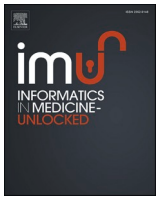
Informatics in Medicine Unlocked 48 (2024) 101514



Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu



Elucidating B4GALNT1 as potential biomarker in hepatocellular carcinoma using machine learning models and mutational dynamics explored through MD simulation

Rohit Kumar Verma [a], Kiran Bharat Lokhande [a], Prashant Kumar Srivastava [b,**], Ashutosh Singh [a,*]

[a] Translational Bioinformatics and Computational Genomics Laboratory, Department of Life Sciences, School of Natural Sciences (SONS), Shiv Nadar Institution of Eminence (SNIoE), Delhi, NCR, India
[b] National Heart and Lung Institute, Imperial College London, London, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Feature selection methods
Beta-1,4-N-Acetyl-galactosaminyltransferase 1
Molecular dynamics simulation
Survival analysis

ABSTRACT

Liver hepatocellular carcinoma (LIHC) is considered one of the primary contributors to cancer-related mortality on a global scale. The identification of new biomarkers is of utmost importance due to the fact that patients with LIHC are frequently detected at advanced stages, leading to an increased mortality rate. The study utilized TCGA-LIHC gene expression datasets to identify biomarkers and to address the complexity of datasets. A combination of feature selection (FS) techniques was used, and the performance of this strategy was assessed using ten machine learning classifiers. The findings were integrated, revealing biomarkers identified through at least five FS techniques. Through our proposed approach, we identified 55 potential biomarkers for LIHC. The Gaussian Naive Bayes Classifier (AUC = 0.99) was found to be the most effective classifier, achieving 98.67% accuracy when utilizing the 55 identified biomarkers in the test dataset. Additionally, we conducted differential gene expression, survival analysis, and enrichment analysis for all the identified biomarkers. Subsequently, Lasso-penalized Cox regression further refined the identified biomarkers to thirteen. Out of thirteen genes, we singled out B4GALNT1 because of its statistical significance in differential expression analysis and increasing importance across various cancer types, including LIHC. We carried out comprehensive bioinformatics and molecular dynamics simulation studies along with other structural analysis of B4GALNT1 in LIHC. In LIHC, six mutations (P64Q, S131F, A311S, R340Q, D478H, and P507Q) have been predicted to be probably damaging by evaluating in-silico prediction algorithms. In comparison to the wild type, the B4GALNT1 variations, specifically P64Q and S131F, demonstrate increased stability. However, these mutations lead to decreased atomic fluctuations, indicating a rigid protein structure. Again, mutations like A311S and P507Q induce increased flexibility, highlighting their structural impact on B4GALNT1. The study demonstrated the combination of various feature selection methods effectively reveals new biomarkers, thereby directly impacting their biological significance. Furthermore, our findings indicate a link between increased B4GALNT1 expression in individuals with liver cancer and a poorer prognosis, highlighting its potential as a promising therapeutic target.

1. Introduction

Cancer, including breast, lung, colorectal, prostate, and stomach cancers, continues to be a major cause of death worldwide [1]. Liver cancer (LIHC) stands as the sixth most widespread cancer globally and is the third leading cause of cancer-related deaths worldwide. According to the Globacan report, around 830,180 new cases of liver cancer were reported in 2020, making up about a third of all cancer-related deaths [1]. Globally, it is projected that the number of deaths from liver cancer (LIHC) will increase to 1.24 million by 2030. The primary reason for liver cancer mortality is its diagnosis at advanced stages, highlighting the durgent need for early detection methods to curb the rising mortality rates [2]. Common symptoms of LIHC include abdominal discomfort and weight loss. Imaging investigations, such as computed tomography (CT)

* Corresponding author.
** Corresponding author.
E-mail addresses: prashant.srivastava@imperial.ac.uk (P.K. Srivastava), ashutosh.singh@snu.edu.in (A. Singh).

https://doi.org/10.1016/j.imu.2024.101514
Received 27 February 2024; Received in revised form 25 April 2024; Accepted 2 May 2024
Available online 7 May 2024
2352-9148/© 2024 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

or magnetic resonance imaging (MRI), are commonly used to diagnose LIHC [3].

Despite advances in screening and discoveries, LIHC demonstrates a rapid clinical course with increased mortality rates. Patients with LIHC are often diagnosed at later stages, limiting treatment options and resulting in early death. Additionally, there is a substantial recurrence rate of 60–70% post-curative resection treatment. Thus, achieving an absolute cure for this disease is challenging, underscoring the urgent requirement for more sensitive novel diagnostic and prognostic markers for LIHC [4].

Alpha-fetoprotein (AFP) is widely utilized as a biomarker for liver hepatocellular carcinoma (LIHC), with detectable levels indicating advanced tumor stages [5]. CTNNB1 and TP53 are prominently mutated genes in HCC. In contrast, several other genes (genetic alteration), including MET, MYC, TERT, CCND1, FGF19, PTEN, TP53, CDKN2A, and CDKN2B, have been strongly linked to the initiation and progression of hepatocellular carcinoma [6]. Furthermore, the emergence of next-generation sequencing (NGS) technology and advancements in bioinformatics analysis have facilitated the identification of potential diagnostic and prognostic biomarkers for cancer. NGS generates extensive data, posing a significant computational challenge in pinpointing the crucial genes associated to a specific disease. Despite the multitude of genes identified through differential expression analysis, only a select few genes exert a significant influence on disease advancement and response to treatment. Hence, feature selection (FS) methods have been employed to identify the most important subset of features whose expression profiling aids in classifying between Cancer and Normal conditions [7]. This process aids in identifying genes that exhibit significant differences across sample conditions. Feature selection (FS) operates by utilizing a small input dataset before constructing a predictive model, ensuring that relevant information is retained. Furthermore, it enhances both performance and computational efficiency. A larger feature may introduce redundancy and potentially noisy data. Handling a larger number of features from higher dimensions can result in overfitting. Thus, employing multiple feature selection techniques and extracting consensus features from them can serve as an alternative strategy to find the best subset.

Different research investigations have employed feature selection techniques to identify the transcriptomic signature in LIHC. Yangyang et al. demonstrated the efficacy of the MMPSO algorithm in efficiently extracting features from high-dimensional datasets. 18-gene signatures were identified that effectively distinguished normal samples from tumors in LIHC [8]. Harpreet et al. utilized computational models designed to categorize LIHC patients according to RNA expression and DNA methylation data, effectively discerning between early and late stages and distinguishing cancerous from normal samples. The study employed a hybrid methodology, merging RNA expression characteristics with DNA methylation probe identifiers [9]. The studies exclusively utilized the primary TCGA-LIHC dataset, consisting of 374 cancer and 50 normal samples. Such data may lead to biased outcomes when classifying cancer and normal samples. In our study, we adopted an approach that involves up-sampling the minority class and subsequently employing a machine learning-based method for better classification. We also employed an approach that combines various feature selection techniques and a prediction model to identify a concise set of features capable of distinguishing between cancerous and normal samples, utilizing the TCGA-LIHC dataset. Our model development involved the utilization of filter, wrapper, and embedded feature selection methods [10]. Filter approaches work independently of models, using statistical tests to determine the relevance of features to the target variable. Wrapper approaches, on the other hand, evaluate feature subsets by training numerous models and picking the subset with the highest performance. Embedded techniques are selection procedures implemented into specific machine learning models, penalizing features with low contributions while training to choose the most important ones effectively. All available approaches were employed to identify set of features capable enough to differentiate cancerous samples from normal ones with enhanced precision. We intend to fully harness the strengths of various feature selection (FS) strategies. We hypothesized that discovering consensus features across all FS approaches would reveal the genes of utmost significance.

An integrated approach combining machine learning and multilayered analysis uncovered 55 gene features serving as potential biomarkers for LIHC. These 55 gene features were identified in at least five feature selection methods. Interestingly, we haven't observed any common features in all the feature selection methods. Differential gene expression analysis confirmed the statistical difference of all identified gene features. Further, using lasso-penalized Cox regression yielded a subset of 13 genes, establishing them as prognostic indicators. Among the 13 genes, we conducted a comprehensive investigation of B4GALNT1 to elucidate its association and impact on LIHC. Expression levels of B4GALNT1 across various cancer types were assessed using TCGA, TIMER2, and UALCAN databases. A recent investigation revealed that B4GALNT1 plays an important role in promoting carcinogenesis by regulating epithelial-mesenchymal transition (EMT) in liver cancer [11], suggesting its potential utility as a biomarker. B4GALNT1, also called β-1,4-N-acetyl-galactosaminyltransferase 1, is a transferase protein crucial for glycosphingolipid biosynthesis, essential for normal cellular functioning. Dysregulated expression of specific GSLs and their associated enzymes is implicated in tumor initiation and malignant transformation across various cancer types. B4GALNT1, or GM2/GD2 synthase, synthesizes sialic acid-containing glycosphingolipids (GSLs) [12]. The change in expression or mutation of B4GALNT1 has been associated with tumor progression in melanoma [13] by inducing gangliosides GM2/GD2, clear cell renal cell carcinoma (ccRCC) [14], colorectal cancer [15], lung adenocarcinoma [16], oral squamous cell carcinoma [17], breast cancer [18], and cervical cancer [19]. Apart from cancer, B4GALNT1 has been associated with Hereditary Spastic Paraplegia [20], Parkinson's disease [21], and Axonal Charcot-Marie-Tooth Disease [22]. Additionally, we examined the correlation between B4GALNT1 and immune cell infiltration. Structural analysis was performed through molecular dynamic simulation to investigate the effects of mutations on LIHC. Through cBioportal [23], Cosmic Database [24], and Human Variant Database (HuVarBase) [25], we identified 16 missense mutations, of which 6 (P64Q, S131F, A311S, R340Q, D478H, and P507Q) were predicted to affect protein function and possibly contribute to disease using in-silico prediction algorithms like SIFT, PolyPhen 2, and CADD_PHRED scores. We conducted thorough molecular dynamics simulations to investigate the structural dynamics of the wild type and different genetic variants at the amino acid level of B4GALNT1. By utilizing RMSD and RMSF analyses, our objective was to assess the stability and flexibility of B4GALNT1 and impact of variations on protein structures. MD simulation studies enabled us to elucidate the impact of specific mutations on the structural integrity and dynamic behavior of B4GALNT1. These findings underscore the significance of B4GALNT1 as an important indicator in cancer progression, with its aberrant expression potentially influencing individual prognosis.

2. Materials and methods

The workflow followed in this study is shown in Fig. 1, which provides a concise overview and description, outlining the process from obtaining LIHC datasets to conducting machine learning and comprehensive analysis.

2.1. Data acquisition and pre-processing

In this study, we accessed publicly available datasets for LIHC from TCGA and downloaded the transcriptome dataset (HTSeq–Counts). The dataset comprises 424 samples (Normal = 50 and Cancer = 374). Furthermore, we retrieved the dataset containing protein-coding genes (Protein coding = 19,962). Subsequently, the protein-coding genes were

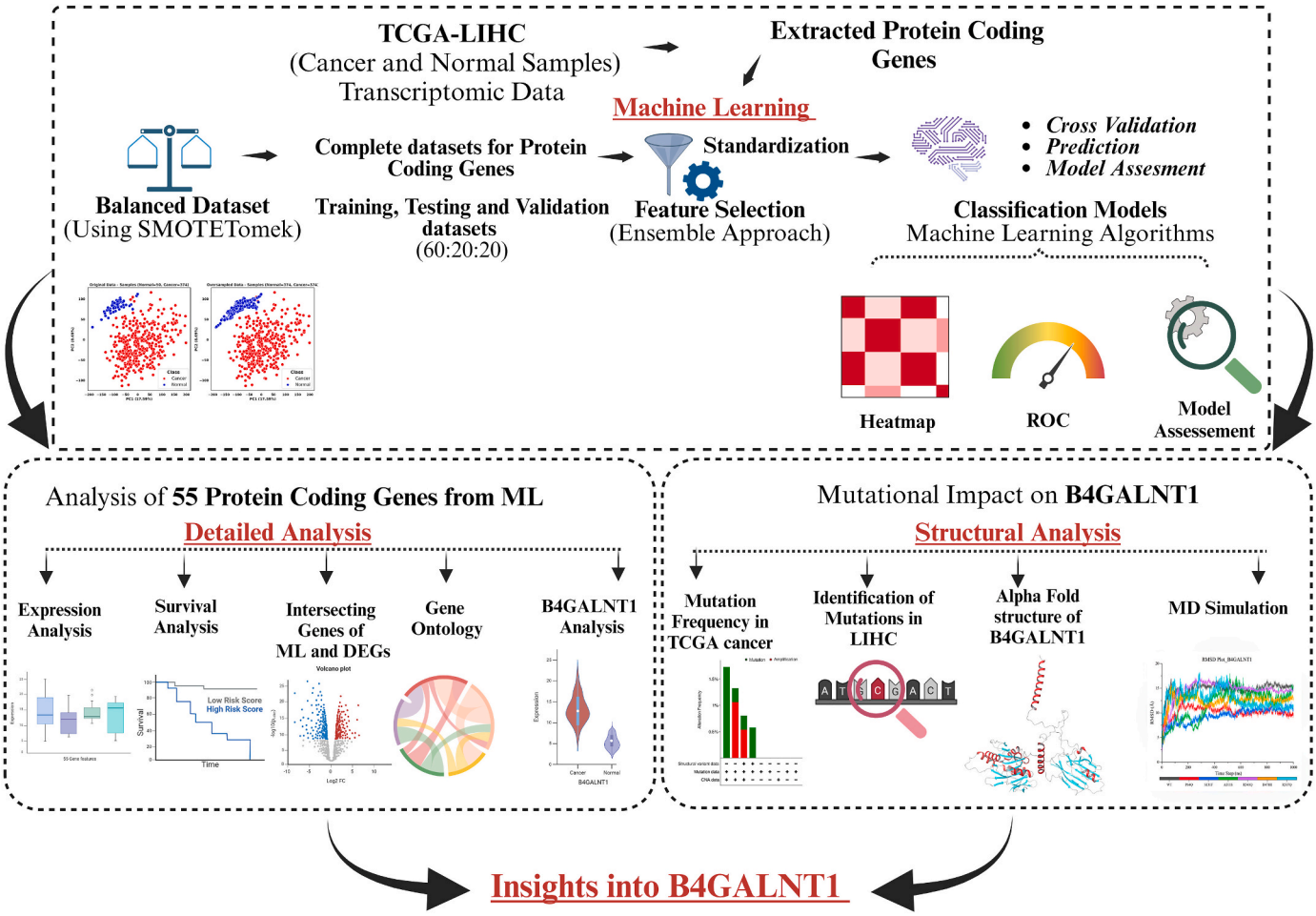


Fig. 1. Overview of the bioinformatics analysis using TCGA-LIHC. This study utilized TCGA-LIHC transcriptome datasets comprising 50 normal and 374 cancer samples. Initially, protein-coding genes (19,962) were extracted from the LIHC dataset. To address class imbalance (Normal = 374, Cancer = 374), the SMOTETomek technique was employed. The dataset was then randomly divided into training, testing, and validation sets. Machine learning was subsequently applied for classification and performance assessment. Furthermore, analyses were conducted for differential gene expression, survival, and pathway enrichment. Additionally, from a subset of 55 gene features, B4GALNT1 was singled out for comprehensive and structural analysis.

