Integrated multi-omics analysis and machine learning based on O\_linked\_glycosylation genes refine molecular subtypes and prognosis for hepatocellular carcinoma

Hongxu Li  
The First Affiliated Hospital of Zhengzhou University

Jie Gao  
The First Affiliated Hospital of Zhengzhou University

Minghao Li  
The First Affiliated Hospital of Zhengzhou University

Bowen Hu  
The First Affiliated Hospital of Zhengzhou University

Zhihui Wang  
The First Affiliated Hospital of Zhengzhou University

Wenzhi Guo  
The First Affiliated Hospital of Zhengzhou University

Yi Zhang  
The First Affiliated Hospital of Zhengzhou University

ShuiJun Zhang (✉ zhangshujun@zzu.edu.cn)  
The First Affiliated Hospital of Zhengzhou University

Research Article

Keywords: Hepatocellular carcinoma, Multi-omics, Single cell, Subtype, O-Glycosylation modification, Machine learning

Posted Date: December 6th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3700057/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.
Abstract

O-glycosylation exerts significant influence on cellular physiological processes and disease regulation by modulating the structure, function, and stability of proteins. However, there is still a lack of research focusing on O-glycosylation in relation to the prognosis of HCC patients. Here we explored expression and function of O-glycosylation gene in HCC from both bulk and single-cell perspectives. The multi-omics data associated with O-glycosylation, identified through the Weighted Gene Co-expression Network Analysis (WGCNA), combined with ten distinct clustering algorithms to define the molecular subgroups of HCC. CS1 was characterized by significant genomic variation, moderate immune cell infiltration and immune function enrichment. CS2 performed a better prognosis, and was featured by stable genomic structure, an immune-hot phenotype with rich immune cell infiltration and sensitive to immunotherapy. CS3 was characterized by a poor prognosis, outstanding genomic instability, an immune-cold phenotype, but can benefit more from treatment with drugs such as sorafenib, cisplatin, paclitaxel, and gemcitabine. Ultimately, we re-emphasized O-glycosylation genes in individual HCC patients, deploying 59 types of machine learning to construct and evaluate the prognostic signature. The microarray results indicated a pronounced upregulation of O-glycosylation hub genes involved in HCC stratification and modeling within HCC tumorous tissues. In conclusion, we have highlighted the significant impacts of O-glycosylation on HCC by redefining the subtypes of HCC as well as constructing the CMLS. This research has established an optimized decision-making platform that enables precise stratification of HCC patients, refines tumor treatment plans, and predicts patient survivability holding broad clinical implications.

Introduction

Liver cancer represents the second leading cause of death in Asia and the third worldwide(1, 2). Hepatocellular Carcinoma (HCC) constitutes 80% of all liver cancer cases, ranking sixth in global cancer incidence and second in mortality rates(3–5). Given the presence of risk factors such as viral hepatitis, alcoholic and non-alcoholic liver disease, coupled with the objective reality of difficulty in early diagnosis of HCC, the median survival of late-stage patients are less than one year(6, 7). The burgeoning field of immunotherapy, particularly immune checkpoint inhibitors (ICIs), has demonstrated inspiring advancements in enhancing the prognosis for patients suffering from solid tumors in recent years(8). Owing to the marked heterogeneity of HCC, an immediate necessity arises to devise novel biomarkers and explore new treatment strategies customized according to individual needs, for achieving precision treatment in the modern era. However, traditional classification systems cannot achieve this well. In addition to classical molecular subtypes, the continual emergence of newer classification, including ferroptosis, cuproptosis, and CD8 + T cells, signifies the colossal potential of developing novel classification systems in addressing the heterogeneity of HCC and towards individualized treatment(9–11).

O-glycosylation is a post-translational modification of proteins, anchoring glycans onto tyrosine (Tyr) residues of proteins(12). The common instances of O-glycosylation modifications entail the addition of N-acetyl galactosamine, N-acetyl glucosamine, mannose, etc. to proteins, which is an essential step in the
biosynthesis of glycoproteins(13). O-glycosylation modifications participate in regulating protein folding, stability, and signal transduction, affecting cell-cell interactions, and thus play an important role in the antigenic properties of viruses and the occurrence as well as development of diseases, including tumors(14–16). N-acetylgalactosamine-transferases (GalNAc-Ts) promotes the spread of tumors by modifying cell surface proteins, enhancing the migration and invasion capabilities of tumor cells(17). Concurrently, O-GlcNAcylation aids in regulating the activation of oncogenic pathways, including Wnt and PI3K/Akt, thereby augmenting the proliferation and metastasis of HCC(16). Previous studies have targeted specific functions such as m6A methylation, lactylation modifications, establishing new molecular subtypes and prognostic signatures, thereby enabling precise stratification and individualized treatment of cancer patients(18, 19). However, given the inherent limitations of expression spectra as well as the singularity of the cohort, the clinical value of these new models cannot be fully realized.

Here, we focused on O-glycosylation modification, using large-scale multi-omics data, consensus ensemble clustering, and advanced machine learning algorithms to decode the heterogeneity of HCC. Additionally, we also provided the assistance in the stratification, precision treatment, and prognosis prediction for HCC patients.

Materials And Methods

Acquisition and processing of multi-omics data and multiple center cohorts for HCC.

We utilized the “TCGABIOLINKS” R package to acquire multi-omics data of the TCGA-LIHC cohort, encompassing mRNA, lncRNA, miRNA, DNA methylation, single nucleotide variations, copy number variations, and corresponding clinical information(20). Furthermore, we collected complete information from four HCC cohorts as the validation set, including three from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo; GSE76427, GSE14520, GSE54236), and one from the International Cancer Genome Consortium (https://dcc.icgc.org/; ICGC_LIRI). Subsequently, we employed the limma package to perform background correction, log2 transformation, and quantile normalization on the data(21). To evaluate the homogeneity of the distribution of expression levels among samples, box plots were utilized as a visualization tool. When multiple probes were assigned to the same gene symbol, we designated the probe with the highest expression level as the gene. For the integration of different datasets, we employed the sva R package to remove batch effects(22). For transcriptomic data, we downloaded FPKM values and performed a log2(FPKM + 1) transformation. Samples that were exclusively presented in a single omics group and patients with an overall survival (OS) of less than one month have been excluded to strengthen downstream robustness(23).

