Discover Oncology

Analysis

Characteristic genes and immune infiltration analysis of gastric cancer based on bioinformatics analysis and machine learning

Chengwei Xia[1] · Yini Liu[2] · Xin Qing[3]

Received: 24 January 2025 / Accepted: 8 May 2025
Published online: 23 May 2025

Abstract

Background Gastric cancer (GC), a common and deadly malignancy worldwide, is a serious burden on society and individuals. However, available diagnostic biomarkers for GC are very limited. The current study aimed to identify potential diagnostic biomarkers for GC and analyze the activity of infiltrating immune cells in this pathology.

Methods Microarray data for GC were acquired from the Gene Expression Omnibus (GEO) database. The limma package was utilized to normalize these data, thus identifying differentially expressed genes (DEGs). For normalized data of samples, we established a weighted gene co-expression network (WGCNA) to reveal key genes in the significant module. Afterward, we obtained overlapping genes by intersecting the DEGs and the key genes from the WGCNA module. Next, after applying the three algorithms (LASSO, RandomForest, and SVM-RFE) to analyze these overlapping genes and take the intersection, we established a GC diagnosis. The diagnostic significances of these identified genes were evaluated with receiver operating characteristic (ROC) curves and validated in the external dataset. Furthermore, ssGSEA and CIBERSORT were employed for evaluating the infiltrating immune cells and the association of the immune cells and diagnostic biomarkers.

Results Herein, we identified 49 overlapping genes, and the results of enrichment analysis demonstrated that these genes may be involved in the signaling transduction-related process. Finally, BANF1, DUSP14, and VMP1 were regarded as key biomarkers in GC patients based on the overlapping genes that we found, and these three biomarkers demonstrated great diagnostic significance. Additionally, the hub biomarkers had different levels of association with macrophages, neutrophils, memory B cells, and plasma cells.

Conclusions BANF1, DUSP14, and VMP1 are promising diagnostic biomarkers for GC, and infiltrating immune cells may dramatically affect gastric carcinogenesis and progression.

Keywords Gastric cancer · Biomarker · WGCNA · Machine learning algorithm · Immune cell infiltration

Chengwei Xia and Yini Liu contributed equally to this work.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12672-025-02624-x.

✉ Xin Qing, qingxin1997@126.com; Chengwei Xia, 1076878246@qq.com; Yini Liu, 2453900253@qq.com | [1]Department of Thyroid and Breast Surgery, Chengdu Seventh People's Hospital (Affiliated Cancer Hospital of Chengdu Medical College), Chengdu, China. [2]Department of Anesthesiology, The People's Hospital of Zhongjiang, Deyang, China. [3]Department of Hepatobiliary Vascular Surgery, Chengdu Seventh People's Hospital (Affiliated Cancer Hospital of Chengdu Medical College), Chengdu, China.

Abbreviations

GC	Gastric cancer
GEO	Gene expression omnibus
WGCNA	Weighted gene co-expression network analysis
DEGs	Differentially expressed genes
GO	Gene ontology
KEGG	Kyoto Encyclopedia of Genes and Genomes
DO	Disease oncology
BP	Biological process
CC	Cellular component
MF	Molecular function
LASSO	Least absolute shrinkage and selection operator
RF	Random forest
SVM-RFE	Support vector machine-recursive feature elimination
ROC	Receiver operating characteristic
AUC	Area under the ROC curve
OS	Overall survival
PPS	Post progression survival
ssGSEA	Single sample gene set enrichment analysis

1 Introduction

Gastric cancer (GC) is one of the most common magnificence in the digestive system [1]. GC at an early stage has an excellent prognosis with a 5-year survival probability of over 90%, while the 5-year survival probability of progressive GC is only about 30% [2, 3]. Therefore, [4, 5]. Until now, the early diagnosis of GC has relied on gastroscopy and pathological judgment of biopsy tissue, and these methods are limited by their complexity and invasiveness [6, 7]. In recent years, with the development of high-throughput sequencing technologies, systematically identifying the diagnostic biomarkers of GC is a promising approach from multi-omics levels, thereby providing reliable support for the prevention of GC.

Currently, the primary methods for diagnosing GC include gastroscopy and histopathological examination of biopsy tissues. Although gastroscopy is considered the gold standard, its widespread application is limited by factors such as its invasiveness, high cost, and dependence on the operator's expertise [8]. Additionally, while histopathological examination offers high specificity, its sensitivity can vary, and false-negative results may occur due to sampling errors and tumor heterogeneity. Research reports indicate that the sensitivity of traditional gastroscopy for screening GC ranges from 69 to 89%, with specificity ranging from 69 to 96% [9]. Furthermore, serological tests based on biomarkers, such as carcinoembryonic antigen (CEA) and carbohydrate antigen 19-9 (CA19-9), lack sufficient sensitivity and specificity for reliable early diagnosis [10, 11]. Hence, there is an urgent need for novel, non-invasive, and highly accurate biomarkers to improve the early detection rate of GC.

Bioinformatics analysis is a notable and prospective visual modality for detecting potential biomarkers in various diseases [12, 13]. Data extraction and analysis from public databases enable personalized diagnosis and treatment for patients at the molecular level [14]. Specifically, the combination of weighted gene co-expression network analysis (WGCNA) and machine learning algorithms has substantially contributed to the precise detection of disease-associated biomarkers [15–17]. WGCNA, a biological network, can reveal the correlations between genes and biological functions, and specific genes and pathways associated with diseases are further identified [18]. Meanwhile, machine learning algorithms serve a promising role in exploring potential relationships, and making reliable predictions based on complex data [19]. The combination of WGCNA and machine learning algorithms can broaden the horizons into pathological and molecular mechanisms of diseases, providing valuable therapeutic targets for clinical treatment. For instance, SLC2A6 was regarded as an immunological biomarker of sepsis by combining WGCNA and least absolute shrinkage and selection operator (LASSO) logistic regression analysis [20]. IL1R2, IRAK3, and THBD were recognized to be key genes as promising biomarkers and therapeutic objectives for acute myocardial infarction by combining LASSO regression and support vector machine-recursive feature elimination (SVM-RFE) algorithms [21]. Undoubtedly, combining key approaches of machine learning (such as feature identification and classification) with WGNCA can effectively determine the diagnostic value of biomarkers [22, 23]. However, few reports on combining machine learning and WGCNA have explored valuable biomarkers of GC, especially the simultaneous application

of different machine learning methods. Therefore, there is a novel thinking to investigate the GC-related biomarkers based on multiple machine learning algorithms and WGCNA, thus providing reliable options for early diagnosis and treatment of GC.

Hence, the purpose of this study is to identify promising biomarkers of GC by combining WGCNA and machine learning algorithms (LASSO logistic regression, RF algorithm, and SVM-RFE). First, differentially expressed genes (DEGs) were identified between GC and control samples in the integrated dataset (GSE54129 and GSE65801). Next, WGCNA was executed based on the expression profiles of the integrated dataset to determine key genes in the key module. Machine learning algorithms were further utilized to determine hub biomarkers based on overlapping genes from DEGs and key genes. The diagnostic values of hub biomarkers were further evaluated with the additional dataset (GSE66229). Furthermore, we explored the correlation between infiltrating immune cells and hub biomarkers. This research may introduce novel biomarkers for the diagnosis and management of GC and facilitate the clarification of the pathogenesis of gastric cancer.

2 Methods

2.1 Data collection and preprocessing

The datasets included in this study were obtained from the Gene Expression Omnibus (GEO) public database [24], including GSE54129, GSE65801, and GSE66229. These datasets focused on the gene sequencing results of GC patients, and each dataset contains more than 60 samples (Table 1). GSE54129 and GSE65801 were further integrated, and the batch effects between different reports and platforms were eliminated with combat algorithm in the sva package version 3.46.0 [25]. The integrated dataset was regarded as a training cohort, and GSE66229 was regarded as a validation cohort. Additionally, the TCGA-GTEx dataset was further extracted to valitae the expression levels of hub genes. All analyses were performed using R software (version 4.2.3).

2.2 Identification of differently expressed genes

In the training cohort, the "limma" package version 3.54.0 was utilized to normalize the expression matrix and identify differentially expressed genes (DEGs) between the GC samples and control samples [26]. The limma package was applied to perform genomic analysis of inter-specimen discrepancies, and diverse hypothesis examination and correction were further carried out. The p-value threshold was determined by controlling for the false discovery rate, and the corrected p-value was adjusted p-value [27]. The cut-off value was set as $|\log2FC| > 1$ and adjusted p-value < 0.05. Then, the volcano plot and heatmap were displayed to show the expression status of DEGs.