merged with samples mapped to their corresponding Ensembl IDs. This selection results in a protein-encoding gene matrix of 19,962. Additionally, the resultant matrix underwent filtration to exclude genes with counts less than ten and those absent in at least three samples, resulting in 19,098 gene features. The dataset normalized to counts per million (CPM) and was subsequently transformed using log2(CPM+1) through an in-house R script. Following preprocessing, the resulting dataset containing solely protein-coding genes comprises 424 samples and 19,098 features, designated as the complete dataset. The data underwent z-score normalization using equation (1) [26]. In this context, z signifies the normalized expression value, x represents the expression value of individual genes, u denotes the mean expression value across all gene samples, and s signifies the standard deviation.

$$z = \frac{x - u}{s} \tag{1}$$

2.2. Machine learning

To improve the classification performance of our model, we addressed the issue of class imbalance by employing SMOTETomek [27], which combines under-sampling and over-sampling. Given the significantly lower number of normal samples (50) compared to cancer samples (374), we utilized an in-house Python script using imbalanced-learn library (imblearn.combine) that included SMOTE function for over-sampling as well as a Tomek Link function for under-sampling. This approach ensured a balanced dataset without any loss of information. SMOTETomek is a hybrid approach in machine learning for handling uneven class distributions. It combines two algorithms: SMOTE (Synthetic Minority Over-sampling Technique) and Tomek links. SMOTE oversamples the minority class, while Tomek links identify and remove potential noisy instances from the dataset. A balanced dataset was obtained after applying the SMOTETomek algorithm (Normal = 374 and Cancer = 374).

2.3. Feature selection

One of the major challenges in this study is identifying the crucial set of features for machine learning models. Next-generation sequencing can detect thousands of genes within a single sample, yet only a small subset of these genes may be pertinent to a particular disease [28]. Identifying the essential genes from this pool of thousands poses a significant challenge. Feature selection is the act of automatically selecting those features in the data that contribute the most to the prediction variable. It is an important stage in machine learning since it helps to limit the amount of data that the model must learn, speeding up training and improving performance. It also enables the discovery and deletion of features that are irrelevant to the task at hand, which might enhance the model's interpretability. The detailed feature selection method has been provided in the **Supplementary Methodology section.**

2.4. Machine learning models

After preprocessing and employing feature selection methods, we implemented widely used ML algorithms, including K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Classifier (SVC), AdaBoost (AB), Decision Tree (DT), XGBoost (XGB), ExtraTrees (ET), Gaussian Naive Bayes (GNB), Logistic Regression (LR), and Multi-layer Perceptron (MLP), to develop binary classification models. Python was employed to implement all the machine learning algorithms, utilizing the scikit-learn package (scikit-learn version 1.2.2, Python version 3.8.3). The performance of the classifiers was examined to select the best-performing algorithm, which was then evaluated further. Fig. 2 represents the overall workflow of machine learning and the implementation of feature selection methods [29,30]. The dataset was randomly split into 60:20:20 ratios, with 60% of the data allocated to the training dataset and the remaining 20% each assigned to the test and validation datasets. Classification models were built using the identified consensus features, and 5-fold Cross-Validation was performed on the training dataset. In this process, the model trains on four folds and predicts on the fifth fold, and this iteration runs on every fold and is averaged out. Internal validation is conducted on both the test and validation datasets, and results are noted to keep track of models that yield promising results. Additionally, performance measures are calculated for both the test and validation datasets.

2.5. Differential gene expression analysis

Differential gene expression was assessed using the DESeq2 package in R/Bioconductor for LIHC samples and normal tissues, considering the 424 samples and 19,962 gene features. The statistical threshold to identify differentially expressed genes (DEGs) was defined as a $|\log2\text{-fold change}| > 1$ and a p-value < 0.05. Furthermore, a volcano plot was generated to visualize the DEGs between LIHC and normal samples [31].

2.6. Survival analysis

Kaplan-Meier survival analysis was conducted for all ML-identified genes using KM plotter, an open-source web-based tool [32]. For the survival analysis, patient samples in LIHC were categorized into high and low mRNA expression groups for each gene. Subsequently, Lasso-penalized Cox regression analysis was utilized to further reduce the gene features from the identified set and assess the association between prognostic genes and patient survival outcomes using Sangerbox (http://vip.sangerbox.com/, Retrieved October 2023) [33]. A p-value < 0.05 was considered statistically significant for both the survival and Cox regression analyses. The risk score was calculated for the reduced gene features for TCGA-LIHC samples to ensure that the median value of each patient's risk score was categorized into high and low groups. The Kaplan-Meier curve and log-rank test were used to examine overall survival between the groups.

2.7. Association of B4GALNT1 in different cancers

To study the differential expression of B4GALNT1 in cancer and normal samples, we utilized CancerLivER (https://webs.iiitd.edu.in/raghava/cancerliver, Retrieved October 2023) [34], TIMER2 (http://cistrome.Shinyapps.io/timer/, Retrieved October 2023) [35], and the UALCAN database (http://ualcan.path.uab.edu, Retrieved September 2023) [36]. The UALCAN portal was also used to examine the relative expression of B4GALNT1 in LIHC and normal tissues, as well as in other sub-groups such as cancer stages, tumor grades, age, gender, weight, historical subtypes, TP53 mutation status, promoter methylation, nodal metastasis status, and race. Furthermore, TNMplot (https://tnmplot.com/analysis/, Retrieved October 2023), which provides differential gene expression of 22 cancers in tumor and normal samples, was employed [37]. All these online tools provide the analysis of differential expression of genes in different tumor and normal samples, ultimately providing insight into understanding B4GALNT1 gene expression in LIHC and different types of cancers. To investigate the genes that are positively and negatively correlated with B4GALNT1 expression in LIHC, we used LinkedOmics (https://linkedomics.org/login.php, Retrieved November 2023), an online web server [38].

2.8. Association of B4GALNT1 with prognosis in 33 different cancers

We performed a survival analysis in different cancers using data obtained from UCSC Xena for all 33 cancer types [39]. The analysis encompassed overall survival, disease-free survival, disease-specific survival, and progression-free survival. The primary objective was to elucidate the correlation between the expression levels of the B4GALNT1 gene (with high and low median cut-off) and disease progression in cancer patients across various cancer types. Forest plots (generated using R packages 'survival' and 'forestplot') and survival plots (created with R packages 'survival' and 'survminer') were plotted using the Kaplan-Meier method. The log-rank test was employed, with significance set at p-value < 0.05, and months were utilized as the unit

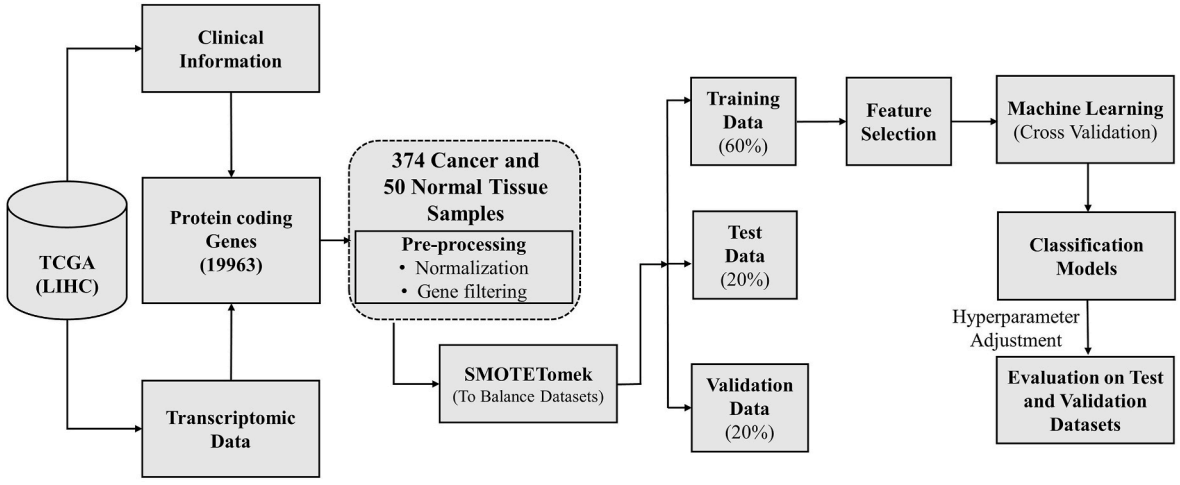


Fig. 2. Representation of the workflow for machine learning analysis. Initial processing involves the normalization of a count matrix. Then, the dataset undergoes upsampling for the minority class using SMOTETomek. Subsequently, it is randomly divided into training (60%), validation (20%), and test (20%) sets. The training sets are utilized for model training. In contrast, the test set is used for initial validation to assess the accuracy of the trained models and fine-tune the model's hyperparameters. After optimizing the hyperparameters, the validation set is evaluated through each model for further assessment.

of survival time.

2.9. Gene alteration analysis

The cBioportal (http://cbioportal.org, Retrieved September 2023) database was accessed to retrieve mutation type, alteration frequency, and copy number alteration (CNA) data from the TCGA database for B4GALNT1. Additionally, analysis was conducted on data concerning the prognosis of B4GALNT1 in LIHC, with or without genetic alteration, utilizing overall, disease-specific, disease-free, and progression-free data. Furthermore, various mutational sites for B4GALNT1 were investigated across different resources, such as Cosmic and the Human Variant Database, an open-access web-based resource for accessing multidimensional genomics cancer datasets.

2.10. Construction of PPI networks of B4GALNT1

We employed GeneMANIA to investigate the connections between B4GALNT1 and other interacting proteins, delving into genetic interactions, co-expression patterns, co-localization, physical interactions, and pathway interactions [40]. Utilizing the STRING tool, we identified the top 50 B4GALNT1 interacting proteins through a comprehensive approach, including text mining, experimental evidence, database queries, co-expression, neighborhood, gene fusion, and co-occurrence information [41]. Subsequently, we conducted analyses to elucidate the impact of B4GALNT1 gene expression on various pathways. Enrichment analyses were conducted on the B4GALNT1 interacting proteins to deepen our understanding. GO and pathway enrichment analyses were performed using the ShinyGO tool, focusing on the 50 identified interacting proteins from the STRING database [42].

2.11. Full atomistic molecular dynamic simulation

For the structural analysis of the B4GALNT1 model, we used different in-silico-based modelling. Still, we found the alpha-fold structure more suitable for studying atomic details for the mutational aspect of structural impact. We downloaded the B4GALNT1 structure from Uniport (https://www.uniprot.org/, Retrieved September 2023), database, and protein were subjected to residue mutation using Maestro [43].

We conducted a comprehensive molecular dynamics simulation for 1 μs (1000 ns) to assess the effects of mutations on the B4GALNT1 enzyme. This simulation encompassed the wild-type B4GALNT1 and its mutant variants, specifically P64Q, S131F, A311S, R340Q, D478H, and P507Q, utilizing the Desmond software [44]. We conducted a total of seven individual simulations, each spanning 1 μs, covering seven different structural configurations. These molecular dynamics simulations enable the calculation of forces and the tracking of the motion of amino acid atoms. It's worth noting that Desmond offers enhanced capabilities, with a more comprehensive system for managing temperature, pressure, and volume, along with built-in functionalities for modelling protein dynamic behaviour [45]. The proteins were prepared within the Maestro program using Desmond's system builder. In this process, they were inserted into an orthorhombic box filled with 1 Å spacing of water molecules. The water molecules were modelled using a three-point water model known as TIP3P. Additionally, a POPC lipid bilayer was introduced into the system, and periodic boundary conditions were applied. The placement of the POPC lipid bilayer was specifically on the N-terminal helix of the B4GALNT1, which encompassed residues Met 1 to Arg 28. The overall charge of the solvent system was balanced by introducing suitable counter ions at random within the solvated system. Energy minimization, a critical stage in MD, was performed using the steepest descent method. An orthorhombic box configuration was employed to enable edge effects in a finite system and enable the application of periodic boundary conditions. This involved placing the atoms of the system to be simulated inside a space-filling box, which was then enveloped by replicated copies of itself. In this study, the OPLS_2005 force field, specifically designed for molecular dynamics simulations of proteins, was employed. Molecular dynamics simulations were executed with constraints set as "all-bonds" and utilized the MD integrator. Temperature control (300 K) was achieved using the Nose-Hoover chain thermostat, while pressure was regulated through the Martyna-Tobias-Klein barostat methods. The simulations were conducted within the NPT ensemble, ensuring constant particle number, pressure, and temperature. These parameters were carefully selected to accurately model the dynamic behaviour of proteins in the system.

Once the system reaches equilibrium, the MD simulation trajectories are recorded and analyzed to investigate conformational alterations. Specifically, we examined the conformational changes in the C-alpha backbone of both the B4GALNT1 wild type and its mutant variants by comparing them to their initial conformations. This comparison was conducted using Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF) plots. Additionally, we calculated changes in the side chains of the residues to provide a more comprehensive understanding of the structural dynamics.