Bulk RNA-seq combined with single-cell RNA-seq analysis.

We assessed the prognostic risk and functional enrichment of O-glycosylation genes with the help of the BEST (https://rookieutopia.com/app_direct/BEST/). The GSCA database (http://bioinfo.life.hust.edu.cn/GSCA/#/expression) was used to study the expression of O-glycosylation genes and their relationships with the stage and prognosis of HCC patients(24).
Single-cell data (GSE166635) were processed using the Seurat (v4.3.0.1) package, and subsequent analyses such as data quality control, normalization, de-batching, dimension reduction and cell clustering were completed. Samples with over 15% mitochondrial gene content or less than 200 transcripts/cell were excluded, completing the filtering of single-cell expression profiles. The “RunHarmony” function in harmony R package removed batch effects. PCA linear dimensional reduction and clustering visualization were completed using the “RunPCA” function and “RunUMAP” function in Seurat. The SingleR package was applied to annotate the different cell clusters, and then the unique characteristic genes of each cell subtype were identified via the “FindAllMarkers” function of Seurat(9). GSVA was used to assess the cells’ O-glycosylation score(25). The interplay of single-cell transcriptome features was scrutinized for high and low-score groups with the aid of “CellChat” R package(26).

Identification of O-glycosylation-related multi-omics data.

Co-expression gene modules were typically identified using the weighted gene co-expression network analysis (WGCNA)(27). We utilized the WGCNA package to construct the TCGA-LIHC co-expression O-glycosylation related multi-omics network. After removing outliers from the samples, the expression matrix of the first 5,000 genes were converted into an adjacency matrix. Subsequently, an unsupervised co-expression relationship was further constructed. By sequentially completing power β (soft threshold) selection, adjacency matrix transformation, gene clustering, dynamic cut module identification, and merging of similar modules, we obtained multi-omics modules significantly associated with O-glycosylation. Subsequently, we selected the most relevant mRNA, IncRNA, miRNA, and DNA methylation modules for further analysis.

Multi-omics consensus ensemble analysis

To enable a thorough analysis, we initially matched multi-dimensional omics through the sample ID (n = 335). Here, we utilized the "getElite" function of the Multi-Omics Integration and Visualization in Cancer Subtyping (MOVICS) package to screen gene features(28). Continuous variables were screened using the "getElitee" function, with the "method" parameter set to "mad", to select the top 500 genes exhibiting maximum variance. The "method" parameter value was then changed to "cox" to integrate the resulting genes with clinical data and identify the prognostic genes for each dimensional dataset (p < 0.05). We organized the filtered muti-dimensional data into a list and applied it to subsequent multi-omics consensus ensemble clustering analysis. We utilized the "getClustNum" function from the MOVICS package to determine the optimal number of clusters. Subsequently, By defining parameter N.clust = 3 and useing 10 clustering algorithms ("iClusterBayes", "SNF", "PINSPlus", "NEMO", "COCA", "LRAcluster", "ConsensusClustering", "IntNMF", "CIMLR", "MoCluster") as input parameter for the "methodslst", we applied the "getMOIC" function for cluster analysis. After computing clustering results for 10 different methods, we integrated the outcomes of various algorithms through the "getConsensusMOIC" (CMOIC) function based on the notion of consensus clustering, aiming to improve the robustness of clustering. Through the above efforts, we obtained the final clustering outcome.

Stability and specific molecular characteristics of CMOIC subtypes
We initially validate the clustering outcomes with subtype-specific features in the validation cohorts. The consistency between the NTP and PAM classifiers regarding consensus ensemble clustering CMOIC was then compared(29). The “compSurv” function in the MOVICS R package was utilized to visually represent the differences in prognoses among subtypes in the validation cohorts. To compare the clinical characteristics of COMIC subtypes in distinct cohorts of HCC, we conducted multivariate COX regression analysis using the “coxph” function from the survival R package. To evaluate the potential relationship between our consensus ensemble clustering subtypes and traditional classifications, five conventional HCC molecular subtype systems were examined, encompassing SURVIVAL_UP and SURVIVAL_DN classifications proposed by Ju-Seog Lee et al; G1-G6 classifications suggested by Sandrine Boyault et al; S1-S3 classifications advanced by Yujin Hoshida et al; Proliferation, CTNNB1, INTERFERON, POLYSOMY, UNANNOTATED classifications put forth by Derek Y Chiang et al, as well as iCluster1-iCluster3 classifications offered by the Cancer Genome Atlas Research Network(30–34).

The evaluation of the immune landscape and Immunotherapy Efficacy of CMOIC subtypes.

Based on the provided cellular metabolism and immune gene sets, we utilized the GSVA R package to further elucidate the differences in biological characteristics among distinct clusters. The R package "ESTIMATE" was utilized for the evaluation of immune scores in tumor tissues. Additionally, GSVA was employed to detect the infiltration of 24 types of immune cells in tumors microenvironment. We additionally assessed the expression of 39 immune checkpoints involved in co-stimulation and co-inhibition to provide further characterization of the tumor immune microenvironment. To gauge the prospective response of HCC patients to ICI treatment, we employed Tumor Immune Dysfunction and Exclusion (TIDE) analysis. The application of the TIDE algorithm facilitates the simulation of tumor immune evasion(35). Patients presenting with elevated TIDE scores are indicative of an enhanced potential for tumor immune evasion and a less favorable immunotherapeutic response.

The genomic variation landscape of the CMOIC subtype.

