Clinics of Oncology

Research Article ISSN: 2640-1037 |Volume 8

BANF1 and LY6E May Be Involved in Causal Relationships between Immune Cells, Circulating Metabolites, and Gastric Cancer: Multivariate Mendelian Studies Combined with Machine Learning and Experiments to Validate Characteristic Genes

Xuetong Ren[1], Shaowei Liu[1], Yuhua Wang[1], Haoyu Chen[1], Tianyu Gao[1], Pingping Zhou[2], Haiyan Bai[2*] and Yangang Wang[3*]

[1]Graduate School of Hebei University of Traditional Chinese Medicine, Shijiazhuang, 050000, China

[2]Department of Spleen and Stomach Diseases, Hebei Provincial Hospital of Traditional Chinese Medicine, Shijiazhuang, 050000, China

3Department of Spleen and Stomach Diseases, The Third Affiliated Hospital of Beijing University of Chinese Medicine, Beijing, 100020, China

***Corresponding author:**

Haiyan Bai,

Department of spleen and stomach diseases, The Third Affiliated Hospital of Beijing University of Chinese Medicine (11 North Third Ring Road East, Chaoyang District, Beijing), Beijing, 100020, China

Co-Corresponding author

Yangang Wang,

Department of Spleen and Stomach Diseases, Hebei Provincial Hospital of Traditional Chinese Medicine (389 Zhongshan East Road, Shijiazhuang City), Shijiazhuang, China

Received: 10 Nov 2024

Accepted: 17 Dec 2024

Published: 21 Dec 2024

J Short Name: COO

Citation:

Yangang Wang, BANF1 and LY6E May Be Involved in Causal Relationships between Immune Cells, Circulating Metabolites, and Gastric Cancer: Multivariate Mendelian Studies Combined with Machine Learning and Experiments to Validate Characteristic Genes.
Clin Onco. 2024; 8(5): 1-10

Keywords:

Multivariate Mendelian Randomization; gastric cancer; machine learning; forecast; immune infiltration; causal effect

1. Abstract

Gastric cancer is the fifth most prevalent cancer globally, and most patients miss the critical window for early diagnosis and treatment due to invasive diagnostic procedures and poor patient adherence. Thus, identifying highly sensitive and feasible biomarkers is crucial. A Mendelian Randomization (MR) approach was utilized to investigate the causal relationships between immune cells, metabolites, and gastric cancer. Four MR methods—Inverse Variance Weighted (IVW), MR Egger, Weighted Median, and Weighted Mode—were employed, with IVW as the primary method. MR Egger and Cochrane's Q-test helped detect heterogeneity and pleiotropy, while the left method was used to assess the stability of MR results. Additionally, machine learning methods were applied to screen for characteristic genes using gastric cancer data from the Gene Expression Omnibus (GEO). This included Gene Ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG), Disease Ontology (DO) enrichment, Gene Set Enrichment Analysis (GSEA), LASSO regression, and Support Vector Machine-Recursive Feature Elimination (SVM-RFE). Validation involved Receiver Operating Characteristic (ROC) curve analysis, immune cell infiltration, and qRT-PCR experiments. The study found that the exposure factor, CD25 on Ig D-CD24-B cells, and the mediating factor, plasma-free asparagine levels, may contribute to gastric cancer. Three key genes—BANF1, LY6E, and VSTM2A—were identified through LASSO and SVM-RFE analyses, which showed strong performance (ACU > 95%) and were validated via various methods, including PCR experiments. BANF1 and LY6E emerge as robust potential biomarkers for gastric cancer.