3. Results

3.1. A precision gene panel for accurate cancer and normal sample discrimination in LIHC

Our primary goal was to uncover potential biomarkers from protein-coding genes capable of distinguishing between cancerous and normal samples. All 424 LIHC samples, featuring 19,098 protein-coding gene attributes, were utilized for training supervised machine learning classifiers, enhancing the classification of cancer and normal tissue samples. Moreover, we employed ten FS, including Fast Correlation-Based Filter Method (FCBF), Spike and Slab "spikeslab", Univariate statistical tests (F-test), Boruta, Recursive Feature Elimination (RFE), SelectFromModel with class from scikit-learn XGBoost, SVC linear, Random Forest, Extra Tree, and LASSO to determine the most significant subset of gene features. Subsequently, we extracted features that were common in at least five FS techniques. These 55 genes were also utilized as a training dataset for ten different classifiers: AB, DT, NB, KNN, LR, RF, XGB, SVC, MLP, and ET. Across both the test and validation datasets, all classifiers performed well with the 55 gene features except DT (AUC = 0.83) for the test dataset and (AUC = 0.80) for the validation dataset. The dimensionality of the LIHC gene expression dataset was reduced and subsequently subjected to Principal Component Analysis (PCA) following the application of SMOTETomek. This was done to gain deeper insights into the clustering patterns within both the original and up-sampled datasets Fig. 3 (A-B). Furthermore, a heatmap was utilized to visualize the 55 genes capable of accurately classifying LIHC and normal samples, as depicted in Fig. 3 (C).

On the test datasets, we further examined the performance of all classifiers. The Gaussian Naive Bayes and K-nearest neighbor classifier attained the highest performance, achieving an accuracy of 98.67% for the 55 gene features and an AUC of 0.99 (Fig. 3 (D) and Supplementary Table S1). Furthermore, for the validation dataset, the maximum accuracy was reached with AUCs of 0.99 for Logistic Regression, Gaussian Naive Bayes, and Support Vector Classifier, with accuracies of 99.67%, 99.00%, and 98.67%, respectively (Fig. 3 (E) and Supplementary Table S2). The differentially expressed genes (DEGs) were obtained (3408 up and 1218 downregulated) and intersected with 55 ML-gene features. These features were visualized using a volcano plot to display up and downregulated genes in Fig. 3 (F). Among the 55 feature subsets, 37 genes were upregulated, 15 genes were found to be downregulated, and three genes were found not significant (ZNF526, SH3BP5L, and PRCC) when the statistical threshold to identify differentially expressed genes (DEGs) was defined as a $|\log2\text{-fold change}| > 1$ and a p-value < 0.05. Additionally, the half violin plot demonstrates the differential expression of these 55 gene features in both cancer and normal samples

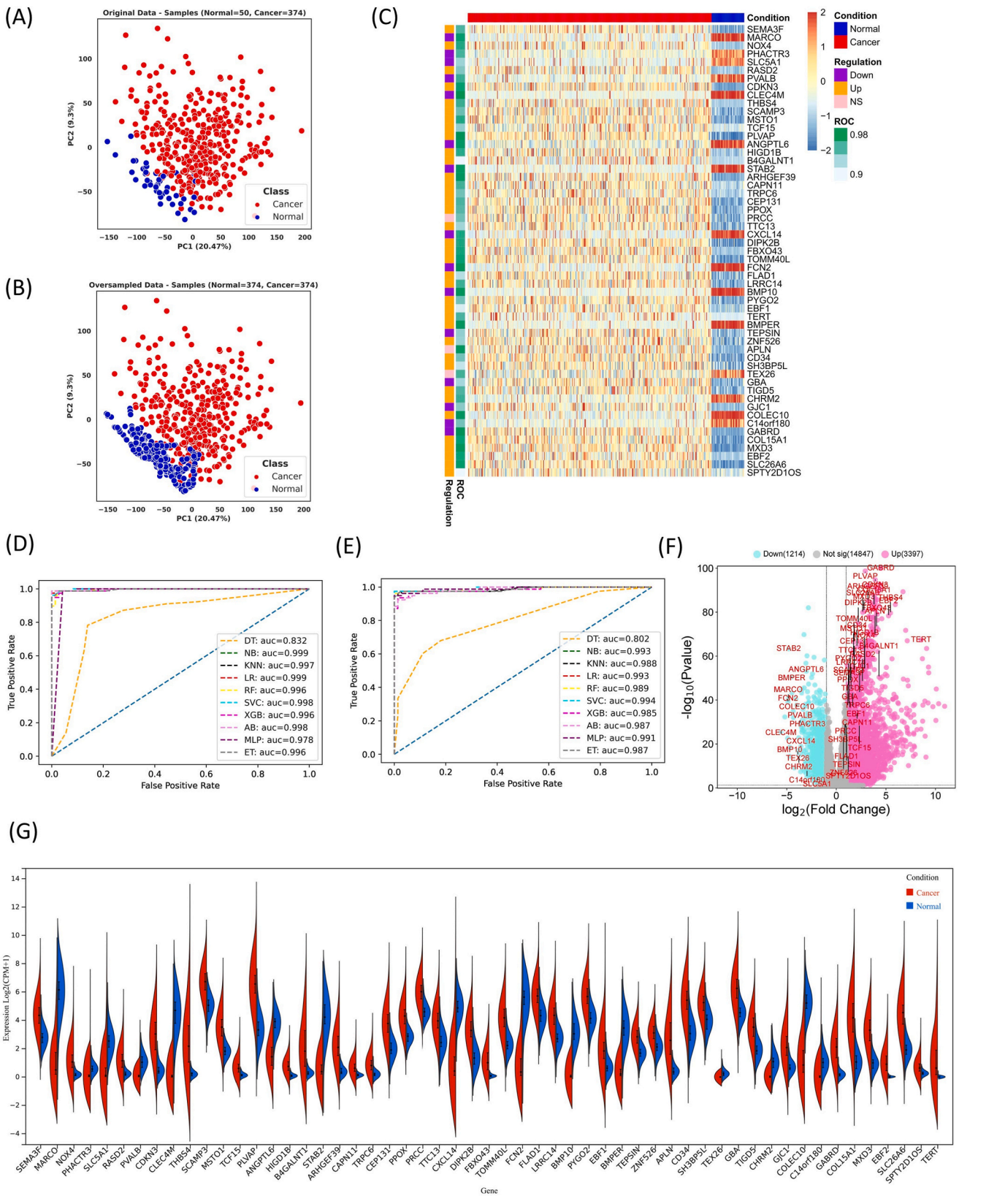


Fig. 3. PCA plot depicting the two different clusters of cancer and normal samples (A–B) before and after applying SMOTETomek. (C) The heat map shows the differential expression pattern in cancer and is normal with regulation and ROC value. (D–E) ROC curve of the 55 PC genes for test and validation datasets. (F) Volcano plot for patients with LIHC compared with the normal with a threshold of log2FC > 1 in purple are upregulated, and log2FC < 1 in cyan is downregulated. The ML-obtained genes are represented in the volcano plot. (G) Complete differential expression of ML-obtained genes is represented in the violin plot. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Fig. 3 (G). A detailed report of 55 gene features in LIHC samples, including their log2 Fold Change, AUC, p-value, padj, and Kaplan-Meier Log Rank P, is provided in Supplementary Table S3.

3.2. Protein-protein interaction and survival analysis

To study the physical interactions among 55 protein-coding genes, the STRING online webserver tool was accessed with a minimum required interaction score set at a confidence level of 0.150. We observed 132 edges, with an average node degree of 4.8. Additionally, we found that 51 gene features exhibited physical interactions, while four did not show any interaction, among which B4GALNT1 is one Fig. 4 (A). Furthermore, using lasso-penalized Cox regression on the 55 gene features, the list was reduced to only 13 genes with $\lambda = 0.04$ (SEMA3F, CDKN3, SCAMP3, PLVAP, B4GALNT1, CEP131, TOMM40L, PYGO2, APLN, GBA, COLEC10, C14orf180, SPTY2D1OS) Fig. 4 (B). These 55 genes were further used for KEGG pathway analysis, and these gene features were involved in pathways such as phagosome, mineral absorption, glycosphingolipid biosynthesis, hepatocellular carcinoma, pathways in cancer, and others, as shown in Fig. 4 (C). The KM curve represents those patients with high risk had a poor prognosis compared to low-risk patients (p-value < 0.05) Fig. 4 (D). For the individual model, a risk score was calculated for all LIHC patients based on the expression value of the 13 genes and based on the risk scores; the patients were divided into high and low-risk Fig. 4 (E). Furthermore, for the prediction of 3 years of survivability, the ROC curve was generated with its AUC value in Fig. 4 (F). For each gene, hazard ratios (HRs) and 95% confidence intervals (CIs) were analyzed in a forest plot Fig. 4 (G).

3.3. B4GALNT1 expression in other cancers

To study the expression of B4GALNT1 in various malignancies, we began our investigation with the TIMER2 webserver. We discovered that the expression of B4GALNT1 was considerably higher in 19 tumour types than in corresponding TCGA normal tissues. (breast invasive carcinomas (BRCA), cholangiocarcinoma (CHOL), head and neck squamous cell carcinoma (HNSC), colon adenocarcinoma (COAD), kidney renal clear cell carcinomas (KIRC), liver hepatocellular carcinoma (LIHC), lung squamous cell carcinomas (LUSC), lung adenocarcinoma (LUAD), prostate adenocarcinoma (PRAD), thyroid carcinoma (THCA), uterine corpus endometrial carcinoma (UCEC), stomach adenocarcinoma (STAD), rectum adenocarcinoma (READ), pheochromocytoma and paraganglioma (PCPG), mesothelioma (MESO), kidney renal papillary cell carcinoma (KIRP), kidney chromophobe (KICH), glioblastoma multiforme (GBM), esophageal carcinoma (ESCA)). Data collected from the TCGA database are shown in Fig. 5 (A). The same observable difference was noted in Fig. 5 (B–C), obtained from ULACAN and TNMplot expression plots compared to cancer and normal samples. The expression of B4GALNT1 was significantly increased in LIHC tissues compared with normal samples (p-value < 0.05) Fig. 5 (D). The survival plot obtained from KMplot showed higher expression correlating with poor survival (p-value < 0.05 and Kaplan-Meier Log-rank P $= 0.002$) Fig. 5 (E), suggesting an association between B4GALNT1 expression in LIHC and survival. When comparing B4GALNT1 expression with different cancer stages (1, 2, 3, and 4), a significant difference was observed for stages 1, 2, and 3 compared to normal samples, but no significant difference was observed with stage 4 Fig. 5 (F). The promoter methylation level of B4GALNT1 in cancer compared to normal samples showed a significant difference (Fig. 5 (G)), suggesting the role of B4GALNT1 in methylation (p-value < 0.05). Additionally, we examined the relationship between B4GALNT1 expression and different clinical aspects using UALCAN, including patients' race, age, weight, gender, histological subtypes, tumor grades, nodal status, and TP53 mutation **(Supplementary Fig. S1 (A–H))**, with significant differences marked in red.

3.4. Expression of B4GALNT1 in other datasets, survival analysis (OS, RFS, PFS, DSS) and B4GALNT1 binding proteins

The expression of B4GALNT1 was examined through the CancerLivER database for Affymetrix, Illumina, and Agilent (set 1 and set 2) and cross-checked with TCGA samples. It was found to be upregulated in all datasets. Fig. 6 (A) demonstrates the number of samples utilized for the analysis in both tumor and adjacent normal tissue, alongside the expression of B4GALNT1 across various NGS methods. We observed a difference in mean expression between cancer and normal samples in different platforms. Furthermore, B4GALNT1 survival analysis was determined using a single gene-based method in KMplot for OS, RFS, PFS, and DSS, showing poor survival when the expression is high Fig. 6 (B-E). To investigate the genes that positively and negatively correlated with B4GALNT1 in LIHC, we used LinkedOmics to identify them. The heatmap displays the top 50 positively and negatively associated genes in Fig. 6 (F). Further, we found the top five most mutated genes higher than 3% (AXIN1, DNAH10, SYNE2, TP53, and LRRI1Q) associated with B4GALNT1 (p-value < 0.05) are depicted in the boxplot of mutated and wild-type samples Fig. 6 (G). To evaluate the model's accuracy regarding B4GALNT1 in tumors, we depicted the AUC values across various TCGA tumors. The results indicate a high predictive capability for several cancer types, including CHOL (AUC $= 0.95$), ESCA (AUC $= 0.80$), GBM (AUC $= 0.87$), HNSC (AUC $= 0.85$), KIRC (AUC $= 0.86$), LIHC (AUC $= 0.90$), LUSC (AUC $= 0.93$), MESO (AUC $= 0.92$), PCPG (AUC $= 0.99$), and SARC (AUC $= 0.87$), all-surpassing an AUC threshold of 0.8 Fig. 6 (H). These results suggest that B4GALNT1 serves as a good predictor of tumor prognosis. Additionally, Fig. 6 (I) represents the ROC curve for LIHC, exhibiting an AUC of 0.90.

