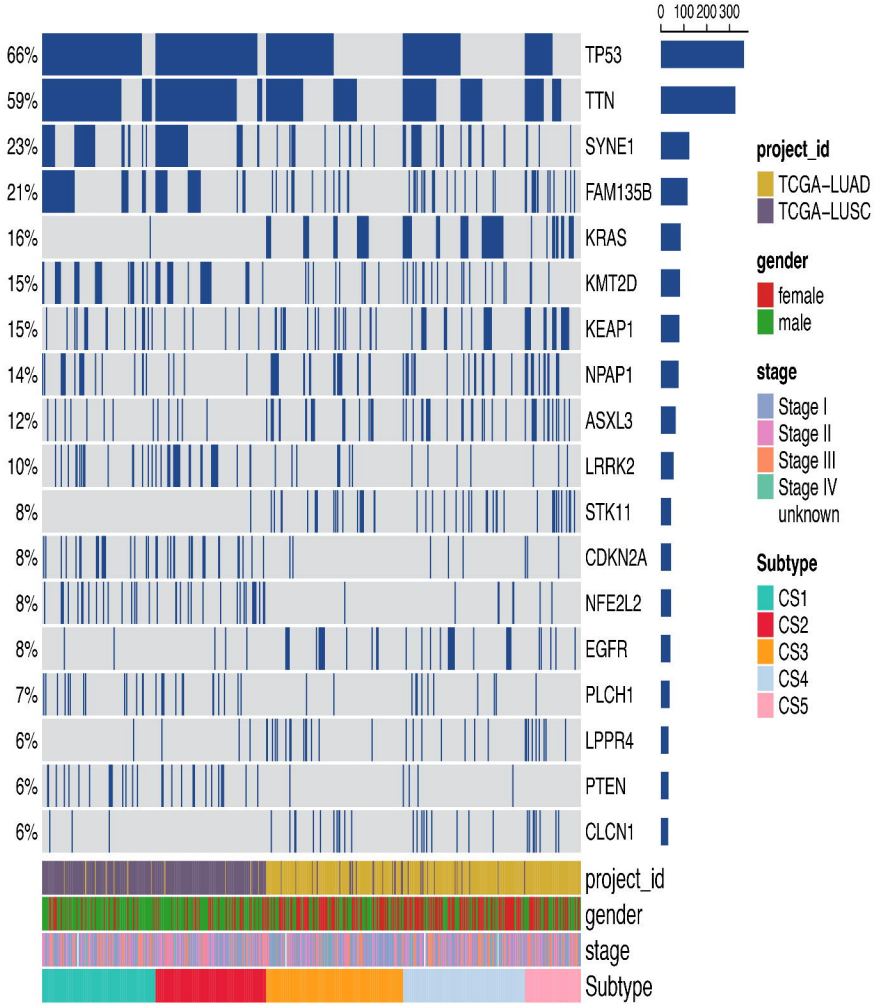

MOVICS Clusters Composition by TCGA_Project
Percentage (%)
100
75
50
25
0
6.0%
94.0%
3.5%
96.5%
87.9%
12.1%
91.2%
8.8%
100.0%
1
2
3
4
5
MOVICS Cluster
Project_id
TCGA_LUAD
TCGA_LUSC

