Discover Oncology

ANALYSIS Open Access

Integrated multiomics analysis and machine learning refine molecular subtypes and prognosis for thyroid cancer

Peng Zhang[1†], Meizhong Qin[1†], Fen Li[3†], Kunpeng Hu[1*], He Huang[3*] and Cuicui Li[2*]

†Peng Zhang, Meizhong Qin and Fen Li have contributed equally to this work.

*Correspondence:
Kunpeng Hu
hukpeng@mail.sysu.edu.cn
He Huang
huangh85@mail.sysu.edu.cn
Cuicui Li
hnlicuicui@126.com
[1]Department of Thyroid And Breast Surgery, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, Guangdong Province, P.R. China
[2]Department of Nephrology, the Fifth Affiliated Hospital of Guangzhou Medical University, 621 Gangwan Road, Huangpu District, Guangzhou 510730, Guangdong Province, P.R. China
[3]Department of Gastrointestinal Surgery, the Third Affiliated Hospital of Sun Yat-sen University Lingnan Hospital, Guangzhou 510530, Guangdong Province, P.R. China

Abstract

Background Thyroid cancer (THCA) exhibits high molecular heterogeneity, posing challenges for precise prognosis and personalized therapy. Most existing models rely on single-omics data and limited algorithms, reducing robustness and clinical value.

Methods We integrated five omics layers from THCA patients using eleven clustering algorithms to identify molecular subtypes. Based on stable prognosis-related genes (SPRGs), we applied 99 combinations of ten machine learning methods to construct a robust prognostic model—Consensus Machine Learning-Driven Signature (CMLS). The model was validated across multiple internal and external cohorts. Immunogenomic characteristics and drug sensitivity were also evaluated.

Results Three molecular subtypes (CS1–CS3) with distinct clinical outcomes and molecular features were identified; CS2 showed the worst prognosis. A nine-gene CMLS was established, demonstrating strong prognostic performance across cohorts. Patients in the low-CMLS group had better outcomes, stronger immune infiltration, higher TMB/TNB, and greater predicted responsiveness to immunotherapy. Conversely, the high-CMLS group exhibited poor prognosis and lower immunotherapy sensitivity. Drug screening identified six candidate agents for high-CMLS patients.

Conclusion Our study provides a robust multiomics-based classification of THCA and develops a clinically relevant CMLS model for prognostic prediction and therapy guidance. These findings may facilitate risk stratification and inform personalized treatment strategies in clinical practice.

Keywords Thyroid cancer, Multiomics, Machine learning, Immunotherapy, Prognostic signature

1 Introduction

THCA is the most common malignancy of the endocrine system, and its incidence has been steadily rising in recent years [1]. Approximately 90% of THCA cases are derived from epithelial cells and are classified into three major subtypes: papillary thyroid carcinoma (PTC), follicular thyroid carcinoma (FTC), and anaplastic thyroid carcinoma (ATC). In contrast, medullary thyroid carcinoma (MTC), which arises from

parafollicular C-cells, accounts for less than 5% of all cases [2]. According to statistics from 2022, THCA affected 11,860 men and 31,940 women in the United States, underscoring a pronounced gender disparity in incidence [3]. While the overall mortality rate of THCA remains relatively low, 10–30% of patients experience recurrence or metastasis, often leading to more aggressive disease progression [4].

Recent advancements in immunotherapy and targeted treatments have shown promise for managing advanced and treatment-refractory thyroid cancers [5, 6]. Nevertheless, clinical responses to these therapies remain highly variable, and substantial challenges persist in optimizing treatment strategies for high-risk patients [7, 8]. Although immunotherapy has demonstrated efficacy in specific subgroups, many patients fail to derive significant benefit, highlighting the need for more precise therapeutic approaches [9]. One of the key obstacles lies in the marked molecular heterogeneity of THCA, which complicates prognosis and therapeutic response prediction [10]. Therefore, there is a critical need to refine molecular subtypes and develop robust predictive models to guide clinical decision-making.

To address this gap, the integration of large-scale multi-omics data with machine learning offers a promising avenue for discovering reliable biomarkers and improving personalized treatment strategies [11]. By leveraging comprehensive genomic, transcriptomic, and epigenomic profiles, it is possible to identify molecular signatures that can more accurately predict patient outcomes and responses to immunotherapy and other treatments [12].

In this study, we conducted an integrated multi-omics analysis of THCA, incorporating genomic alterations, DNA methylation, and expression profiles of mRNA, microRNA (miRNA), and long non-coding RNA (lncRNA). Using ten multi-omics clustering algorithms, we established a consensus-based molecular subtype classification system. We further identified nine stable prognostic-related genes (SPRGs) and developed a Consensus Machine Learning-Driven Signature (CMLS) through the integration of ten machine learning algorithms. This model demonstrated robust prognostic performance and effectively predicted therapeutic responses in both training and validation cohorts [13]. Our findings provide valuable insights into the molecular heterogeneity of THCA and present a practical framework for improving patient stratification, prognosis prediction, and personalized therapy selection [14].

2 Methods and materials

2.1 Dataset source and data pre-processing

Multi-omics data for THCA patients were obtained from The Cancer Genome Atlas (TCGA-THCA) via the TCGAbiolinks R package [15], including whole transcriptome expression (mRNA, lncRNA), DNA methylation, somatic mutations, and clinical information. Mature microRNA (miRNA) expression profiles were identified using the miRBaseVersions.db package. DNA methylation and clinical annotations were retrieved from UCSC Xena (https://xena.ucsc.edu/), and mutation data were processed using the maftools package.

Patient inclusion criteria were as follows:

- Confirmed THCA diagnosis.
- Complete clinical information.

- Availability of all multi-omics profiles (mRNA, miRNA, lncRNA, methylation, and mutation).
- Overall survival (OS) or progression-free survival (PFS) > 30 days.

Patients failing to meet any of these criteria were excluded. Missing expression or methylation values (< 5% missingness per feature) were imputed using K-nearest neighbors (KNN) imputation. Expression data were normalized using log2(TPM + 1) transformation. For methylation, β-values were used and quantile normalization was applied across samples. All omics matrices were scaled (z-score standardization) prior to integration.

To validate the immunotherapy predictive performance, external datasets (GSE91061, GSE78220, GSE135222 from GEO, and IMvigor210 from the IMvigor210CoreBiologies R package) were incorporated [16]. Only patients with available PFS and treatment outcome annotations were included.

2.2 Multiomics consensus analysis

We employed the MOVICS R package [17] to conduct integrative multiomics clustering and feature selection. For transcriptomic data (mRNA, lncRNA, and miRNA), the top 1,000 genes with the highest median absolute deviation (MAD) were selected to capture the most variable expression features across samples. To identify features associated with patient survival, univariate Cox proportional hazards regression was applied, and genes with significant associations ($p < 0.05$) were retained. For somatic mutation data, the top 5% most frequently mutated genes were selected using the freq method within MOVICS.

Prior to integration, each omics layer (mRNA, lncRNA, miRNA, DNA methylation, and mutation) was independently normalized via z-score transformation to account for differences in measurement scales. All selected features were concatenated into a unified matrix (samples × features), which served as input for downstream clustering.

Multiomics integration and clustering were performed using ten state-of-the-art algorithms provided in MOVICS: CIMLR, ConsensusClustering, Similarity Network Fusion (SNF), iClusterBayes, PINSPlus, moCluster, NEMO, Integrative Non-negative Matrix Factorization (IntNMF), Cluster-Of-Clusters Analysis (COCA), and Latent Representation Analysis (LRA). To determine the optimal number of clusters (k), we utilized the getClustNum and getMOIC functions. These functions evaluate clustering quality based on multiple stability metrics, including the cluster prediction index (CPI) and silhouette width. The CPI is a resampling-based metric that quantifies the reproducibility of clustering by assessing the consistency of sample assignments across multiple iterations. A higher CPI value indicates more stable and robust clustering. The optimal cluster number was defined as the local maximum of the CPI curve.