3.5. Analysis of high and low B4GALNT groups in LIHC, TME score, AFP, and exploring correlations with immune cells

We examined the relationship between B4GALNT1 and alpha-fetoprotein (AFP) in LIHC patients and discovered that LIHC patients with high AFP expression also exhibited higher B4GALNT1 expression Fig. 7 (A). These findings suggested that higher B4GALNT1 expression in cancer is closely associated with AFP levels in LIHC. Next, we utilized Sangerbox to explore the role of B4GALNT1 and immune genes in different cancers Fig. 7 (B). The results indicated a robust interaction of B4GALNT1 with both stimulatory and inhibitory immune genes across various TCGA cancers. Surprisingly, CD276, TGFB1, VEGFA, HAVCR2, PDCD1, and CTLA demonstrated strong correlations with most TCGA cancers, notably LIHC, with statistical significance (p-value < 0.05). Concerning stimulatory immune genes, CD70, CD27, CD28, IL1B, and CD40 exhibited significant associations with most TCGA cancers (p-value < 0.05). Additionally, in TCGA-LIHC, B4GALNT1 expression showed a positive correlation with ImmuneScore ($r = 0.17$, p-value < 0.01), ESTIMATEScore ($r = 0.14$, p-value < 0.01), and StromalScore ($r = 0.08$, p-value $= 0.13$). As ImmuneScore and ESTIMATEScore were significantly positively associated, except for StromalScore, we concluded that B4GALNT1 plays an important role in the tumor immune microenvironment in liver cancer Fig. 7 (C-E). The high-B4GALNT1 group exhibited elevated TME scores in TCGA-LIHC ImmuneScore (p-value < 0.05) Fig. 7 (F). In LIHC, we assessed the difference in immune cell infiltration proportions between the high and low B4GALNT1 groups, revealing that certain immune cell fractions were higher in the high B4GALNT1 group Fig. 7 (G). In LIHC, the correlation of B4GALNT1 with immune cells showed that it is positively correlated with B cells memory, B cells naive activated, Dendritic cells activated, T cells CD4 memory activated, and negatively correlated with NK cells activated Fig. 7 (H). In LIHC, B4GALNT1 showed the strongest correlation with B cell memory. We concluded that B4GALNT1 is associated with immune cells and plays an important role in regulating the immune microenvironment in LIHC. Subsequently, we analyzed the TIMER database to determine if B4GALNT1 expression levels in LIHC correlated with

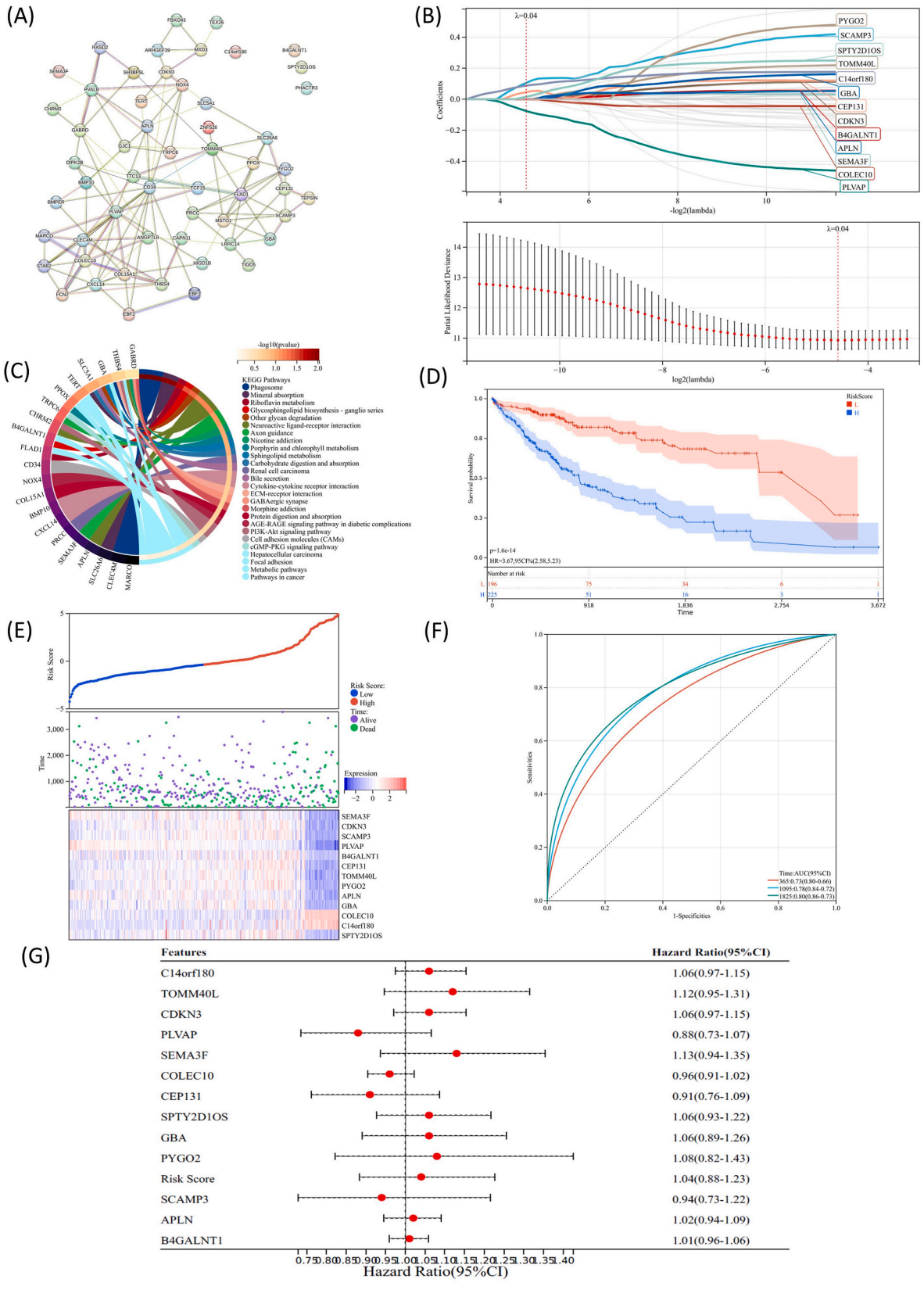


Fig. 4. PPI interaction, KEGG pathways, and prognostic value of 55 protein-coding genes. (A) Protein-protein interaction of 55 protein-coding genes. (B) Lasso-penalized Cox regression (λ = 0.04). (C) The KEGG pathways. (D)The KM-curve for high and low-risk score impacts on survival analysis. (E) The 13 genes with risk scores were stratified into high and low-risk score. (F) ROC curve (G) Hazard ratios with 95% CI.

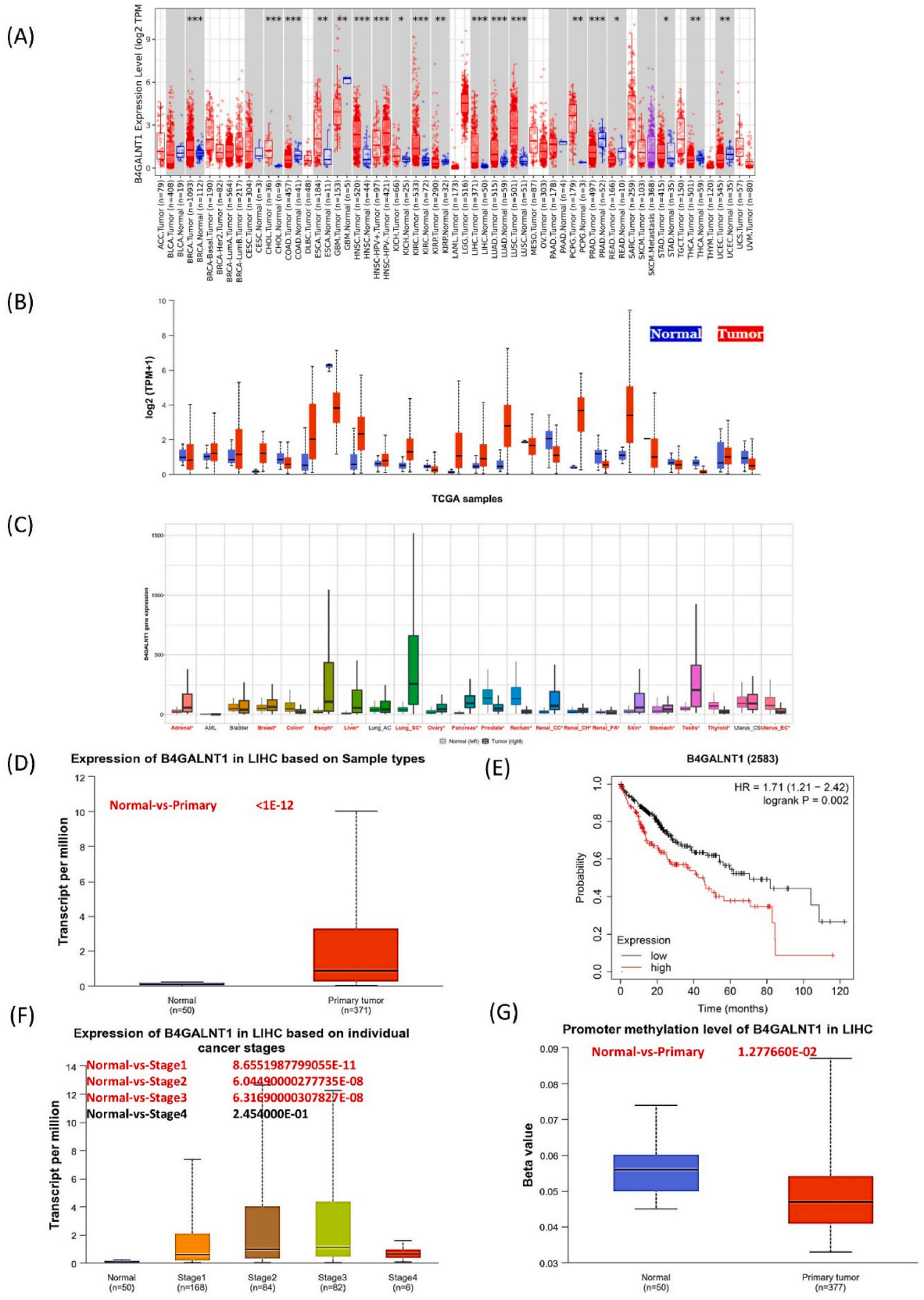


Fig. 5. The expression of B4GALNT1 in different TCGA cancers. (A–C) Comparison of B4GALNT1 expression between cancer and normal samples in pan cancer using TIMER2, ULACAN, and TNM plotter, (D) B4GALNT1 expression in LIHC (p-value < 0.05) (E) Association of B4GALNT1 in overall survival in LIHC (F) B4GALNT1 expression in LIHC in different stages of cancer, and (G) Methylation pattern of B4GALNT1 in LIHC.

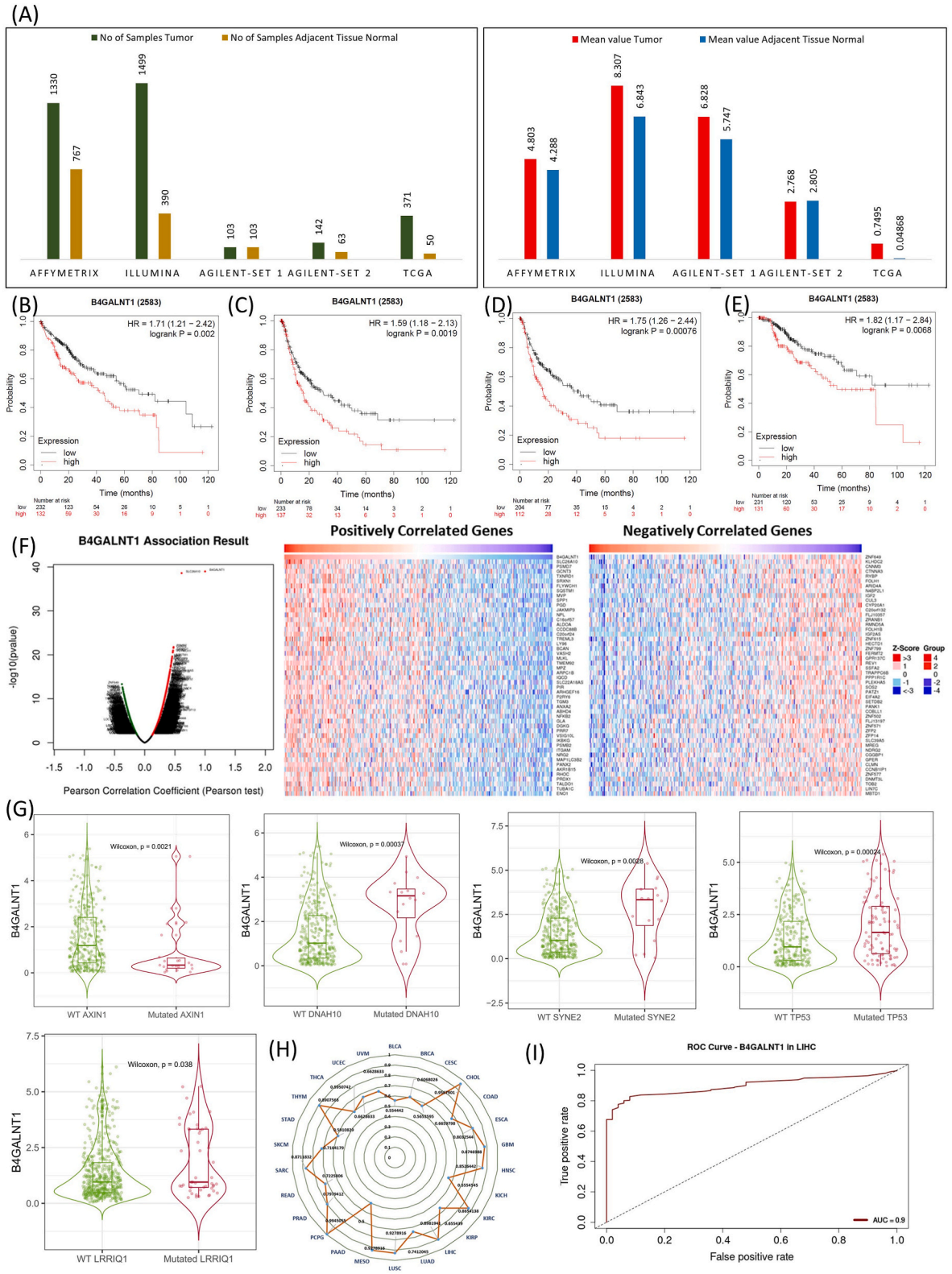


Fig. 6. Expression of B4GALNT1 in different datasets using CancerLivER. (A) Different datasets for HCC and the mean expression of B4GALNT1 in tumor and adjacent tissue normal. (B–E) Survival probability of B4GALNT1 samples in OS, RFS, PFS, and DSS. (F) Volcano plot showing correlated genes with B4GALNT1, positively and negatively 50 correlated genes. (G) Association of B4GALNT1 genes with wt AXIN1, DNAH10, SYNE2, TP53, LRRIQ1 genes and mutated genes with (p-value < 0.05). (H) The pan-cancer AUC values for B4GALNT1 across the curve. (I) ROC plot for B4GALNT1 in LIHC with AUC $= 0.9$.