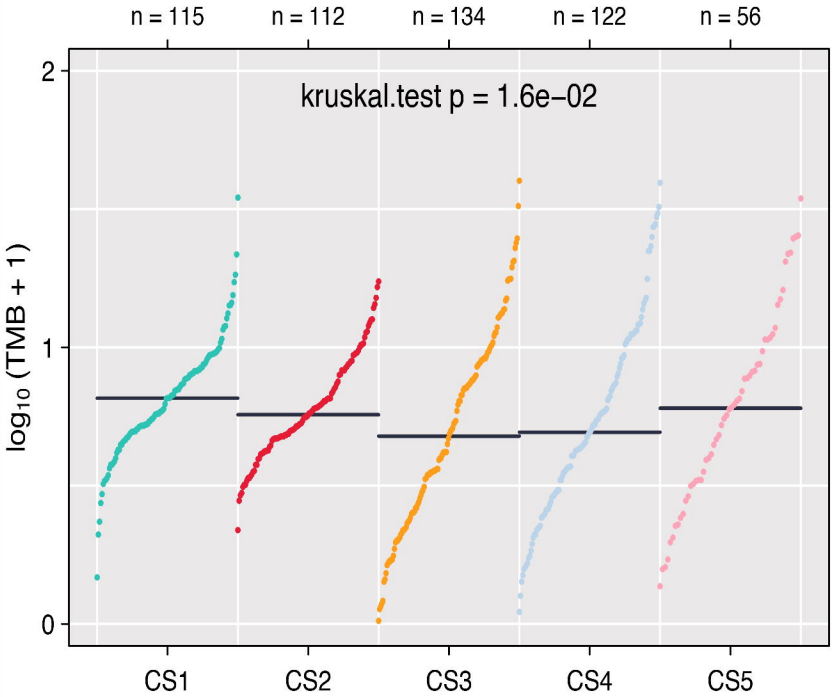

MET Clusters Composition by TCGA_Project and Sex
Percentage (%)
100
75
50
25
0
93.0%
6.4%
0.6%
2.6%
88.1%
9.3%
4.2%
95.8%
0.6%
2.8%
1.7%
95.0%
1
2
3
4
MET Cluster
Group
LUAD_male
LUAD_female
LUSC_male
LUSC_female

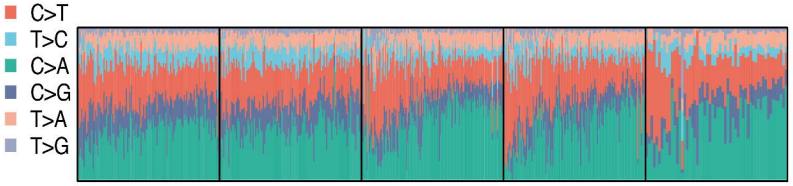

CS1 113 88 58 38 27 20 12 6 6 4 2 2 1 0 0 0 0 0 0 0 0
CS2 112 86 63 41 31 25 11 8 6 4 3 2 1 0 0 0 0 0 0 0 0
CS3 137 103 57 36 23 14 9 7 4 2 2 1 1 1 0 0 0 0 0 0 0
CS4 121 95 54 29 19 12 11 7 5 4 4 3 2 2 1 0 0 0 0 0 0
CS5 56 43 26 18 11 9 8 7 4 4 2 1 1 1 1 1 1 1 1 1 0
CS1 114 78 50 32 24 16 9 4 4 2 2 2 1 0 0 0 0 0 0 0 0
CS2 112 84 52 35 25 17 9 5 3 2 2 1 1 0 0 0 0 0 0 0 0
CS3 137 92 44 26 15 9 6 3 1 1 1 0 0 0 0 0 0 0 0 0 0
CS4 121 85 40 23 14 9 9 5 3 2 2 2 2 2 1 0 0 0 0 0 0
CS5 56 38 23 14 9 8 7 6 4 3 2 1 1 1 1 1 1 1 1 1 0