2.3 Establishment of weighted gene co-expression network analysis (WGCNA)

WGCNA is a bioinformatics approach for introducing patterns of gene correlations between different specimens, and revealing gene module information with biological significance [28, 29]. First, the Pearson correlation coefficient between paired genes was computed to establish the correlation matrix. Next, this matrix was transferred into a weighted neighborhood matrix based on the soft threshold feature. Afterward, the neighborhood matrix was further converted into a topological overlap matrix (TOM) revealing the correlative levels between genes. 1-TOM was regarded as the distance for clustering the genes, and the dynamic tree chopping was constructed to determine the module. The least gene number in the modules was set to 50. Finally, we determined 13 modules by adjusting the merging threshold to 0.25. After the individual module was identified according to the key gene expression data and the sample classification, the association of the module key genes with sample classifications was also identified.

Table 1 Information on microarray datasets obtained from GEO database

Dataset	Platform	GC	Control
GSE54129	GPL570	111	21
GSE65801	GPL14550	32	32
GSE66229	GPL570	300	100

2.4 Functional enrichment analysis

The overlapping genes from DEGs and the key module were selected to perform functional enrichment analysis based on the "clusterProfiler" package version 4.10.1 [30]. Gene ontology (GO) enrichment analysis was conducted to explore gene-associated biological processes (BP), encompassing cellular components (CC) and molecular functions (MF). Kyoto encyclopedia of genes and genomes (KEGG) enrichment analysis was performed to identify gene-related signaling pathways. Disease Ontology (DO) enrichment analysis was executed to identify gene-related diseases. Adjusted P value < 0.05 was considered statistically significant.

2.5 Screening of hub biomarkers

The abovementioned genes were further utilized to identify significant feature genes, thus diagnosing GC. The feature identification approach is a procedure of limiting the number of factors, specifically vital for establishing a predictive model [31]. The (LASSO, random forest (RF) algorithm, and SVM-RFE were included in this study to explore feature genes. The "glmnet" package version 4.1-8 was applied to conduct minimum LASSO regression, thus choosing the linear model and keeping the reliable variables [32]. Binomial distribution variables were further presented in the LASSO categorization, with a standard error value as the minimum parameter. Next, according to various dependent decision trees from a training pool, the RF algorithm promotes the precision of the model by randomly limiting the overfitting of individual decision trees [33, 34]. SVM-RFE can identify optimal parameters by removing the SVM-derived eigenvectors [35]. An SVM module based on the "e1071" package version 1.7-14 was created to further evaluate the diagnostic value of the selected biomarker in GC [36]. The intersected genes, as the most significant feature genes from these three algorithms, were identified for subsequent analysis.

2.6 Diagnostic nomogram construction and validation

The expression difference of hub biomarkers was presented in box plots based on the limma package and ggpubr package version 0.6.0 [8, 26], and the receiver operating characteristic (ROC) curve was displayed to indicate the area under the curve (AUC) for identifying hub genes, and access the diagnostic significance of these genes in GC with the "ROCR" package version 1.0-11 [37]. The nomogram was also established with the "rms" package version 6.7-0 [38]. Similarly, the expression difference and diagnostic significance of the hub biomarkers were also evaluated with a validation cohort (GSE66229) based on differential analysis and ROC curves. Additionally, the prognostic significance of these biomarkers in GC was investigated on the Kaplan–Meier online website (https://kmplot.com/analysis/) [39].

2.7 Immune cells infiltration analysis

The single sample gene set enrichment analysis (ssGSEA) and CIBERSORT algorithms were applied to analyze the normalized gene expression data in the integrated dataset, and the fraction of immune cells was identified [40, 41]. For ssGSEA, we utilized the immune cell gene signatures curated from 24 human hematopoietic cell types published by Bindea et al.[42], which include 18 adaptive/innate immune cell subtypes (e.g., T helper cells, cytotoxic T cells, dendritic cells) and 6 non-immune stromal components. CIBERSORT analysis was performed with the LM22 leukocyte gene matrix containing 22 functionally defined immune cell types [43]. The relative fractions of immune cells were calculated using the GSVA R package (v1.46.0) for ssGSEA and the CIBERSORT web tool (https://cibersort.stanford.edu) with 1000 permutations.Violin plots were displayed to present the expressional difference of the immune infiltrating cells. And Spearman correlation analysis was executed to investigate the association between immune infiltrating cells and hub biomarkers [44]. These results were visualized with the "ggplot2" package version 3.5.1 [45]. P-values < 0.05 demonstrated statistical significance.

3 Results

3.1 Identification of DEGs between GC and control samples

The integrated dataset composed of GSE54129 and GSE65801 was applied to determine DEGs between GC and control samples. Totally, 913 DEGs were determined with a criterion of |log2FC|> 1 and adjusted p-value < 0.05. 415 DEGs were

up-regulated in the GC samples in comparison to control samples, while 498 DEGs were down-regulated (Fig. 1A). The top 30 enriched DEGs in samples were presented in Fig. 1B. To investigate the specific developmental mechanism of GC, these up-regulated genes in GC were selected for subsequent analysis.

3.2 Establishment of a co-expression network and hub module

WGCNA was further performed to identify the key gene correlated with the biological process of GC. The outliers and missing values were corrected by sample clustering, and the soft threshold was set to 4 (scale-free $R^2 = 0.88$; slope = −1.64) to align with the scale-free network (Fig. 2A, B). The co-expression matrix was established based on a one-step approach, and a total of 13 gene modules were acquired from dynamic hybrid shearing (Fig. 2C). The relationship of these gene modules with GC and adjacent controls was displayed in the heatmaps (cyan module as a hub module with 701 genes) presenting the strongest correlation (cor) with GC (cor = 0.77; p-value = 2e−39) (Fig. 2D). The genes from the cyan module were considered key genes involved in the development of GC.

3.3 Functional enrichment analysis of overlapping genes based on DEGs and WGCNA

Based on the abovementioned results, the intersection of DEGs and key genes identified 49 overlapping genes (Fig. 3A). To explore the molecular function of overlapping genes correlated with GC, enrichment analyses were conducted. As shown in Fig. 3B, the overlapping genes were markedly enriched in signaling transduction and immune-related pathways, such as receptor-ligand activity and regulation of cytokines. The findings of KEGG enrichment analyses demonstrated that the DEGs were significantly associated with immune- and metabolism-related pathways (Fig. 3C). Additionally, DO enrichment analysis was also conducted, and these overlapping genes were correlated with the development of stomach cancer (Fig. 3D). In summary, these findings ascertained the physiological procedures and abnormal signaling pathways participating in the development of GC.

3.4 Screening of hub genes via machine learning algorithms

To determine the potential hub genes in GC samples from 49 overlapping genes, machine learning algorithms were chosen. The LASSO regression analysis was first conducted. Employing the LASSO model with a minimum of λ, 14 of

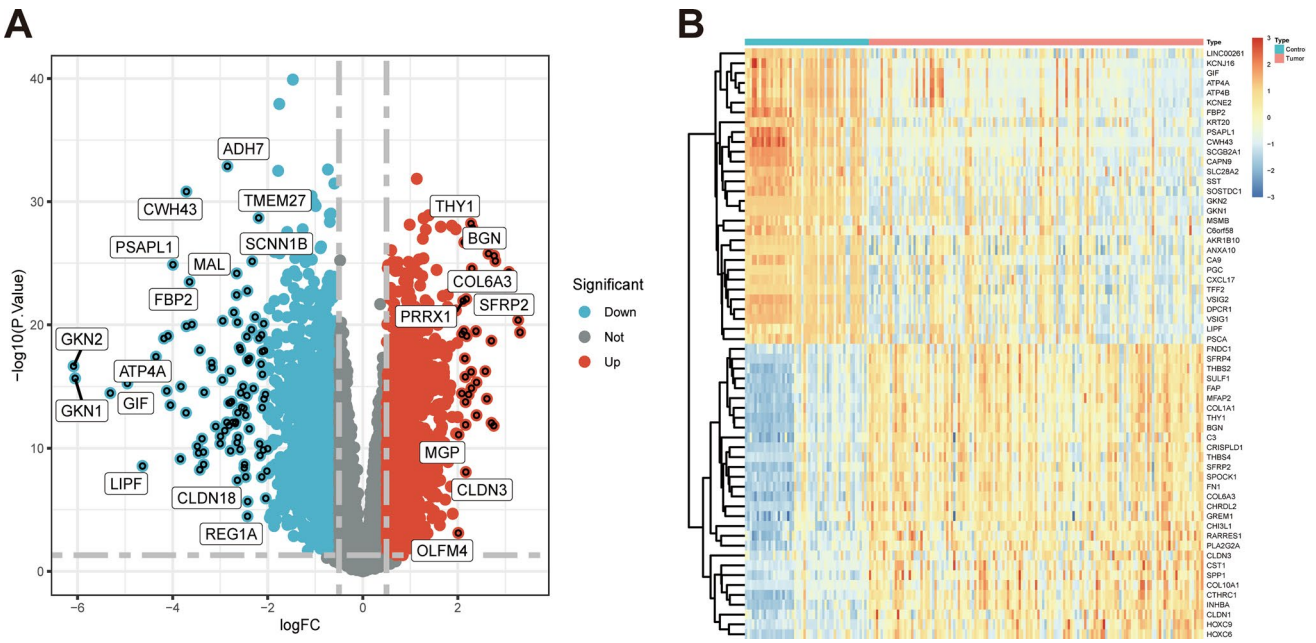

Fig. 1 Identification of DEGs. **A** Volcano plot of DEGs between the GC samples and control samples. The red plots represent upregulated genes, the black plots represent nonsignificant genes, and the green plots represent downregulated genes. **B** Heatmap of DEGs between the GC samples and control samples. Red rectangles represent a high expression, and green rectangles represent a low expression