2. Introduction

Gastric cancer is the fifth most common malignant tumor in the world [1] and one of the most common malignant tumors in China, with a mortality rate that is among the highest in the world and a 5-year survival rate of less than 30% for patients with progressive gastric cancer [2-4]. Endoscopic pathology is still the "gold

standard" for gastric cancer diagnosis, but it is invasive, and patient compliance is poor. Although the sensitivity of various carcinoembryonic antigens is good, they lack specificity in diagnosing early gastric cancer. Therefore, it is crucial to find simple and feasible biomarkers with high detection rates to improve the early detection rates of gastric cancer patients and to study the mechanism of gastric cancer. Although increasing evidence suggests a crucial link between immune cells and the onset and progression of gastric cancer, the relationship between immune cells and gastric cancer is still not fully understood. Research has found that T lymphocytes, depleted natural killer cells, regulatory T cells, and other cells dominate the tumor microenvironment of esophageal cancer [5]. T cell infiltration may be a prognostic marker for gastric cancer, while CD45RO (memory) and FOXP3 (regulation) T cells may provide personalized follow-up and adjuvant therapy strategies for gastric cancer prognosis [6]. A retrospective study showed that higher levels of CD4+T cells, CD4+/CD8+ratio, NK cells, and Treg cells in peripheral blood were more effective in treating advanced gastric cancer patients before and after immunotherapy and correlated with good prognoses [7]. Furthermore, it is worth noting that immunotherapy has become an effective clinical strategy for treating cancer. However, most studies have elucidated the correlation between immune cells and gastric cancer without elucidating the causal relationship between the two. Recently, metabolomics has been used to analyze small endogenous metabolites, showing great potential in cancer diagnosis/screening, with plasma metabolites exhibiting very high diagnostic sensitivity and specificity. Therefore, this study utilized Mendelian randomization to investigate the causal relationship between immune cells, plasma metabolites, and gastric cancer. Mendelian randomization (MR) is an epidemiological survey tool that evaluates the causal relationship between genetic variations strongly correlated with exposure factors and outcomes based on these genetic variations as instrumental variables. The association obtained by MR is not affected by causal inversion and confounding factors. With the increasing availability and integration of multiple data types, including genomic, transcriptomic, and histopathological data, machine learning algorithms are becoming increasingly common to automate these tasks and assist in cancer detection (identifying the presence of cancer) and diagnosis. Machine learning can use data generated by other genomic assays as input data, for example, microarray or RNA-seq expression data, chromatin accessibility assays (e.g., DNase-seq, MNase-seq, and FAIRE), or histone modification, transcription factor binding ChIP-seq data, etc. Gene expression data are often helpful in differentiating between different disease phenotypes and identifying potentially valuable disease biomarkers [8,9]. In this study, we combined multivariate Mendelian randomization (MVMR) analysis of immune cells and cancer with machine learning to screen for cancer-characteristic genes and experimentally validate them, exploring potential targets for cancer diagnosis and treatment (Figure 1).

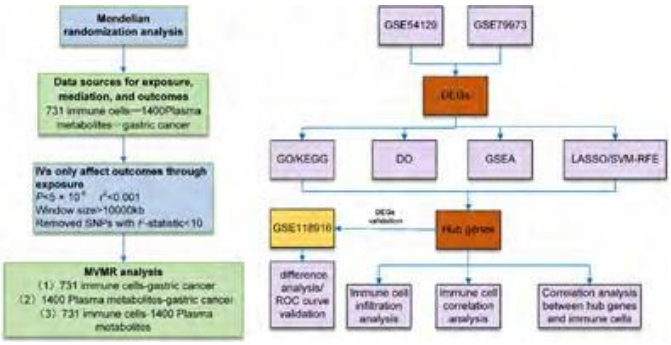

Figure 1: Flow chart illustrating the study design and methodology.

3. Materials and Methods

3.1. Data Sources for Exposure, Mediation, and Outcomes

Seven hundred thirty-one immune cells were selected as exposure factors. The GWAS statistics of immune cells mainly come from a cohort of 3,757 Sardinians. This queue involves the impact of approximately 22 million variants and 122 significant (P<1.28×10-11) independent association signals for 459 cell traits at 70 loci (53 of their novels) [10]. Select 1400 blood metabolites as mediating factors. This study involved 1,091 metabolites and 309 metabolite ratios from 8,299 Canadian participants [11]. This study is considered a result of gastric cancer. Gastric cancer statistical data comes from the Finnish database

https://r11.finngen.fi/pheno/C3_STOMACH_EXALLC）

It involved 2,384 participants, 807 females and 1,577 males.