Utilizing the maftools R package, we can describe characteristics such as the tumor mutation burden (TMB), single nucleotide variations (SNV), and copy number variations (CNV). This provided deeper insight into the genomic landscape differences among CMOIC subtypes. Amplification and homozygous deletions (HOMDEL) in the chromosomal region (AMP), as well as the top 20 high-frequency mutation genes, were visualized using the ComplexHeatmap R package to intuitively display these differences.

Screening for sensitive drugs

Paul developed the commonly used pRRophetic R package for predicting sensitive drugs(36). In this study, we utilized the “compDrugsen” function of the MOVICS R package and compared the sensitivity of our consensus ensemble subtypes to various clinical drugs by leveraging the built-in pRRophetic. It is important to note that IC50 is indicative of the drug’s half-maximal inhibitory concentration. A lower IC50 value implies greater sample sensitivity to the drug under consideration. The drugs sensitivity prediction of CMOIC subtype were based on public databases CTRP2.0 and CTRP2.0, which include sensitivity data.
for 481 and 1448 compounds, respectively. The Area under the receiver operating characteristic curve (AUC) was calculated to assess drug sensitivity in these two datasets, and a lower AUC value indicates increased sensitivity to treatment.

Establishment and assessment of the CMLS

We followed the procedure below to complete the construction and evaluation of the HCC consensus prognosis model. First, we integrated ten machine learning algorithms and combined them to form 59 kinds of machine learning algorithm combinations. In the pursuit of an enhanced construction of CMLS and to streamline subsequent evaluations, we employed TCGA-LIHC, endowed with robust clinical data, as the training cohort. A prognosis signature was constructed employing an amalgamation of 59 algorithms and 23 prognostic-related genes (PRGs). The CMLS in the validation cohorts including GSE14520, GSE54236, ICGC_LIRI, and Merge were calculated using the signature obtained in the training cohort, to augment the precision and generalization of our model. Relying on the average C-index from the four validation cohorts, we ultimately selected SuperPC as the most consensus prognosis model for HCC. The SuperPC package constructed the SuperPC model.

The survival and survminer R packages were used to plot the survival curves for patients in each cohort, classified by high and low scores. The timeROC R package was deployed to forecast the survival rates for 1, 3, and 5 years, subsequently assisting in the drawing of the ROC curve. For the correlation analysis related to the CMLS and CMOIC subtypes, we employed the “ggboxplot” function of the ggplot2 R package for boxplot rendering, and the “to_lodes_form” function from the ggalluvial package to generate the Sankey diagram. The multivariate COX regression analysis for clinical factors was created using the “comph” function from the Survival R package, and we utilized the “forestplot” function of the forestplot R package to finalize the risk forest plot illustration.

Prognostic value of CMLS and potential clinical application

To compare the predictive performance difference between CMLS and these published signatures, we systematically searched for articles on prognostic models published within the past 5 years on PubMed. Utilizing the genes and coefficients provided in the articles, we computed the risk scores for each signature across five different cohorts and ranked them based on the C-index. To escalate the prospective clinical application value of CMLS, we have utilized the independent prognostic hazard factors of HCC, obtained from multivariate COX analysis, to construct a nomogram. Calibration curves were served to delineate the precision of the nomogram, where decision curves were employed to compute the clinical benefits for HCC patients.

Immunohistochemistry (IHC)

134 pairs of HCC and adjacent normal tissue samples were collected in the First Affiliated Hospital of Zhengzhou University. The results of IHC were evaluated and scored by calculating the H-score. H-SCORE
= percentage of weak intensity cells ×1 + percentage of moderate intensity cells ×2 + percentage of strong intensity cells ×3(40).

**Statistical analysis**

All data cleaning, statistical analysis, and visualization work in this study were completed using R 4.2.3 software. We employed the limma R package for the differential expression analysis and simultaneously relied on the MOVICS R package for the consensus ensemble clustering of multi-omics data. In the comparison of two continuous variables, t-tests or Wilcox rank-sum tests were applied to evaluate differences. For comparisons involving more than two groups, one-way ANOVA and the Kruskal-Wallis test were respectively applied to parametric or non-parametric variables. Categorical variables were statistically compared using Chi-square tests or Fisher's exact tests. The survival R package was utilized for both univariate and multivariate Cox, as well as Kaplan-Meier survival analyses. The packages of timeROC and rms were respectively used to plot the ROC and calibration curves. We applied the Cox proportional hazards model combined with a 95% confidence interval (95% CI) to evaluate the Hazard Ratios (HR). All statistical analyses were two-tailed, and statistical significance was determined at a level of p < 0.05.

**Results**

Single-cell combined with bulk-seq analysis to explore the expression and function of O-glycosylation genes.

The workflow of this study was shown in Fig. 1. According to the TCGA-LIHC transcriptome data, a majority of O-glycosylation genes presented high expressions in cancer tissue (Fig. S1) and concurrently served as risk factors for multiple prognostic indicators (DFI, DSS, OS, PFS), providing guidance for patients' prognosis (Fig. S2). Our research demonstrated that the Gene Set Variation Analysis (GSVA) score, grounded on O-glycosylation, was an independent risk factor across multiple HCC cohorts (Fig. S3A). There was a significant correlation between elevated O-glycosylation scores and advanced tumor stages of HCC patients (Fig. S3B). The Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis also revealed that the O-glycosylation gene set was significantly enriched in the cell cycle-related pathway (Fig. S3C).