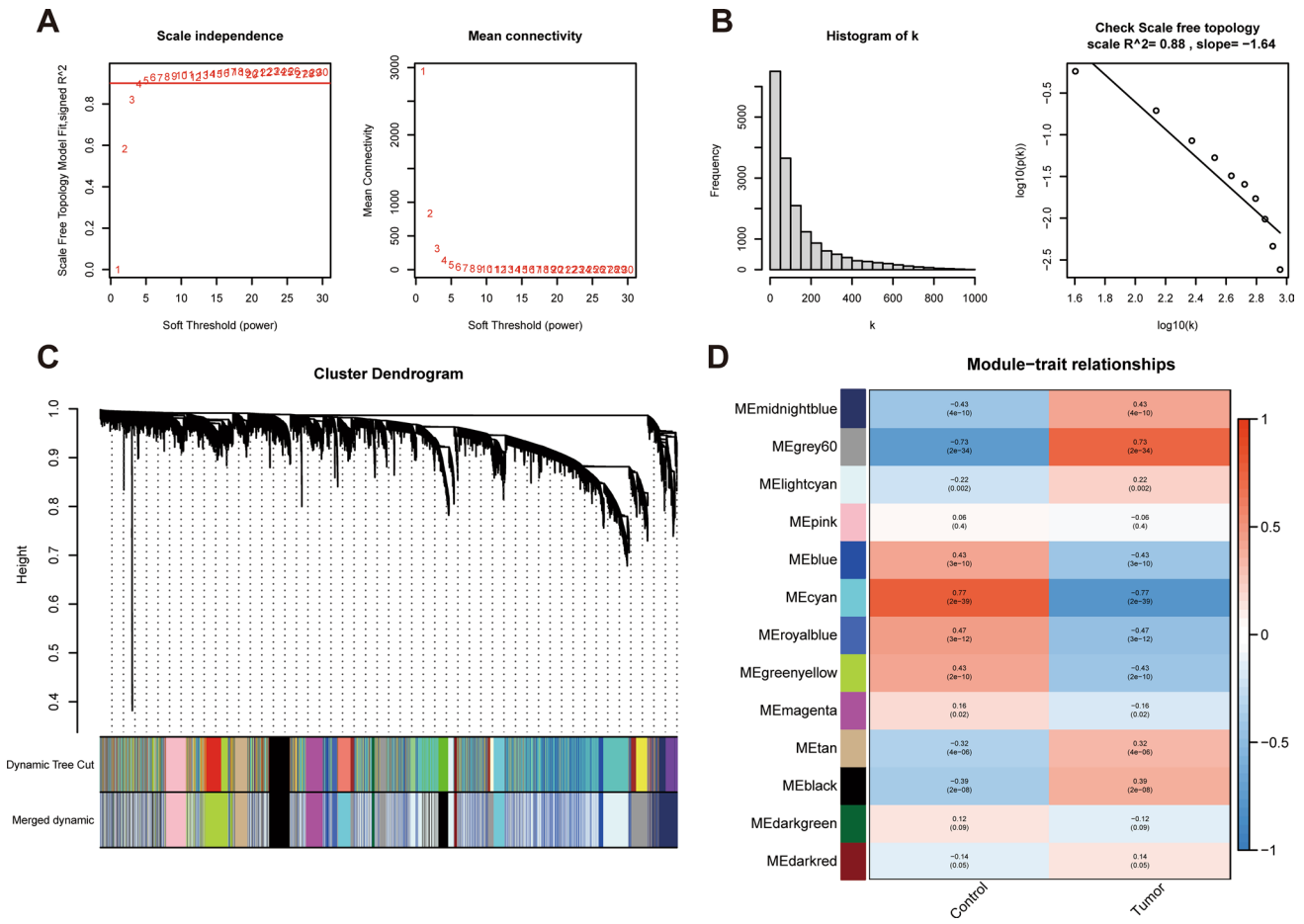

Fig. 2 Construction of the weighted gene co-expression network analysis (WGCNA). **A** Determination of soft-thresholding power in the WGCNA. **B** Histogram of connectivity distribution and checking the scale-free topology when β = 4. **C** The cluster dendrogram of the genes. Each branch in the figure represents one gene, and every color below represents one co-expression module. **D** Heatmap of the module-trait relationships. The cyan module was significantly correlated with GC

the 49 DEGs were considered to further establish a model sufficient to quantify the individual (Fig. 4A). Afterward, the RF algorithm was also applied. Based on the plot of the model error versus the number of decision trees, 500 trees were selected as the variables of the definite model, demonstrating the great reliability of this model (Fig. 4B). The importance of the top 20 genes from overlapping genes was displayed, and the most significant 10 genes in importance were selected to further analysis. Additionally, SVM-RFE analysis demonstrated that the SVM model based on the overlapping genes presented the best accuracy rate (0.939) and error rate (0.0611, Fig. 4C). Finally, the intersection genes of three machine learning algorithms were found and further considered as hub genes (Fig. 4D). These 3 hub genes (BANF1, DUSP14, and VMP1) were used for subsequent analysis.

3.5 Evaluation of diagnostic values for hub genes

The expression differences of the 3 hub genes were presented in Fig. 5A. We observed that these genes demonstrated an higher expression level in the GC samples compared to control samples. Meanwhile, the ROC curves were utilized to assess the diagnostic robustness of these genes. The AUC values of these 3 genes were over 0.700, implying that these genes had an excellent diagnostic value for GC (Fig. 5B). Afterward, a nomogram model based on the integrated dataset was established to predict the incidence of GC patients (Fig. 5C), and ROC curves demonstrated that the predictive power of the nomogram model was great (Fig. 5D).

The expression status of these 3 hub genes was also evaluated in GSE66229. The results from GSE66229 were consistent with the abovementioned differential expression analysis (Fig. 6A). To further validate their diagnostic reliability, the clinical value of these hub genes was also confirmed. The VMP1 had an AUC value of 0.642, while other genes had an AUC