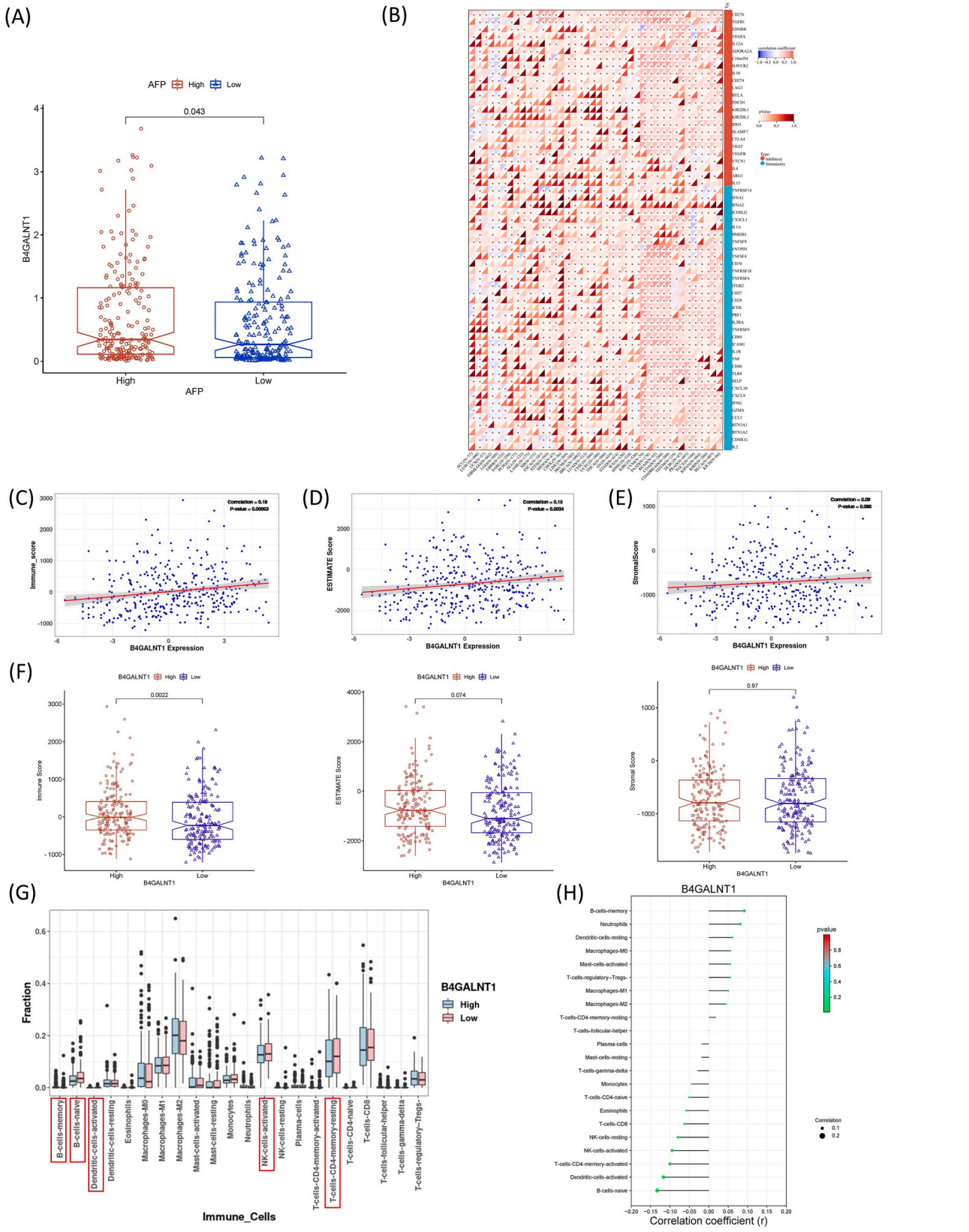


Fig. 7. Examining the correlation between B4GALNT1 and the tumor immune microenvironment and immune cells in HCC. (A) Investigating the relationship between alpha-fetoprotein (AFP-High and Low) and B4GALNT1 expression. (B) Importance of pan-cancer to immune stimulatory and inhibitory genes. (C) Correlation of B4GALNT1 with ESTIMATEScore. (D) Correlation of B4GALNT1 with ImmuneScore. (E) Correlation of B4GALNT1 with StromalScore. (F) Boxplots of TME scores for ESTIMATEScore, ImmuneScore, and StromalScore between high−B4GALNT1 and low−B4GALNT1 groups. (G) The difference in the fraction of immune cells between high−B4GALNT1 and low−B4GALNT1 groups, with a higher immune fraction observed in the high−B4GALNT1 group. (H) Exploring the correlation of B4GALNT1 with various cell types (p-value < 0.05).

immune infiltration. B4GALNT1 was positively correlated with all immune cells (B cells, macrophages, neutrophils, dendritic cells, $CD8^+$, and $CD4^+$ T cells) with p-value < 0.05 **(Supplementary Figure S2 (A)).** In LIHC (**Supplementary Figure S2 (B)**), no immune cells showed any significant survival analysis except when cumulative survival analysis was done, which in turn demonstrated that immunological infiltrates were strongly correlated (p-value < 0.05) with B4GALNT1. Finally, GISTIC 2.0 was used to characterize somatic copy number variations such as deep deletion (−2), arm-level deletion (−1), diploid/normal (0), arm-level gain (1), and strong amplification (2). In LIHC (**Supplementary Figure S2 (C)**), box plots depict the distribution of each immune subgroup at each copy number status of B4GALNT1, where macrophages showed significant results for high amplification. Different algorithms such as TIMER, EPIC, MCPCOUNTER, XCELL, TIDE, CIBERSORT, and CIBERSORT-ABS were used to investigate correlations between the infiltration level of various immune cells and the expression level of the B4GALNT1 gene in different TCGA tumors. LIHC was favorably correlated with cancer-associated fibroblasts and other immune cells (**Supplementary Figure S2 (D)**).

3.6. Methylation analysis of B4GALNT1

We employed the SMART (Shiny Methylation Analysis Resource Tool) (http://www.bioinfo-zs.com/smartapp, Retrieved September 2023) App for DNA methylation study, which is a user-friendly and simple-to-use online web server for extensively analyzing DNA

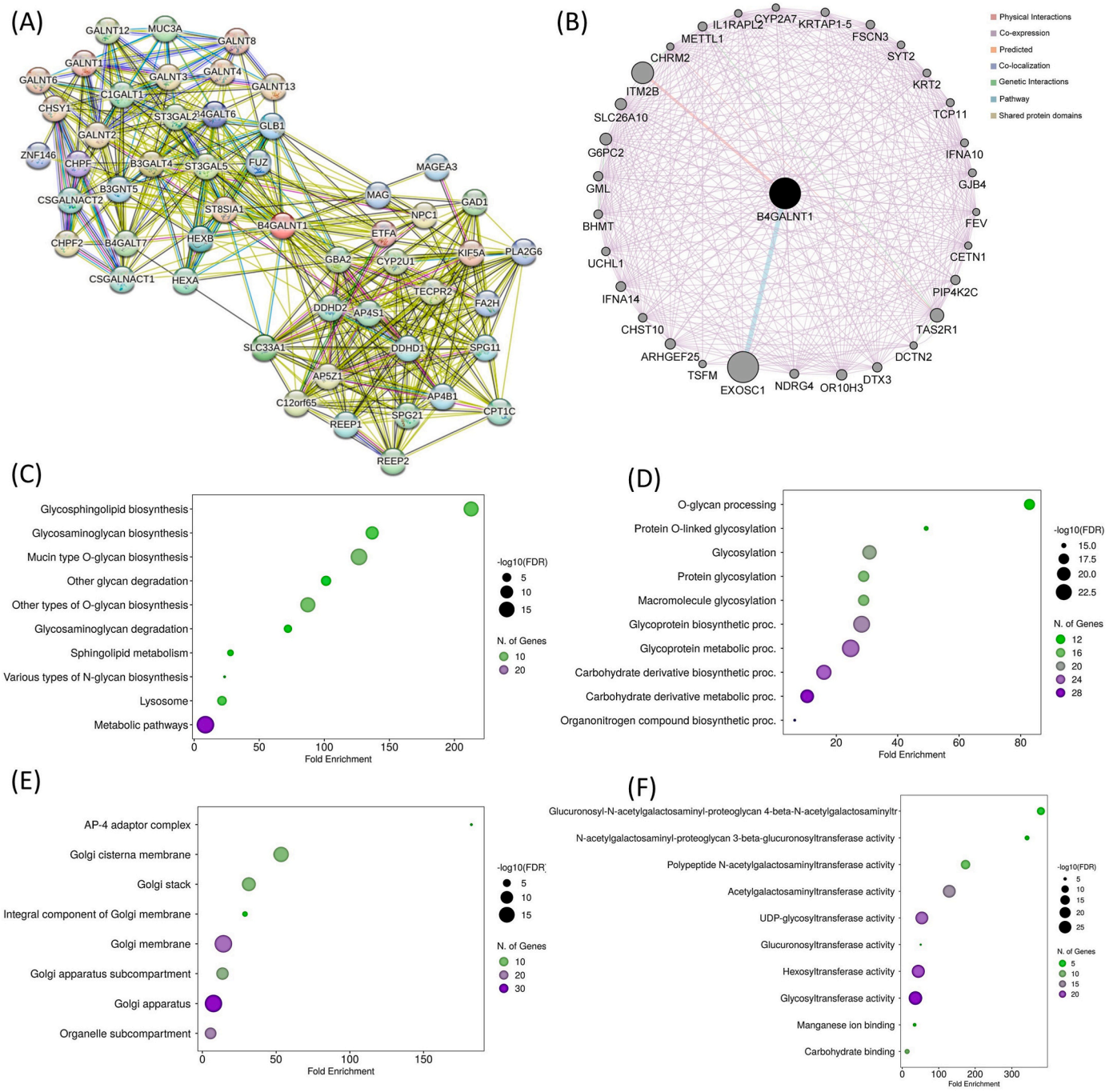


Fig. 8. B4GALNT1related genes enrichment analysis (A) Top 50 interacting proteins with B4GALNT1 PPI network, (B) Genes interacting with B4GALNT1 in tumors by PPI network from GeneMANIA (C–F) KEGG, BP, CC, MF of top 50 B4GALNT1 associated proteins.

methylation data for the B4GALNT1 in LIHC investigation, such as CpG visualization, differential methylation analysis, correlation analysis, and survival analysis [46]. The methylation pattern was observed for B4GALNT1 in LIHC for the exon, gene body, and promoter region for cancer and normal samples (**Supplementary Figure S3 (A)**). The CpG probes were further visualized in the promoter region (**Supplementary Figure S3 (B)**). The differential methylation pattern among different TCGA cancers for B4GALNT1 revealed that there is a significant difference in cancer and normal samples in 15 TCGA cancers, including BRCA, CHOL, COAD, ESCA, HNSC, KIRC, LIHC, LUAD, LUSC, PAAD, PRAD, READ, SKCM, THCA, and UCEC (p-value < 0.05) (**Supplementary Figure S3 (C)**). Among CpG probes, only 5 CpG probes (cg01723148, cg25392692, cg12230728, cg05771369, and cg21361094) showed significant differences when compared to LIHC and normal samples (**Supplementary Figure S3 (D)**), indicating these CpG probes might play important roles in methylation. These probes were further analyzed for stage-wise difference, correlation, and survival analysis (**Supplementary Fig. S4 (A–C)**).

3.7. B4GALNT1-interacting proteins and enrichment analysis

We utilized the STRING tool to investigate 50 interacting proteins with B4GALNT1 expression to gain insights into how the B4GALNT1 gene influences various biological processes. The protein-protein interaction network of proteins interacting with B4GALNT1 is depicted in Fig. 8 (A-B), utilizing STRING and GeneMANIA. Furthermore, using these 50 interacting genes, we examined KEGG pathways, biological processes, cellular components, and molecular functions. KEGG data indicated significant involvement in pathways such as glycosphingolipid biosynthesis, glycosaminoglycan biosynthesis, and other glycan-related pathways. GO enrichment analysis revealed that most of these genes are associated with O-glycan processing, glycosylation, protein glycosylation, and other processes. These GO and KEGG pathways were visualized using a bubble plot in Fig. 8 (C–F).

3.8. Gene alteration analysis

The genetic alteration analysis of B4GALNT1 was performed using cBioPortal. We observed the genetic alteration of B4GALNT1 in LIHC with the maximum mutational frequency reaching 2.4% Fig. 9 (A).