The Single cell sequencing data GSE166635, encompasses 22,631 cells of two HCC samples were sorted into 21 distinct clusters (Fig. 2A). Upon characteristic genes of different cell subgroups, we identified ten cell clusters: Monocyte.Macrophage, CD4_T, Hepatocytes, NK_cell, Malignant, B_cell, Fibroblast, Endothelial_cell, CMP, and gamma_delta_T (Fig. 2B). Subsequently, the expression of O-glycosylation genes across these identified cell groups were examined (Fig. S4) and visualized according to their expression levels (Fig. 2C). Based on the median value of O-glycosylation score, we categorized cell subpopulations into high-score groups and low-score groups (Fig. 2D). By analyzing the compositional ratios within these distinct cell groups, we observed a heightened presence of both hepatocyte and
malignant cells in the high-score group. Conversely, the low-score group showed a significant increase in the proportion of immune cells (Fig. 2E). Furthermore, we employed CellChat, a tool grounded on ligand-receptor signaling, to investigate the disparities in cellular interactions between high and low scoring groups. In comparison to the low-scoring group, there were significant enhancement in both the number and strength of cellular interactions found within the high-score group (Fig. 2F-G). However, such enhancements demonstrated distinctions across different cellular populations. Specifically, the increase of cellular interactions within high-score groups were predominantly exhibited among the hepatocytes, the malignant cells, and fibroblasts. Conversely, the interactions between tumor cells and immune cells such as T cells, B cells, and NK cells, were comparatively reduced or even declined.

Multi-omics consensus molecular subtypes of O-glycosylation in HCC.

We employed the WGCNA algorithm to screen various omics modules that exhibited the highest level of correlation with the O-glycosylation gene set. The modules with the highest score were the most relevant mRNA, IncRNA, miRNA, and methylation expression profiles associated with O-glycosylation, participating in subsequent multi-omics integration analysis (Fig. S5A-D). Then, we utilized a combination of ten multi-omics integrated clustering algorithms, integrating various indicators such as Cluster Prediction Index, Gap Statistical Analysis, and Silhouette Score, as well as referring to previous research studies (Fig. S5E-F). Eventually, we identified three distinct subtypes within the TCGA-LIHC cohort. Multiple omics data, including mRNA, IncRNA, miRNA, and epigenetic DNA methylation, were involved in the consensus ensemble approach to reduce biases caused by single omics or clustering methods, making the results more reliable (Fig. 3A-B). The results revealed a strong correlation between our classification system and the overall survival (OS) of HCC patients (Fig. 3C). Notably, CS3 subtype exhibited the poorest survival outcomes among all evaluated subtypes, warranting special attention.

Validation of consensus ensemble clustering in multi-cohort.

Here, we utilized differential analysis to identify the characteristic genes of each subtype. Subsequently, we selected 100 upregulated genes from each subtype as classifiers and conducted validation in multiple external datasets to demonstrate the reliability of the ensemble clustering results of consensus-MOVICS (CMOIC). Fig. S6 assessed the consistency between cancer subtypes (CSs) and nearest template prediction (NTP) or Partition around medoids (PAM) in the TCGA-LIHC training dataset. As anticipated, CS3 exhibited the poorest prognosis among all subtypes in multiple external cohorts (p-value < 0.005), with findings aligned with those of NTP and PAM, highlighting the universality of our consensus ensemble clustering (Fig. 4A-F, Fig.S6C-H).

As anticipated, our consensus integrated clustering results revealed, CS1 and CS2 acted as prognostic protective factors (Hazard Ratio < 1) within the TCGA-LIHC cohort. Conversely, CS3, aligning with the poorer prognosis, represented an independent risk factor for HCC patients (HR > 1) (Fig. 4G). In addition, a substantial correlation persisted between our CMOIC subtypes and five published subtypes. Precisely, the CS1 and CS2 subtypes, which performed a favorable prognosis, were considerably correlated with SURVIVAL_UP, Hoshida_S3, Boyault_G12/G56, Chiang_CTNNB1, and iCluster_2/3 subtypes. Conversely,
there existed a notable correlation among the CS3 subtype and SURVIVAL_DOWN, Hoshida_S1/S2, Boyault_G3, and iCluster_1 subtypes (P < 0.001, Fig. 4H). Consistently, the results from the validation cohort including GSE14520, GSE54236, ICGC_LIRI, and Meta were in agreement with this (Fig. 4I-J, Fig. S7A-F). In general terms, the three subgroups identified by our multi-omics consensus ensemble clustering demonstrated significant clinical heterogeneity. Their robust and reproducible performance in HCC aligned well with studies reported previously.

The multiple landscape differences among the CMOIC subtypes.

To decipher the genomic alterations in CMOIC subtypes, we delved deeper into the characteristic mutations and CNVs of these three subtypes. Notably, the mutation frequency of TTN, CTNNB1, ALB, APOB, and ABCA13 in subtypes CS1 and CS2 had prominently escalated. Furthermore, a pronouncedly high mutation of TP53 was a significant feature of the CS3 subtype. In the CS3 subtype, extensive chromosomal alterations were observed, including but not limited to 11q13.3-Amp, 1q21.3-Amp, and 9p21.3-Del (Fig. S8A). The results from FGL/FGG measuring the proportion of genome loss or gain and FGA indicating the rate of genome alterations, highlighted notable variations in the genomes of CS1 and CS3. Conversely, the CS2 subtype displayed fewer genomic changes, suggesting a higher level of genome stability in patients (Fig. S8B).

In consideration of the potential biological functional discrepancies among distinct subtypes, we employed Gene Set Variation Analysis (GSVA) to assess the distinct molecular characteristics of these CSs. Intriguingly, our findings highlighted a salient enrichment of cell cycle and Senescence in CS3, whilst immune cell functions comprising B cells, T cells, NK cells, and antigen processing capacity were significantly aggregated in CS2 (Fig. 5A). Given the pivotal role of tumor immunity in tumor initiation and progression, we evaluated the degree of immune cell infiltration in the microenvironment. Notably, subtypes CS1 and CS2 demonstrated a significantly high level of immune cell infiltration, tending towards an immune hot phenotype. Meanwhile, CS3, characterized by both a high expression of immune checkpoints and low immune cell infiltration, was likely deemed as an immune suppressive phenotype (Fig. 5B-C). The TIDE algorithm was utilized to evaluate and forecast the immune treatment responses of distinct subtypes. The results indicated that CS3 subtype possessed weak immunogenicity, significant markers of immune evasion and exclusion, resulting in a poor response to immunotherapy (Fig. 5D-G). In contrast, the CS1 and CS2 subtypes appeared to yield better benefits from immunotherapy. In clinical practice for HCC, common therapeutic agents were used, including the target drug Sorafenib; chemotherapy drugs such as Cisplatin, Paclitaxel, and Gemcitabine. Notably, the CS3 subtype had demonstrated increased sensitivity, indicating a potential for higher therapeutic gains from these treatments (Fig. S9A-D). Next, we predicted the sensitive drugs of the CS3 subtype based on the public databases CTRP2.0 and PRISM, hoping to provide theoretical support for the development of new clinical treatment. The results showed that the CS3 subtype was more sensitive to BI-2536, clofarabine, dasatinib, gemcitabine, paclitaxel, SB-743921 from the CTRP2.0 database, and dolastatin-10, ispinesib, vincristine, vindesine, volasertib, YM-155 from the PRISM database (Fig. S9E-F).
Development of a consensus machine learning-driven prognostic signature (CMLS) in HCC.

