Heliyon 9 (2023) e16873

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

# The sub-molecular characterization identification for cervical cancer

XinKai Mo [a,1], Na Wang [b,1], Zanjing He [b], Wenjun Kang [b], Lu Wang [b], Xia Han [b,**], Liu Yang [a,*]

[a] Department of Clinical Laboratory, Shandong Cancer Hospital and Institute, Shandong First Medical University and Shandong Academy of Medical Sciences, Jinan 250117, Shandong, PR China
[b] Department of Medical Laboratory Science, Xinjiang Bayingoleng Mongolian Autonomous Prefecture People's Hospital, Xinjiang, China

ARTICLE INFO

*Keywords:*
Cervical cancer
Non-negative matrix factorization
Classifier
GALNT2

ABSTRACT

*Background:* The efficacy of therapy in cervical cancer (CESC) is blocked by high molecular heterogeneity. Thus, the sub-molecular characterization remains primarily explored for personalizing the treatment of CESC patients.

*Methods:* Datasets with 741 CESC patients were obtained from TCGA and GEO databases. The NMF algorithm, random forest algorithm, and multivariate Cox analysis were utilized to construct a classifier for defining the sub-molecular characterization. Then, the biological characteristics, genomic variations, prognosis, and immune landscape in molecular subtypes were explored. The significance of classifier genes was validated by quantitative Real-Time PCR, cell transfection, cell colony formation assay, wound healing assay, cell proliferation assay, and Western blot.

*Results:* The CESC patients were classified into two subtypes, and the high classifier-score patients with significant differences in ECM-receptor interaction, PI3K-Akt signaling pathway, and MAPK signaling pathway showed a poorer prognosis in OS ($p < 0.001$), DFI ($p = 0.016$), PFI ($p < 0.001$) and DSS ($p < 0.001$), and with high the M0 Macrophage and resting Mast cells infiltration and low HLA family gene expression. Moreover, the constructed classifier owns a high identified accuracy in the tumor/normal groups (AUC: 0.993), the tumor/CIN1–CIN3 groups (AUC: 0.963), and normal/CIN1–CIN3 groups (AUC: 0.962), and the total prediction performance is better than currently published signatures in CESC (C−index: 0,763). The combined prediction performance further indicated that Nomogram (AUC = 0.837) is superior to the classifier (AUC = 0.835) and Stage (AUC = 0.568), and the C−index of calibration curves is 0.784. The potential biological function of classifier genes indicated that silencing GALNT2 inhibited the cancer cell's proliferation, migration, and colony formation; Conversely, the cancer cell's proliferation, migration, and colony formation were increased after the upregulation of GALNT2. The Epithelial-Mesenchymal Transition Experiment showed that GALNT2 knockdown might reduce the levels of Snail and Vimentin proteins and increase E-cadherin; Conversely, the levels of Snail and Vimentin proteins were increased, E-cadherin was reduced by GALNT2 upregulation.

* Corresponding author.
** Corresponding author.
*E-mail addresses:* moxinkai@163.com (X. Mo), 274696822@qq.com (N. Wang), he_zanjing@163.com (Z. He), 2206145368@qq.com (W. Kang), 451918554@qq.com (L. Wang), 1653475851@qq.com (X. Han), 2295722106@qq.com (L. Yang).
[1] Co-author.

https://doi.org/10.1016/j.heliyon.2023.e16873
Received 9 March 2023; Received in revised form 28 May 2023; Accepted 31 May 2023
Available online 7 June 2023

*Conclusion:* The classifier we constructed may help improve our understanding of subtype characteristics and provide a new strategy for developing CESC therapeutics. Remarkably, GALNT2 may be an option to directly target drivers in CESC cancer therapy.

## 1. Introduction

Cervical cancer, the second most common cancer in women in terms of incidence and mortality [1], occurs mainly in less developed areas where effective screening and HPV vaccination programs are lacking [2] and results in more than 604,127 new cases and 341, 831 deaths worldwide each year [1]. The extent of the disease at the time of diagnosis determines the patient's treatment method, which may involve surgical treatment, adjuvant, and/or combined treatment. However, they had limited efficiency for advanced or/and recurrent cervical cancers [3]. Moreover, while some clinical trials have shown promising results, immunotherapies exhibit efficiency for multiple solid tumors. They have been approved by FDA for treating recurrent or metastatic cervical cancers; for the patients with advanced stage, less than 20% of them show a positive response to ICI, the 5-year survival rate is only 16.5%, the median survival time is only 8–13 months [4], which may be attributed to the high heterogeneity of tumors.

The emergence of large-scale high-throughput detection technology is accompanied by the formation of multiple expression datasets, which makes it possible to further identify key molecular features to better provide personalized treatment plans [5]. Accumulating evidence suggests that the identification of tumor molecular subtypes with high heterogeneity has emerged as a promising strategy in the personalized treatment of CESC patients, and this heterogeneity is manifested not only in terms of tumor metabolism and tumor immune infiltration microenvironment [6,7] but also in terms of molecular features [8]. According to tumor molecular heterogeneity, previous studies had identified multiple cervical cancer-related molecular subtypes, such as HPV-A9, HPV-A7, HPV-negative subtypes, keratin-low/high squamous and adeno-enriched subtypes, hormonal, epithelial-mesenchymal transition, and PI3K-AKT-associated subtypes [8], and HPV + G1 and HPV + G2 subtypes [9].

However, although the previous studies had defined different molecular subtypes, there were still crossovers in those subtypes [9]. The predicting accuracy and the molecular included of these subtypes are different [10–13]. Therefore, are there other ways to hierarchically cluster CESC patients to assess prognosis and response to immunotherapy? How to evaluate the accuracy of defined subtypes? Given the molecular subtypes were particularly important for providing potential therapeutic targets and guiding the clinical decision-making of personalized treatment, in this study, we performed batch correction on multiple datasets in the GEO database to identify differentially expressed genes, multivariate Cox analysis and NMF algorithms were used to identify molecular subtypes and verify in an external verification, the random forest tree algorithms was utilized to evaluate the superiority of molecular subtypes. The transcription levels of classifier genes were verified in H8, SiHA, Casiki, HeLa, and C33a cell lines. And then, quantitative

**Table 1**
The characteristics of TCGA-CESC cohort.

| Category | Type | Total | Test | Train | Pvalue |
|---|---|---|---|---|---|
| Age | ≤65 | 228 (89.41%) | 119 (92.25%) | 109 (86.51%) | 0.1985 |
| | >65 | 27 (10.59%) | 10 (7.75%) | 17 (13.49%) | |
| T | T1 | 119 (46.67%) | 58 (44.96%) | 61 (48.41%) | 0.7356 |
| | T2 | 60 (23.53%) | 33 (25.58%) | 27 (21.43%) | |
| | T3 | 16 (6.27%) | 8 (6.2%) | 8 (6.35%) | |
| | T4 | 8 (3.14%) | 5 (3.88%) | 3 (2.38%) | |
| | Tis | 1 (0.39%) | 1 (0.78%) | 0 (0%) | |
| | unknow | 51 (20%) | 24 (18.6%) | 27 (21.43%) | |
| N | N0 | 106 (41.57%) | 55 (42.64%) | 51 (40.48%) | 0.7783 |
| | N1 | 50 (19.61%) | 24 (18.6%) | 26 (20.63%) | |
| | unknow | 99 (38.82%) | 50 (38.76%) | 49 (38.89%) | |
| M | M0 | 93 (36.47%) | 47 (36.43%) | 46 (36.51%) | 1 |
| | M1 | 10 (3.92%) | 5 (3.88%) | 5 (3.97%) | |
| | unknow | 152 (59.61%) | 77 (59.69%) | 75 (59.52%) | |
| Stage | Stage I | 142 (55.69%) | 69 (53.49%) | 73 (57.94%) | 0.3599 |
| | Stage II | 55 (21.57%) | 34 (26.36%) | 21 (16.67%) | |
| | Stage III | 34 (13.33%) | 17 (13.18%) | 17 (13.49%) | |
| | Stage IV | 18 (7.06%) | 8 (6.2%) | 10 (7.94%) | |
| | unknow | 6 (2.35%) | 1 (0.78%) | 5 (3.97%) | |
| Grade | G1 | 14 (5.49%) | 3 (2.33%) | 11 (8.73%) | 0.0434 |
| | G2 | 117 (45.88%) | 66 (51.16%) | 51 (40.48%) | |
| | G3 | 98 (38.43%) | 45 (34.88%) | 53 (42.06%) | |
| | G4 | 1 (0.39%) | 0 (0%) | 1 (0.79%) | |
| | unknow | 25 (9.8%) | 15 (11.63%) | 10 (7.94%) | |
| Therapy | Chemotherapy | 30 (11.76%) | 15 (11.63%) | 15 (11.9%) | 0.0312 |
| | Chemotherapy/Radiation_therapy | 99 (38.82%) | 57 (44.19%) | 42 (33.33%) | |
| | Radiation_therapy | 42 (16.47%) | 14 (10.85%) | 28 (22.22%) | |
| | unknow | 84 (32.94%) | 43 (33.33%) | 41 (32.54%) | |

Real-Time PCR, Cell transfection, Cell colony formation assay, Wound healing assay, Cell proliferation assay, and Western blot were performed to validate the Possibility of GALNT2 as a potential biomarker and potential therapeutic target for CESC.