Finally, consensus clustering was conducted by aggregating subtype labels across the ten algorithms through majority voting, resulting in a robust molecular classification of THCA into three consensus subtypes (CS1–CS3).

2.3 Characterization of subtypes and functional annotation

To evaluate subtype-specific molecular characteristics, we performed gene set variation analysis (GSVA) [18] using hallmark oncogenic and immune-related pathways. Tumor immune and stromal infiltration scores were estimated using the ESTIMATE and IOBR packages, while immune checkpoint expression and tumor-infiltrating lymphocyte

signatures were compared across subtypes. Tumor methylation-inferred lymphocyte (MeTIL) scores were computed using established methods. Transcriptional regulatory networks were reconstructed using the RTN package, focusing on 23 cancer-related transcription factors and 35 chromatin-modifying regulators [19]. To evaluate subtype reproducibility, we selected the top 100 most distinctive genes from each cluster and assessed classification consistency between PAM (Partitioning Around Medoid) and NTP (Nearest Template Prediction) classifiers.

2.4 Development of the machine Learning-Based prognostic signature (CMLS)

Subtype-specific genes were further filtered by univariate Cox regression in the training cohort. A total of ten machine learning algorithms were used for prognostic signature construction: random survival forest (RSF), LASSO, Ridge, elastic net (Enet), CoxBoost, generalized boosted regression modeling (GBM), survival support vector machine (survival-SVM), supervised principal components (SuperPC), partial least squares regression for Cox (plsRcox), and stepwise Cox regression.

For each method, models with fewer than two genes were excluded. The average concordance index (C-index) across five-fold cross-validation was computed for each algorithm. The best-performing model (highest average C-index) was selected to define the Consensus Machine Learning-Driven Signature (CMLS). The final score for each patient was calculated as a linear combination of gene expression weighted by model-derived coefficients:

$$\mathrm{CMLS}_i = \sum_{j=1}^{n} \beta_j \cdot x_{ij}$$

where β_j is the weight of gene j and x_{ij} is its normalized expression in patient i.

2.5 Evaluation of prognostic and therapeutic value

Patients were stratified into high- and low-CMLS groups based on the optimal cutoff identified by the survminer package. Prognostic value was assessed using Kaplan–Meier survival analysis, log-rank test, and multivariate Cox regression adjusting for clinical covariates. To evaluate robustness, the performance of CMLS was compared against 22 previously reported signatures using the C-index.

Immunotherapy response was predicted using multiple approaches. Tumor microenvironment profiles were characterized using the IOBR package with published immune-related gene sets [20]. Tumor mutational burden (TMB), neoantigen burden (TNB), and M1 macrophage levels were compared across CMLS groups. Immune checkpoint blockade response was predicted using TIDE (http://tide.dfci.harvard.edu), SubMap (https://www.genepattern.org), and the TIP pipeline (http://biocc.hrbmu.edu.cn/TIP/).

2.6 Drug sensitivity and pathway analysis

The Oncogenic signaling pathways were explored using GSEA [21]. Drug response prediction was performed using the PRISM and CTRP v2.0 pharmacogenomic datasets, matched to the CMLS expression profile. Drug sensitivity was quantified by area under the curve (AUC) values, and differential responses between CMLS groups were assessed using the pRRophetic and oncoPredict packages.

2.7 Statistical analysis

All analyses were conducted in R (version 4.3.1). Statistical comparisons between groups were performed using Student's t-test or Wilcoxon rank-sum test for continuous variables, and chi-square or Fisher's exact test for categorical variables. Survival differences were assessed using log-rank tests and Cox proportional hazards regression. For all analyses involving multiple testing (e.g., GSVA, drug screening), false discovery rate (FDR) correction was applied using the Benjamini–Hochberg method. Results were considered statistically significant at FDR < 0.05 unless otherwise stated.

3 Results

3.1 Multiomics consensus clustering and prognostic stratification

Using an integrative approach that combined ten ensemble clustering algorithms across multiple omics layers—including somatic mutation profiles, epigenetic methylation, and transcriptomic expression data (mRNA, lncRNA, and miRNA)—we identified three robust molecular subtypes of thyroid cancer (designated CS1, CS2, and CS3). The optimal number of clusters (k = 3) was determined based on the cluster prediction index, a metric reflecting cluster stability and reproducibility (Fig. 1A). These subtypes displayed distinct molecular landscapes (Fig. 1B–D) and were significantly associated with overall survival outcomes ($p = 0.001$, Fig. 1E). Notably, CS2 and CS3 were characterized by worse prognoses compared to CS1 and showed frequent somatic mutations in the BRAF gene, highlighting their aggressive biological behavior.

3.2 Biological and molecular characterization of cancer subtypes

To elucidate the biological underpinnings of the identified subtypes, we applied the IOBR package to assess tumor immune infiltration patterns. CS2 and CS3 exhibited significantly elevated expression of immune checkpoint genes (Fig. 2A), implying their potential responsiveness to immune checkpoint inhibitors (ICIs). Immune cell deconvolution revealed that these subtypes were enriched for macrophages and dendritic cells (DCs), which are crucial components of the tumor immune microenvironment. Further pathway analysis using single-sample gene set enrichment analysis (ssGSEA) showed enrichment of epithelial-mesenchymal transition (EMT) pathways in CS2 and CS3, indicating enhanced metastatic and invasive capabilities. Moreover, these subtypes were associated with a hypoxic tumor microenvironment, which may promote anti-apoptotic mechanisms and contribute to their survival advantage over CS1. Immune-related pathways also showed differential activation, reinforcing the immune heterogeneity among subtypes.

Additionally, hormone response analysis revealed subtype-specific sensitivity: CS1 was primarily responsive to androgen stimulation, whereas CS2 and CS3 were more responsive to estrogen, suggesting a gender-related heterogeneity in thyroid cancer biology (Fig. 2B).

We also examined 23 transcription factors and putative regulators implicated in cancer progression (Fig. 2C). CS1 demonstrated significant activation of androgen receptor (AR) and fibroblast growth factor receptors FGFR3 and FGFR1, while CS2 and CS3 showed elevated activity of HIF1A and epidermal growth factor receptor (EGFR). Moreover, FOXM1 and FOXA1 expression was higher in CS2 relative to CS3, underscoring subtle regulatory differences. Patterns of regulon activity related to chromatin