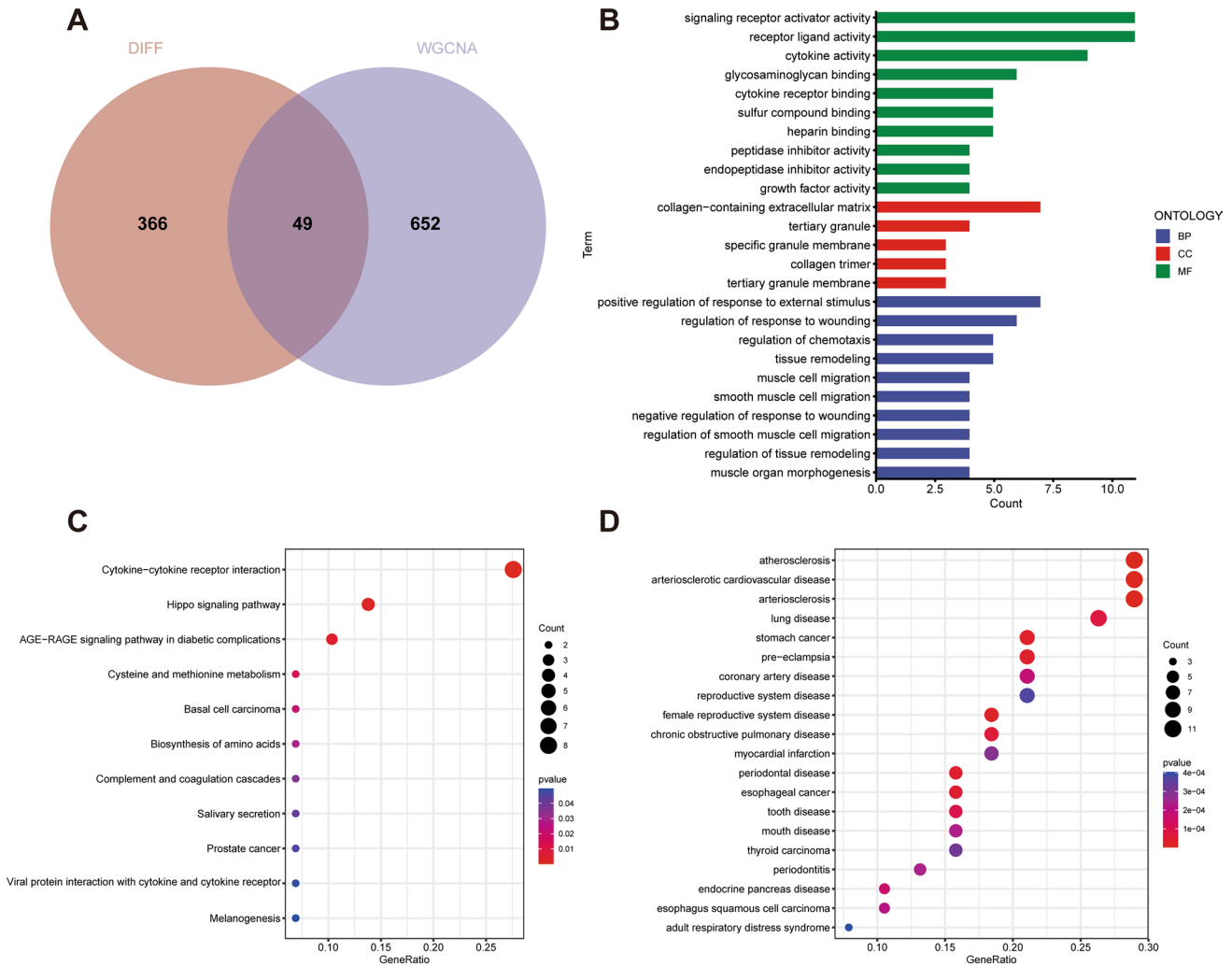

Fig. 3 Functional enrichment analysis of overlapping genes. **A** Venn diagram for intersections between DEGs and the cyan module. **B** GO enrichment analysis from overlapping genes. **C** KEGG enrichment analysis from overlapping genes. **D** DO enrichment analysis from overlapping genes

value of over 0.700 (Fig. 6B). Additionally, the prognostic significance of these 3 hub genes for GC patients was discussed by the Kaplan–Meier online tool, and the prognostic values of these genes were displayed in Fig. 6C. The low expression of BANF1 and DUSP14 were correlated with the superior prognosis of GC patients, while this correlation was not discovered between VMP1 and prognosis. Meanwhile, the expression of BANF1 and DUSP14 were negatively correlated with the PPS, while the opposite performance was observed in VMP1 (Fig. 6D). We further analyzed the expressive levels of these hub genes in TCGA-GTEx dataset, and the results was presented in Figure S1.

3.6 Analysis of immune cell infiltration

To explore the discrepancy in immune cell infiltration between GC and control samples, the ssGSEA algorithm was applied to evaluate their association in the integrated dataset. The layout of 28 immune cells in the GC samples was shown in Fig. 7A, B. The finding of the immune infiltrating analysis indicated a markedly greater level of CD4 + T cells, CD8 + T cells, natural killer (NK) cells, and dendritic cells in the GC samples compared to control samples, indicating that these immune cells are significantly associated with the development of GC. Moreover, the relationship of hub genes with 28 immune cells was explored. As shown in Fig. 7C, DUSP14 was positively correlated with NK T cell, mast cell, macrophage, and γ δ T cell. And VMP1 was positively correlated with activated dendritic cells. These findings provided further evidence for the essential role of these immune cells in the development of GC.

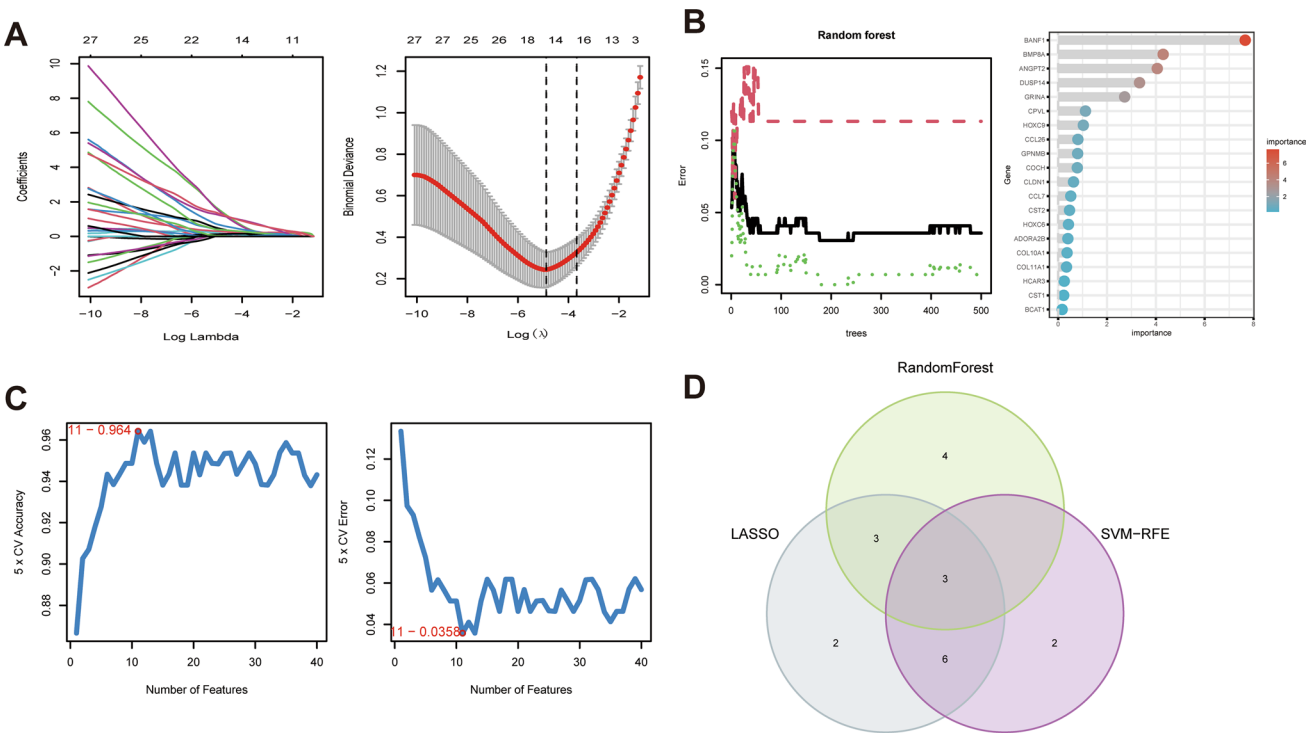

Fig. 4 Identification of potential biomarkers for GC based on machine learning algorithms. **A** Establishment of the LASSO model. **B** Screening biomarkers based on random forest (RF) machine learning algorithm. **C** Results of screening biomarkers based on SVM-RFE algorithm. **D** Venn diagram shows the intersected genes in LASSO, RF, and SVM-RFE algorithms

3.7 Correlation between hub biomarkers and infiltrating immune cells

The CIBERSORT algorithm was also utilized to generate an enrichment abundance index of 22 immune cell species in the integrated dataset. The fraction of naïve B cell, activated CD4 T cell, activated NK cell, M0/M1 macrophage, and neutrophil was markedly higher in the GC samples than in the healthy control, while the fraction of plasma cell, resting memory CD4 T cell, and resting NK cells was markedly lower in the GC tissues compared with the control tissues (Fig. 8A). The correlation of 22 kinds of immune cells was further discussed (Fig. 8B). Meanwhile, the correlation between these hub genes and infiltrating immune cells was analyzed. As displayed in Fig. 8C, BANF1 was positively associated with M0/M1 macrophage, activated NK cells, naïve B cells, monocytes, activated memory CD4 T cells, and follicular helper T cells, while the opposite correlation was observed in memory B cells, memory resting CD4 T cells, and plasma cell (Fig. 8C). A similar correlation between other biomarkers (DUSP14 and VMP1) and infiltrating immune cells was also observed (Fig. 8D, E)

4 Discussion

The available machine learning algorithms enable bioinformatic technology to more accurately and rapidly identify hub biomarkers correlated with disease initiation and development, allowing for disease diagnosis, treatment, and therapeutic agents' investigation. Given the lack of obvious symptoms in the early stage of gastric cancer [46], sensitive tumor biomarkers are essential to maximizing the benefits of personalized treatment. Therefore, applying machine learning algorithms to investigate tumor biomarkers is a promising application direction.