### 1.1. Data obtaining and processing

Cervical cancer (CESC) related expression files, somatic mutation (SNPs and small INDELs), copy number variations profiles, DNA methylation, and clinical data were obtained from The Cancer Genome Atlas (TCGA, https://cancergenome.nih.gov/) database. The lncRNA and protein-coding gene files were further classified based on gene annotations from the GENCODE project (https://www.gencodegenes.org/). Tumor mutation burden (TMB) values were calculated for each tumor sample based on the number of tumor mutations per megabase. The four major clinical outcome endpoints, including overall survival event (OS), disease-specific survival event (DSS), disease-free interval event (DFI), and progression-free interval event (PFI) were obtained from the previous study [14]. The TCGA-CESC cohort was defined as a classifier cohort that is mainly for analysis. Based on the SangerBox (http://www.sangerbox.com/) database, batch batch-correction for non-biotech bias was performed on 4 datasets (Including GSE6791, GSE7803, GSE9750, and GSE63514) from the Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) database by the "ComBat" algorithm of the sva package to identify a differentially expressed gene identification cohort (DEG-Cohort). GSE44001 was defined as the external validation cohort. The transcription factors list was downloaded from the Cistrome Cancer database (http://cistrome.org/CistromeCancer/). The identified DEG criteria were $P < 0.05$ and $|log2FC| > 1$. The All the details information were list in Supplement Table 1. In this study, we only included patients with complete clinical information and a survival period of over 30 days, regardless of the treatment they received. Table 1 summarized the characteristics of the study cohort. Given all the information involved in our study was publicly available, the ethics committee has no specific ethical approval.

### 1.2. Between-group analysis of tumor and normal samples

The "ggplot2" and "pheatmap" packages were used to visualize the volcano plots and heatmaps of DEGs. And then, we explored the difference between Gene Ontology (GO) [15] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [16] through the "clusterProfiler" package to investigate the most significantly enriched pathways and biological processes in tumor and normal groups.

### 1.3. The classifier construction and evaluation

Based on the "survival" package [17], the univariate Cox regression analysis and Kaplan–Meier methods ($P < 0.01$) were utilized to attest to the prognostic value of DEGs. And then, those prognosis-DEGs were further considered for inclusion in the multivariate Cox regression analysis to calculate the regression coefficients for constructing the classifier score. The prognosis in DFI, PFI, DSS, and OS between the two subgroups of the classifier was compared by Kaplan-Meier survival curves. Random forest, as a popular classification and regression method, has been proven powerful for biological studies in various prediction problems [18]. Therefore, we ranked the classifier genes based on the mean decrease accuracy of this algorithm to quantify the contribution of each classifier gene for the classification accuracy of tumor and normal groups and evaluate the classifier performance with the area under the curve (AUC) by cross-validation. Similarly, this classification accuracy of the classifier was also evaluated in the normal/cervical intraepithelial neoplasia (CIN1– CIN3) groups and tumor/CIN1– CIN3 groups. To further demonstrate the molecular heterogeneity of CESC, we performed unsupervised classification with non-negative matrix factorization (NMF) [19] based on the expression matrix of the classifier genes in the classifier-Cohort, by setting the number of clusters k to determine the average contour width, and then determine the optimal number of clusters of CESC. The Sankey diagram was utilized to assess and visualize for crossover between the two subtypes. The prognostic potential of the classifier was evaluated in each of the two cohorts created by randomly dividing the TCGA cohort into a training cohort and a test cohort in a 1:1 ratio. To further assess the performance advantage in the constructed classifier, the ROC value and C-index were considered in previously published signatures [10–13]. The changes in DNA methylation play a crucial role in malignant transformation, causing the overexpression of oncogenes and silencing of tumor-suppressor genes [20]. Therefore, we further evaluated the methylation status of poor prognosis genes.

### 1.4. Comprehensive analysis between two molecular subtypes

Based on the somatic mutation files obtained from the TCGA database, the mutation landscapes and the top 20 genes with the highest mutation frequency across all samples were further evaluated and visualized in the two subtypes by the "Maftools" package. The "clusterProfiler" package investigates the significantly enriched pathways in two subtypes. Given that the efficacy of immunotherapy in malignant tumors is indeed related to the tumor microenvironment, we also evaluated the Immune infiltration profile of two subtypes based on CIBERSORT algorithm [21]. TMB and classifier scores were included for further stratified analysis for CESC patients. Immunotherapy data of CESC patients, obtained from TCIA (https://tcia.at/) websites, was utilized to predict the sensitivity of clinical drug treatment in two subtypes.

### 1.5. Nomogram construction and evaluation

Based on the results of Cox regression analysis of clinical files (Stage and Grade) and classifier genes, we further constructed a nomogram for evaluating cancer prognosis and evaluated the accuracy and consistency of the nomogram by decision curve analysis

(DCA). Finally, the calibration plots and time-dependent ROC curves of this Nomogram were further visualized through the "survival" and "rms" packages.

### 1.6. Cell culture

The immortalized human normal cervical cell line (H8) and cervical cancer cell lines CaSki, SiHa, HeLa, and C33A cell lines were purchased from the Chinese Academy of Sciences (Shanghai, China). The SiHa, HeLa, and C33A cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM) (Gibco, USA), and the H8 and CaSki cell lines were maintained in Roswell Park Memorial Institute-(RPMI-) 1640 medium (Gibco, USA), containing 10% FBS (Gibco, USA) and 1% antibiotics. All cells were grown at 37 °C with 5% CO2.

### 1.7. RNA extraction and quantitative real-time PCR

Total RNA was extracted with Trizol reagent (Invitrogen, USA) and reverse transcribed with ReverTra Ace qPCR RT kit (Toyobo, China) according to the manufacturer's regulations. The amplification was performed by Light Cycler 480 for qRT-PCR. GAPDH was used as an internal control. Fold changes were calculated by the 2-ΔΔCt method. The primer information is shown in Supplementary Table 2.

### 1.8. Cell transfection

For the overexpression of GALNT2 in HeLa cells and SiHa cells, the GALNT2 cDNA sequence was cloned into pEnter vector, which was obtained from WZ Biosciences Inc. The negative control siRNA (NC) and siRNA-GALNT2 (Supplementary Tables 2 and 3) were synthesized by GenePharma and transfected into HeLa cells and SiHa cells using Lipofectamine 3000 (Invitrogen, USA) following the protocol from the manufacturer.

### 1.9. Cell proliferation assay

The 96-well plate cultured 2 × 103 cells per well and the cell viability was calculated using the cell counting kit - 8 (CCK-8) system. The optical density (OD) value of each hole was measured at 450 nm by a microplate reader.

### 1.10. Wound healing assay

The migration ability was tested by wound healing assay. When the confluence of cells reaches 90%, scratch with the tip of the pipette. At 0 h and 48 h, image J software 2.0 was used to take photos of the wound area for analysis.

### 1.11. Cell colony formation assay

For the colony formation experiment, the cells were seeded in a 6-well plate with a density of (500 cells/well). The cells were cultured for 2 weeks to form colonies. The colonies were fixed with 4% paraformaldehyde buffer and stained with 0.01% crystal violet. Colonies containing 50 or more cells were evaluated as positive for statistical analysis.

### 1.12. Western blot

Western blot was performed according to the standard protocol. The cells were washed with precooled PBS and lysed with lysate. The total protein extract was separated on 10% SDS-PAGE gel and transferred to PVDF membrane. After sealing PVDF membrane with 5% skimmed milk powder, it was incubated with the first antibody overnight at 4 °C. Subsequently, the membrane was incubated with goat anti rabbit secondary antibody at room temperature for 1 h and detected with an enhanced chemiluminescent substrate (ECL) kit (Vazyme, Nanjing, China). Finally, use Image J software 2.0 to quantify the protein bands and normalize them with their respective GAPDH bands.