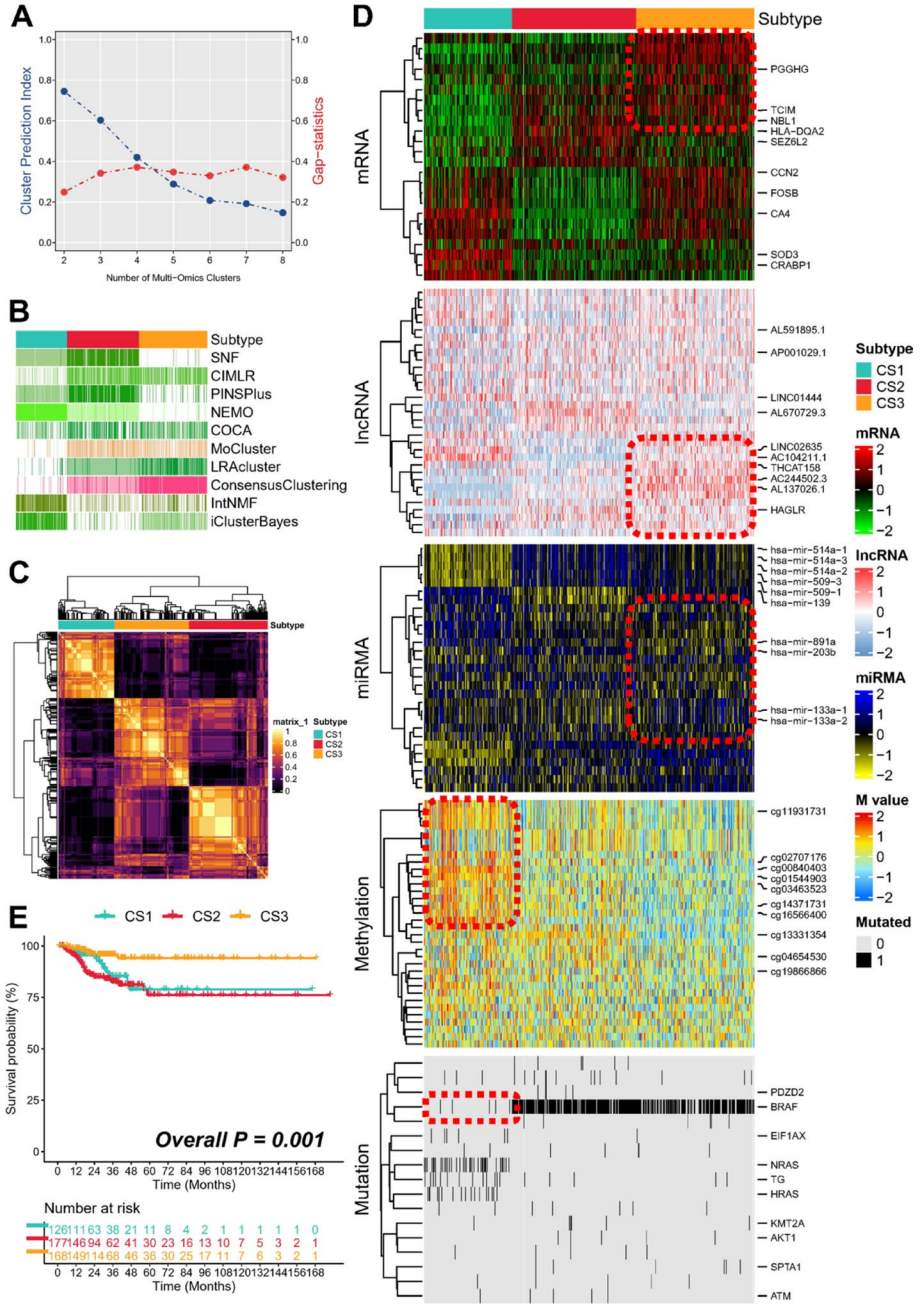

Fig. 1 Multiomics characterization of thyroid cancer prognostic subtypes. A) Cluster prediction index and gap statistic analysis determining optimal cluster number for multiomics classification. B) Consensus clustering heatmap integrating results from 10 state-of-the-art multiomics clustering algorithms. C) Consensus matrix heatmap demonstrating robust sample assignment into three novel prognostic subtypes. D) Integrated multiomics heatmap displaying coordinated patterns across mRNA, lncRNA, miRNA, DNA methylation, and somatic mutations. E) Kaplan-Meier curves showing significant differences in overall survival among the three cancer subtypes (CS1-3)

remodeling suggested that epigenetic mechanisms contribute to the transcriptional distinctions among subtypes.

3.3 Validation of subtype stability

To validate the reproducibility of the molecular subtypes, we selected 100 subtype-specific genes derived from differential expression analyses as classifiers (Fig. 3A).

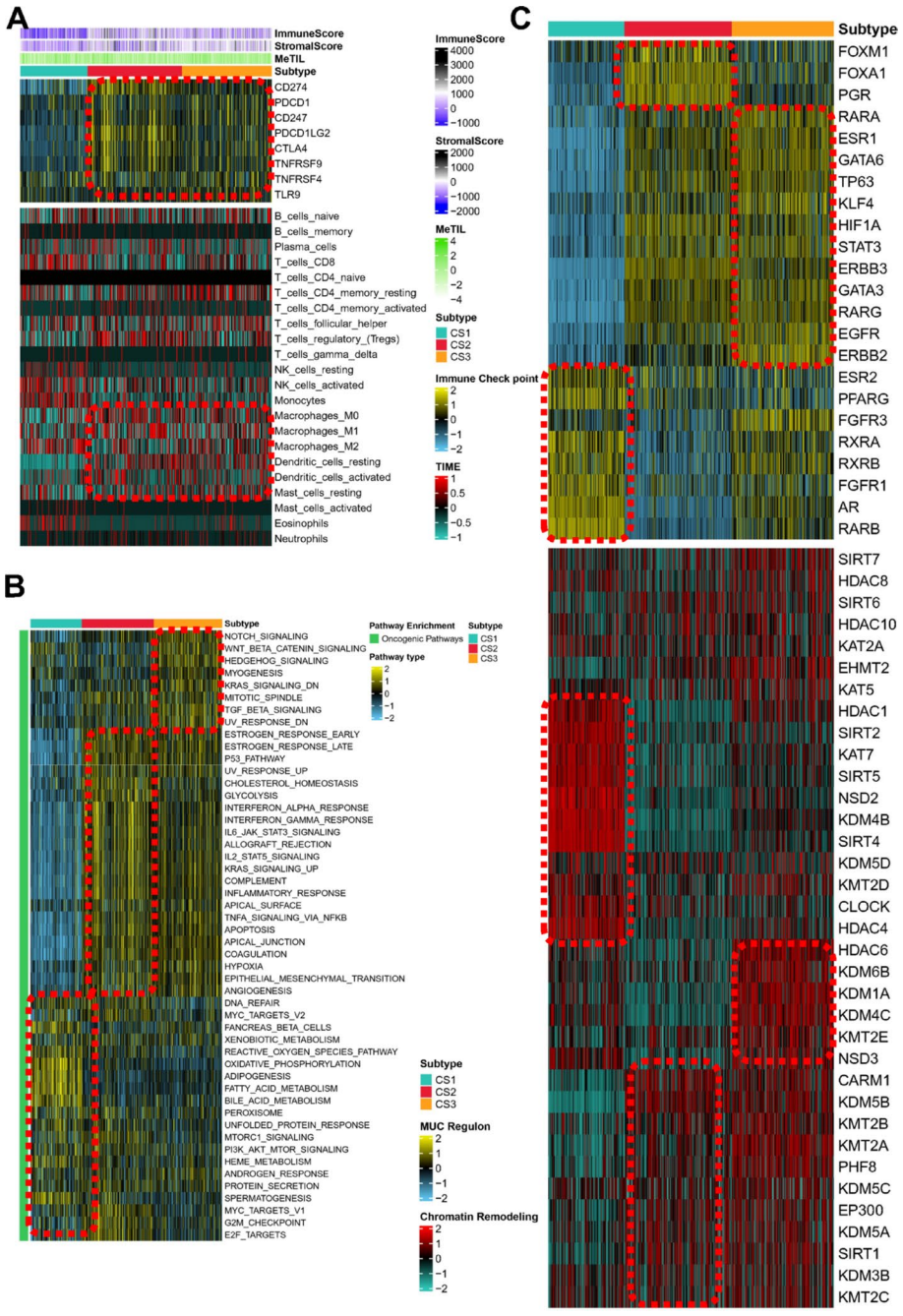

Fig. 2 Tumor microenvironment and molecular features of thyroid cancer subtypes. A) Immune landscape heatmap with annotations for immune/stromal scores, DNA methylation patterns, immune checkpoint expression (top), and immune cell infiltration (bottom). B) Pathway activity heatmap showing differential oncogenic pathway activation across subtypes. C) Transcription factor regulon activity heatmap highlighting chromatin remodeling regulators distinguishing subtypes

Validation was performed using Nearest Template Prediction (NTP) and Partitioning Around Medoids (PAM) algorithms, integrated with clinical data. Both algorithms confirmed the prognostic stratification, with CS1 showing the best overall survival (Fig. 3B–C). The consistency between subtype assignments by NTP and PAM was statistically significant ($p < 0.001$, Fig. 3D–E), supporting the robustness of the subtyping framework.