In this research, we first determined 415 up-regulated DEGs in the GC samples compared with healthy control in the integrated dataset. Next, 49 genes of up-regulated genes were considered intersection genes based on DEGs and key genes from WGCNA. These genes were observed to be associated with signaling transduction-related processes,

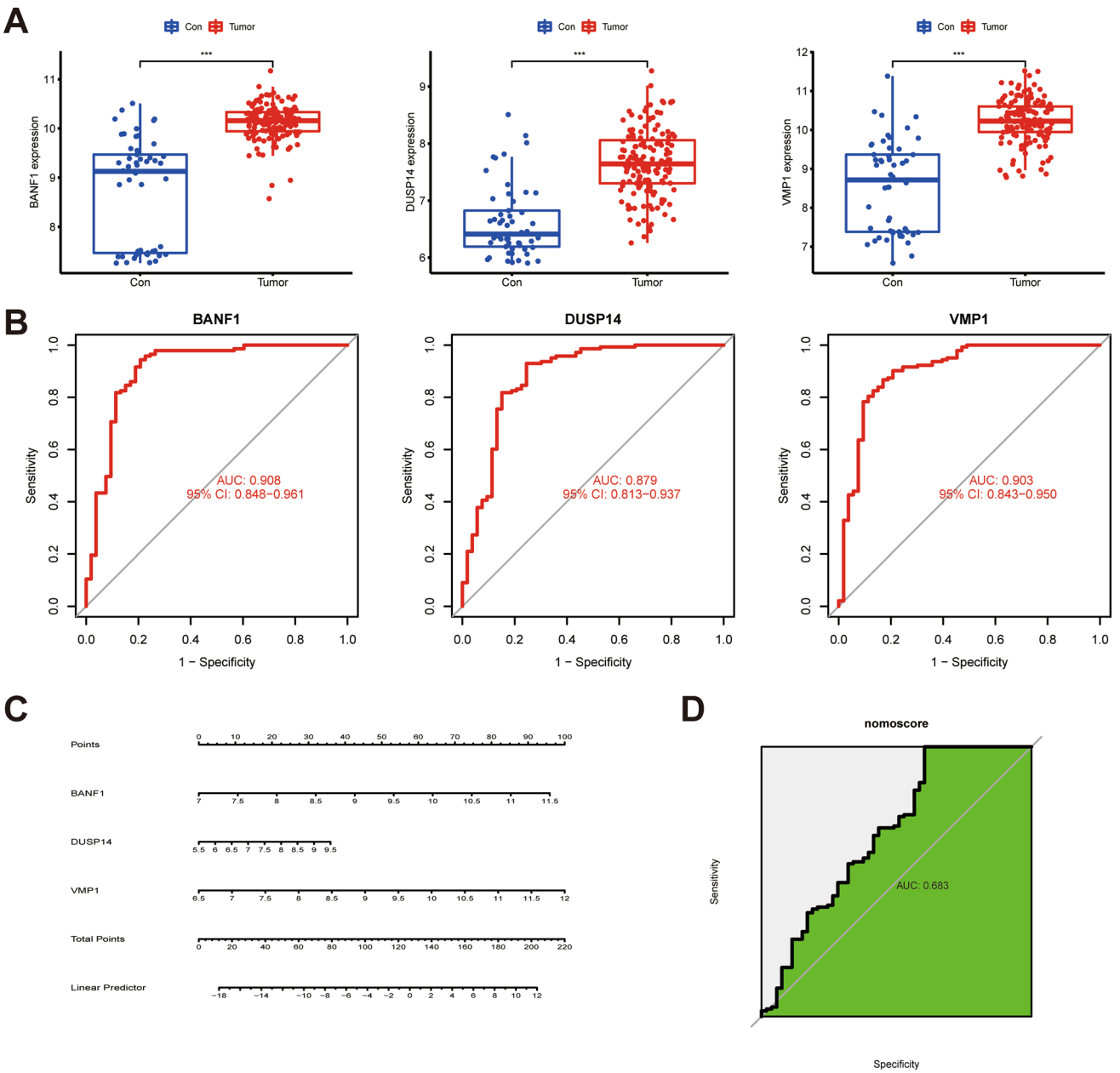

Fig. 5 Expression analysis of 3 hub biomarkers (**A**) expressional difference of 3 hub biomarkers between GC samples and control samples in the integrated dataset. **B** ROC curves of 3 hub biomarkers in the integrated dataset. **C** Establishment of a nomogram based on 3 hub biomarkers. **D** The ROC curve of the nomogram ($p < 0.001$***)

which may be correlated with the progression of GC. We hypothesized that these genes could serve a pivotal part in GC by modulating signal transduction and stress response. Emerging reports have also demonstrated that stress response was responsible for the development of tumors [47–49]. Hence, this research may facilitate the elucidation of the molecular mechanism of GC.

To further optimize the availability of GC-related biomarkers for pre-screening objectives, diverse machine-learning algorithms were performed, including LASSO logistic regression, RF algorithms, and SVM-RFE. LASSO logistic regression identifies variables by looking for λ based on the minimum incidence of categorical error [50]. RF is composed of an integrative decision tree, where an individual internal node represents a test of the categorical property [51]. SVM-RFE is consistent with the statistical learning principle and determines the optimal parameters by subtracting the developed feature vectors [52]. Based on the intersection of feature genes from these algorithms, BANF1, DUSP14, and VMP1 were identified as promising diagnostic biomarkers for GC.

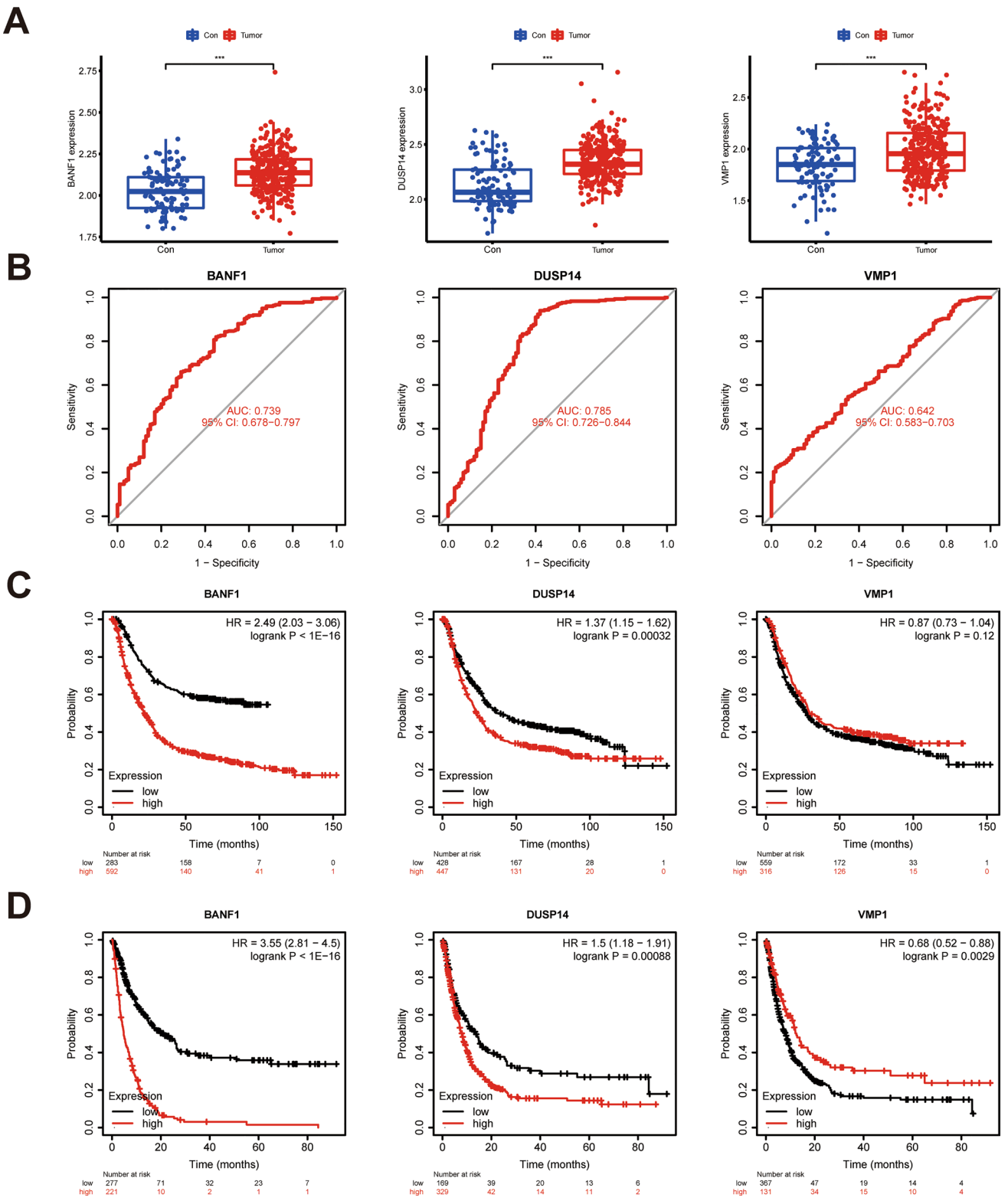

Fig. 6 Validation of hub biomarkers in the external dataset. **A** Expressional difference of 3 hub biomarkers between GC samples and control samples in the GSE66229 dataset. **B** ROC curves of 3 hub biomarkers in the GSE66229 dataset. **C**, **D** Survival curves (OS and PPS) were established by the Kaplan–Meier plotter online database based on the low and high expression of the hub genes in GC patients ($p < 0.001$***)