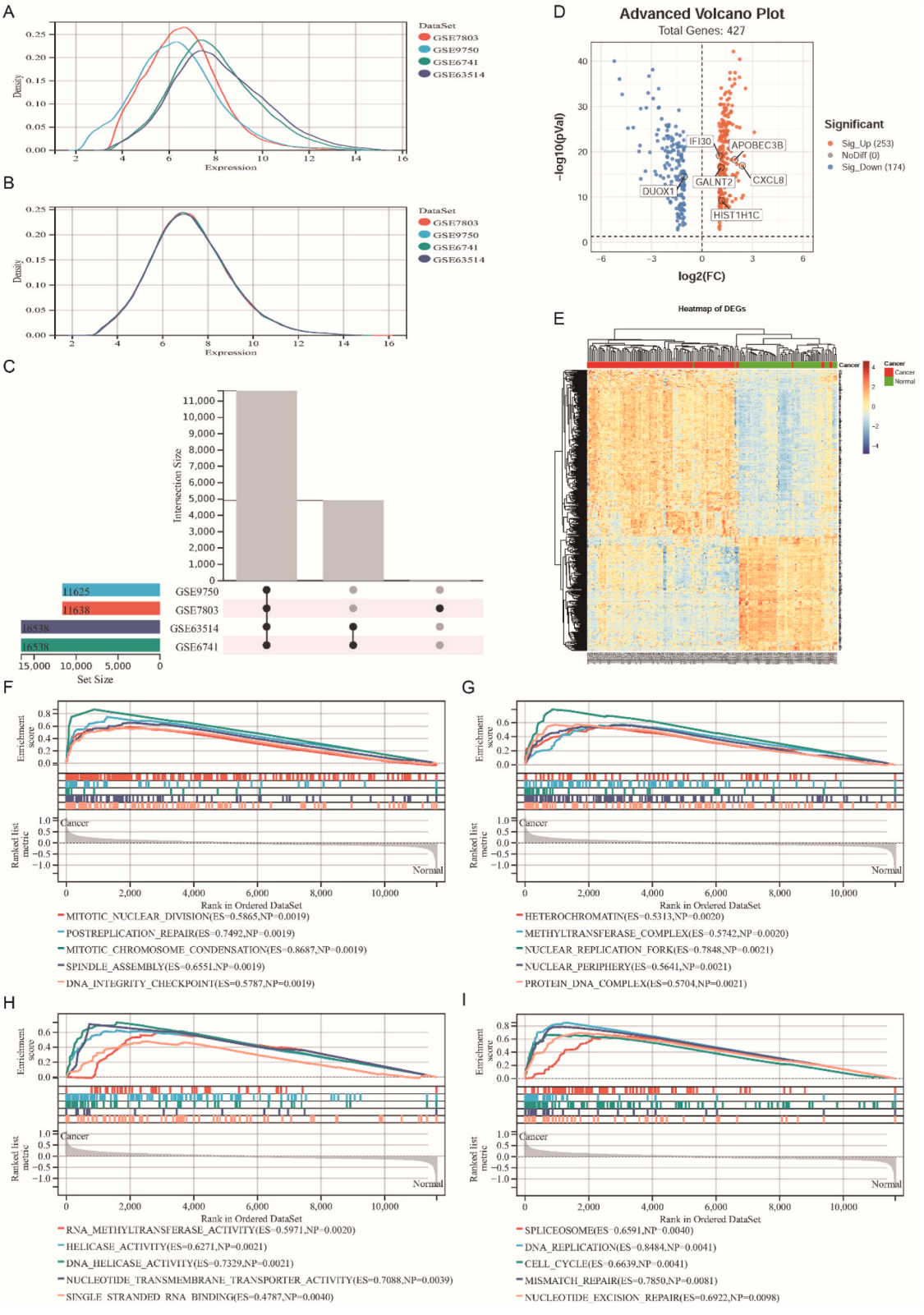
### 1.13. Statistical analysis

GraphPad Prism (Version 8.0, USA) was used for statistical analysis. Student's t-test was used for statistical comparison between two samples, and a one-way analysis of variance was used for differences between three or more groups. $P < 0.05$ was considered statistically significant (*$P < 0.05$).

## 2. Result

### 2.1. Comprehensive analysis of normal/tumor groups

After batch rectification on 4 datasets (Fig. 1A–C), a DEG cohort was defined. And then, a total of 427 DEGs were obtained from DEG-Cohort based on the "limma" package (Fig. 1D–E). KEGG analyses indicated that the tumor group was mainly enriched in the




*(caption on next page)*

**Fig. 1.** The batch-effect correction and functional pathway analysis of datasets in DEG-Cohort. (A) The density plots before batch-effect correction in DEG-Cohort. (B)The density plots after batch-effect correction. (C) The Venn diagram of DEG-Cohort. (D) The Volcano Plots of DEGs. (E) The heatmap of DEGs. (F) The KEGG pathway analysis of DEGs. (G, H, I) The GO analysis of DEGs. DEGs means differentially expressed gene; DEG-Cohort means differentially expressed gene identification cohort; KEGG means Kyoto Encyclopedia of Genes and Genomes; GO means Gene Ontology.

spliceosome, DNA replication, cell cycle, mismatch repair, and nucleotide excision repair pathways compared with the normal group (Fig. 1F). GO analysis shows that the occurrence of tumors may be related to the dysregulation of biological processes such as mitotic nuclear division, post-replication repair, mitotic chromosome condensation, spindle assembly, and DNA integrity checkpoint (Fig. 1G–I).

### 2.2. The classifier construction and evaluation

We finally identified a robust classifier formula in the classifier-cohort that can divide the CESC patients into two specific subgroups, the calculation formula of classifier-score = (−0.078 * expression of DUOX1) + (0.691 * expression of GALNT2) + (0.241 * expression of CXCL8) + (−0.034 * expression of APOBEC3B) + (−0.175 * expression of HIST1H1C) + (−0.156 * expression of IFI30). Compared with the low classifier-score patients, the high classifier-score patients showed a poorer prognosis across the four primary clinical outcome endpoints (Fig. 2A–D), including OS ($p < 0.001$), DFI ($p = 0.016$), PFI ($p < 0.001$) and DSS ($p < 0.001$). To evaluate the accuracy of the classifier, we confirmed the performance of this classifier by NMF analysis (Fig. 2E). According to the results, the two subtypes identified by NMF analysis highly overlap with subtypes identified by Classifier and similarly had a poor prognosis (Fig. 2F–G). And then, in the predicted performance comparison of multiple signatures, the performance of our classifier (The C-index is 0.763) is superior to the previous multiple signatures (Fig. 2I-N), and the C-index and ROC value was 0.763 and 0.780, respectively. However, the accuracy of the constructed classifier is not significant in the GSE44001 dataset (Fig. 2H). In addition to subtype identification in tumor patients, we further evaluated the classifier's identified performance on tumor and normal samples based on the random forest method (Fig. 3A) and found that IFI30, GALNT2, CXCL8, APOBEC3B, and DUOX1 (The Mean Decrease Gini>10) contributed more to the classifier accuracy (Fig. 3B), and this classifier has high accuracy in the tumor/normal groups (AUC: 0.993, 95% CI: 0.983−1.00) (Fig. 3C), and the expression of classifier-genes was significantly different between tumor and normal groups (Fig. 3D). Given that revealing molecular changes associated with progression contributes to risk assessment, diagnosis, prognosis, and treatment of CESC patients, we also evaluated the identified performance of classifier-genes in tumor/CIN1–CIN3 groups (Fig. 3E) and normal/CIN1–CIN3 groups (Fig. 3I), and found that DUOX1, GALNT2 and IFI30 contributed more to the classifier accuracy in tumor/CIN1–CIN3 groups (Fig. 3F), conversely, APOBEC3B, IFI30 and HIST1H1C contributed more to the classifier accuracy in normal/CIN1–CIN3 groups (Fig. 3J). More interestingly, the identified performance of the classifier was superior in tumor/CIN1–CIN3 groups

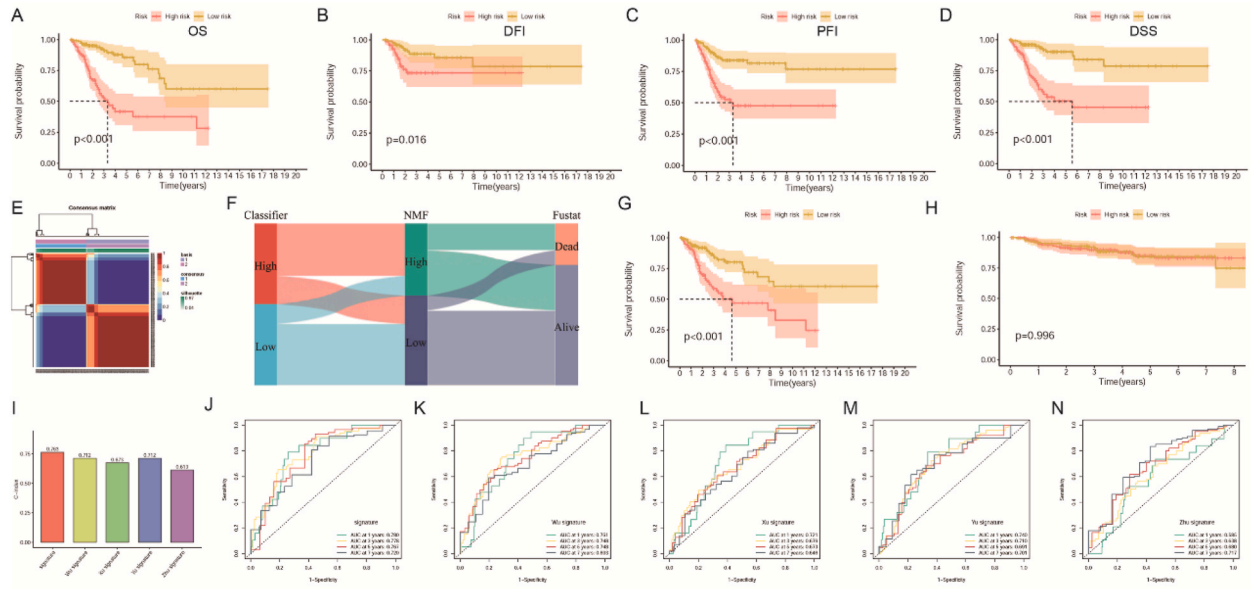



**Fig. 2.** The performance evaluation of Classifier. Compared with the low classifier-score patients, the high classifier-score patients showed a poorer prognosis in OS (A), DFI (B), PFI (C) and DSS (D). The subtypes identified by NMF algorithm (E). Sankey diagram shown that the subtypes identified by NMF algorithm had highly overlap with subtypes identified by Classifier (F), and similarly had a poor prognosis (G). The accuracy of constructed classifier is not significant in the GSE44001 dataset (H). In the predict performance comparison of multiple signatures, the performance of our classifier (J) is superior to Wu et al. (K), Xu et al. (L), Yu et al. (M) and Zhu et al. (N) signatures, and the C-index and ROC value of our classifier was 0.763 and 0.780 respectively (I).