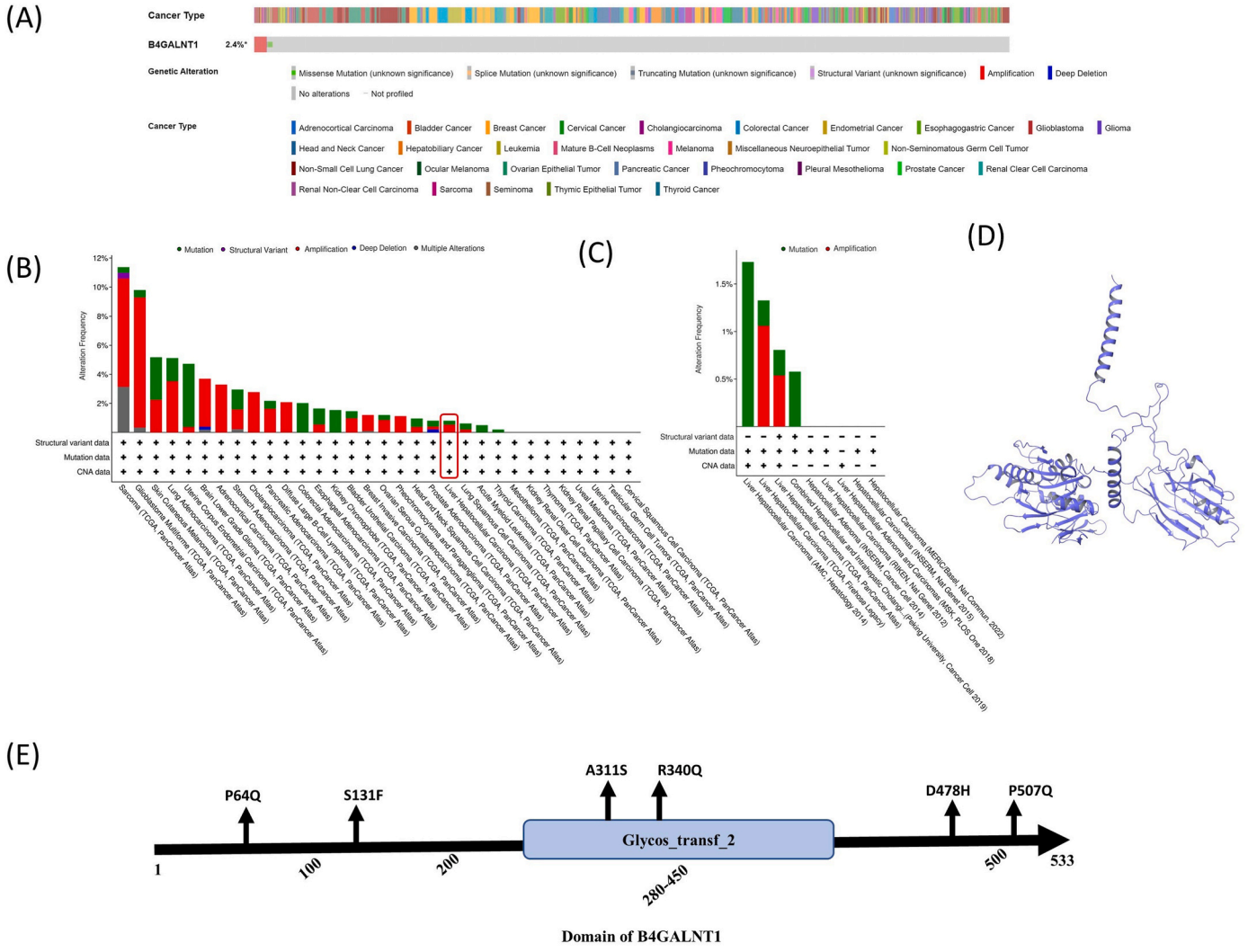


Fig. 9. Mutational landscape of B4GALNT1 in different cancers (A) The frequency of B4GALNT1-genetic alterations in different cancer samples. (B) Copy number alteration of B4GALNT1 across different cancers. Green denotes mutations, purple signifies structural variants, red indicates amplifications, blue represents deep deletions, and grey reflects multiple alterations. (C) B4GALNT1 mutational level in Liver hepatocellular carcinoma (D) Predicted 3D structure of Alpha fold of B4GALNT1 (E) Mutations of B4GALNT1 in liver cancer profile. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Sarcoma exhibited the highest alteration frequency of B4GALNT1, around 11%. The mutation landscape in different cancers for B4GALNT1 was depicted in Fig. 9 (B), along with the frequency of mutations observed in liver cancers in Fig. 9 (C). The alpha fold structure available for B4GALNT1 was utilized for structural analysis with mutations in liver cancers. Fig. 9 (D) represents the 3D structure of B4GALNT1 from the alpha-fold model accessed from UniProt. Furthermore, missense mutations in Fig. 9 (E) identified from cBioPortal, Cosmic Database, and Human Variant Database were filtered out from 16 different mutations, which were found to be affected when analyzed using *in-silico* approaches such as SIFT, Polyphen2, and CADD-Phred score along with I-Mutant, Mutpred2 score, and FATHMM methods, resulting in 6 mutations (P64Q, S131F, A311S, R340Q, D478H, and P507Q)—the Supplementary Table S4, which represents the outcomes of computational impact analysis. The table demonstrates how single nucleotide variants (SNVs) affect protein function, as assessed by various *in-silico* tools. Specifically, the analysis incorporates SIFT to determine whether SNVs are tolerated or affect protein function, Polyphen2 to classify them as benign or damaging, and CADD_Phred score, which identifies substitutions among the most deleterious changes to the human genome. A score of 10 or higher indicates the top 10% most deleterious changes, while a score of 20 or higher signifies the top 1% [47–49]. Additionally, I-mutant is employed to assess structural stability, where a negative change in free energy (DDG < 0) indicates decreased stability, whereas a positive change (DDG > 0) indicates increased stability [50]. MutPred2 scores, ranging from 0 to 1, are interpreted such that scores closer to 0 are considered benign, while those closer to 1 are pathogenic, with a mean cutoff of 0.5 [51]. Furthermore, FATHMM is utilized to predict the oncogenic status based on scores indicating benign or oncogenic traits [52]. Moreover, patients with B4GALNT1 amplification exhibited considerably poorer progression-free survival than those with no alteration, as shown in Supplementary Fig. S5 (A-D).

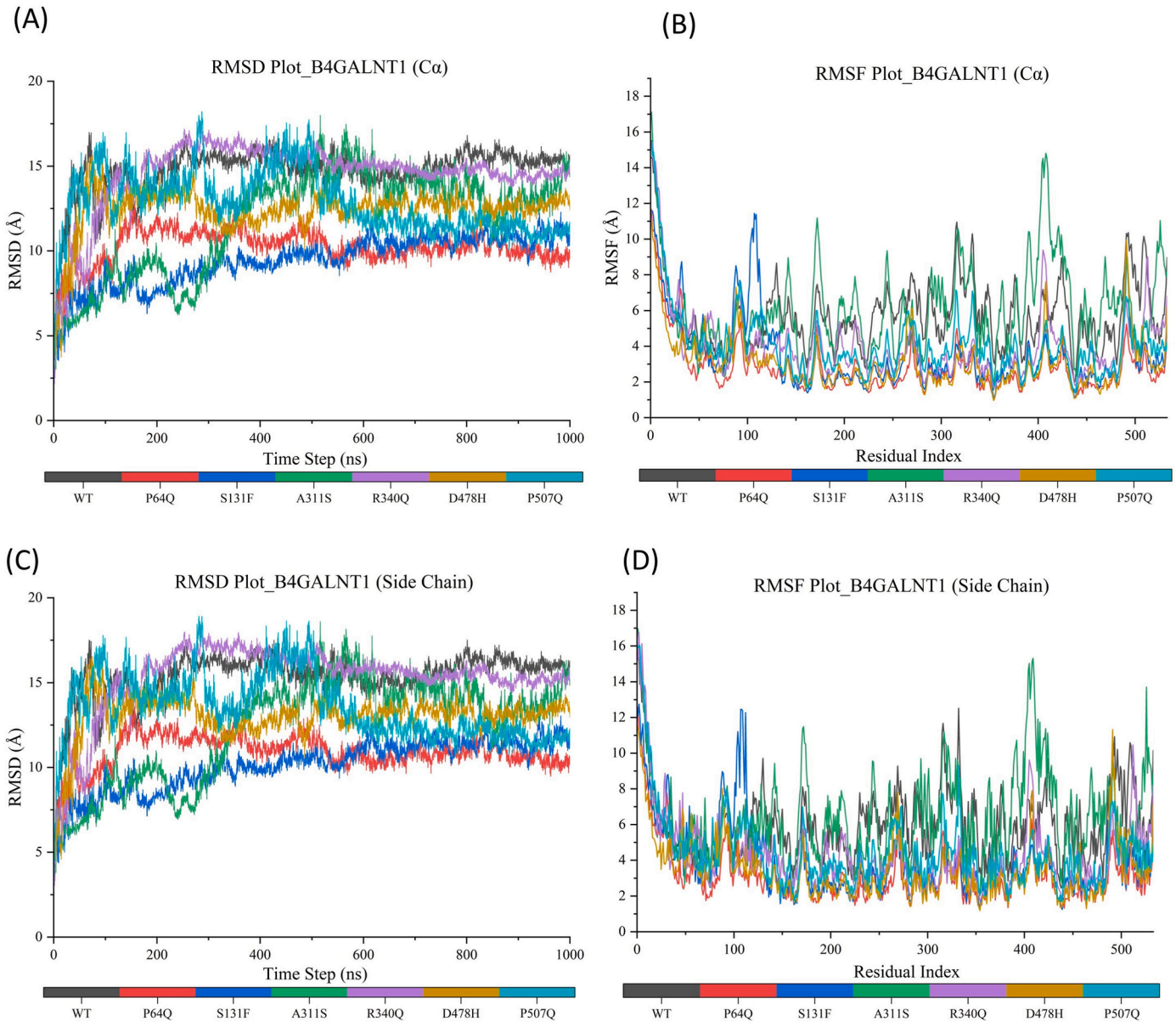


Fig. 10. MD simulation study of B4GALNT1 WT and different mutation sites (A) Root-Mean-Square Deviation (RMSD) Analysis of B4GALNT1 and Its Variants (C-α atoms) during the 1μs of MD simulation. (B) Root-Mean-Square Fluctuation (RMSF) Analysis of B4GALNT1 and Its Variants (C-α atoms) during the 1μs of MD simulation. (C) Root-Mean-Square Deviation (RMSD) Analysis of B4GALNT1 and Its Variants (Side Chain atoms) during the 1μs of MD simulation, and (D) Root-Mean-Square Fluctuation (RMSF) Analysis of B4GALNT1 and Its Variants (Side Chain atoms) during the 1μs of MD simulation.

3.9. B4GALNT1 structural analysis upon mutation

RMSD: Through extensive all-atom molecular dynamics simulations conducted on the B4GALNT1 wild type and several of its genetic variants, the RMSDs were computed. These RMSD values were determined by aligning the molecular trajectories with their respective initial conformations, which served as the input for the MD simulations. Based on the RMSD plots for both C-α atoms Fig. 10 (A) and side-chain atoms Fig. 10 (C) of B4GALNT1, it is evident that the various B4GALNT1 variants exhibit greater conformational stability when compared to the wild-type B4GALNT1 structure. These plots indicate that the structural deviations in the variants are significantly minimized, highlighting the enhanced structural stability inherent in these genetic variants in contrast to the wild-type B4GALNT1 structure. Supplementary Table S6 summarizes the average RMSDs for the B4GALNT1 wild type and its various variants. The RMSD data demonstrates that mutations at specific amino acid positions within B4GALNT1 significantly impact the protein's structural stability, surpassing that of the wild type. This observation suggests that these genetic mutations lead to a more stable conformation in the B4GALNT1 protein, underlining the structural consequences of these specific amino acid alterations.

The Wild Type B4GALNT1 has an RMSD of 14.705 Å with a standard deviation of approximately 1.50 Å, indicating that the structure of the protein exhibits some degree of deviation during the simulation. When comparing this to the B4GALNT1 variants, several noteworthy observations can be made. The P64Q variant shows the lowest RMSD at 10.241 Å, suggesting that this mutation results in a more stable protein structure. Similarly, the S131F variant exhibits a low RMSD of 9.417 Å, indicating a high degree of structural stability. However, variants A311S, D478H, R340Q, and P507Q all show RMSD values higher than the Wild Type, with notable differences in RMSD for A311S (11.752 Å) and R340Q (14.53 Å).

Fig. 10 (A) in the analysis indicates that the B4GALNT1 protein with the P64Q variant exhibits a distinctive structural behaviour. It appears to adopt two convergent equilibrium states during the course of the MD simulation. The first equilibrium phase is noticeable within the time frame of 170 ns–500 ns. Subsequently, a second equilibration phase begins at 550 ns and persists throughout the production run, which extends to 1 μs. This second phase is characterized by lower structural deviation, as evidenced by the averaged C-α Root-Mean-Square Deviation (RMSD) of 10.241 ± 1.01 Å. Furthermore, a noteworthy observation in the analysis is that after the initial 500 ns of the molecular dynamics (MD) simulation, all the B4GALNT1 variants, including the wild type, demonstrate a similar level of conformational stability. The structural behaviour appears to converge among the variants during this phase. Particularly, the B4GALNT1 R340Q variant exhibits a thermodynamic behaviour resembling that of the wild type B4GALNT1, with no significant changes in its structural characteristics observed. This suggests that, after an initial period, the R340Q variant returns to a conformational state comparable to the wild type, indicating that this specific mutation does not induce substantial alterations in the protein's stability. These findings suggest that specific mutations in the B4GALNT1 gene can have a significant impact on the protein's structural integrity, which can, in turn, have implications for its function and interactions within biological systems.

RMSF: Supplementary Table S6 also provides RMSF data for different variants of the B4GALNT1 protein, specifically focusing on the fluctuations in the C-α atoms and side chain atoms. RMSF is a metric used to measure the variability or flexibility of individual atoms within a protein's structure. The RMSF data reveals interesting insights into the dynamic behaviour of these B4GALNT1 variants. The wild-type B4GALNT1 exhibits a C-α RMSF of 5.72 Å ± 1.91 Å, indicating a moderate level of atom fluctuation within the protein structure. Comparing this to the variants, it is clear that several of them display reduced RMSF values, signifying decreased atomic fluctuations and increased structural stability. The P64Q variant, for instance, demonstrates the lowest C-α RMSF at 2.831 Å, while the S131F variant shows a C-α RMSF of 3.535 Å. These lower RMSF values in P64Q and S131F suggest that these mutations lead to a more rigid protein structure with less atomic movement.

On the other hand, the A311S, D478H, R340Q, and P507Q variants have higher RMSF values compared to the Wild Type. Among them, the A311S variant has the highest RMSF at 6.099 Å. These higher RMSF values indicate increased atomic fluctuations and potentially a less stable structure in these variants. Comparing C-α RMSF to side chain RMSF values, it's evident that in most cases, the C-α RMSF values are relatively consistent with the side chain RMSF values. This suggests that the fluctuations in the backbone atoms (C-α) are in line with the fluctuations in the side chain atoms, indicating that structural stability or flexibility affects the entire protein structure consistently.

The RMSF data highlights variations in the flexibility and stability of the B4GALNT1 variants. The P64Q and S131F variants exhibit reduced atomic fluctuations Fig. 10 (B) and 10 (D), suggesting enhanced structural stability. In contrast, the A311S, D478H, R340Q, and P507Q variants display increased atomic fluctuations, indicating a higher degree of structural flexibility. These findings shed light on the impact of specific mutations on the dynamic behaviour of the B4GALNT1 protein.