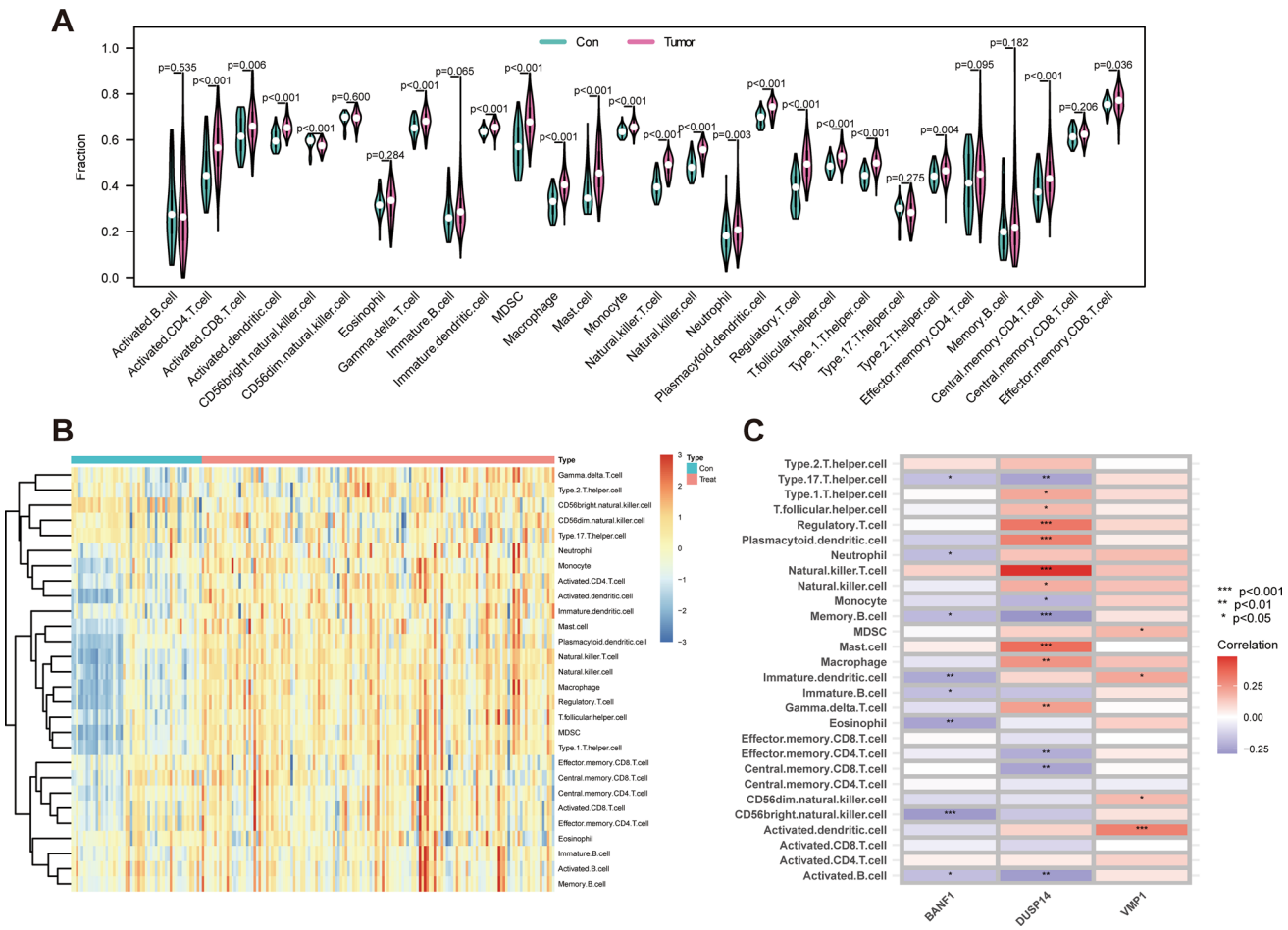

Fig. 7 Analysis of GC-related immune landscape with ssGSEA algorithm in the integrated dataset. Violin plot (**A**) and Heatmap (**B**) displays the distribution of 28 types of immune cells in GC samples and control samples. **C** The relationship between 3 hub genes and immune cell infiltration

The expression levels of the three hub biomarkers were different between GC and control samples. Certainly, these genes were a significantly higher expression in the GC samples. The specificity and sensitivity for the diagnosis of GC were also accessed, and the results demonstrated that these genes were reliable diagnostic biomarkers. Subsequently, the external dataset was applied to evaluate the expression status of three hub biomarkers in GC and control samples, and the findings were consistent with the previous conclusion from the integrated dataset. Meanwhile, the prognostic significance of these biomarkers was explored on the Kaplan–Meier website, and we observed that high expression of BANF1 and DUSP14 were significantly associated with poorer prognosis in GC patients. However, no significant association was noted between VMP1 expression levels and survival outcomes. The varying prognostic performance of VMP1 across different cohorts may reflect the evolution of the tumor microenvironment during tumor progression. VMP1 is closely associated with early outcomes in the GEO cohort and EMT features across all datasets, suggesting its relevance to the treatment of localized disease.

The expression of three hub biomarkers was disordered in GC samples in comparison to control samples, implying that these biomarkers could be essential to the occurrence and development in GC. Despite diverse reports that have discussed three hub biomarkers associated with the management and prognosis of several tumors, their potential roles in initiation and development in GC are not yet adequately known [53, 54]. BANF1 (BAF Nuclear Assembly Factor 1) has been discovered to be correlated with the integration of retroviral DNA [55]. Moreover, BANF1 can regulate various cellular processes, such as protein dimerization, it's binding to DNA, and subcellular localization of the protein [56]. Thus, BANF1 may be critical to supporting an inherent cellular genome. DUSP14 (Dual Specificity Phosphatase 14), a widely expressed phosphatase, is involved in the negative regulation of apoptosis in gastrointestinal tumor cells [57]. More importantly, DUSP14 can regulate the immune response by modulating the phosphatase activity of MAPK substrates