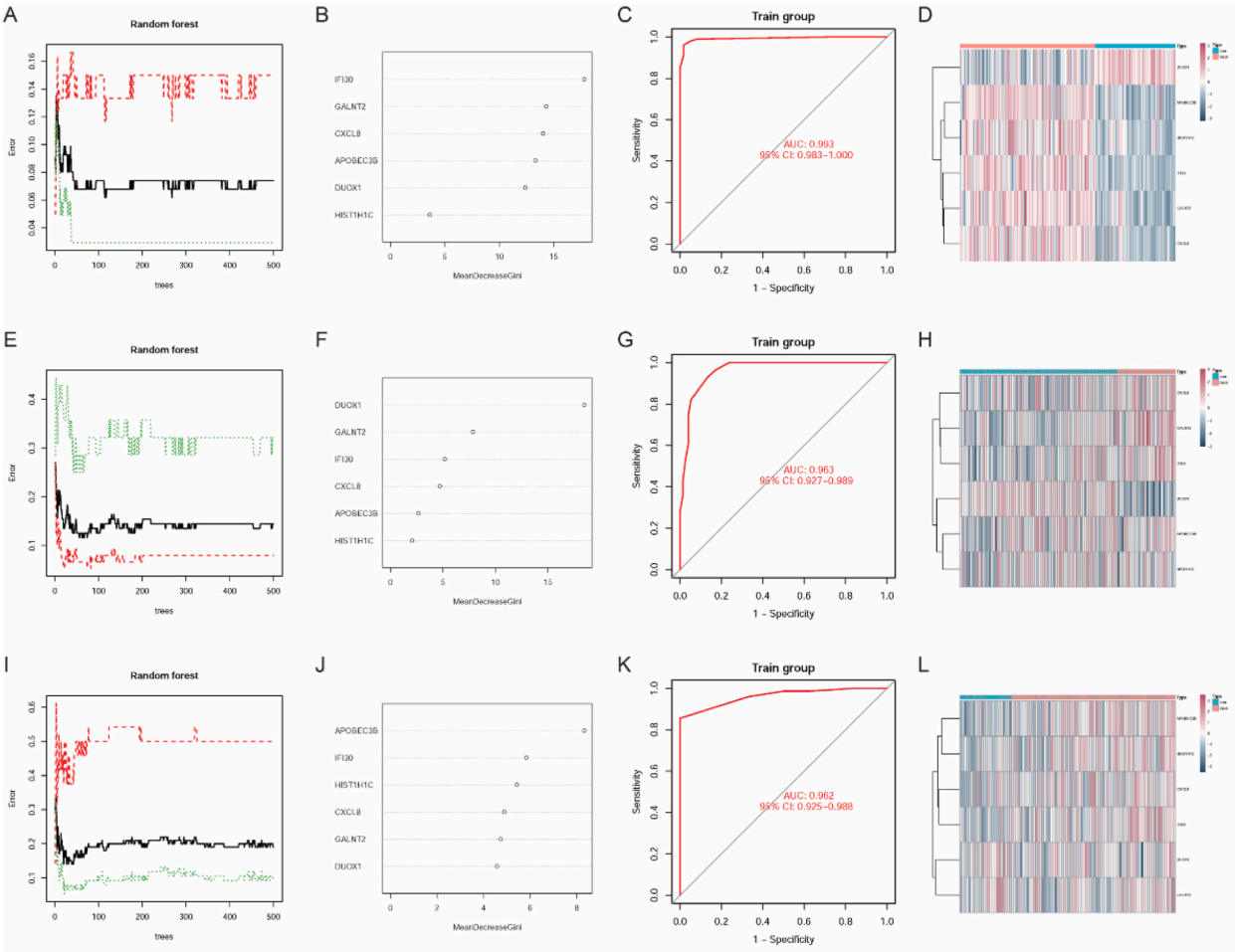



**Fig. 3.** The classifier's identified performance on tumor and normal samples based on the random forest method (A). The IFI30, GALNT2, CXCL8, APOBEC3B and DUOX1 (The Mean Decrease Gini>10) contributed more to the classifier accuracy (B). The AUC of classifier-genes in the tumor/normal groups is 0.993 (C). The heatmap of classifier-genes in tumor/normal groups (D). The identified performance of classifier-genes in tumor/CIN1–CIN3 groups (E), and DUOX1, GALNT2 and IFI30 contributed more to the accuracy in tumor/CIN1–CIN3 groups (F), the AUC of Identification performance of classifier was 0.963 (G), and heatmap shown that classifier-gene expression was significantly different in tumor/CIN1–CIN3 groups (H). The identified performance of classifier-genes in normal/CIN1 groups (I). The Contribution rank of classifier-genes in normal/CIN1 groups (J), the AUC of Identification performance of classifier was 0.962 (K), and the heatmap of classifier-genes in tumor/CIN1–CIN3 groups (L).

(AUC: 0.963, 95% CI: 0.927−0.989), and normal/CIN1–CIN3 groups as well (AUC: 0.962, 95% CI: 0.925−0.988) (Fig. 3G and K). Similarly, the expression of classifier genes was significantly different between tumor/CIN1–CIN3 groups and normal/CIN1–CIN3 groups (Fig. 3H, L).

### 2.3. Comprehensive analysis between two subtypes

The results of mutation between the two subtypes showed that the high classifier-score group has more missense mutations, the variant types are mainly SNPs and DELs, SNV types are predominantly C > T, and the mutation rates of TTN and PIK3CA were >20% in two subtypes (Fig. 4A–B). The results of enrichment pathway analysis showed significant differences in Cytokine-cytokine receptor interaction, IL-17 signaling pathway, ECM-receptor interaction, TNF signaling pathway, PI3K-Akt signaling pathway, and MAPK signaling pathway between the two subtypes ($p < 0.001$) (Fig. 4C). The immune cell infiltration of two subtypes showed that the low-risk group with higher immune cell infiltration of $CD8^+$ T cell, M2 Macrophage, and activated Mast cell, while the M0 Macrophage and resting Mast cells were higher in the high-risk group (Fig. 4E). Most of these immune cells were closely associated with classifier genes (Fig. 4H). Moreover, as a prerequisite of T cell immune activation, the HLA family gene expression in the low-score group was similarly higher than that in the high-score group (Fig. 4F). The stratified analysis of TMB and classifier score further showed that the patients, even with a similar TMB state had a better prognosis in the low classifier score (Fig. 4D). The cellular characteristics of immune invasion indicate that oncogene determines the immune phenotype and tumor escape mechanism. Charoentong Pornpimol et al. developed a quantitative scoring scheme called immunophenoscore (IPS) for predicting the efficacy of CTLA-4 and PD-1/PD-L1

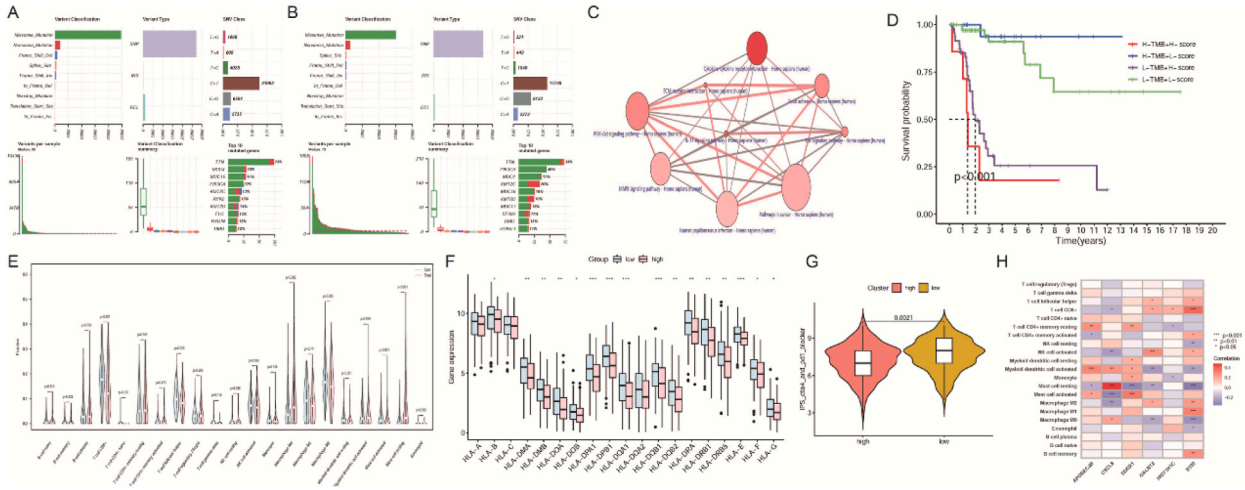


**Fig. 4.** Characterization of mutation profiles in high classifier-score and low classifier-score groups (A, B). The different KEGG pathway analysis in high classifier-score and low classifier-score groups (C). The stratified survival analysis of TMB and classifier (D). The low classifier score group with higher immune cell infiltration of $CD8^+$ T cell, M2 Macrophage, and activated Mast cell, while the M0 Macrophage and resting Mast cells were higher in the high-risk group (E). The difference of HLA family gene expression in the two subtypes (F). The benefit difference of combination with CTLA-4 and PD-1/PD-L1 inhibitors between the two subtypes (G). The co-relationship between immune cells and classifier-genes (H).

inhibitors [22]. In this study, we found that Patients in the low classifier score group have higher IPS than those in the high classifier score group, which means that the low classifier-score group patients may benefit from the combination of CTLA-4 and PD-1/PD-L1 inhibitors (Fig. 4G). The ROC of the test cohort and train cohort in 5 years were 0.832 and 0.801 (Fig. 5). The methylation status of GALNT2 indicated that multiple methylation sites of GALNT2 are closely associated with poor prognosis in patients ($P < 0.05$; Fig. 5).