Utilizing univariate Cox regression analysis, we screened for 23 prognostic-related genes (PRGs) from the O-glycosylation gene set, based on the TCGA-LIHC cohort. Subsequently, these 23 PRGs were incorporated into our integrated projects to develop a CMLS. In the TCGA-LIHC training cohort, we applied 59 algorithm combinations and 10-fold cross-validation to construct prediction models and evaluated their predictive ability by calculating the average C-index of each model in all cohorts (Fig. 6A). According to the average C index in all validation queues (GSE14520, GSE54236, ICGC_LIRI, and Merge), we ultimately chose SuperPC (0.631) as our final model. Among the 23 feature genes included in SuperPC, we elected to display those with an HR greater than 1.5 as significant risk factors (Fig. 6B). Initially, within the TCGA-LIHC cohort, the CMLS scores for each sample were calculated. It was found that patients with high CMLS scores showed signs of poorer prognosis, with an AUCs of 1-, 3-, and 5-year OS were 0.759, 0.677, and 0.656, respectively (Fig. 6C-D). The multivariate Cox analysis further indicated that the CMLS score, with a higher Hazard Ratio (HR) than stages (III + IV), was a significant risk factor for patients' prognosis (Fig. 6F). Moreover, we also evaluated the performance of CMLS scores across all external cohorts. The findings suggested that the CMLS score consistently functioned as an independent risk factor in predicting patients' prognosis in multiple cohorts (Fig. S10-11). In conjunction with the previous multi-omics consensus ensemble clustering, the results showed a significant consistency between the poor prognosis C subtype and the high CMLS score group.

Comparison with previously 60 published signatures in HCC

Accompanied by the vigorous development of high-throughput sequencing technology in recent years, numerous prognostic signatures have been constructed grounded on LASSO and stepwise COX algorithms with the aim to stratify HCC patients and guide their prognosis(41, 42). After a system retrieval of the relevant literature published over the past three years, we finally included 60 unique signatures to compare with our CMLS. Our model displayed superior accuracy compared to other models across multiple cohorts (Fig. 7). It should be noted that prior research primarily relied solely on TCGA-LIHC for modeling and validation. As a result, each signature in this cohort exhibited a higher C-index, while generally perform poorly in other HCC cohorts, which may be caused by overfitting. The CMLS which was built based on integrated machine learning algorithms, had the advantage of strong generalization capability, not only has a high C-index in the TCGA-LIHC cohort, but also ranks among the top in multiple external cohorts.

Clinical evaluation of CMLS and validation of HUB genes.

We integrated the independent risk factors of HCC obtained from multivariate COX analysis earlier, namely the stage, CMOIC subtype, with our newly constructed CMLS score. By building a comprehensive nomogram, we tried to decode the clinical application prospects of the CMLS (Fig. S12A). For instance, a patient with clinical stage IV, CMOIC subtype C, and CMLS score 0.4, had an approximate total point of 90, leading to 3-year survival rate of about 32% and a 5-year survival rate of about 17%. The calibration curve verified that the accuracy of the nomogram was consistent with the actual situation (Fig. S12B). At
the same time, the Decision Curve Analysis (DCA) also indicated that the nomogram brought about vastly superior clinical benefits for patients compared to the singular CMLS (Fig. S12C-D). Finally, it's essential to assess the intrinsic relationship between CMLS and the prior multi-omics consensus ensemble clustering. The O-glycosylation genes involved in modeling have a higher C-index in identifying CS3 subtypes (Fig. S13). Considering both the hazard ratio (HR) of O-glycosylated genes in the HCC cohorts and the AUC value of the identified CMOIC C subtype, we ultimately selected two hub genes: POMT2 and DPM1 (Fig. S14A-G). Survival analysis indicated poorer prognosis in patients exhibiting high expressions of either DPM1 or POMT2 (Fig. S14H-I). Moreover, our microarray data validated that the O-glycosylation genes, POMT2 and DPM1, were significantly upregulated in a comparison of 134 paired HCC tissues, as opposed to the normal tissues (Fig. 8A-B, Fig. S15-16). Moreover, the follow-up study of 134 cases demonstrated that individuals with elevated levels of either POMT2 or DPM1 expression exhibited poorer survival rates overall (Fig. 8C-D).

Discussion

Post-translational modifications typically take place subsequent to the synthesis and folding of proteins. These modifications alter the structure, function, and activity of proteins, thus regulating an organism's biological activities(43). O-glycosylation, a form of post-translational modification, influences the function or stability of proteins through the incorporation of glycan. Extensive research indicates the significant role of O-glycosylation in disease onset, progression, and tumor regulation(13). Specifically, Receptor Tyrosine Kinases (RTKs), which are types of transmembrane cell surface receptors, play a significant role in the regulation of cell signaling(44). The aberrant activation of RTKs primarily roots in the alteration of their glycosylation status. Over O-glycosylation modification can potentially result in hyperactivation of RTKs and amplified signaling, promoting the tumor cells' proliferation and survival(13). This work, we explored the expression and function of O-glycosylation genes based on Bulk seq and single-cell seq, clarifying their impact on tumor microenvironment regulation, cell interaction, and prognosis of HCC patients, among other aspects. Simultaneously, we combined integrated clustering, machine learning and other research methods for the first time to complete the stratification of HCC population and prediction of individual patient's survival in multi-center cohorts, in hopes of aiding future O-glycosylation research.