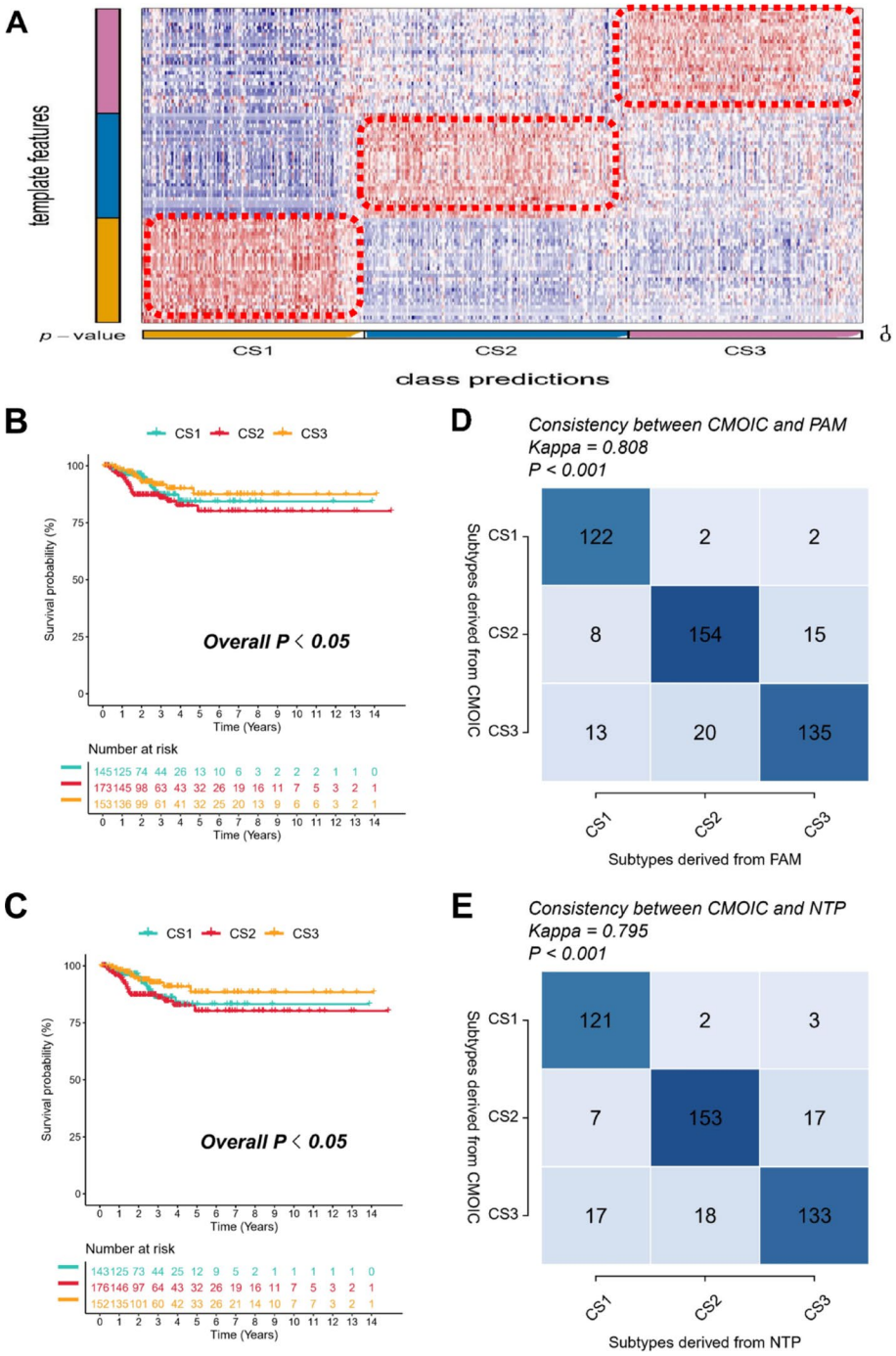

Fig. 3 Validation of thyroid cancer subtypes across independent cohorts. A) Nearest template prediction heatmap validating subtype classification in the testing cohort. B-C) Kaplan-Meier survival curves confirming prognostic significance in both training (B) and testing (C) cohorts. D-E) Heatmaps demonstrating concordance between consensus subtypes and PAM50 (D) or NTP (E) classification methods

3.4 Construction of the consensus machine Learning-Driven signature (CMLS)

The TCGA-THCA cohort was randomly split into training and testing subsets at a 7:3 ratio. Candidate genes were first filtered through NTP and then subjected to univariate Cox regression, retaining those with p-values < 0.05. These genes were fed into an ensemble machine learning framework comprising 99 algorithmic combinations of

feature selection and model building methods (Fig. 4A). Models including fewer than two genes were excluded from further analysis. The final prognostic signature comprised nine hub genes that collectively yielded the highest mean concordance index (C-index) among all tested models (Fig. 4B–C). The derived CMLS score stratified patients into three risk groups, with the high-CMLS group consistently exhibiting poorer survival across training, testing, and the entire TCGA cohort (Fig. 4D–F).