### *2.4. Nomogram construction and evaluation*

Based on the results of Cox regression analysis, we found that the Stage (HR:1.584, 95% CI:1.231–2.038, $p < 0.001$) and classifier score (HR:1.293, 95% CI:1.198–1.397, $p < 0.001$) were closely related to prognosis (Fig. 6A–B). And then, we constructed a robust nomogram to forecast the prognosis of CESC patients in the 1–5 years (Fig. 6C), and the combined prediction performance of the Nomogram (AUC = 0.837) is superior to that of the classifier score (AUC = 0.835) and Stage (AUC = 0.568) (Fig. 6D). This performance advantage is further confirmed by the calibration curves (C−index: 0.784, 95% CI: 0.755−0.813) and DCA (Fig. 6E–F).

### *2.5. In vitro validation of gene expressions and function of GALNT2*

Given the overexpression of GALNT2 and CXCL8 was an unfavorable prognostic factor of OS, DFI, PFI, and DSS for CESC patients in the TCGA-CESC cohort and GSE44001 dataset (Fig. 7A–J). Therefore, we confirmed the expression level of GALNT2 and CXCL8 in CaSki, SiHa, HeLa, and C33A cell lines and H8 cell line by qRT-PCR, and found that GALNT2 was indeed highly expressed in cervical cancer cell CaSki, SiHa, HeLa and C33A cell lines compared with H8 cell line, and CXCL8 was different to some extent (Fig. 8A–B). And then, we assessed overexpression or knockdown efficiency of GALNT2 in HeLa cells and SiHa cells by real-time PCR and western blotting, respectively (Fig. 8C, J). We further explored the potential biological function of GALNT2 in cervical cancer cells, and results indicated that silencing GALNT2 inhibited the proliferation, migration, and colony formation of HeLa cells and SiHa cells in vitro (Fig. 8D–I). Conversely, cell proliferation, migration and colony formation were increased after the upregulation of GALNT2 in HeLa cells and SiHa cells (Fig. 8D–I). Therefore, we further investigated the influence of GALNT2 on EMT by Western blot. The results showed that the levels of Snail and Vimentin proteins were reduced, E-cadherin was increased by GALNT2 knockdown in HeLa cells and SiHa cells (Fig. 8J). Conversely, the levels of Snail and Vimentin proteins were increased, E-cadherin was reduced by GALNT2 upregulation in HeLa cells and SiHa cells (Fig. 8J). KEGG pathway analysis revealed a high co-expression between GALNT2 and important genes of proteoglycan and Hippo signaling pathways in cancer (P value < 0.001; Fig. 7K). The full and non-adjusted images as supplementary Figure 1 provided.

## 3. Discussion

Cervical cancer is the second most common cancer in women, and although the current variety of treatments for cervical cancer is exciting, the identification of highly heterogeneous molecular subtypes is expected to further develop personalized treatment strategies and increase the benefiting proportion of patients.

In this study, we constructed a prognostic classifier by Cox regression analysis and NMF methods, which consists of DUOX1,

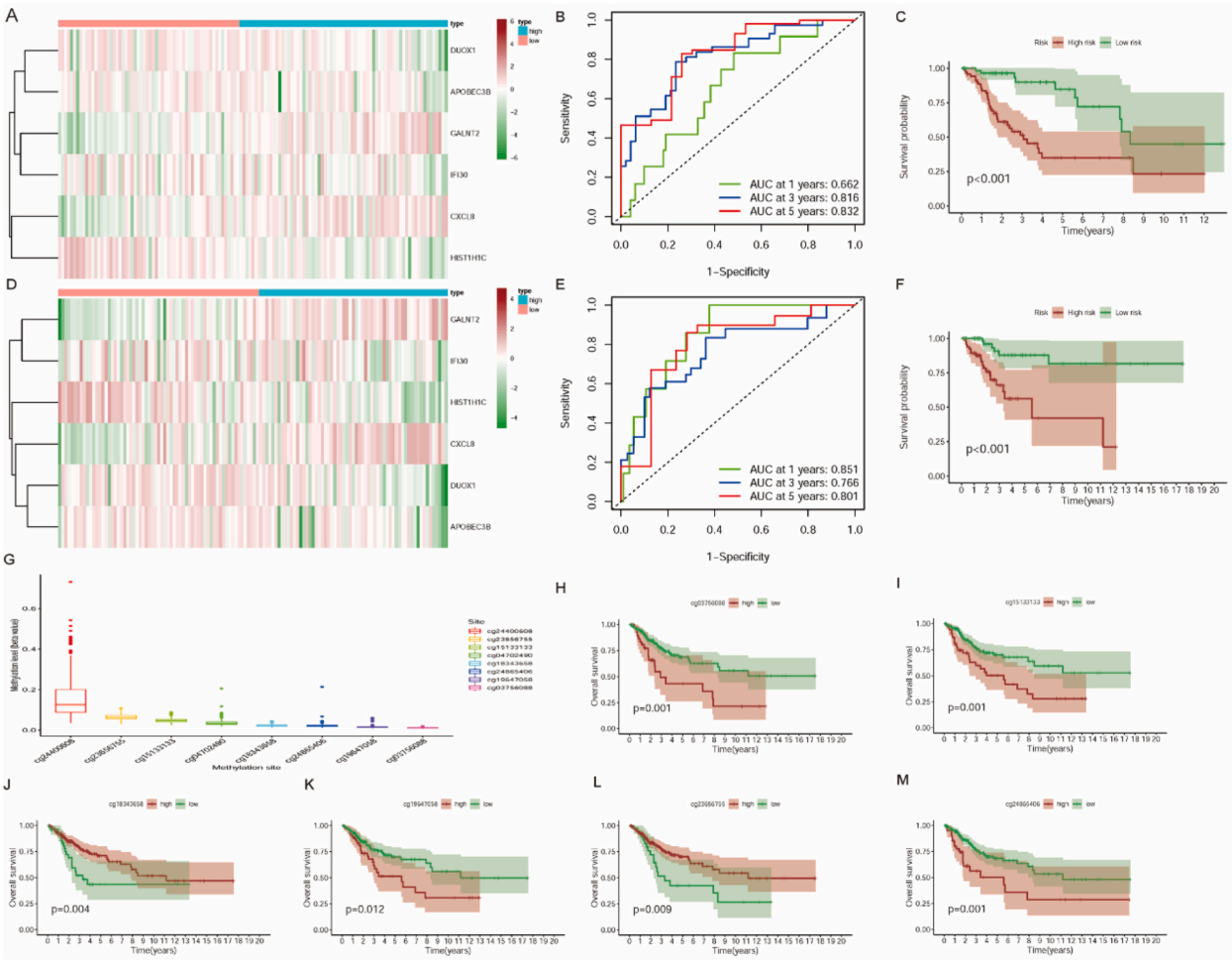
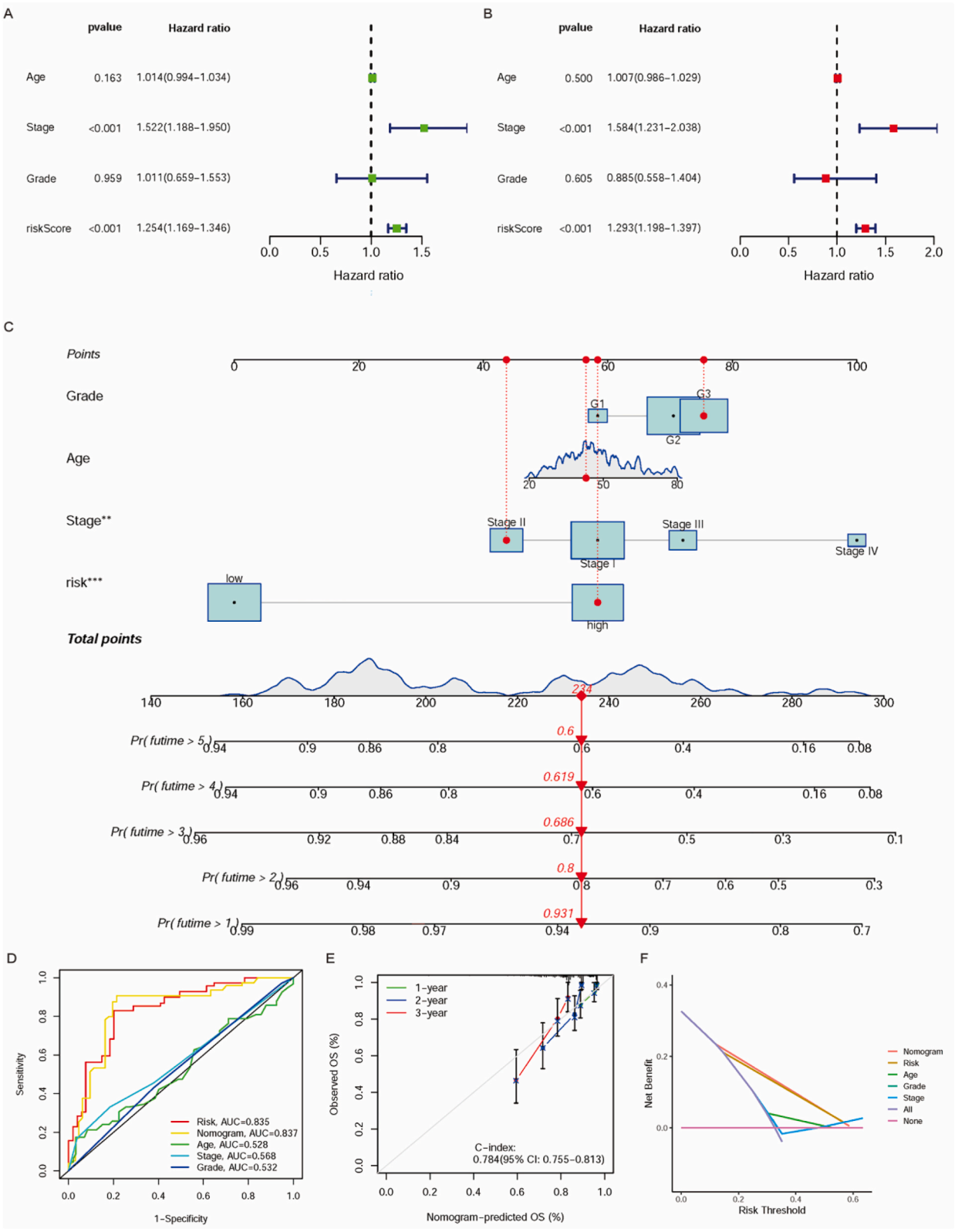