Supplementary Table S5 presents a comparison between the RMSF values of both the C-α atoms and the side chain atoms for the wild-type B4GALNT1 and the respective mutant residues. The RMSF comparison of single mutated residues with their corresponding wild-type residues provides valuable insights into the impact of these mutations on protein flexibility. Variants like P64Q, S131F, and R340Q exhibit reduced RMSF values, indicating enhanced structural stability. On the other hand, A311S and P507Q show increased RMSF values, suggesting greater flexibility and potential instability. D478H demonstrates lower RMSF values, pointing toward increased stability. These findings help in understanding the structural consequences of specific mutations in the B4GALNT1 protein.

4. Discussion

There is a significant morbidity and mortality associated with liver hepatocellular carcinoma (LIHC). Therefore, improving the clinical diagnosis and treatment of LIHC is important. Several additional studies have reported using gene expression data to find LIHC biomarkers [53]. The main disadvantage is the lack of a large sample size in publicly available datasets. As a result, we sought to solve the issue of utilizing the same set of samples for machine learning classification by up-sampling TCGA-LIHC data using SmoteTomek. Machine learning techniques emerge as a better method for unraveling deep patterns in gene expression datasets to uncover possible new biomarkers [54–56]. However, a large number of protein-coding genes compared to a smaller subset of samples is a barrier to improving ML classifier performance. To overcome this constraint, we used a variety of feature selection algorithms. The reason for using different feature selection methods is twofold: first, it aids in identifying the core representative and functional genes that can potentially distinguish between cancer and normal samples, and second, one feature selection method can provide a set of genes that may differ from another feature selection technique. We used a polling approach to identify the best features shared by at least five feature selection methods. The polling-based method yielded genes, the majority of which are known to be associated with LIHC.

Here, we evaluated the performance of ten ML-based classification algorithms: Random Forest (RF), K-Nearest Neighbors (KNN), Decision Tree (DT), Support Vector Classifier (SVC), XGBoost (XGB), Naive Bayes (NB), AdaBoost (AB), Multi-layer Perceptron (MLP), ExtraTrees (ET), and Logistic Regression (LR). This study used these algorithms to develop binary classification models for the classification of LIHC and normal. The analysis found that ML algorithms could more accurately classify LIHC. Almost every ML algorithm improved sensitivity and specificity. As indicated in the results section, Gaussian Naive Bayes

Classifiers outperform all other classifiers in terms of accuracy and AUC for the whole dataset trained with the 55 features in Fig. 3 (D). Several genes have established roles in LIHC out of the 55 gene features given in the results section, validated using the above-mentioned pipeline. Differential gene expression (DGE) analysis was performed on the complete dataset, and 51 PC genes were identified to be differentially expressed from ML-identified genes Fig. 3 (F). After that, using lasso-penalized Cox regression analysis, these 55 PC genes were reduced to 13 genes. B4GALNT1, CEP131, PYGO2, SEMA3F, COLEC10, APLN, TOMM40L, SCAMP3, PLVAP and GBA have been reported in the literature, while CAPN11, SPTY2D1OS, and C14orf180 have not been studied in LIHC so far. These 13 genes from survival analysis are also predicted to be prognostic genes for OS in LIHC cases. Among them, COLEC10 and C14orf180 were downregulated in our study, while other genes were upregulated.

CEP131, also known as centrosomal protein 131, has oncogenic activity in different cancers, including liver. CEP131 has demonstrated oncogenic potential by activating the PI3K/AKT signaling pathway and could serve as a promising prognostic biomarker in HCC [57]. PYGO2, also known as Pygopus Family PHD Finger 2, and the overexpression of PYGO2 are found in cancers such as ovarian, lung, breast, and liver [58]. One of the studies reported that knockdown of EFNA4 suppressed cell proliferation, invasion, angiogenesis, and Wnt/β-catenin signaling in hepatocellular carcinoma (HCC) by lowering PYGO2 expression [59]. SEMA3F belongs to the semaphorin protein family and is known for its activities in cell migration, axon guidance, and angiogenesis. SEMA3F, in particular, has been investigated for its possible role in cancer. It is classified as a tumor suppressor gene in several cancers. It promotes hepatocellular carcinoma metastasis in the liver by activating the focal adhesion pathway [60]. COLEC10, also known as Collectin Subfamily Member 10, encodes a protein-coding gene, and the decreased expression level of this gene has been studied as poor OS in liver cancer patients [61]. Apelin, which is APLN, was found to be upregulated and promotes cancer in the liver through the PI3/Akt signaling pathway [62]. TOMM40L (Translocase of Outer Mitochondrial Membrane 40 Like) is a mitochondrion-related gene. The role of this gene in LIHC has not been well studied, but the function is reported in cellular homeostasis, apoptosis, lipid synthesis, and energy metabolism. However, a study found through qRT-PCR analysis that TOMM40L is notably upregulated in hepatocellular carcinoma [63]. SCAMP3, a membrane-trafficking protein of the secretory carrier membrane proteins (SCAMPs) family, is involved in endosome transport. A higher expression level indicates poor progression in hepatocellular carcinoma, and our study also found that SCAMP3 is upregulated in LIHC [64]. In the liver, by decreasing E-cadherin expression, it promotes metastasis and invasion in liver cancer. Plasmalemmal Vesicle Associated Protein (PLVAP) is found to be a potential therapeutic target for the treatment of HCC [65]. One study identified GBA as a direct target of Artesunate (ART), indicating its potential to effectively suppress lysosomal autophagy as a therapeutic strategy for HCC. Interestingly, our study also found that the expression of GBA was upregulated [66].

Additionally, B4GALNT1 emerged among the 13 genes that exhibited upregulation in our study. It plays an important role in the development and normal function of the central nervous system. The major KEGG pathways associated with B4GALNT1 and its interacting proteins were predominantly glycosphingolipid biosynthesis, glycosaminoglycan biosynthesis, mucin-type O-glycan biosynthesis, glycosaminoglycan degradation, sphingolipid metabolism, and diverse types of N-glycan biosynthesis Fig. 8 (C). Furthermore, recent studies have examined the involvement of B4GALNT1 expression in the progression and prognosis of liver hepatocellular carcinoma (LIHC). Hence, a comprehensive and extensive structural analysis of B4GALNT1 was conducted in the present study. Our analysis revealed a significant increase in the expression of B4GALNT1 in LIHC, as determined by analysing the Timer, UALCAN, and HCCDB databases. Moreover, while examining subgroups according to stages, tumour grade, age, gender, race, and weight, it was observed that the expression level of B4GALNT1 was significantly elevated in patients with liver hepatocellular carcinoma (LIHC) compared normal individuals. Through long-scale molecular dynamics simulations, we have investigated the dynamic behavior of B4GALNT1 and its mutants. The RMSD analysis reveals that genetic variants generally exhibit increased conformational stability compared to the wild type, with mutations like P64Q The RMSD analysis reveals that genetic variants generally exhibit increased conformational stability compared to the wild type, with mutations like P64Q and S131F displaying notably lower RMSD values. The unique structural behaviour of the P64Q mutant, adopting two dynamic equilibrium states, and the convergence of stability among mutants after an initial period. The RMSF data further underscores the impact of mutations on flexibility, with P64Q and S131F showing reduced atomic fluctuations, indicative of enhanced structural stability, while A311S, D478H, R340Q, and P507Q exhibit increased fluctuations, suggesting greater flexibility. MD simulation results collectively contribute to a comprehensive understanding of how specific mutations influence the dynamic behaviour of B4GALNT1 protein, shedding light on potential functional implications within biological systems.

Our findings proved the feasibility of employing a polling-based feature selection technique with machine learning to identify biomarkers that discriminate cancer samples from normal samples.

5. Conclusion

Despite notable progress in cancer research, early detection of Hepatocellular carcinoma remains a challenge. An important challenge is to discover specific genes that can differentiate between cancer and normal samples by analysing RNA-seq data. The considerable number of genes found in these analyses poses a significant challenge in precisely isolating disease associated genes. Hence, integrating an ensemble of feature selection (FS) techniques has demonstrated robustness and reliability in identifying precise and biologically relevant biomarker genes in LIHC. Our approach incorporates filter, wrapper, and embedded FS methods, showcasing a forward-looking perspective on the crucial integration of machine learning. Through our approach, we identified 55 gene features across at least five FS methods, effectively discriminating between cancer and normal samples with higher accuracy. We have identified promising biomarkers by analyzing transcriptomic data from LIHC, achieving higher performance by employing upsampling techniques for the minority class of data. The initial dataset was highly imbalanced, posing a risk of biased results. Moreover, our study primarily utilized TCGA high-throughput datasets for both internal testing and validation. These potential biomarkers require further validation on external datasets from various studies to assess their clinical utility. Moreover, differential expression analysis also confirmed the dysregulation of these 55 feature genes in LIHC. This gene set holds promise for targeted sequencing, potentially restricting sequencing costs and aiding in cancer type identification, thus facilitating cancer progression study and personalized treatment.

Furthermore, from the initial 55 gene features, a subset of 13 gene features was reduced using lasso-penalized Cox regression. The log-rank test indicated that high-risk patients exhibited poorer prognosis than those with low-risk scores (p-value < 0.05), as depicted by Kaplan-Meier Curve analysis. Notably, the identified genes are predominantly involved in pathways such as phagosome, mineral absorption, glycosphingolipid biosynthesis, bile secretion, and PI3K-Akt signaling pathway. Molecular dynamics simulations revealed notable differences in the structural stability and flexibility of B4GALNT1 caused by specific mutations. Notably, variants such as P64Q and S131F displayed increased stability, while mutations like A311S and P507Q led to heightened flexibility. These findings strengthen our understanding of how genetic alterations affect the structural dynamics of B4GALNT1, providing valuable insights into its functional significance and broader implications in biological systems.

Employing various bioinformatics approaches, we uncovered novel signatures that enhance our understanding of diseases by elucidating molecular mechanisms, potentially informing therapeutic strategies for improved cancer management. Additionally, our results highlight B4GALNT1 among the 55 gene features, suggesting its role in cancer progression and proposing it as a potential therapeutic target in LIHC. However, thorough research and clinical investigations are imperative to validate the roles of other gene features in LIHC. The study has demonstrated the strength of machine learning to construct modelsfor classifying cancer and normal samples using transcriptomic data, indicating the potential applicability of our approach to other complex diseases for identifying novel biomarkers.

Availability of data and materials

In conducting this study, we have used mainly datasets from the publicly available TCGA projects https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga and all the available online bioinformatics tools outlined in the materials and methods.

Ethical statement

The results of this study are original and have not been previously published, are not under review elsewhere, and have received approval from all authors. All the authors have approved final version of the manuscript.

CRediT authorship contribution statement

Rohit Kumar Verma: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation. **Kiran Bharat Lokhande:** Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Formal analysis. **Prashant Kumar Srivastava:** Writing – review & editing, Validation, Supervision, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Ashutosh Singh:** Writing – review & editing, Validation, Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

Authors declared no competing interest.

Acknowledgments

The researchers extend their gratitude to the Shiv Nadar Institution of Eminence (SNIoE) for furnishing the essential research framework and infrastructure vital to the successful execution of this study. The authors would like to express their gratitude to the SNIoE for providing access to the high-performance computing resource (HPC-magus 02). Dr. Kiran Bharat Lokhande acknowledges SNIoE for the Post-Doctoral Fellowship. One image was created using BioRender.com.

Abbreviations

Aa	Amino acid
ACC	Adrenocortical carcinoma
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
GBM	Glioblastoma multiforme
HNSC	Head and Neck squamous cell carcinoma
KICH	Kidney Chromophobe
KIRC	Kidney renal clear cell carcinoma
KIRP	Kidney renal papillary cell carcinoma
LAML	Acute Myeloid Leukemia
LGG	Brain Lower Grade Glioma
LUAD	Lung adenocarcinoma
LUSC	Lung squamous cell carcinoma
OV	Ovarian serous cystadenocarcinoma
PCPG	Pheochromocytoma and Paraganglioma
PRAD	Prostate adenocarcinoma
SARC	Sarcoma
SKCM	Skin Cutaneous Melanoma
TGCT	Testicular Germ Cell Tumors
THCA	Thyroid carcinoma
THYM	Thymoma
UCEC	Uterine Corpus Endometrial Carcinoma
UCS	Uterine Carcinosarcoma
CHOL	Cholangiocarcinoma
COAD	Colon adenocarcinoma
ESCA	Esophageal carcinoma
LIHC	Liver hepatocellular carcinoma
PAAD	Pancreatic adenocarcinoma
READ	Rectum adenocarcinoma
STAD	Stomach adenocarcinoma
GI	Gastrointestinal
BP	Biological process
CC	Cellular component
MF	Molecular Function
GEPIA	Gene expression profiling interactive analysis
GO	Gene ontology
KEGG	Kyoto encyclopedia of genes and genomes
TIMER	Tumor immune estimation resource

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2024.101514.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J Clin 2021;71(3):209–49. https://doi.org/10.3322/caac.21660.

[2] Singal AG, Lampertico P, Nahon P. Epidemiology and surveillance for hepatocellular carcinoma: new trends. J Hepatol 2020;72(2):250–61. https://doi.org/10.1016/j.jhep.2019.08.025.

[3] Brancato V, Garbino N, Salvatore M, Cavaliere C. MRI-based radiomic features help identify lesions and predict histopathological grade of hepatocellular carcinoma. Diagnostics 2022;12(5):1085. https://doi.org/10.3390/diagnostics12051085.

[4] Agarwal PD, Lucey MR. Management of hepatocellular carcinoma recurrence after liver transplantation. Ann Hepatol 2022;27(1):100654. https://doi.org/10.1016/j.aohep.2021.100654.

[5] Lee CW, Tsai HI, Lee WC, Huang SW, Lin CY, Hsieh YC, Kuo T, Chen CW, Yu MC. Normal alpha-fetoprotein hepatocellular carcinoma: are they really normal? J Clin Med 2019;8(10):1736. https://doi.org/10.3390/jcm8101736.

[6] Ho DW, Lo RC, Chan LK, Ng IO. Molecular pathogenesis of hepatocellular carcinoma. Liver Cancer 2016;5(4):290–302. https://doi.org/10.1159/000449340.

[7] Wang L, Wang Y, Chang Q. Feature selection methods for big data bioinformatics: a survey from the search perspective. Methods 2016;111:21–31. https://doi.org/10.1016/j.ymeth.2016.08.014.