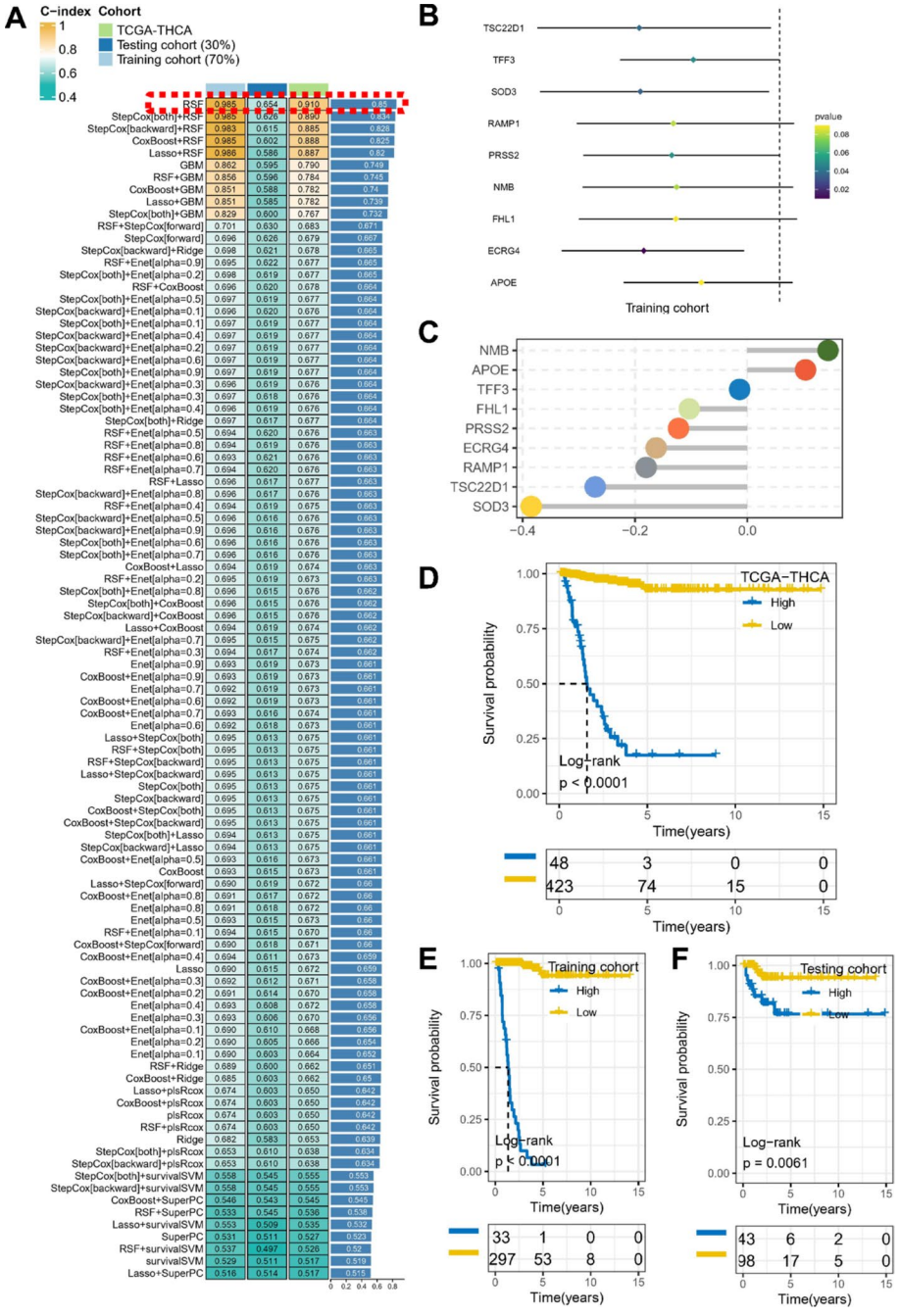

Fig. 4 Machine learning-based prognostic model development. A) Performance heatmap of 99 machine learning algorithms ranked by average C-index across TCGA, training, and testing sets. B) Forest plot of univariate Cox regression results for candidate hub genes. C) Variable importance plot from Random Survival Forest (RSF) algorithm identifying top prognostic genes. D-F) Survival curves stratified by CMLS (high vs. low) in TCGA (D), training (E), and testing (F) cohorts

3.5 Immune microenvironment characteristics associated with CMLS

We investigated the immune landscape differences between high- and low-CMLS groups. Patients with high CMLS had significantly elevated tumor mutational burden (TMB) and tumor neoantigen burden (TNB) (Fig. 5A–B). However, immune cell profiling revealed that high-CMLS tumors were immunologically "cold," characterized by reduced infiltration of key anti-tumor immune cells, notably CD8 + T cells and M1 macrophages (Fig. 5C). Quantitative analysis via the IOBR package confirmed a negative

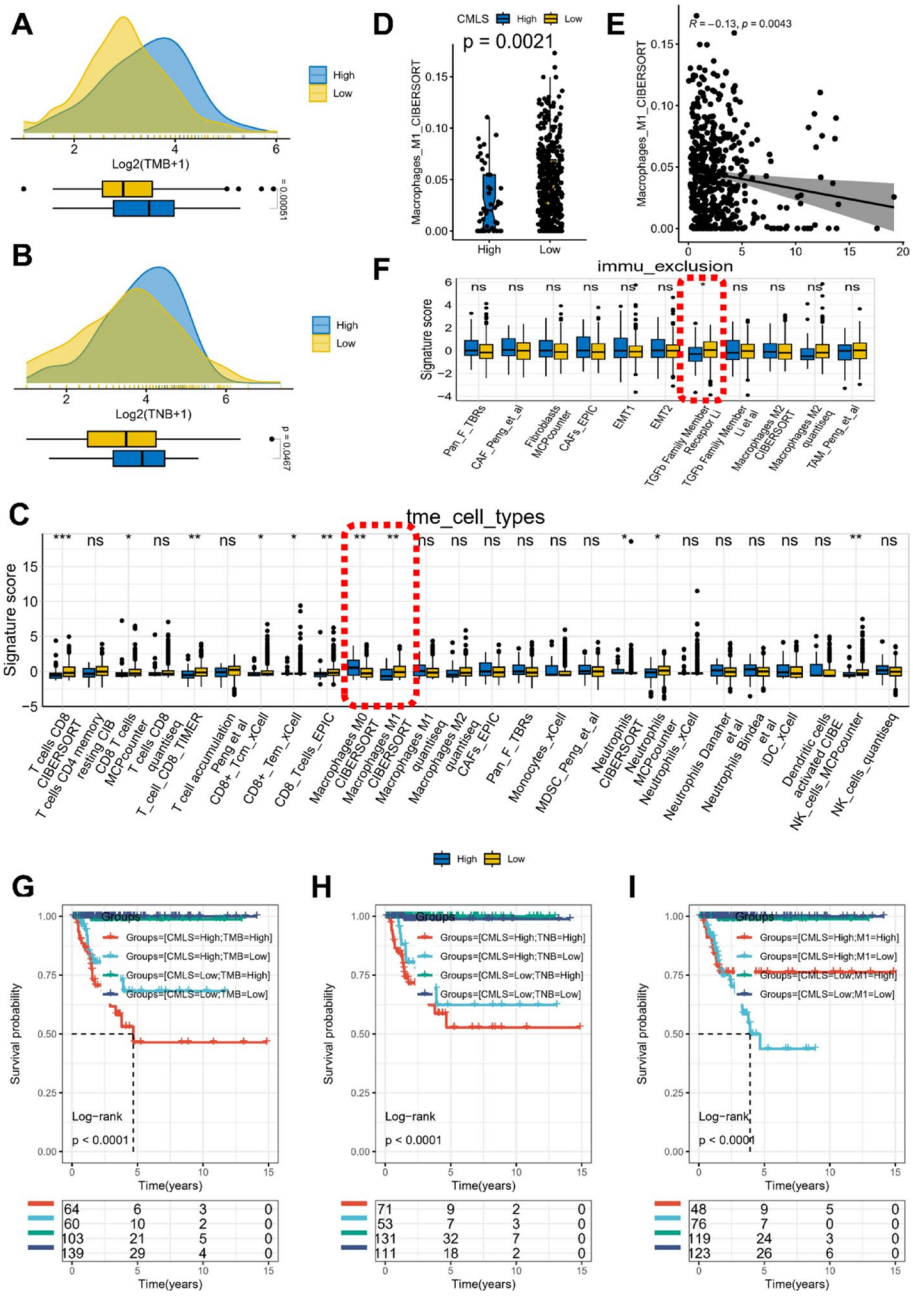

Fig. 5 Immunogenomic characteristics of CMLS subgroups. A-B) Box plots comparing tumor mutational burden (TMB, A) and neoantigen burden (TNB, B) between CMLS groups. C) Heatmap of immune cell infiltration differences between CMLS groups. D-E) Box plots (D) and correlation scatter plot (E) showing M1 macrophage association with CMLS. F) Violin plots comparing immune exclusion scores between CMLS groups. G-I) Stratified survival analyses combining CMLS with TMB (G), TNB (H), and M1 macrophages (I)

correlation between CMLS scores and M1 macrophage abundance (Fig. 5D–E). Differential expression of TGFβ family receptors between groups (Fig. 5F) suggested altered immunomodulatory signaling. Stratified survival analyses further indicated that CMLS, when combined with TMB, TNB, and M1 macrophage infiltration, effectively distinguished patient outcomes (Fig. 5G–I). Collectively, these data demonstrate that CMLS reflects the immune contexture within the tumor microenvironment, which in turn impacts prognosis.