**Fig. 5.** The expression heatmap, prognosis, and ROC analysis of classified genes in the test group (A–C); The expression heatmap, prognosis, and ROC analysis of classified genes in the test group (D–F); The methylation status of GALNT2 and prognostic analysis of methylation sites (G–M).

GALNT2, CXCL8, APOBEC3B, HIST1H1C, and IFI30. Two subtypes were identified by this classifier in the tumor groups, and the low classifier score subtype patients with higher abundance of $CD8^{+}$ T cell, M2 Macrophage, and activated Mast cell, while the M0 Macrophage and resting Mast cells were superior in the high classifier score subtype. As a prerequisite of T cell immune activation, the HLA family gene expression in the low-score group was higher than that in the high-score group, which suggests that these patients are expected to benefit from immunotherapy. And then, in the comparison of the performance in several signatures, we found that the predicted performance of this classifier is indeed superior to previous multiple signatures in the CESC, and the C-index is 0.763. Nomogram results further showed a significant advantage of subtype classifier (AUC = 0.835) and classifier-related-Nomogram (AUC = 0.837) in the prognostic prediction aspects compared to conventional factor stage (AUC = 0.568) and grade (AUC = 0.532). Therefore, this constructed classifier is robust and promising to be used as potential biomarkers for the individualized treatment of CESC.

As the unfavorable prognostic genes in this prognostic prediction classifier, previous studies have shown that CXCL8, as an inflammatory marker, plays an important role in the development of various cancers. As a chemokine secreted by activated tumor cells, CXCL8 can act on CXCR1/2 in the tumor microenvironment to regulate the proliferation and self-renewal of inflammatory factors and tumor stem cells [23]. Moreover, CXCL8 may induce VEGFA to participate in tumor proliferation and metastasis in HCC cells [24], and CXCL8 receptor antagonists may become a potential targeted therapy for HCC [25]. In addition, CXCL8 can also inhibit the expression of ER in endometrial cancer cells and promote tumor invasion [26]. Interestingly, in this study, we found for the first time that CXCL8 is closely related to the patient's poor prognosis, and it may be one of the key molecules in improving patient survival rate, which can be further explored in future studies.

Furthermore, as another unfavorable prognostic gene, we observed that GALNT2 overexpression is not only a poor prognostic factor in CESC patients but also affects the regulation of specific cytokines, chemokines, and immune molecules [27]. Thus, the expression levels of GALNT2 were initially confirmed in CaSki, SiHa, HeLa, and C33A cell lines and H8 cell lines by qRT-PCR. And we found that GALNT2 was indeed highly expressed in cervical cancer cell lines compared with the H8 cell line. As a member of the




*(caption on next page)*

**Fig. 6.** The forest plots of the univariate and multiple Cox analysis in Stage, Grade and classifier score in CESC patients (A, B). Based on the item of nomogram to predict the survival rate of CESC patients for 1-, 3-year and 5-years (C). The time-dependent ROC curve of nomogram in the CESC patients is 0.837 (D). The C-index of the nomogram calibration curve is 0.784 (E). The decision curve analysis of nomogram and classifier risk score (F).

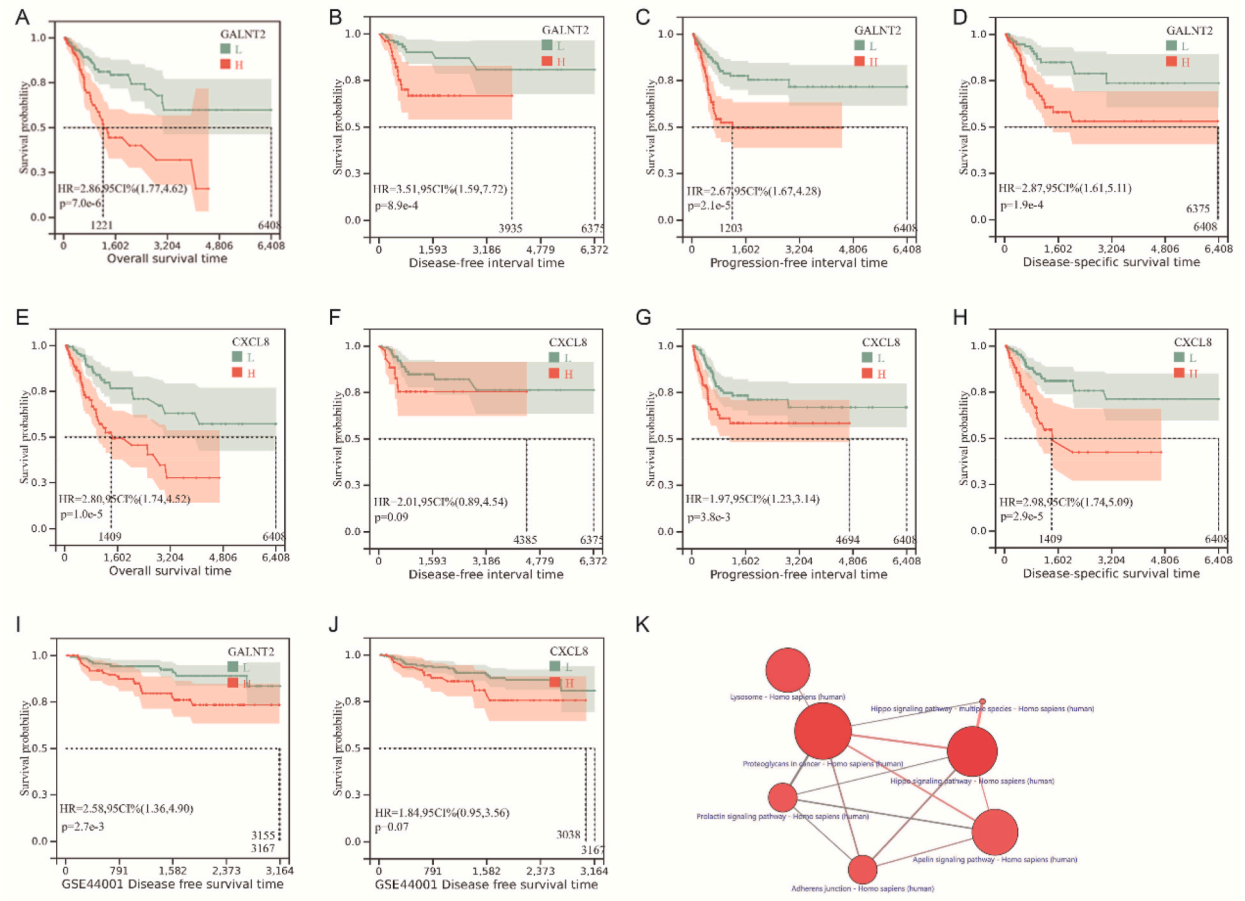



**Fig. 7.** Overexpression of GALNT2 and CXCL8 suggested poor prognosis of CESC patients. The poor prognosis of GALNT2 in OS (A), DFI (B), PFI (C), DSS (D) in TCGA-CESC cohort, and DFS in GSE44001 dataset (I). The poor prognosis of CXCL8 in OS (E), DFI (F), PFI (G), DSS (H) in TCGA-CESC cohort, and DFS in GSE44001 dataset (J). KEGG pathway analysis revealed a high co-expression between GALNT2 and important genes of proteoglycan and Hippo signaling pathways in cancer (K). OS means overall survival event; DSS means disease-specific survival event; DFI means disease-free interval event; PFI means progression-free interval event; DFS means disease free survival event.