To date, research commonly focuses on specific biological pathways, such as immunity, m6A methylation, and ferroptosis, to stratify HCC populations or construct prognostic survival models for patients(45–47). However, most of these studies were biased towards a single-omics and employed singular analysis methods such as ConsensusClustering, NMF clustering, or LASSO-Cox, which significantly limited the scope and adversely affected the accuracy of the final outcomes. In this study, we employed WGCNA-based screening to filter multi-omics data correlated with O-glycosylation, inclusive of mRNA, lncRNA, miRNA, and methylation, contributing collectively to the construction of an optimized molecular subtype. Additionally, we have forsaken the single clustering algorithm approach in favor of a method that integrated the consensus of 10 distinct clustering algorithms, facilitating the construction of a new molecular subtype of HCC. Subsequently, HCC patients from multiple cohorts including TCGA-
LIHC, GSE14520, GSE54236, ICGC-LIRI, Merge, were included within the subtype research and ensuing validation. The nearest template prediction (NTP) algorithm classifies samples across numerous external cohorts, based on previously identified classifiers(48). The Partition around medoids (PAM) algorithm clusters the dataset optimally through continuous iterations driven by the similarity of the data points(49). Both NTP and PAM worked in tandem to carry out an external validation of the consensus ensemble clustering. Here, CS1 subtype exhibited better survival, while CS3 showed the worst prognosis, with CS2 falling in between. The multivariate COX regression analysis, based on clinical survival, indicated that CS3 was an independent risk factor for the prognosis of HCC. We further explored the potential associations between these three subtypes and previously published subtypes to unveil the potential characteristics of different HCC subtypes. In an analysis of 91 HCC expression spectra, Ju-Seog Lee et al. delineated two HCC subtypes: SURVIVAL_UP and SURVIVAL_DOWN. Tumors associated with the SURVIVAL_DOWN subtype exhibited salient characteristics of cell proliferation and apoptosis-resistant gene expression, while the SURVIVAL_UP subtype correlates with a high patient survival rate(34). Similarly, Yujin Hoshida defined three HCC subtypes: S1, S2, and S3. The S1 subtype was characterized by aberrant activation of the WNT oncogenic signal, S2 was typified by elevated plasma alpha-fetoprotein (AFP) levels and concurrent AKT activation, and S3 was associated with increased expression of genes related to hepatic function and improved patients’ prognosis(33). A further six molecular subtypes of HCC were identified by Sandrine Boyault et al., in which the G3 subtype exhibited TP53 mutations and overexpression of cell cycle regulation genes, indicating a poorer prognosis for these patients(31). It was observed that the significant correlation between our consensus subtype CS3 and SURVIVAL_DOWN, Hoshida_S1/S2, as well as Boyault_G3, confirming the depressing prognostic outcomes of CS3 and emphasizing the need for more attention on these patients’ clinical treatments.

Single nucleotide variations and copy number variations are important components of the mutation landscape which is key to the heterogeneity of tumors(50). The mutation frequency of TP53 was markedly elevated in the CS3, while the mutational rates of TTN, CTNNB1, and ALB increased significantly in the CS2. TP53 mutation in various cancers including breast cancer, liver cancer, leads to abnormal proliferation of tumor cells and escapes immune surveillance, thereby promoting tumor development(51). Patients with high TP53 mutation in HCC have a rather pessimistic prognosis, and the effect of immunotherapy are poor(52). HCC with CTNNB1 mutation display a non-proliferative phenotype, which is often accompanied by high expression of genes involved in liver cell differentiation and function, such as APOB and ALB(53). In recent years, research related to copy number variation (CNV) has gradually deepened. Researchers have found that DNA amplification of 6p21 (VEGFA) and 11q13 (FGF19/CNND1), as well as the homozygous deletion of chromosome 9 (CDKN2A), promote the progression of HCC. VEGFA displays oncogenic effects by promoting angiogenesis and inducing overexpression of HGF, and patients with VEGFA upregulated have poor survival. In addition, the homozygous deletion of CDKN2A inhibits the retinoblastoma pathway, which regulates the transition of the cell cycle from the G1 to the S phase(54). Here, significant genomic variations were noted in the CS1 and CS3 subtype, whereas CS2 exhibited robust stability.
The Tumor Microenvironment (TME) is constituted by a series of immune cells, tumor-associated fibroblasts, and endothelial cells, playing a significant role in cancer cells proliferation and immune evasion(55, 56). The CS2 subtype was characterized by high infiltration of immune cells. CD8 + T cells exert anti-tumor immune response by secreting cytotoxins and IFNγ(57). Dendritic cells pass the auxiliary signals from CD4 + T cells to CD8 + T cells to enhance the response of cytotoxic T lymphocytes (CTL) (58). Furthermore, Natural Killer (NK) cells, upon identifying a carcinogenic transformation on the cell surface marker, possess the ability to directly exterminate the transformed cells(59). Activated Carcinoma Associated Fibroblasts (CAFs) may facilitate tumor growth, angiogenesis, invasion, and metastasis. CAFs interact with other immune cells in the microenvironment by secreting various cytokines and growth factors, thereby forming an immune-suppressive TME that promotes cancer cells to evade immune surveillance(60). Immune checkpoints are a type of membrane-binding protein involved in the regulation of immune response, including co-stimulation (such as CD27, CD28) and co-inhibition (CTLA-4, PD1, LAG3) checkpoints. The high expression of immune checkpoints can lead to a reduction in T-cell co-stimulation factors, inducing a decrease in T-cell effector function(61). The TIDE score can be utilized to evaluate evasion of the immune system within tumors and forecast the rate of response to immune therapy. MSI is an indicator describing the instability of the tumor genome. Tumors with high MSI usually have higher immune activity and can benefit more from immune therapy(62). In this study, the CS2 subtype showed a lower TIDE score and a higher MSI score, predicting a higher response rate to immune therapy. Therefore, we defined CS2 with high infiltration of immune cells and high immune response rate as an immune-hot phenotype. Although CS3 had significantly high expression of immune checkpoints, it lacked activatable immune cells and the presence of immune evasion, thereby being defined as the immune-suppressive phenotype. Sorafenib manifests its anti-tumor properties by blocking the RAF/MEK/ERK-mediated cell signaling pathways(63). Our research indicated patients in CS3 subtype who are unsuitable for immunotherapy may receive more benefits from treatments involving Sorafenib, as well as other chemotherapy drugs such as Cisplatin and Paclitaxel. Combination therapy with Atezolizumab (anti-pd1) and Bevacizumab (anti-VEGF) significantly improved the OS of HCC patients compared to Sorafenib(64). In the future, the development and application of new immunotherapies have promising prospects in improving the HCC patients' prognosis.