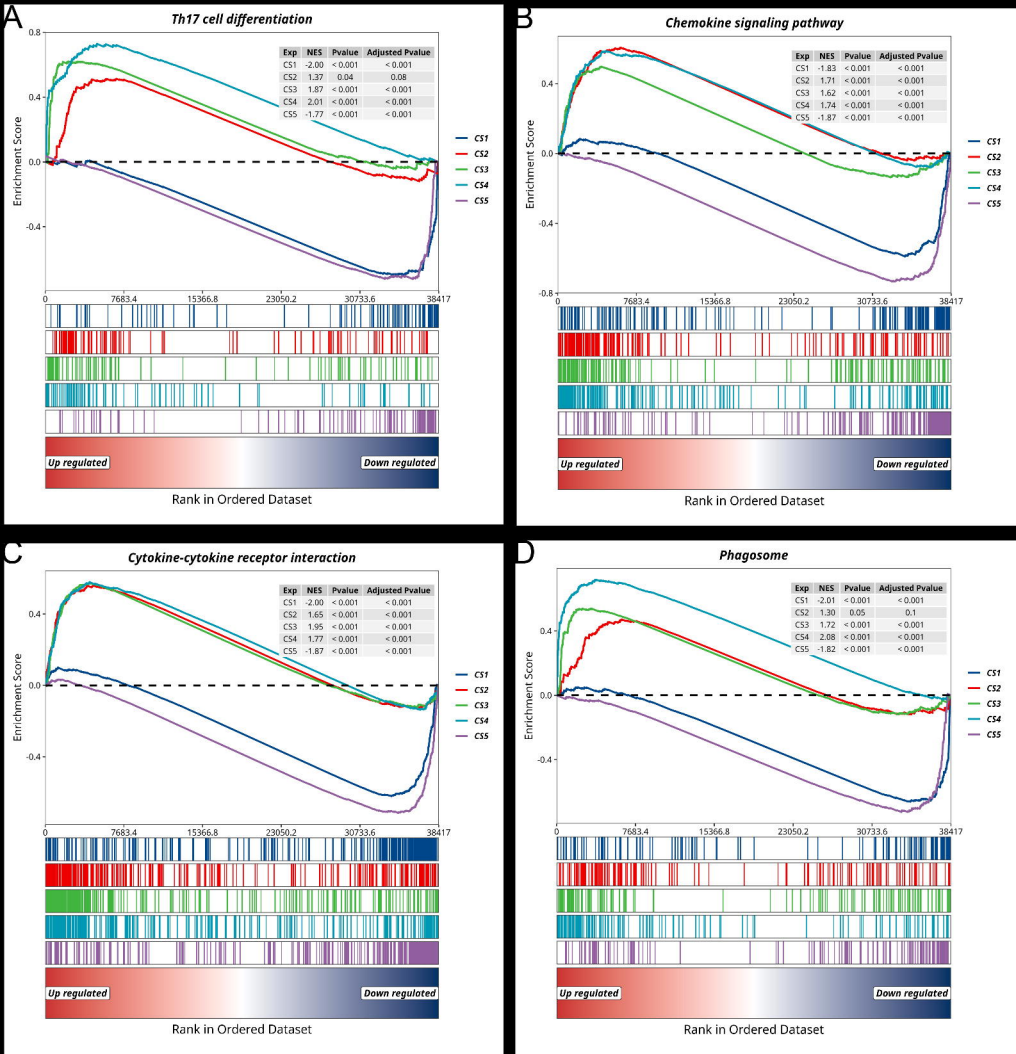

C
Hazard Ratio (95% CI) for Cluster and Stage
cluster
CS1 (N=116) reference
CS2 (N=113) 0.954 (0.627 – 1.45) 0.8278
CS3 (N=140) 1.086 (0.714 – 1.65) 0.7007
CS4 (N=125) 1.336 (0.892 – 2.00) 0.1603
CS5 (N=57) 1.124 (0.695 – 1.82) 0.6338
stage
I (N=276) reference
II (N=163) 1.672 (1.203 – 2.32) 0.0022 **
III (N=93) 2.346 (1.651 – 3.33) <0.001 ***
IV (N=15) 3.967 (2.138 – 7.36) <0.001 ***
Events: 213; Global p–value (Log–Rank): 8.1996e–06
AIC: 2272.82; Concordance Index: 0.63
1
2
5