3.2. Ivs only affect Outcomes Through Exposure

Our goal regarding immune cell data is to maintain reliability and ensure sufficient SNPs for exposure analysis. We set the genome-wide threshold for exposing relevant SNPs to 1 × 10-5. We applied standard GWAS thresholds (P<5 × 10-8) to SNPs associated with gastric cancer and blood metabolites. In addition, we also removed chain disequilibrium (r2<0.001, window size>10000kb) and removed SNPs with F-statistic<10 (F=Beta2/SE2).

3.3. Mendelian Randomization

We used MVMR to evaluate the causal relationship between immune cells and gastric cancer. We use inverse variance weighted (IVW), MR Egger, and weighted median to test causality. IVW is the primary analytical method. Evaluate targeted pleiotropy using intercept values from MR Egger regression. When the regression intercept is non-zero, P<0.05, we consider this a statistically significant indicator of genetic pleiotropy. Heterogeneity was tested using Cochran's Q test. We use MVMR to evaluate the causal relationship between 731 types of immune cells and gastric cancer. We use inverse variance weighted (IVW), MR Egger, and weighted median to test causality. IVW is the primary analytical method. Evaluate targeted pleiotropy using intercept values from MR Egger regression. When the regression intercept is non-zero, P<0.05, we consider this a statistically significant indicator of genetic pleiotropy. Heterogeneity was tested using Cochran's Q test.

When heterogeneity exists, we first perform IVW analysis. Use error occurrence rate (FDR) to correct IVW results. FDR q>0.05 is considered to have a potential causal relationship with gastric cancer. The mediation ratio is the ratio of mediation effect to total effect, used to quantify the mediating role of circulating metabolites in the relationship between immune cells and gastric cancer. We used the R software package (version 4.3.1), which includes "Two Sample MR," "Mendelian randomization," and "MVMR."

3.4. Collection of GEO Data

Searched the GEO database (https://www.ncbi.nlm.nih.gov/geo/) for the keyword "gastric cancer" and downloaded the microarray data "GSE54129", which uses the GPL570 [HG-U133_Plus_2] expression profile. Affymetrix Human Genome U133 Plus 2.0 Array platform, containing 132 gastric tissue samples, including 111 gastric cancer samples and 21 normal gastric tissues. Download microarray data "GSE79973" for this expression profile using the GPL570 [HG-U133_Plus_2] Affymetrix Human Genome U133 Plus 2.0 Array platform. Twenty gastric tissues were included, including ten cancer samples and ten normal gastric tissues. Download microarray data "GSE118916." This expression profile was performed using the GPL15207 [Prime View] Affymetrix Human Gene Expression Array platform. Thirty gastric tissues were included, including 15 cancer samples and 15 normal gastric tissues.

3.5. Screening for Differential Genes

Using the R software packages "limma," "heatmap," "ggplot2", "ggrepel, "and "dplyr" software packages, using |log2FC|>2 and corrected q-value <0.05 as filtering criteria, the data from the gastric cancer group and the control group were extracted and combined, and the analysis of variance was performed, and the results were corrected using the " FDR" method was used to correct for the differences, and the obtained data were used to draw heat maps and volcano maps.

3.6. LASSO Regression to Screen for Trait Genes

Using the "glmnet" package in R, we constructed the model, screened the signature genes, and plotted the cross-validation graphs.

3.7. SVM-RFE Analysis Screening of Signature Genes and Data Merging

Using the "e1071", "kernlab" and "caret" packages in R, we built the models, screened the feature genes, plotted the cross-validation graphs and combined the data with the LASSO regression algorithm model, took the intersection genes from the machine learning algorithm and planned the Venn diagrams.

3.8. Correlation Analysis of Genes and Immune Cells

Using the "limma," "reshape2", "ggpubr," "ggExtra " packages to obtain the expression of target genes and plot correlation scatter plots and lollipop plots.