Machine learning currently serves as an effective method for multi-omics data analysis, participating in constructing prognostic signatures to predict patients' survival(65). Overfitting is a common problem encountered by machine learning in solving biomedical problems, often due to the model being too complex or the training data being too scarce(66). While overfitted models often demonstrate an impressive capability for fitting the training set well, their ability to generalize are limited(21). Therefore, we have constructed an HCC prognosis signature in 5 HCC cohorts based on 59 combinations of machine learning algorithms to overcome the limitations brought by the algorithm or cohort selection. Based on the average C-index of the model in four validation cohorts, we ultimately chose SuperPC to build the O-glycosylation prognosis signature. The models built by LASSO or Enet had a good C-index in TCGA-LIHC, but they performed poorly in the validation cohorts. Supervised principal component is a generalization of principal component regression, which can be used in regression models to predict
outcome variables). Further, we have collected a total of 60 articles on prognostic signatures published over the past three years for model prediction comparisons with our consensus signature, using C-index as a measurement standard. Although some signatures showed a better C-index in the training cohort, they all performed poorly in the four external cohorts without exception. The CMLS we screened showed better accuracy than other models in almost all cohorts, showing strong robustness and generalizability.

Our research had the following advantages: (1) The expression and function of O-glycosylation genes were elucidated by combining bulk-seq with single-cell-seq. (2) We conducted an in-depth study on the stratification and prognosis prediction of HCC patients, using multi-omics data, 10 clustering algorithms, and 59 machine learning algorithms. (2) Our investigation spanned across five HCC cohorts and has been validated through numerous methods. (3) Utilizing the original HCC data from our center, we completed the verification of O-glycosylation hub gene expression and survival. Of course, we should also pay attention to certain limitations. First, our multicenter queues were mostly retrospective studies, lacking validation of prospective research. Second, the intrinsic mechanism of O-glycosylation impact on HCC requires more fundamental research for explanation. Finally, further clinical trials are needed to confirm the sensitivity of the potential drugs to CS3 subtypes, as well as the accuracy of prognosis prediction for patients in the high score group.

Conclusion

This study, based on the O-glycosylation gene set, identified three molecular subtypes of HCC through the multi-omics consensus ensemble clustering, revealing differences in clinical characteristics, genomic variation, tumor microenvironment, immunotherapy and drug treatment and provided new insights for stratification of HCC populations. Additionally, we have selected prognosis-associated genes and developed CMLS combined with machine learning, which has great potential in evaluating patients' prognosis and predicting survival rates. Our work helps to understand the heterogeneity of HCC and aids in fostering early and precise clinical decisions for tumor patients.

Abbreviations

HCC, Hepatocellular Carcinoma; CMOIC, Consensus-MOVICS; CSs, Cancer subtypes; NTP, Nearest template prediction; PAM, Partition around medoids; HR, Hazard Ratio; SNV, Single nucleotide variations; CNV, Copy number variations; TMB, Tumor mutation burden; FGG/FGL, Fraction of Genome Lost or Gained; FGA, Fraction of Genome Altered; CMLS, Consensus machine learning-driven prognostic signature; PRGs, Prognostic-related genes; TCGA, The Cancer Genome Atlas; GEO, Gene Expression Omnibus; FPKM, Fragments Per Kilobase Million; GSCA, Gene Set Cancer Analysis; MsigDB, Molecular Signatures Database; GSVA, Gene Set Variation Analysis; KEGG, Kyoto Encyclopedia of Genes and Genomes; DEGs, Differentially Expressed Genes; ssGSEA, single-sample Gene Set Enrichment Analysis; GSEA, Gene-set Enrichment Analysis; IHC, Immunohistochemistry; TIDE, Tumor Immune Dysfunction And Exclusion; OS, Overall Survival; DFI, Disease-Free Interval; DSS, Disease-Specific Survival; PFS,
Progression-Free Survival; ICI, Immune Checkpoint Inhibitor; MSI, Microsatellite instability; ROC, Receiver Operating Characteristic; AUC, Area Under the Curve.

**Declarations**

**Funding information**

This work was supported by the Medical Science and Technology Program of Henan (SB201901045), National Natural Science Foundation of China (81971881).

**Author contributions**

HL: Data curation, Formal analysis, Methodology, Writing—original draft, Visualization, Software, Validation. JG: Conceptualization, Writing – review & editing, Methodology. ML: Data curation, Methodology. BH and ZW: Investigation. WG, YZ and SZ: Conceptualization, Funding acquisition, Project administration, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

**Declaration of competing interest:**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**Acknowledgments:**

It is with great gratitude that we acknowledge the TCGA, GEO, and GSCA databases for providing data.

**Ethics approval and consent to participate:** Not applicable.

**Consent for publication:**

All authors have seen and approved the manuscript and consent publication.

**Availability of data and materials:**

The bioinformatics data used in support of the findings of this study are public. Further inquiries can be directed to the corresponding author.