3.6 Predictive power of CMLS for immunotherapy response

We assessed the clinical utility of CMLS in predicting response to immunotherapy. Analysis of the IMvigor210 cohort showed that patients with low CMLS exhibited significantly improved restricted mean survival (RMS) at 3, 6, and 12 months post-treatment, consistent with delayed immunotherapy effects ($p < 0.05$, Fig. 6A–B). Low-CMLS patients demonstrated superior long-term survival and were more likely to respond to therapy (complete or partial response), while higher CMLS scores were associated with progressive or stable disease ($p = 0.015$, Fig. 6C).

These findings were corroborated across additional immunotherapy-treated cohorts (GSE78220, GSE135222, GSE91061), where low-CMLS was consistently linked to better survival and treatment outcomes (Fig. 6D–F). Interestingly, SubMap analysis using a melanoma immunotherapy dataset indicated that high-CMLS patients might be more sensitive to PD-1 blockade (nominal $p < 0.01$, Fig. 6G), a result that warrants cautious interpretation given cohort differences.

Complementary analyses via the Tumor Immune Dysfunction and Exclusion (TIDE) algorithm showed poorer predicted immunotherapy response in high-CMLS patients (Fig. 6H). Additionally, the Tracking Tumor Immune Phenotype (TIP) framework revealed that high-CMLS tumors had reduced recruitment of CD8 + and CD4 + T cells (Fig. 6I), further supporting the immunologically cold phenotype in this group.

3.7 Comparison with published prognostic signatures and therapeutic implications

To benchmark the CMLS, we compared its prognostic performance against 22 previously published gene expression-based signatures in THCA. CMLS demonstrated superior concordance indices in both TCGA and training cohorts, indicating improved predictive accuracy (Fig. 7A–B).

Differential pathway analysis revealed significant enrichment of the epithelial-mesenchymal transition (EMT) pathway in high-CMLS tumors, consistent with a more aggressive phenotype (Fig. 7C).

Potential therapeutic vulnerabilities were explored by integrating data from the Cancer Therapeutics Response Portal (CTRP) and the PRISM drug sensitivity databases. Notably, high-CMLS patients showed decreased sensitivity to cisplatin, a chemotherapeutic commonly used for solid tumors. Previous studies have linked the transcription factor ZEB1 to cisplatin resistance; our analyses suggest that targeting ZEB1 expression may enhance chemotherapy efficacy (Fig. 7D). Moreover, low-CMLS patients exhibited increased sensitivity to six chemotherapeutic agents relative to high-CMLS patients, indicating potential for personalized treatment strategies (Fig. 7E).

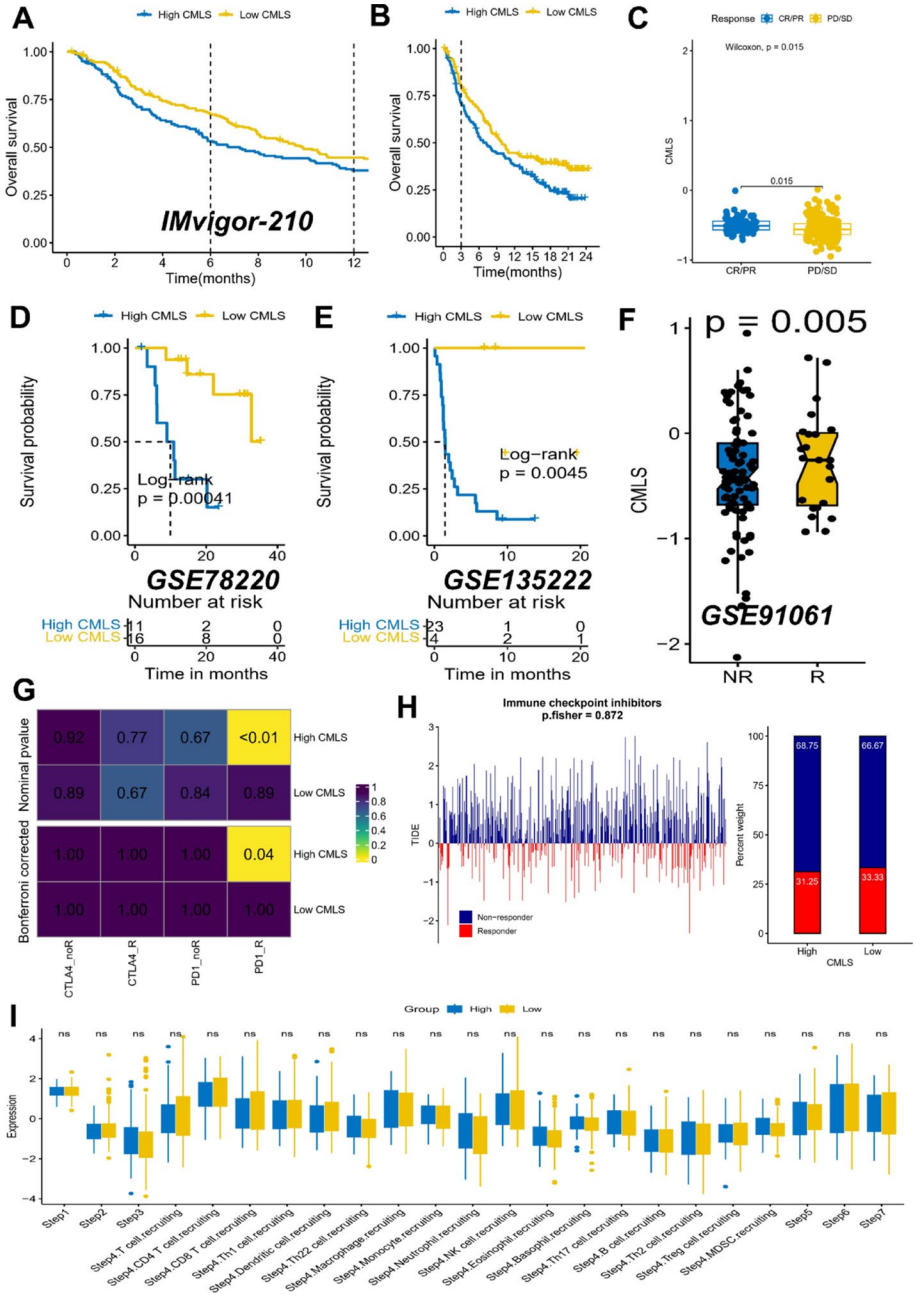

Fig. 6 Immunotherapy response prediction by CMLS. A-B) Restricted mean survival (A) and long-term survival (B) differences between CMLS groups post-treatment. C) Box plot showing CMLS distribution across immunotherapy response groups. D-E) Validation of CMLS prognostic value in immunotherapy cohorts GSE78220 (D) and GSE135222 (E). F) CMLS distribution across response groups in GSE91061 cohort. G-H) Submap (G) and TIDE (H) algorithm predictions of immunotherapy response. I) Box plots comparing tumor-immune cycle activity steps between CMLS groups

4 Discussion

In this study, we identified three molecular subtypes of THCA with distinct clinical outcomes by integrating five omics dimensions using a consensus clustering framework built on ten state-of-the-art algorithms. Based on these subtypes, we developed a robust and generalizable prognostic model—the Consensus Machine Learning-Driven Signature (CMLS)—from 99 combinations of machine learning algorithms. The CMLS