glycosyltransferase family, GALNT2 is an enzyme that mediates the initial step of mucin type-O glycosylation and plays a key role in regulating cellular and molecular interactions as a widespread modification of proteins and lipids [28], such as cell signaling and communication, tumor cell dissociation and invasion, cell-matrix interactions, tumor angiogenesis, immune modulation and metastasis formation [29]. Subsequently, the potential biological function further certificated that silencing GALNT2 were indeed inhibited the proliferation, migration, and colony formation of HeLa cells and SiHa cells in vitro. Conversely, cell proliferation migration and colony formation were increased after the upregulation of GALNT2 in HeLa cells and SiHa cells. Moreover, we further investigated the influence of GALNT2 on EMT by Western blot and found that the levels of Snail and Vimentin proteins were reduced, E-cadherin was increased by GALNT2 knockdown in HeLa cells and SiHa cells. Conversely, the levels of Snail and Vimentin proteins were increased, E-cadherin was reduced by GALNT2 upregulation. Similarly, in gliomas, we observed that increased GALNT2 expression was related to an unfavorable prognosis and advanced tumor grade. GALNT2 knockdown decreased the level of phosphorylated EGFR and the expression of the Tn antigen on EGFR and affected the expression levels of p21, CDK4, cyclinD1, MMP2, and MMP9 through the EGFR/PI3K/Akt/mTOR pathway to inhibit glioma cell proliferation, migration, and invasion [30]. Similarly, overexpression of GALNT2 may activate the Notch/Hes1-PTEN-PI3K/Akt signaling axis to enhance cell line proliferation, migration, and invasion abilities in lung adenocarcinoma [31].

Interestingly, the clinical significance of malfunctioning GALNT2 expression observed in different cancers is inconsistent. The

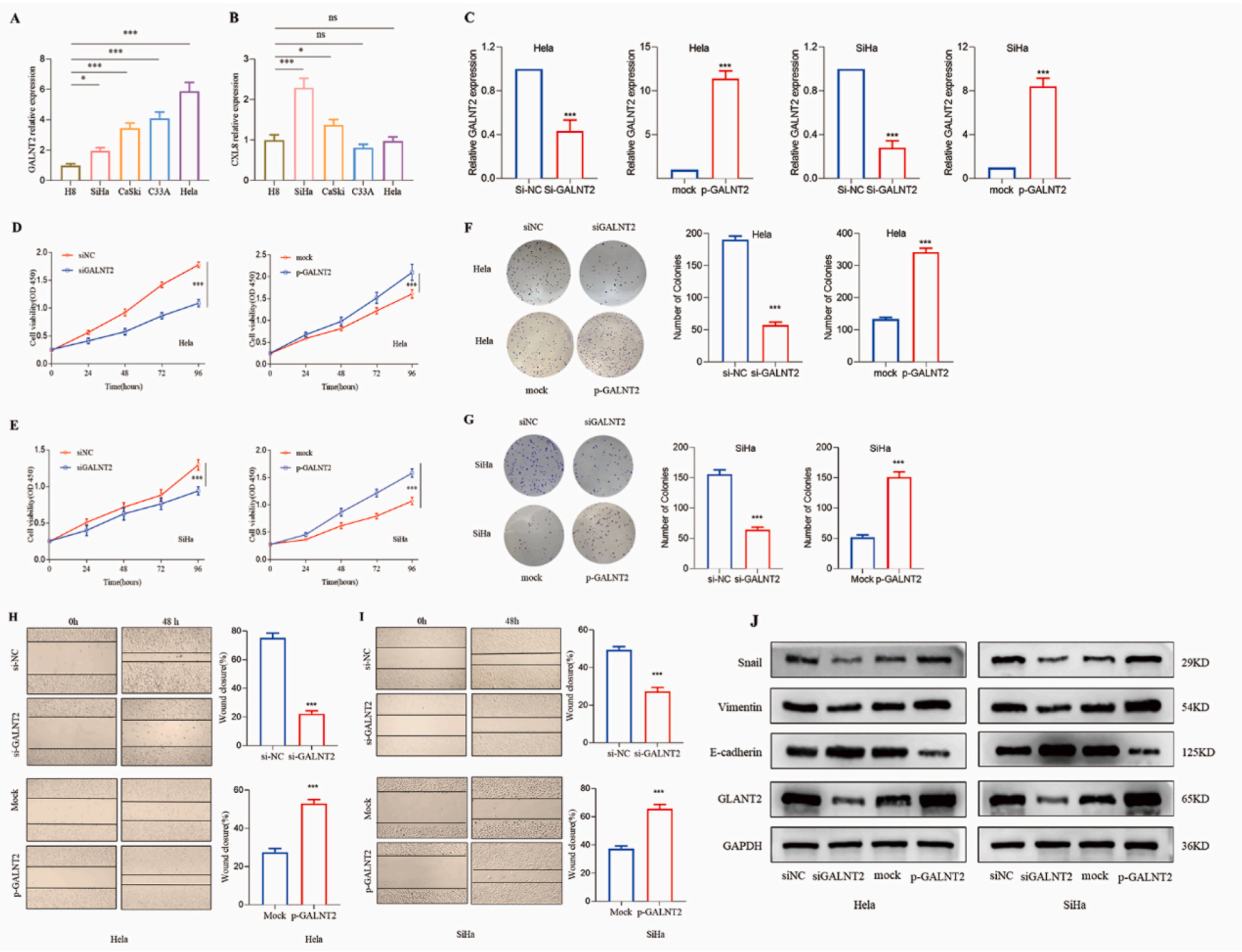



**Fig. 8.** In vitro validation and function of GALNT2. (A, B) qRT-PCR analysis of GALNT2 and CXCL8 mRNA levels in cell lines. (C, D) Cell proliferation of HeLa and SiHa cells was measured by the CCK8 assay. (E, F) The effect of GALNT2 knockdown or overexpression on colony formation in HeLa and SiHa cells. (G, H) The effect of GALNT2 knockdown or overexpression on cell migration ability in HeLa and SiHa cells. (I) Western blot analysis of Snail, Vimentin, E-cadherin in GLANT2 knockdown HeLa cells and overexpression SiHa cells compared to control cells. All data were presented as mean ± SD. *P < 0.05; **P < 0.01; ***P < 0.001.

previous study showed that mRNA and protein of GALNT2 are downregulated in gastric cancer, and this downregulation is related to more advanced disease and shorter RFS. Inhibition of GALNT2 expression increases tumor cell growth, migration, invasion, and tumor metastasis [32]. Moreover, GALNT2 knockdown increases epidermal growth factor receptor (EGFR) phosphorylation and decreases O-glycosylation of EGFR and expression of the Tn antigen on EGFR, enhancing the malignant phenotype of gastric cancer in vitro by increasing EGFR phosphorylation and activating the EGFR-Akt pathway [33]. Remarkably, we also observed a higher expression consistency between GALNT2 and important genes in the Proteoglycans in cancer and the Hippo signaling pathway (p-value <0.001). While Hippo/YAP pathway may interact with EGFR signaling and HPV oncoproteins E5 and E7 to regulate cervical carcinogenesis and progression [34].

Given that multiple oncogenic drivers usually coexist in cancers, which calls for multiple targeting in therapy. Therefore, understanding the effects and mechanisms of GALNT2 and O-glycosylation on cMET, EGFR or other receptor tyrosine kinase activity may be an option to directly target drivers in cancer therapy, which may provide a novel strategy for the development of CESC therapeutic agents. Remarkably, glycosylation inhibitors can reverse cancer-moderated inhibition of the immune system, which may have enormous implications for treating CESC patients on PD-1/PD-L1 immune checkpoint inhibitors [35].

Although the constructed classifier shows superiority in predicting the prognosis of CESC patients compared to previous features, as this study is only based on the TCGA database, and there are no suitable datasets to validate risk prediction features, therefore, a comprehensive in vitro experiment is needed further to explore the regulatory mechanism of these classifier-genes.

## 4. Conclusion

Based on Cox regression analysis, Random Forest and non-negative matrix factorization algorithms, we attempted to identify

molecular subtypes in 741 cervical cancer patients, and the constructed classifier will help to improve our understanding of subtype characteristics and may be utilized for patient prediction of prognosis and response to immunotherapy. Remarkably, GALNT2 may be an option to directly target drivers in CESC cancer therapy.

## Declarations

### *Acknowledgments*

This workwas supported by grants from State Key Laboratory of Causes and Prevention ofHigh Morbidityin Central Asia Project (SKL-HIDCA-2021-BZ2).

### *Data availability statement*

All dataset information were obtained from the public databases.

### *Ethics statement*

Not applicable.

### *Author contributions*

Liu Yang; Xia Han: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data.
XinKai Mo; Na Wang: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.
Lu Wang; Wenjun Kang; Xia Han: Analyzed and interpreted the data.