3.9. Animals Specimens

Animal care and experimental procedures in this study were approved by the Animal Care and Use Committee of Hebei College of Traditional Chinese Medicine (DWLL202212021).20 male Sprague-Dawley (SD) rats (weighing 150-180 g) were supplied by Beijing Viton Lever Laboratory Animal Technology Co. The rats were housed in a temperature-controlled environment (24 ℃ ± 4 °C) with a 12/12 h light-dark cycle and free access to water and food. After one week of adaptation, the rats were randomly divided into a normal group (n = 10) and a GC model group (n = 10). This GC rat model was produced by administration of MNNG (200 μg/ml) and irregular fasting (1 day of fasting and one day of feeding) with a single fasting gavage of 2% sodium salicylate on each day of fasting as previously described. Thereafter, at week 24, two rats were randomly selected from the model group and euthanized by intraperitoneal injection of 2% sodium pentobarbital (140 mg/kg). The stomachs were removed for pathological examination to confirm the successful establishment of the GC model.

3.10. Quantitative Real-Time PCR (qRT-PCR) Assay

Gastric tissue was processed to extract total RNA. Then, 1ug of total RNA was denatured in a mixture containing the Oligo dT primer and random primers for 5 minutes at 70°C and immediately cooled on ice to be used as template RNA for qRT-PCR, and the processed RNA was added to a reverse transcription reaction solution containing reaction buffer, MgCl2, PCR Nucleotide Mixture, Ribonuclease Inhibitor, Reverse Transcriptase, and Ribonuclease-Free Water to a volume of 20 μl. Reaction solution to a volume of 20 μl. Reactants are mixed slowly, centrifuged briefly, and then placed into a PCR cycle: 25°C for 5 min (annealing), 42°C for 15 min (extension), 85°C for 5 min (inactivation), 15 min (extension), 5 min at 85°C (inactivation) and the cycle continues. The reactions were cooled on ice and stored at -20°C for later use. The Cq values were obtained for each target gene and internal reference gene b-actin. The relative quantitative value of the expression of each target gene was expressed as the Q-value of each target gene/Q-value of the first sample, RQ==2-DDCc value, and statistically analyzed using the RQ-values. The primers are as follows: GAPDH: F-5′-GCAAGTTCAACGGCACAG-3′, R-5′-CTCGCTCCTGGAAGATGG-3′;BANF1:F-5′-CGTGGTAGTTCCTAACGGGG-3′,R-5′-GTGGTGCACGTAAGGCA ATC-3′;VSTM2A:F-5′-GCCGGAGGAGACACTTCAC-3′,R-5′-CCCCATCATGCCAAGGGAAT-3′;LY6E:F-5′-GAGAGTCTTCTTGCCTGT-3′,R-5′-ATTGTTCTTCTGATCGGT-3′.

3.11. Statistical Analysis

Gene expression differences between specimens from gastric cancer and normal specimens were compared using Student's t-test. A one-way ANOVA test was used to compare the groups. Statistical analysis was performed using R software (version 4.2.1) and GraphPad Prism (version 6.0) software, *indicates p<0.05, **indicates p<0.01.

4. Results

4.1. Genetic Variable Tools Related to Immune Cells and Metabolites

We used F>10 as the screening criterion to eliminate defective instrumental variables and obtained 12,790 SNP records of immune cells and 32,131 SNP records of metabolites. (Supplementary 1)

4.2. Mendelian Randomization Analysis of Immune Cells and Gastric Cancer

IVW as the main analysis method, we found that CD25 on Ig D−CD24−B cell (OR=0.893, P=0.029), CD20-CD38-B cell% lymphocyte (OR=0.887, P=0.027), T cell% lymphocyte (OR=0.846, P=0.009), CD8+and CD8dim T cell% leukocyte (OR=0.811, P=0.008) may be protective factors for gastric cancer. Naive CD8+ T cell %CD8+ T cell(OR=1.024, P=0.004)、CD28−CD4−CD8−T cell %T cell(OR=1.097, P=0.033)、CD45 on CD4+ T cell(OR=1.288, P=0.003), CCR2 on CD14-CD16+monocyte (OR=1.059, P=0.028), and SSC-A on myeloid dendritic cell (OR=1.040, P=0.039) may be risk factors for gastric cancer (Figure 2). It should be noted that in reverse MR analysis, CD4+/CD8+T cells, CD8+and CD8dim T cell% leukocytes, SSC-A on HLA DR+ Natural Killer have a reverse causal relationship with gastric cancer (Supplementary 1). In addition, we conducted a sensitivity analysis, and the results were robust (Supplementary 1). It is worth noting that the exposure factor Terminally Differentiated CD4+T cell% CD4+T cell has pleiotropy, and the results are unreliable (Supplementary 1).