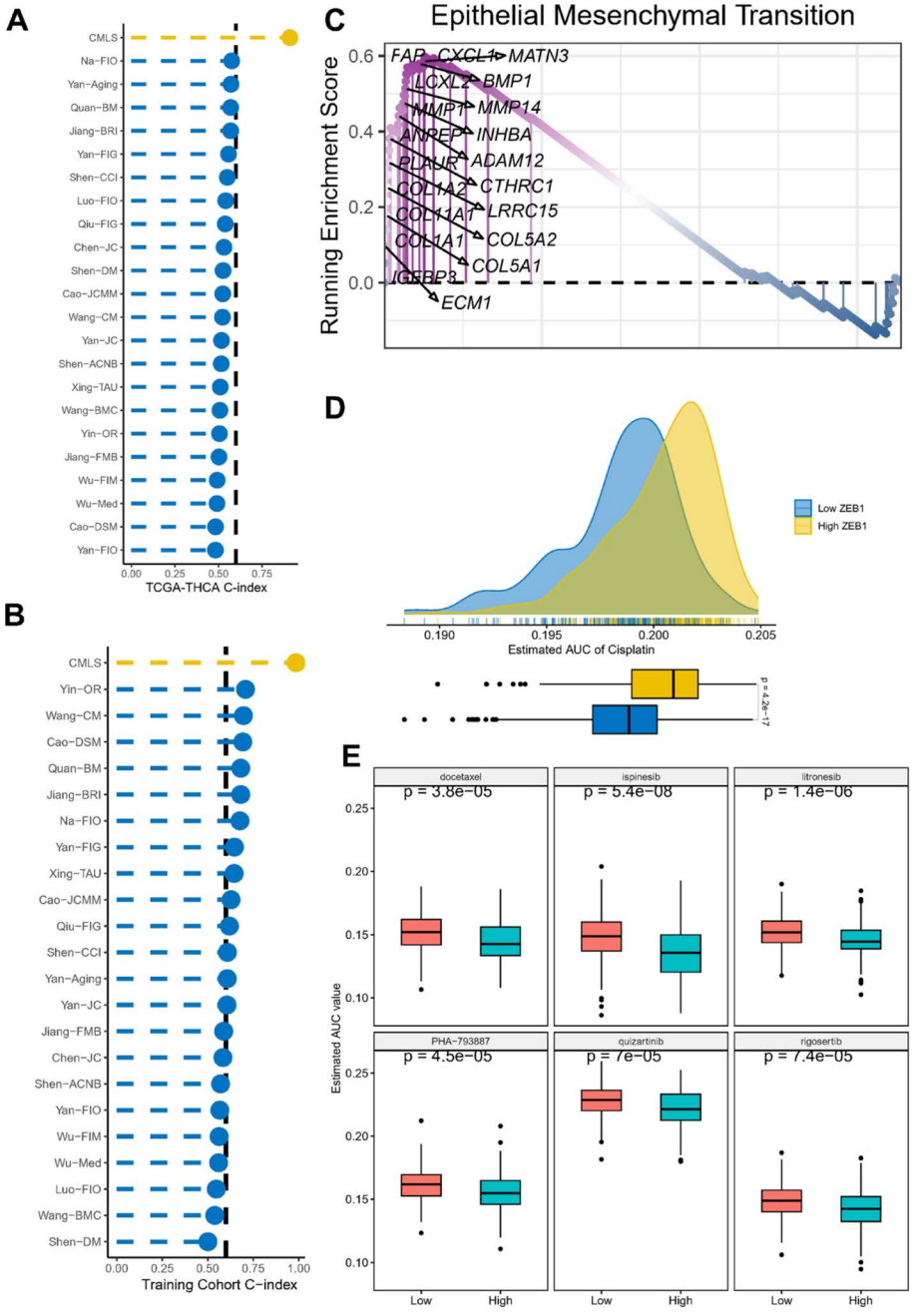

Fig. 7 Comparative analysis and therapeutic implications. A-B) Bar plots comparing CMLS performance against 22 published models in training (A) and TCGA (B) cohorts. C) GSEA enrichment plot showing pathways activated in high-CMLS patients. D) Box plot comparing predicted cisplatin sensitivity between CMLS groups. E) Heatmap displaying drug sensitivity correlations and differential analysis from PRISM datasets

effectively stratified patients across multiple cohorts and showed strong predictive value. Furthermore, we explored the immune landscape associated with CMLS and discovered its relevance to immunotherapeutic responsiveness. For high-CMLS patients with poorer prognosis and limited immunotherapy benefit, we proposed several candidate chemotherapeutic agents. Collectively, our work provides a comprehensive molecular stratification strategy and a clinically applicable prognostic tool to support precision treatment in THCA.

The heterogeneity of THCA poses a major challenge for effective classification and treatment. While most prior studies focused on single-omics layers or relied on a limited number of clustering algorithms, our approach combined five complementary omics data types with ten clustering methods, yielding a robust subtype classification less susceptible to algorithm selection bias [22, 23]. This integrative strategy helps to more faithfully capture the multi-layered regulation of gene expression in cancer, which involves genetic and epigenetic mechanisms such as mutation, methylation, and histone modification [23, 24]. The three subtypes identified in our study demonstrated consistent reproducibility and prognostic value across validation cohorts, suggesting their potential to complement or improve upon traditional THCA classification systems.

To further translate these subtypes into clinically actionable information, we constructed the CMLS using 99 machine learning algorithm combinations. This exhaustive strategy was guided by the average C-index across multiple validation cohorts, effectively mitigating overfitting—a common pitfall in high-dimensional survival modeling [25]. While random survival forest (RSF) performed well on training data, its generalization to testing data was suboptimal [26]. In contrast, the selected CMLS consistently outperformed existing prognostic signatures [27], demonstrating excellent risk stratification ability across cohorts.

We also evaluated the immunogenomic context of CMLS-defined risk groups. Although patients in the high-CMLS group had higher tumor mutational burden (TMB) and neoantigen burden (TNB), they did not show superior immunotherapy response based on TIDE scores [28–31]. This suggests that immune evasion mechanisms may dominate despite high immunogenicity, consistent with the complexity of immune-tumor interactions. In contrast, the low-CMLS group showed better predicted responses to immunotherapy, highlighting CMLS as a potential tool for identifying immunotherapy-sensitive patients. It is worth noting that the observed discrepancy with SubMap analysis, which indicated high-CMLS sensitivity to PD-1 blockade based on a melanoma cohort, likely reflects fundamental differences in the immune microenvironments of thyroid cancer and melanoma. Melanoma is a highly immunogenic tumor with abundant tumor-infiltrating lymphocytes and well-established responsiveness to immune checkpoint inhibitors [32], whereas thyroid cancer typically exhibits a more immunosuppressive and heterogeneous microenvironment. Moreover, SubMap predictions are limited by their reliance on reference cohorts from different tumor types, underscoring the need for cautious interpretation. Consequently, while CMLS shows promise in thyroid cancer, its predictive value in other cancers, including melanoma, requires further validation in tumor-specific contexts.

Given the unfavorable immune profile of the high-CMLS group, we explored alternative therapeutic strategies using a pharmacogenomic screening approach [33, 34]. Six chemotherapeutic drugs, including docetaxel, were identified as potential candidates. Previous studies have shown that docetaxel, especially in combination with anti-PD-1 therapy, can reduce tumor burden in preclinical models [35], providing a rationale for further investigation in high-CMLS THCA patients [36].