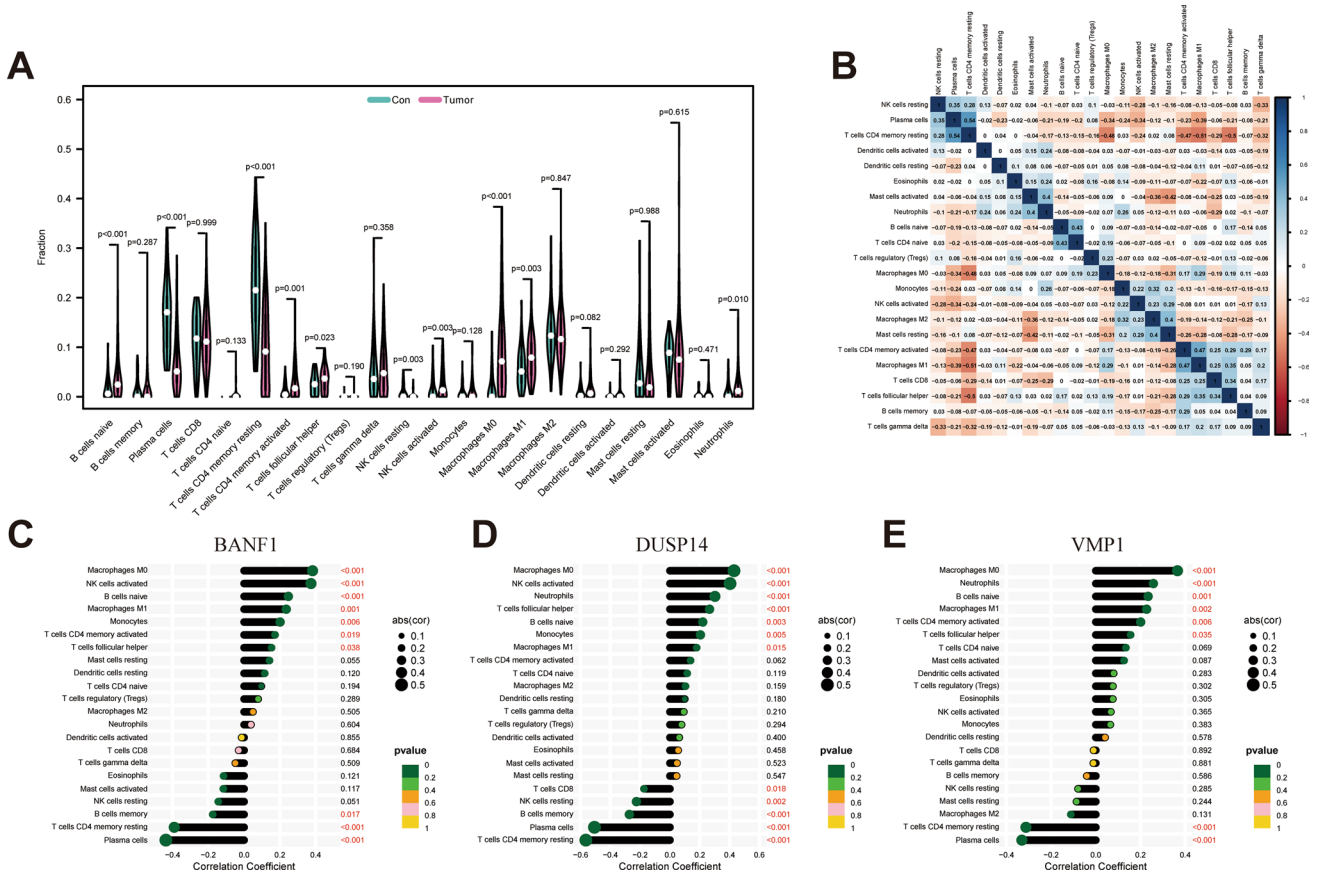

Fig. 8 Analysis of GC-related immune landscape with CIBERSORT algorithm in the integrated dataset. **A** Violin plot of the estimated fraction of 22 types of immune cells between GC and control samples. **B** Correlation heat map of 22 types of immune cells. A positive and negative correlation was respectively shown in blue and red color, whereas the number represents the correlation parameters. **C**–**E** Correlation map of 22 types of immune cells and 3 hub biomarkers

[58]. Therefore, DUSP14 may serve a crucial function in GC by modulating apoptosis and immunological activity. VMP1 (Vacuole Membrane Protein 1), endoplasmic reticulum-resident transmembrane proteins for the regulation of autophagy, has been observed to be associated with lipids distribution by modulating cholesterol and phosphatidylserine [59, 60]. Hence, VMP1 may serve a decisive role in GC by modulating autophagy and lipid metabolism.

Then, this study adopted ssGSEA and CIBERSORT to assess the role of the infiltrating immune cells in GC. Based on the finding of this study, an obvious enrichment of macrophages and neutrophil might be correlated with GC occurrence and progression. As is well known, although they have a typical protective role in the immune system, macrophages can be selected by tumor cells to promote tumor growth [61]. Tumor-associated macrophage is one of the elements of the immunosuppressive myeloid microenvironment and promotes local immune escape when polarized by diverse signaling pathways [62]. Similar to macrophages, Tumor-associated neutrophils are also an important component of the immunosuppressive myeloid microenvironment [63]. The ratio of different subtypes of neutrophils has important implications for tumor development, both killing tumor cells and promoting tumor growth in different situations. Neutrophils can also directly promote tumor development, metastasis, and angiogenesis [64]. Similarly, other immune cells serve a crucial part in the tumor microenvironment of GC, such as follicular helper T cells and NK cells [64]. These findings reveal a significant correlation between the immune microenvironment and the development of GC. We also explored the relationship between three hub biomarkers and immune cells, and the findings displayed a significant correlation with multiple immune cells. This correlation reflects the unfavorable role of hub genes in the immune microenvironment, thus affecting the progression of GC.

Inevitably, there are certain deficiencies and limitations in this research. The findings of our study should be further confirmed in vivo or in vitro studies. Also, the prognostic value of these hub biomarkers is pending to be evaluated in the external database with complete survival information. Therefore, further experimental and prospective studies are

necessary in the future, and future studies employing single-cell sequencing and pathway perturbation models will clarify how these biomarkers orchestrate immune evasion through specific molecular cascades.

5 Conclusions

Taken together, based on bioinformatics analysis of WGCNA and three machine learning algorithms (LASSO, Random Forest, and SVM-RFE), three hub biomarkers (BANF1, DUSP14, and VMP1) were screened, and these biomarkers may participate in the development of GC. Moreover, this research presents promising perspectives for the immune infiltration landscape of GC and its underlying immunomodulatory mechanisms. Further investigation of these biomarkers and corresponding immunological significance will contribute to understanding the pathogenesis of GC and provide guidance for clinical diagnosis and targeted agent exploration.

Acknowledgements Not applicable.

Author contributions All authors contributed to the study's conception and design. XQ performed data collection and analysis. CX and XQ wrote the manuscript. YL and XQ polished and revised the manuscript. All authors commented on previous versions of the manuscript and read and approved the final manuscript.

Funding None.

Data availability The datasets presented in this study can be found in the GEO database (GSE54129, GSE65801, and GSE66229).

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare no competing interests.

References

1. Joshi SS, Badgwell BD. Current treatment and recent progress in gastric cancer. CA Cancer J Clin. 2021;71:264–79.
2. Thrift AP, El-Serag HB. Burden of gastric cancer. Clin Gastroenterol Hepatol. 2020;18:534–42.
3. Smyth EC, Nilsson M, Grabsch HI, van Grieken NC, Lordick F. Gastric cancer. Lancet. 2020;396:635–48.
4. Wang FH, Zhang XT, Li YF, Tang L, Qu XJ, Ying JE, Zhang J, Sun LY, Lin RB, Qiu H, et al. The Chinese Society of Clinical Oncology (CSCO): clinical guidelines for the diagnosis and treatment of gastric cancer, 2021. Cancer Commun (Lond). 2021;41:747–95.
5. Necula L, Matei L, Dragu D, Neagu AI, Mambet C, Nedeianu S, Bleotu C, Diaconu CC, Chivu-Economescu M. Recent advances in gastric cancer early diagnosis. World J Gastroenterol. 2019;25:2029–44.
6. Young E, Philpott H, Singh R. Endoscopic diagnosis and treatment of gastric dysplasia and early cancer: current evidence and what the future may hold. World J Gastroenterol. 2021;27:5126–51.
7. Sasahara M, Kanda M, Kodera Y. Update on molecular biomarkers for diagnosis and prediction of prognosis and treatment responses in gastric cancer. Histol Histopathol. 2021;36:817–32.
8. Zou J, Zheng L, Shuai W, Li Q, Wang Q, Zhang Z, Li D. Comparison of intra-abdominal pressure measurements in critically ill patients using intravesical normal saline at 15 degrees C, 25 degrees C, and 35 degrees C. Med Sci Monit. 2021;27: e932804.
9. Liou JM, Malfertheiner P, Lee YC, Sheu BS, Sugano K, Cheng HC, Yeoh KG, Hsu PI, Goh KL, Mahachai V, et al. Screening and eradication of Helicobacter pylori for gastric cancer prevention: the Taipei global consensus. Gut. 2020;69:2093–112.
10. Jelski W, Mroczko B. Molecular and circulating biomarkers of gastric cancer. Int J Mol Sci. 2022;23:7588.
11. Matsuoka T, Yashiro M. Novel biomarkers for early detection of gastric cancer. World J Gastroenterol. 2023;29:2515–33.
12. Fu Y, Ling Z, Arabnia H, Deng Y. Current trend and development in bioinformatics research. BMC Bioinformatics. 2020;21:538.