## 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.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e16873.

## References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA A Cancer J. Clin. 71 (2021) 209–249, https://doi.org/10.3322/caac.21660.
[2] WHO . Cervical cancer . World Health Organization, Available at: https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer.
[3] B.C. Das, S. Hussain, V. Nasare, M. Bharadwaj, Prospects and prejudices of human papillomavirus vaccines in India, Vaccine 26 (2008) 2669–2679, https://doi.org/10.1016/j.vaccine.2008.03.056.
[4] H. Van Meir, G.G. Kenter, J. Burggraaf, J.R. Kroep, M.J. Welters, C.J. Melief, et al., The need for improvement of the treatment of advanced and metastatic cervical cancer, the rationale for combined chemo-immunotherapy, Anti Cancer Agents Med. Chem. 14 (2014) 190–203, https://doi.org/10.2174/18715206113136660372.
[5] L.H. Zhang, L.Q. Li, Y.H. Zhan, Z.W. Zhu, X.P. Zhang, Identification of an IRGP signature to predict prognosis and immunotherapeutic efficiency in bladder cancer, Front. Mol. Biosci. 8 (2021), 607090, https://doi.org/10.3389/fmolb.2021.607090.
[6] F. Yang, M.A. Thomas, F. Dehdashti, P.W. Grigsby, Temporal analysis of intratumoral metabolic heterogeneity characterized by textural features in cervical cancer, Eur. J. Nucl. Med. Mol. Imag. 40 (2013) 716–727, https://doi.org/10.1007/s00259-012-2332-4.
[7] M.C. Asselin, J.P. O'connor, R. Boellaard, N.A. Thacker, A. Jackson, Quantifying heterogeneity in human tumours using MRI and PET, Eur. J. Cancer 48 (2012) 447–455, https://doi.org/10.1016/j.ejca.2011.12.025.
[8] Integrated genomic and molecular characterization of cervical cancer, Nature 543 (2017) 378–384, https://doi.org/10.1038/nature21386.
[9] X. Zhu, S. Li, J. Luo, X. Ying, Z. Li, Y. Wang, et al., Subtyping of human papillomavirus-positive cervical cancers based on the expression profiles of 50 genes, Front. Immunol. 13 (2022), 801639, https://doi.org/10.3389/fimmu.2022.801639.
[10] F. Xu, C. Zou, Y. Gao, J. Shen, T. Liu, Q. He, et al., Comprehensive analyses identify RIPOR2 as a genomic instability-associated immune prognostic biomarker in cervical cancer, Front. Immunol. 13 (2022), 930488, https://doi.org/10.3389/fimmu.2022.930488.
[11] H. Ji, J.A. Zhang, H. Liu, K. Li, Z.W. Wang, X. Zhu, Comprehensive characterization of tumor microenvironment and m6A RNA methylation regulators and its effects on PD-L1 and immune infiltrates in cervical cancer, Front. Immunol. 13 (2022), 976107, https://doi.org/10.3389/fimmu.2022.976107.
[12] S. Yu, X. Li, M. Ma, R. Yang, J. Zhang, S. Wu, The immunological contribution of a novel metabolism-related signature to the prognosis and anti-tumor immunity in cervical cancer, Cancers (2022) 14, https://doi.org/10.3390/cancers14102399.
[13] S. Yu, X. Li, J. Zhang, S. Wu, Development of a novel immune infiltration-based gene signature to predict prognosis and immunotherapy response of patients with cervical cancer, Front. Immunol. 12 (2021), 709493, https://doi.org/10.3389/fimmu.2021.709493.
[14] J. Liu, T. Lichtenberg, K.A. Hoadley, L.M. Poisson, A.J. Lazar, A.D. Cherniack, et al., An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics, Cell 173 (2018) 400–416.e11, https://doi.org/10.1016/j.cell.2018.02.052.

[15] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat. Genet. 25 (2000) 25–29, https://doi.org/10.1038/75556.
[16] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, M. Tanabe, KEGG: integrating viruses and cellular organisms, Nucleic Acids Res. 49 (2021) D545–D551, https://doi.org/10.1093/nar/gkaa970.
[17] T. Therneau, A Package for Survival Analysis in R. R Package Version 3, 5, 2023, p. 5, https://CRAN.R-project.org/package=survival.
[18] J. Stephan, O. Stegle, A. Beyer, A random forest approach to capture genetic effects in the presence of population structure, Nat. Commun. 6 (2015) 7432, https://doi.org/10.1038/ncomms8432.
[19] Y. Gao, G. Church, Improving molecular cancer class discovery through sparse non-negative matrix factorization, Bioinformatics 21 (2005) 3970–3975, https://doi.org/10.1093/bioinformatics/bti653.
[20] M. Kulis, M. Esteller, DNA methylation and cancer, Adv. Genet. 70 (2010) 27–56.
[21] T.W. Li, J.X. Fu, Z.X. Zeng, D. Cohen, J. Li, Q.M. Chen, B. Li, X.S. Liu, TIMER2.0 for analysis of tumor-infiltrating immune cells, Nucleic Acids Res. 48 (2020) W509–W514, https://doi.org/10.1093/nar/gkaa407.
[22] Charoentong Pornpimol, Finotello Francesca, Angelova Mihaela, et al., Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade, Cell Rep. 18 (2017) 248–262.
[23] H. Ha, B. Debnath, N. Neamati, Role of the CXCL8-CXCR1/2 Axis in cancer and inflammatory diseases, Theranostics 7 (6) (2017) 1543–1588.
[24] B. Zhu, N. Lin, M. Zhang, Y. Zhu, H. Cheng, S. Chen, Y. Ling, W. Pan, R. Xu, Activated hepatic stellate cells promote angiogenesis via interleukin-8 in hepatocellular carcinoma, J. Transl. Med. 13 (2015) 365.
[25] L. Li, M.N. Khan, Q. Li, X. Chen, J. Wei, B. Wang, J.W. Cheng, J.R. Gordon, F. Li, G31P, CXCR1/2 inhibitor, with cisplatin inhibits the growth of mice hepatocellular carcinoma and mitigates high dose cisplatin-induced nephrotoxicity, Oncol. Rep. 33 (2) (2015) 751–757.
[26] H. Tong, J.Q. Ke, F.Z. Jiang, X.J. Wang, F.Y. Wang, Y.R. Li, et al., Tumor-associated macrophage-derived CXCL8 could induce ERalpha suppression via HOXB13 in endometrial cancer, Cancer Lett. 376 (1) (2016) 127–136.
[27] L. Zhou, H. Wu, X. Bai, S. Min, J. Zhang, C. Li, O-glycosylating enzyme GALNT2 predicts worse prognosis in cervical cancer, Pathol. Oncol. Res. 28 (2022), 1610554, https://doi.org/10.3389/pore.2022.1610554.
[28] A. Varki, Biological roles of glycans, Glycobiology 27 (2017) 3–49, https://doi.org/10.1093/glycob/cww086.
[29] S.S. Pinho, C.A. Reis, Glycosylation in cancer: mechanisms and clinical implications, Nat. Rev. Cancer 15 (2015) 540–555, https://doi.org/10.1038/nrc3982.
[30] Z. Sun, H. Xue, Y. Wei, C. Wang, R. Yu, C. Wang, et al., Mucin O-glycosylating enzyme GALNT2 facilitates the malignant character of glioma by activating the EGFR/PI3K/Akt/mTOR axis, Clin. Sci. (Lond.) 133 (2019) 1167–1184, https://doi.org/10.1042/cs20190145.
[31] W. Wang, R. Sun, L. Zeng, Y. Chen, N. Zhang, S. Cao, et al., GALNT2 promotes cell proliferation, migration, and invasion by activating the Notch/Hes1-PTEN-PI3K/Akt signaling pathway in lung adenocarcinoma, Life Sci. 276 (2021), 119439, https://doi.org/10.1016/j.lfs.2021.119439.
[32] S.Y. Liu, C.T. Shun, K.Y. Hung, H.F. Juan, C.L. Hsu, M.C. Huang, et al., Mucin glycosylating enzyme GALNT2 suppresses malignancy in gastric adenocarcinoma by reducing MET phosphorylation, Oncotarget 7 (2016) 11251–11262, https://doi.org/10.18632/oncotarget.7081.
[33] W.T. Hu, C.C. Yeh, S.Y. Liu, M.C. Huang, I.R. Lai, The O-glycosylating enzyme GALNT2 suppresses the malignancy of gastric adenocarcinoma by reducing EGFR activities, Am. J. Canc. Res. 8 (2018) 1739–1751.
[34] C. He, D. Mao, G. Hua, X. Lv, X. Chen, P.C. Angeletti, et al., The Hippo/YAP pathway interacts with EGFR signaling and HPV oncoproteins to regulate cervical cancer progression, EMBO Mol. Med. 7 (2015) 1426–1449, https://doi.org/10.15252/emmm.201404976.
[35] Y.N. Wang, H.H. Lee, J.L. Hsu, D. Yu, M.C. Hung, The impact of PD-L1 N-linked glycosylation on cancer therapy and clinical diagnosis, J. Biomed. Sci. 27 (2020) 77, https://doi.org/10.1186/s12929-020-00670-x.