**References**


Tables

Tables 1 is available in the Supplementary Files section.

Figures
Figure 1

The flowchart of this study (Created with BioRender.com). Bulk-Seq and Single-cell Seq techniques were combined to decode the expression and function of O-glycosylation related genes. Furthermore, the multiomics data corresponding to O-glycosylation were systematically downloaded and organized. Stable HCC prognostic subtypes were obtained by consensus clustering using 10 multiomics clustering algorithms: iClusterBayes, moCluster, Cancer Integration via Multikernel LeaRning (CIMLR), Integrative
Clustering of Multiple Genomic Dataset (IntNMF), ConsensusClustering, Cluster-Of-Cluster-Assignments (COCA), NEighborhood based Multi-Omics clustering (NEMO), PINSPlus, Similarity network fusion (SNF), and Landscape Reconstruction Algorithm (LRA). We adopt a multi-faceted approach to illuminate the variations among patients with different subtypes in multiple aspects, including clinical manifestations, genomic landscape, tumor microenvironment, and drug sensitivity. To further identify prognostic genes, all combinations of machine learning algorithms (stepwise Cox, CoxBoost, ridge regression, RSF, GBM, Survival-SVM, LASSO, SuperPC, plsRcox, and Enet) were used to screen out the hub signature with the highest C-index. Finally, CMLS was constructed based on the SuperPC algorithm. Finally, we compared the differences between CMLS and 60 published prognostic signatures, constructed a nomogram to predict patient’ prognosis, and identified hub genes with experimental verification.
Figure 2

Single-cell landscape of O-glycosylation genes in expression and functional. (A-B) A total of 10 cellular categories were identified from 21 subtypes. (C) Expression of O-glycosylation genes at the single-cell level. (D) Distribution of O-glycosylation gene high and low score groups at the single-cell level. (E) Bar chart illustrates the cells proportion of each subgroup in high and low score groups. (F-G) Differences in
number of inferred interactions (left) and interaction strength (right) of all cells between high- and low-score groups. Circos plots shows putative ligand-receptor interactions between each cell.

Figure 3

The multiomics integrative consensus subtypes of HCC. (A) Comprehensive heatmap of consensus ensemble clustering, including clinical information such as stage, T stage, age, consensus subtype, as well as mRNA, IncRNA, miRNA, and DNA CpG methylation site expression profiles. (B) A consensus clustering matrix for three novel subtypes with ten different algorithms. (C) Kaplan-Meier survival curves for different subtypes of TCGA-LIHC patients.
Figure 4

Validation of HCC CSs based on our consensus ensemble clustering. (A) Validation of HCC CSs in the nearest template of the GSE14520 cohort. (B) Survival analysis of HCC CSs in the GSE14520 cohort. (C) The consistency of NTP with PAM in the GSE14520 cohort. (D) Validation of HCC CSs in the nearest template of the GSE54236 cohort. (E) Survival analysis of HCC CSs in the GSE54236 cohort. (F) The consistency of NTP with PAM in the GSE54236 cohort. (G) Multivariate Cox regression analysis of overall
survival (OS) in TCGA-LIHC dataset. (H) Correlations of HCC subtypes of TCGA-LIHC with clinical features and previous HCC classifications. * P<0.05, **P<0.01, ***P<0.001. (I) Multivariate Cox regression analysis of overall survival (OS) in GSE14520 dataset. (J) Correlations of HCC subtypes of GSE14520 with clinical features and previous HCC classifications.

Figure 5
The immune landscape of HCC subtypes in TCGA cohort. (A) The molecular functional differences among the three subtypes. (B) Immune profiles in the TCGA-LIHC cohort. The top annotation of the heatmap shows the immune enrichment score, stromal enrichment score, and comoic subtype. The top panel shows the expression of canonical immune checkpoint genes, and the bottom panel shows the enrichment levels of 24 TME-related immune cells. (C) Heatmap of expression profiles for co-stimulatory and co-inhibitory checkpoint-related genes of HCC subtypes. (D-F) The TIDE algorithm evaluated the immune exclusion, immune escape, and immunogenicity of HCC subtypes. (G) Immunotherapy response ratio of distinct subtypes in the TCGA-LIHC cohort.
Figure 6

Construction and evaluation of CMLS. (A) we integrated ten machine learning algorithms, including CoxBoost, stepwise Cox, Lasso, Ridge, elastic net (Enet), survival support vector machines (survival-SVMs), generalized boosted regression models (GBMs), supervised principal components (SuperPC), partial least Cox (plsRcox) and RSF. (B) Risk forest plot of SuperPC feature genes (HR>1.5). (C) Survival analysis of high and low CMLS scoring groups in TCGA-LIHC cohorts. (D) Time-dependent receiver-operator characteristic (ROC) analysis for predicting 1-, 3-, and 5-year OS in the TCGA-LIHC cohorts. (E) Multivariate Cox regression analysis of OS in TCG-LIHC cohorts.
Figure 7

C-indexes comparisons of CMLS and 59 Published prognostic signatures in the TCGA-LIHC, GSE14520, GSE54236, ICGC_LIRI and Merge-cohort.
The immunohistochemistry and survival analysis of O-glycosylated HUB genes. (A-B) IHC staining was applied to validate the differential expressions of DPM1 and POMT2 using HCC samples from our clinical center. Scale bar = 50 µm&25 µm. (C-D) The relationship between the expression of DPM1 and POMT2 and patient prognosis were explored using clinical follow-up information from our clinical center.
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Table1.xlsx
- Sup1.pdf
- Sup2.pdf
- Sup3.jpg
- Sup4.jpg
- Sup5.jpg
- Sup6.jpg
- Sup7.jpg
- Sup8.jpg
- Sup9.jpg
- Sup10.jpg
- Sup11.jpg
- Sup12.jpg
- Sup13.jpg
- Sup14.jpg
- Sup15.jpg
- Sup16.jpg
- Supanno.docx