[8] Wang Y, Gao X, Ru X, Sun P, Wang J. A hybrid feature selection algorithm and its application in bioinformatics. PeerJ. Computer science 2022;8:e933. https://doi.org/10.7717/peerj-cs.933.

[9] Kaur H, Bhalla S, Raghava GPS. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. PLoS One 2019;14(9):e0221476. https://doi.org/10.1371/journal.pone.0221476.

[10] Li Y, Mansmann U, Du S, Hornung R. Benchmark study of feature selection strategies for multi-omics data. BMC Bioinf 2022;23(1):412. https://doi.org/10.1186/s12859-022-04962-x.

[11] Liu W, Chen Y, Yang J, Guo M, Wang L. B4GALNT1 promotes carcinogenesis by regulating epithelial-mesenchymal transition in hepatocellular carcinoma based on pan-cancer analysis. J Gene Med 2023;25(11):e3552. https://doi.org/10.1002/jgm.3552.

[12] Groux-Degroote S, Guérardel Y, Delannoy P. Gangliosides: structures, biosynthesis, analysis, and roles in cancer. Chembiochem : a European journal of chemical biology 2017;18(13):1146–54. https://doi.org/10.1002/cbic.201600705.

[13] Yoshida H, Koodie L, Jacobsen K, Hanzawa K, Miyamoto Y, Yamamoto M. B4GALNT1 induces angiogenesis, anchorage independence growth and motility, and promotes tumorigenesis in melanoma by induction of ganglioside GM2/GD2. Sci Rep 2020;10(1):1199. https://doi.org/10.1038/s41598-019-57130-2.

[14] Yang H, Li W, Lv Y, Fan Q, Mao X, Long T, Xie L, Dong C, Yang R, Zhang H. Exploring the mechanism of clear cell renal cell carcinoma metastasis and key genes based on multi-tool joint analysis. Gene 2019;720:144103. https://doi.org/10.1016/j.gene.2019.144103.

[15] Che MI, Huang J, Hung JS, Lin YC, Huang MJ, Lai HS, Hsu WM, Liang JT, Huang MC. β1, 4-N-acetylgalactosaminyltransferase III modulates cancer stemness through EGFR signaling pathway in colon cancer cells. Oncotarget 2014;5(11):3673–84. https://doi.org/10.18632/oncotarget.1981.

[16] Jiang T, Wu H, Lin M, Yin J, Tan L, Ruan Y, Feng M. B4GALNT1 promotes progression and metastasis in lung adenocarcinoma through JNK/c-Jun/Slug pathway. Carcinogenesis 2021;42(4):621–30. https://doi.org/10.1093/carcin/bgaa141.

[17] Jing S, Deng Z, Liang L, Liang J. B4GALNT1 enhances cell proliferation and growth in oral squamous cell carcinoma via p38 and JNK MAPK pathway. Transl Cancer Res 2020;9(4):2340–8. https://doi.org/10.21037/tcr.2020.03.73.

[18] Liang YJ, Ding Y, Levery SB, Lobaton M, Handa K, Hakomori SI. Differential expression profiles of glycosphingolipids in human breast cancer stem cells vs. cancer non-stem cells. Proc Natl Acad Sci USA 2013;110(13):4968–73. https://doi.org/10.1073/pnas.1302825110.

[19] Danolic D, Heffer M, Wagner J, Skrlec I, Alvir I, Mamic I, Susnjar L, Banovic M, Danolić D, Puljiz M. Role of ganglioside biosynthesis genetic polymorphism in cervical cancer development. J Obstet Gynaecol : the journal of the Institute of Obstetrics and Gynaecology 2020;40(8):1127–32. https://doi.org/10.1080/01443615.2019.1692801.

[20] Dad R, Malik U, Javed A, Minassian BA, Hassan MJ. Structural annotation of beta-1,4-N-acetyl galactosaminyltransferase 1 (B4GALNT1) causing hereditary spastic paraplegia 26. Gene 2017;626:258–63. https://doi.org/10.1016/j.gene.2017.05.041.

[21] Wu G, Lu ZH, Seo JH, Alselehdar SK, DeFrees S, Ledeen RW. Mice deficient in GM1 manifest both motor and non-motor symptoms of Parkinson's disease; successful treatment with synthetic GM1 ganglioside. Exp Neurol 2020;329:113284. https://doi.org/10.1016/j.expneurol.2020.113284.

[22] Hong JM, Jeon H, Choi YC, Cho H, Hong YB, Park HJ. A compound heterozygous pathogenic variant in B4GALNT1 is associated with axonal charcot-marie-tooth disease. J Clin Neurol 2021;17(4):534–40. https://doi.org/10.3988/jcn.2021.17.4.534.

[23] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6(269):pl1. https://doi.org/10.1126/scisignal.2004088.

[24] Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Forbes SA. COSMIC: the catalogue of somatic mutations in cancer. Nucleic Acids Res 2019;47(D1):D941–7. https://doi.org/10.1093/nar/gky1015.

[25] Ganesan K, Kulandaisamy A, Binny Priya S, Gromiha MM. HuVarBase: a human variant database with comprehensive information at gene and protein levels. PLoS One 2019;14(1):e0210475. https://doi.org/10.1371/journal.pone.0210475.

[26] Zhao Y, Li MC, Konaté MM, Chen L, Das B, Karlovich C, Williams PM, Evrard YA, Doroshow JH, McShane LM. TPM, fpkm, or normalized counts? A comparative study of quantification measures for the analysis of RNA-seq data from the NCI patient-derived models repository. J Transl Med 2021;19(1):269. https://doi.org/10.1186/s12967-021-02936-w.

[27] Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter 2004;6(1):20–9. https://doi.org/10.1145/1007730.1007735.

[28] Simoneau J, Dumontier S, Gosselin R, Scott MS. Current RNA-seq methodology reporting limits reproducibility. Briefings Bioinf 2021;22(1):140–5. https://doi.org/10.1093/bib/bbz124.

[29] Choi RY, Coyner AS, Kalpathy-Cramer J, Chiang MF, Campbell JP. Introduction to machine learning, neural networks, and deep learning. Translational vision science & technology 2020;9(2):14. https://doi.org/10.1167/tvst.9.2.14.

[30] Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. Nat Rev Mol Cell Biol 2022;23(1):40–55. https://doi.org/10.1038/s41580-021-00407-0.

[31] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.

[32] Lánczky A, Győrffy B. Web-based survival analysis tool tailored for medical research (KMplot): development and implementation. J Med Internet Res 2021;23(7):e27633. https://doi.org/10.2196/27633.

[33] Shen W, Song Z, Zhong X, Huang M, Shen D, Gao P, Qian X, Wang M, He X, Wang T, Li S. Sangerbox: a comprehensive, interaction-friendly clinical bioinformatics analysis platform. Imeta 2022;1(3):e36. https://doi.org/10.1002/imt2.36.

[34] Kaur H, Bhalla S, Kaur D, Raghava GP. CancerLivER: a database of liver cancer gene expression resources and biomarkers. Database : the journal of biological databases and curation 2020;2020:baaa012. https://doi.org/10.1093/database/baaa012.

[35] Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, Li B, Liu XS. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res 2020;48(W1):W509–14. https://doi.org/10.1093/nar/gkaa407.

[36] Chandrashekar DS, Bashel B, Balasubramanya SAH, Creighton CJ, Ponce-Rodriguez I, Chakravarthi BVSK, Varambally S. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia 2017;19(8):649–58. https://doi.org/10.1016/j.neo.2017.05.002.

[37] Bartha Á, Győrffy B. TNMplot.com: a web tool for the comparison of gene expression in normal, tumor and metastatic tissues. Int J Mol Sci 2021;22(5):2622. https://doi.org/10.3390/ijms22052622.

[38] Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. Nucleic Acids Res 2018;46(D1):D956–63. https://doi.org/10.1093/nar/gkx1090.

[39] Goldman MJ, Craft B, Hastie M, Repečka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers D, Brooks AN, Zhu J, Haussler D. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol 2020;38(6):675–8. https://doi.org/10.1038/s41587-020-0546-8.

[40] Franz M, Rodriguez H, Lopes C, Zuberi K, Montojo J, Bader GD, Morris Q. GeneMANIA update 2018. Nucleic Acids Res 2018;46(W1):W60–4. https://doi.org/10.1093/nar/gky311.

[41] Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 2021;49(D1):D605–12. https://doi.org/10.1093/nar/gkaa1074.

[42] Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics 2020;36(8):2628–9. https://doi.org/10.1093/bioinformatics/btz931.

[43] UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res 2015;43(Database issue):D204–12. https://doi.org/10.1093/nar/gku989.

[44] Bowers KJ, Chow DE, Xu H, Dror RO, Eastwood MP, Gregersen BA, Klepeis JL, Kolossvary I, Moraes MA, Sacerdoti FD, Salmon JK, Shan Y, Shaw DE. Scalable algorithms for molecular dynamics simulations on commodity clusters, SC '06 [Paper presentation]. In: Proceedings of the 2006 ACM/IEEE conference on supercomputing. Tampa; 2006. p. 43. https://doi.org/10.1145/1188455.1188544. FL.

[45] Kollar J, Frecer V. How accurate is the description of ligand-protein interactions by a hybrid QM/MM approach? J Mol Model 2017;24(1):11. https://doi.org/10.1007/s00894-017-3537-z.

[46] Li Y, Ge D, Lu C. The SMART App: an interactive web application for comprehensive DNA methylation analysis and visualization. Epigenet Chromatin 2019;12(1):71. https://doi.org/10.1186/s13072-019-0316-3.

[47] Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31(13):3812–4. https://doi.org/10.1093/nar/gkg509.

[48] Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Current protocols in human genetics; 2013. https://doi.org/10.1002/0471142905.hg0720s76 [Chapter 7], Unit7.20.

[49] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46(3):310–5. https://doi.org/10.1038/ng.2892.

[50] Capriotti E, Fariselli P, Casadio R. I-Mutant 2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33(Web Server issue):W306–10. https://doi.org/10.1093/nar/gki375.

[51] Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nat Commun 2020;11(1):5918. https://doi.org/10.1038/s41467-020-19669-x.

[52] Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, Day IN, Gaunt TR. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 2013;34(1):57–65. https://doi.org/10.1002/humu.22225.

[53] Ouyang X, Fan Q, Ling G, Shi Y, Hu F. Identification of diagnostic biomarkers and subtypes of liver hepatocellular carcinoma by multi-omics data analysis. Genes 2020;11(9):1051. https://doi.org/10.3390/genes11091051.

[54] Sinkala M, Mulder N, Martin D. Machine learning and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. Sci Rep 2020;10(1):1212. https://doi.org/10.1038/s41598-020-58290-2.

[55] Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L, Ngom A. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. Front Genet 2019;10:256. https://doi.org/10.3389/fgene.2019.00256.

[56] Yang H, Li G, Qiu G. Bioinformatics analysis using ATAC-seq and RNA-seq for the identification of 15 gene signatures associated with the prediction of prognosis in hepatocellular carcinoma. Front Oncol 2021;11:726551. https://doi.org/10.3389/fonc.2021.726551.

[57] Liu XH, Yang YF, Fang HY, Wang XH, Zhang MF, Wu DC. CEP131 indicates poor prognosis and promotes cell proliferation and migration in hepatocellular carcinoma. Int J Biochem Cell Biol 2017;90:1–8. https://doi.org/10.1016/j.biocel.2017.07.001.

[58] Yuan W, Zhao H, Zhou A, Wang S. Interference of EFNA4 suppresses cell proliferation, invasion and angiogenesis in hepatocellular carcinoma by downregulating PYGO2. Cancer Biol Ther 2022;23(1):1–12. https://doi.org/10.1080/15384047.2022.2149039.

[59] Htun MW, Shibata Y, Soe K, Koji T. Nuclear expression of Pygo2 correlates with poorly differentiated state involving c-myc, PCNA and Bcl 9 in Myanmar hepatocellular carcinoma. Acta Histochem Cytoc 2021;54(6):195–206. https://doi.org/10.1267/ahc.21-00090.

[60] Ye K, Ouyang X, Wang Z, Yao L, Zhang G. SEMA3F promotes liver hepatocellular carcinoma metastasis by activating focal adhesion pathway. DNA Cell Biol 2020;39 (3):474–83. https://doi.org/10.1089/dna.2019.4904.
[61] Zhang B, Wu H. Decreased expression of COLEC10 predicts poor overall survival in patients with hepatocellular carcinoma. Cancer Manag Res 2018;10:2369–75. https://doi.org/10.2147/CMAR.S161210.
[62] Chen H, Wong CC, Liu D, Go MYY, Wu B, Peng S, Kuang M, Wong N, Yu J. APLN promotes hepatocellular carcinoma through activating PI3K/Akt pathway and is a druggable target. Theranostics 2019;9(18):5246–60. https://doi.org/10.7150/thno.34713.
[63] Gu X, Li S, Ma X, Huang D, Li P. Heterogeneity characterization of hepatocellular carcinoma based on the sensitivity to 5-fluorouracil and development of a prognostic regression model. Front Pharmacol 2023;14:1252805. https://doi.org/10.3389/fphar.2023.1252805.
[64] Zhang X, Sheng J, Zhang Y, Tian Y, Zhu J, Luo N, Xiao C, Li R. Overexpression of SCAMP3 is an indicator of poor prognosis in hepatocellular carcinoma. Oncotarget 2017;8(65):109247–57. https://doi.org/10.18632/oncotarget.22665.
[65] Wang YH, Cheng TY, Chen TY, Chang KM, Chuang VP, Kao KJ. Plasmalemmal Vesicle Associated Protein (PLVAP) as a therapeutic target for treatment of hepatocellular carcinoma. BMC Cancer 2014;14:815. https://doi.org/10.1186/1471-2407-14-815.
[66] Chen W, Ma Z, Yu L, Mao X, Ma N, Guo X, Yin X, Jiang F, Wang Q, Wang J, Fang M, Lin N, Zhang Y. Preclinical investigation of artesunate as a therapeutic agent for hepatocellular carcinoma via impairment of glucosylceramidase-mediated autophagic degradation. Exp Mol Med 2022;54(9):1536–48. https://doi.org/10.1038/s12276-022-00780-6.