Our study presents several methodological innovations compared to prior work [37]. First, we comprehensively addressed tumor heterogeneity by integrating five types of omics data. Second, we minimized methodological bias by applying ten independent clustering algorithms. Third, SPRGs derived from multiple cohorts ensured the

robustness and stability of the modeling genes. Fourth, by selecting the model with the best average C-index performance, we further reduced the likelihood of overfitting. Lastly, we emphasized the need for further investigation into the biological functions of CMLS genes to better understand their role in THCA progression [38].

5 Limitations and future directions

Despite these promising results, our study has limitations. The findings are primarily based on retrospective data and require validation in prospective, multicenter cohorts. Moreover, the specific tumorigenic mechanisms of CMLS genes remain unclear. In future research, we will perform in vitro and in vivo experiments to explore the functional roles of these genes and validate the biological significance of our computational predictions (40).

6 Conclusion

In summary, we identified three reproducible molecular subtypes of THCA through integrative multiomics consensus clustering and developed a novel machine learning–derived prognostic signature, CMLS. This signature effectively stratifies patient outcomes and offers insight into immunotherapeutic responsiveness. Moreover, it provides guidance for alternative therapeutic strategies in high-risk patients. Our study lays a foundation for more precise risk assessment and personalized therapy in thyroid cancer by bridging multiomics integration with advanced computational modeling.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1007/s12672-025-02918-0.

Supplementary Material 1.

Author contributions

Peng Zhang, Meizhong Qin, and Fen Li contributed equally to this work. Peng Zhang and Meizhong Qin conceived the study and performed data analysis. Fen Li contributed to algorithm development and model construction. Kunpeng Hu and He Huang supervised the project and provided clinical insights. Cuicui Li designed the overall study, interpreted the results, and revised the manuscript. All authors read and approved the final manuscript.

Funding

No funding.

Data availability

The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Declarations

Competing interests

The authors declare no competing interests.

Received: 8 April 2025 / Accepted: 4 June 2025

Published online: 23 June 2025

References

1. Mao Y, Xing M. Recent incidences and differential trends of thyroid cancer in the USA. Endocrine-related Cancer. 2016;23:313–22.
2. Baloch ZW, et al. Overview of the 2022 WHO classification of thyroid neoplasms. Endocr Pathol. 2022;33:27–63.
3. Grimm D. Recent advances in thyroid Cancer research. IJMS. 2022;23:4631.
4. Grogan RH. A study of recurrence and death from papillary thyroid cancer with 27 years of median follow-up. 2013.
5. Wu Y, Wang Z, Bai H, Gao Y. Thyroid dysfunction during PD–1 inhibitor treatment in patients with cancer: incidence and association with progression–free survival. Oncol Lett. 2022;24.

6. Zhan D, et al. Expanding individualized therapeutic options via genoproteomics. Cancer Lett. 2023;560:216123.
7. Poté N, et al. Borderline hepatocellular adenomas: A practical diagnostic approach based on pathologic and molecular features. Mod Pathol. 2023;36:100211.
8. Li S et al. Protein regulator of cytokinesis 1: a potential oncogenic driver. Mol Cancer. 2023;22.
9. Chmielik E, et al. Heterogeneity Thyroid Cancer Pathobiology. 2018;85:117–29.
10. Xie K. A biomarker and molecular mechanism investigation for thyroid cancer. Cejoi. 2023;48:203–18.
11. Liu W et al. Insight of novel biomarkers for papillary thyroid carcinoma through multiomics. Front Oncol. 2023;13.
12. Kim J, et al. The cannabinoids, CBDA and THCA, rescue memory deficits and reduce Amyloid-Beta and Tau pathology in an alzheimer's Disease-like mouse model. IJMS. 2023;24:6827.
13. Oldberg Å, et al. Collagen-binding proteoglycan fibromodulin can determine stroma matrix structure and fluid balance in experimental carcinoma. Proc Natl Acad Sci U S A. 2007;104:13966–71.
14. Cabanillas ME, McFadden DG, Durante C. Thyroid cancer. Lancet. 2016;388:2783–95.
15. Colaprico A, et al. TCGAbiolinks: an r/bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44:e71–71.
16. Elisei R, et al. Cabozantinib in progressive medullary thyroid Cancer. JCO. 2013;31:3639–46.
17. 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;36:5539–41.
18. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinform. 2013;14.
19. Lu X et al. Multi-omics consensus ensemble refines the classification of muscle-invasive bladder cancer with stratified prognosis, tumour microenvironment and distinct sensitivity to frontline therapies. Clin Trans Med. 2021;11.
20. Zeng D et al. IOBR: Multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. Front Immunol. 2021;12.
21. Subramanian A, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
22. HER2-Positive Breast Cancer. In Her2-Positive Breast Cancer. Elsevier; 2019. pp. 63–74. https://doi.org/10.1016/b978-0-323-58122-6.00004-0
23. Benjamin DI, et al. Multiomics reveals glutathione metabolism as a driver of bimodality during stem cell aging. Cell Metabol. 2023;35:472–e4866.
24. Signor SA, Nuzhdin SV. The evolution of gene expression in cis and trans. Trends Genet. 2018;34:532–44.
25. Demšar J, Zupan B. Hands-on training about overfitting. PLoS Comput Biol. 2021;17:e1008671.
26. Zhang N, et al. Machine learning-based identification of tumor-infiltrating immune cell-associated LncRNAs for improving outcomes and immunotherapy responses in patients with low-grade glioma. Theranostics. 2022;12:5931–48.
27. Liu Z et al. Machine learning-based integration develops an immune-derived LncRNA signature for improving outcomes in colorectal cancer. Nat Commun. 2022;13.
28. Mao J, et al. Graphene aerogels for efficient energy storage and conversion. Energy Environ Sci. 2018;11:772–99.
29. Han Y et al. A risk score combining co-expression modules related to myeloid cells and alternative splicing associates with response to PD-1/PD-L1 Blockade in non-small cell lung cancer. Front Immunol. 2023;14.
30. Lin M et al. An Estrogen response-related signature predicts response to immunotherapy in melanoma. Front Immunol. 2023;14.
31. Zheng H et al. Characterization of stem cell landscape and identification of stemness-relevant prognostic gene signature to aid immunotherapy in colorectal cancer. Stem Cell Res Ther. 2022;13.
32. Long GV, Menzies AM, Scolyer RA. Neoadjuvant checkpoint immunotherapy and melanoma: the time is now. JCO. 2023;41;3236–48.
33. Yang C et al. Prognosis and personalized treatment prediction in *TP53*-mutant hepatocellular carcinoma: an in Silico strategy towards precision oncology. Brief Bioinform. 2021;22.
34. Chu G, Shan W, Ji X, Wang Y, Niu H. Multi-Omics analysis of novel signature for immunotherapy response and tumor microenvironment regulation patterns in urothelial Cancer. Front Cell Dev Biol. 2021;9.
35. Ma L et al. Safety and efficacy of Anti-PD-1/PD-L1 inhibitors compared with docetaxel for NSCLC: a systematic review and Meta-Analysis. Front Pharmacol. 2021;12.
36. Cheng S-Y, et al. Identification of DPP4/CTNNB1/MET as a theranostic signature of thyroid Cancer and evaluation of the therapeutic potential of sitagliptin. Biology. 2022;11:324.
37. Guo Y et al. Corrigendum: Identification and validation of a novel senescence-related biomarker for thyroid cancer to predict the prognosis and immunotherapy. Front Immunol. 2023;14.
38. Sullivan DK, et al. MYC oncogene elicits tumorigenesis associated with embryonic, ribosomal biogenesis, and tissue-lineage dedifferentiation gene expression changes. Oncogene. 2022;41:4960–70.

Publisher's note