13. Zhong Y, Xu F, Wu J, Schubert J, Li MM. Application of next generation sequencing in laboratory medicine. Ann Lab Med. 2021;41:25–43.
14. Zielinski JM, Luke JJ, Guglietta S, Krieg C. High throughput multi-omics approaches for clinical trial evaluation and drug discovery. Front Immunol. 2021;12: 590742.
15. Wang Y, Liu T, Liu Y, Chen J, Xin B, Wu M, Cui W. Coronary artery disease associated specific modules and feature genes revealed by integrative methods of WGCNA, MetaDE and machine learning. Gene. 2019;710:122–30.
16. Zhu YX, Huang JQ, Ming YY, Zhuang Z, Xia H. Screening of key biomarkers of tendinopathy based on bioinformatics and machine learning algorithms. PLoS ONE. 2021;16: e0259475.
17. Zhang C, Feng YG, Tam C, Wang N, Feng Y. Transcriptional profiling and machine learning unveil a concordant biosignature of type I interferon-inducible host response across nasal swab and pulmonary tissue for COVID-19 diagnosis. Front Immunol. 2021;12: 733171.
18. Zhang T, Wong G. Gene expression data analysis using Hellinger correlation in weighted gene co-expression networks (WGCNA). Comput Struct Biotechnol J. 2022;20:3851–63.
19. Deo RC. Machine learning in medicine. Circulation. 2015;132:1920–30.
20. Li Z, Huang B, Yi W, Wang F, Wei S, Yan H, Qin P, Zou D, Wei R, Chen N. Identification of potential early diagnostic biomarkers of sepsis. J Inflamm Res. 2021;14:621–31.
21. Zhao E, Xie H, Zhang Y. Predicting diagnostic gene biomarkers associated with immune infiltration in patients with acute myocardial infarction. Front Cardiovasc Med. 2020;7: 586871.
22. Chai K, Zhang X, Chen S, Gu H, Tang H, Cao P, Wang G, Ye W, Wan F, Liang J, Shen D. Application of weighted co-expression network analysis and machine learning to identify the pathological mechanism of Alzheimer's disease. Front Aging Neurosci. 2022;14: 837770.
23. Bruschi M, Kajana X, Petretto A, Bartolucci M, Pavanello M, Ghiggeri GM, Panfoli I, Candiano G. Weighted gene co-expression network analysis and support vector machine learning in the proteomic profiling of cerebrospinal fluid from extraventricular drainage in child medulloblastoma. Metabolites. 2022;12:724.
24. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991-995.
25. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.
26. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47.
27. Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics. 2003;19:368–75.
28. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
29. Langfelder P, Horvath S. Fast R functions for robust correlations and hierarchical clustering. J Stat Softw. 2012. https://doi.org/10.18637/jss.v046.i11.
30. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012;16:284–7.
31. Ali HEA, Lung PY, Sholl AB, Gad SA, Bustamante JJ, Ali HI, Rhim JS, Deep G, Zhang J, Abd Elmageed ZY. Dysregulated gene expression predicts tumor aggressiveness in African–American prostate cancer patients. Sci Rep. 2018;8:16335.
32. Engebretsen S, Bohlin J. Statistical predictions with glmnet. Clin Epigenetics. 2019;11:123.
33. Yang L, Wu H, Jin X, Zheng P, Hu S, Xu X, Yu W, Yan J. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep. 2020;10:5245.
34. Sapir-Pichhadze R, Kaplan B. Seeing the forest for the trees: random forest models for predicting survival in kidney transplant recipients. Transplantation. 2020;104:905–6.
35. Sanz H, Valim C, Vegas E, Oller JM, Reverter F. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinf. 2018;19:432.
36. Xu N, Guo H, Li X, Zhao Q, Li J. A five-genes based diagnostic signature for sepsis-induced ARDS. Pathol Oncol Res. 2021;27: 580801.
37. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005;21:3940–1.
38. Snipen L, Angell IL, Rognes T, Rudi K. Reduced metagenome sequencing for strain-resolution taxonomic profiles. Microbiome. 2021;9:79.
39. Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res. 2010;1:274–8.
40. Xiao B, Liu L, Li A, Xiang C, Wang P, Li H, Xiao T. Identification and verification of immune-related gene prognostic signature based on ssGSEA for osteosarcoma. Front Oncol. 2020;10: 607622.
41. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. Methods Mol Biol. 2018;1711:243–59.
42. Bindea G, Mlecnik B, Tosolini M, Kirilovsky A, Waldner M, Obenauf AC, Angell H, Fredriksen T, Lafontaine L, Berger A, et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. Immunity. 2013;39:782–95.
43. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, Hoang CD, Diehn M, Alizadeh AA. Robust enumeration of cell subsets from tissue expression profiles. Nat Methods. 2015;12:453–7.
44. Eden SK, Li C, Shepherd BE. Nonparametric estimation of Spearman's rank correlation with bivariate survival data. Biometrics. 2021;78:421.
45. Ito K, Murphy D. Application of ggplot2 to pharmacometric graphics. CPT Pharmacometr Syst Pharmacol. 2013;2: e79.
46. Ono H, Yao K, Fujishiro M, Oda I, Nimura S, Yahagi N, Iishi H, Oka M, Ajioka Y, Ichinose M, Matsui T. Guidelines for endoscopic submucosal dissection and endoscopic mucosal resection for early gastric cancer. Dig Endosc. 2016;28:3–15.
47. Chen X, Cubillos-Ruiz JR. Endoplasmic reticulum stress signals in the tumour and its microenvironment. Nat Rev Cancer. 2021;21:71–88.
48. Ullrich E, Bonmort M, Mignot G, Kroemer G, Zitvogel L. Tumor stress, cell death and the ensuing immune response. Cell Death Differ. 2008;15:21–8.
49. Donohoe C, Senge MO, Arnaut LG, Gomes-da-Silva LC. Cell death in photodynamic therapy: from oxidative stress to anti-tumor immunity. Biochim Biophys Acta Rev Cancer. 2019;1872: 188308.
50. Lian H, Han YP, Zhang YC, Zhao Y, Yan S, Li QF, Wang BC, Wang JJ, Meng W, Yang J, et al. Integrative analysis of gene expression and DNA methylation through one-class logistic regression machine learning identifies stemness features in medulloblastoma. Mol Oncol. 2019;13:2227–45.

51. Buckley SJ, Harvey RJ. Lessons learnt from using the machine learning random forest algorithm to predict virulence in streptococcus pyogenes. Front Cell Infect Microbiol. 2021;11: 809560.
52. Rajapakse JC, Duan KB, Yeo WK. Proteomic cancer classification with mass spectrometry data. Am J Pharmacogenomics. 2005;5:281–92.
53. Shen Y, Liu J, Zhang L, Dong S, Zhang J, Liu Y, Zhou H, Dong W. Identification of potential biomarkers and survival analysis for head and neck squamous cell carcinoma using bioinformatics strategy: a study based on TCGA and GEO datasets. Biomed Res Int. 2019;2019:7376034.
54. Zheng Q, Min S, Zhou Q. Identification of potential diagnostic and prognostic biomarkers for LUAD based on TCGA and GEO databases. 2021. Biosci Rep. https://doi.org/10.1042/BSR20204370.
55. Zhang G. Expression and prognostic significance of BANF1 in triple-negative breast cancer. Cancer Manag Res. 2020;12:145–50.
56. Bolderson E, Burgess JT, Li J, Gandhi NS, Boucher D, Croft LV, Beard S, Plowman JJ, Suraweera A, Adams MN, et al. Barrier-to-autointegration factor 1 (Banf1) regulates poly [ADP-ribose] polymerase 1 (PARP1) activity following oxidative DNA damage. Nat Commun. 2019;10:5501.
57. Jianrong S, Yanjun Z, Chen Y, Jianwen X. DUSP14 rescues cerebral ischemia/reperfusion (IR) injury by reducing inflammation and apoptosis via the activation of Nrf-2. Biochem Biophys Res Commun. 2019;509:713–21.
58. Yang CY, Chiu LL, Chang CC, Chuang HC, Tan TH. Induction of DUSP14 ubiquitination by PRMT5-mediated arginine methylation. FASEB J. 2018;32:6760.
59. Lin W, Sun Y, Qiu X, Huang Q, Kong L, Lu JJ. VMP1, a novel prognostic biomarker, contributes to glioma development by regulating autophagy. J Neuroinflammation. 2021;18:165.
60. Li YE, Wang Y, Du X, Zhang T, Mak HY, Hancock SE, McEwen H, Pandzic E, Whan RM, Aw YC, et al. TMEM41B and VMP1 are scramblases and regulate the distribution of cholesterol and phosphatidylserine. J Cell Biol. 2021. https://doi.org/10.1083/jcb.202103105.
61. Xing Z, Afkhami S, Bavananthasivam J, Fritz DK, D'Agostino MR, Vaseghi-Shanjani M, Yao Y, Jeyanathan M. Innate immune memory of tissue-resident macrophages and trained innate immunity: re-vamping vaccine concept and strategies. J Leukoc Biol. 2020;108:825–34.
62. Halaby MJ, Hezaveh K, Lamorte S, Ciudad MT, Kloetgen A, MacLeod BL, Guo M, Chakravarthy A, Medina TDS, Ugel S, et al. GCN2 drives macrophage and MDSC function and immunosuppression in the tumor microenvironment. Sci Immunol. 2019. https://doi.org/10.1126/sciimmunol.aax8189.
63. Nakamura K, Smyth MJ. Myeloid immunosuppression and immune checkpoints in the tumor microenvironment. Cell Mol Immunol. 2020;17:1–12.
64. Kolaczkowska E, Kubes P. Neutrophil recruitment and function in health and inflammation. Nat Rev Immunol. 2013;13:159–75.