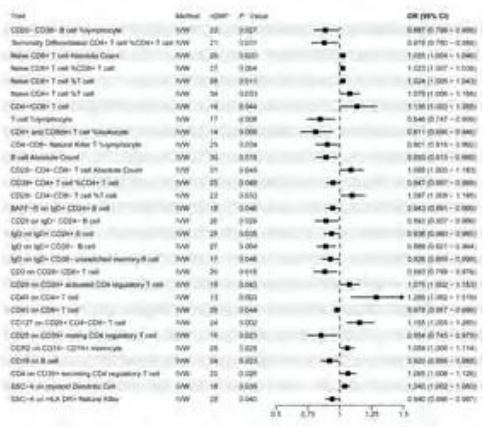

Figure 2: MR analysis of immune cells and gastric cancer.

3.3. Mendelian Randomization Study on Metabolites and Gastric Cancer

IVW analysis showed that EDTA levels (OR=0.784, P=0.050), 4-hydroxy hippurate levels (OR=0.726, P=0.002), Pyrraline levels (OR=0.805, P=0.033), Creatine levels (OR=0.823, P=0.003), and Adenosine 5 ‘- monophosphate (AMP) levels (OR=0.757, P=0.029) may be protective factors for gastric cancer. Plasma free asparagine levels (OR=1.146, P=0.030), N-Methyl-2-pyridone-5-5-carboxamide levels (OR=1.147, P=0.042), 2-hydroxy-3-methylvalerate levels (OR=1.280, P=0.004), 3- (3-hydroxyphenyl) propyl sulfate levels (OR=1.195, P=0.045), and Butyrate/isobutyrate (4:0) levels (OR=1.258, P=0.030) may be risk factors for gastric cancer (Figure 2). It is worth noting that the mediating factor 4-vinylguaiacol sulfate levels have pleiotropy and the results are unreliable. In reverse MR analysis, a reverse causal relationship exists between 5alphapregnan-3beta and 20 alpha-diol disulfide levels. Sensitivity analysis indicates no SNP will affect the observed association between exposure factors and outcomes (Supplementary 2).

3.4. Mendelian Randomization Analysis of Immune Cells and Metabolites

Through multivariate analysis, the exposure factor was CD25 on Ig D−CD24−B cell (GCST90001785, beta=-0.087, P=0.086), the mediating factor was plasma free asparagine levels (GCST9020452, beta=0.131, P=0.022) (Table 1, Supplementary 3), the mediating effect was -0.007, and the mediating effect ratio is 6.2%=-0.007/-0.112×100% (Table 2).

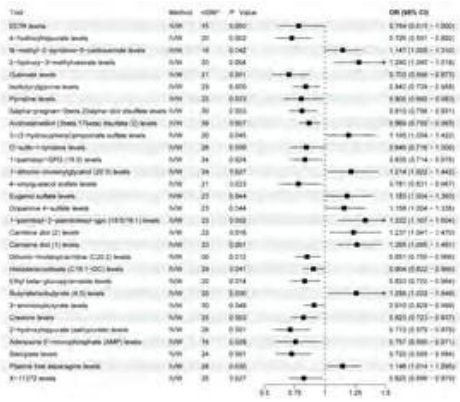

Figure 3: MR analysis of metabolites and gastric cancer.

5.5. Preliminary Screening for Differential Genes

In this study, we found that the difference in gene expression between gastric cancer tissues and normal gastric tissues was compared with a filtering criterion of |log2FC|>2 and a corrected q-value < 0.05. A total of 181 differential genes were screened, including 68 upregulated genes and 113 down-regulated genes in the gastric cancer group. The heat map showed the top 50 most significantly up-and down-regulated genes in the gastric cancer group (Figure 4, A). The volcano map showed 23 upregulated and 40 down-regulated genes with significant differences in the gastric cancer group (Figure 4, B). The above allowed us to speculate that substantial differences exist in transforming normal gastric tissue into gastric cancer.