D
Hazard Ratio (95% CI) for Cluster and Stage
cluster
CS1 (N=116) reference
CS2 (N=113) 0.748 (0.472 – 1.18) 0.2148
CS3 (N=140) 1.326 (0.875 – 2.01) 0.1832
CS4 (N=125) 1.605 (1.073 – 2.40) 0.0213 *
CS5 (N=57) 1.024 (0.607 – 1.73) 0.9302
stage
I (N=276) reference
II (N=163) 1.900 (1.375 – 2.63) <0.001 ***
III (N=93) 2.062 (1.423 – 2.99) <0.001 ***
IV (N=15) 4.188 (2.148 – 8.17) <0.001 ***
Events: 205; Global p–value (Log–Rank): 4.7586e–07
AIC: 2242.71; Concordance Index: 0.65
0.5
1
2
5
10


0 100 200 300
66% TP53
59% TTN
23% SYNE1
21% FAM135B
16% KRAS
15% KMT2D
15% KEAP1
14% NPAP1
12% ASXL3
10% LRRK2
8% STK11
8% CDKN2A
8% NFE2L2
8% EGFR
7% PLCH1
6% LPPR4
6% PTEN
6% CLCN1
project_id
gender
stage
Subtype
project_id
TCGA-LUAD
TCGA-LUSC
gender
female
male
stage
Stage I
Stage II
Stage III
Stage IV
unknown
Subtype
CS1
CS2
CS3
CS4
CS5


n = 115
n = 112
n = 134
n = 122
n = 56
kruskal.test p = 1.6e−02
log10 (TMB + 1)
2
1
0
CS1
CS2
CS3
CS4
CS5


C>T
T>C
C>A
C>G
T>A
T>G


A
Th17 cell differentiation
Exp NES Pvalue Adjusted Pvalue
CS1 -2.00 < 0.001 < 0.001
CS2 1.37 0.04 0.08
CS3 1.87 < 0.001 < 0.001
CS4 2.01 < 0.001 < 0.001
CS5 -1.77 < 0.001 < 0.001
Enrichment Score
0.8
0.4
0.0
-0.4
0
7683.4
15366.8
23050.2
30733.6
38417
CS1
CS2
CS3
CS4
CS5
Up regulated
Down regulated
Rank in Ordered Dataset
B
Chemokine signaling pathway
Exp NES Pvalue Adjusted Pvalue
CS1 -1.83 < 0.001 < 0.001
CS2 1.71 < 0.001 < 0.001
CS3 1.62 < 0.001 < 0.001
CS4 1.74 < 0.001 < 0.001
CS5 -1.87 < 0.001 < 0.001
Enrichment Score
0.4
0.0
-0.4
-0.8
0
7683.4
15366.8
23050.2
30733.6
38417
CS1
CS2
CS3
CS4
CS5
Up regulated
Down regulated
Rank in Ordered Dataset
C
Cytokine-cytokine receptor interaction
Exp NES Pvalue Adjusted Pvalue
CS1 -2.00 < 0.001 < 0.001
CS2 1.65 < 0.001 < 0.001
CS3 1.95 < 0.001 < 0.001
CS4 1.77 < 0.001 < 0.001
CS5 -1.87 < 0.001 < 0.001
Enrichment Score
0.4
0.0
-0.4
0
7683.4
15366.8
23050.2
30733.6
38417
CS1
CS2
CS3
CS4
CS5
Up regulated
Down regulated
Rank in Ordered Dataset
D
Phagosome
Exp NES Pvalue Adjusted Pvalue
CS1 -2.01 < 0.001 < 0.001
CS2 1.30 0.05 0.1
CS3 1.72 < 0.001 < 0.001
CS4 2.08 < 0.001 < 0.001
CS5 -1.82 < 0.001 < 0.001
Enrichment Score
0.4
0.0
-0.4
0
7683.4
15366.8
23050.2
30733.6
38417
CS1
CS2
CS3
CS4
CS5
Up regulated
Down regulated
Rank in Ordered Dataset