Table 1: Mendelian randomization mediated outcomes.

MVMR	Exposure	Outcome	Beta	Se	*P* value
Expose	GCST90001785	Malignant neoplasm of stomach	-0.087	0.05	0.086
Mediation	GCST902004523	Malignant neoplasm of stomach	0.131	0.057	0.022

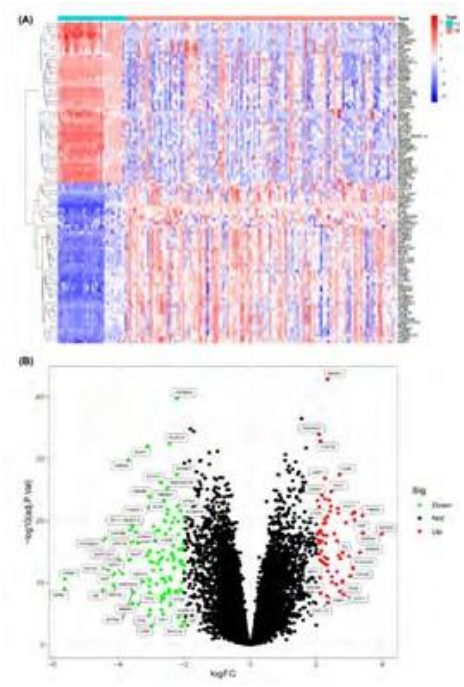

Figure 4. Differential genes expression analysis. (A) Differential genes were filtered using |log2FC| > 2 and corrected *q*-value < 0.05 as the filtering criteria. Blue: down-regulated differential genes. Red: Upregulated differential genes; (B)Red: upregulated genes in the experimental group. Green: down-regulated genes in the experimental group. Black: undifferentiated genes.

5.6. GO Functional Enrichment Analysis and KEGG Pathway Enrichment Analysis

The top ten most enriched GO functions in the analysis of BP (Biological process, BP), CC (Cellular component, CC), and MF (Molecular function, MF) were selected, respectively. In BP, significant enrichment was found for the regulation of angiogenesis and regulation of vascular system development. The enrichment of collagen-containing extracellular matrix and apical cellular components was evident in cc. The enrichment of receptor-ligand activity, signaling receptor activator activity, and structural elements of the extracellular matrix was apparent in MF (Figure 5, A, B). The gene relationship diagram mainly shows the relationship network of regulation of “angiogenesis,”” connective tissue development, “ and” cellular hormone metabolic process”(Figure 5, E). The enrichment of genes in BP and MF was evident by GO circle plots (Figure 5, F). KEGG analysis showed that the enrichment pathway was mainly found in ECM-receptor interaction, Focal adhesion, Gastric acid secretion, Retinol metabolism, and AGE-RAGE signaling pathway in diabetic complications, using q-value <0.05 as a filtering condition. Acid secretion, Retinol metabolism, and AGE-RAGE signaling pathway in diabetic complications (Figure 5, C, D).

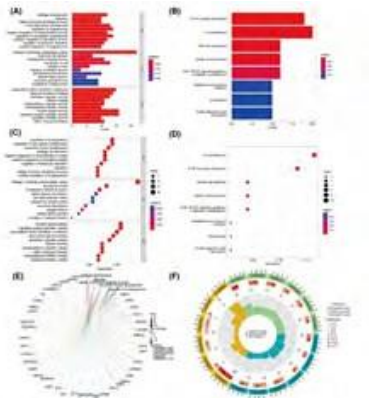

Figure 5: GO functional enrichment analysis and KEGG pathway enrichment analysis. (A, B) GO function enrichment analysis with *q-valve* <0.005 as filtering criteria; (C, D) KEGG pathway enrichment analysis with *q-valve* <0.05 as the filtering criterion; (E)Gene versus function plot; (F)GO circle plot.

5.7. DO Enrichment Analysis

The DO enrichment analysis revealed that the diseases were mainly concentrated in chronic obstructive pulmonary disease, stomach disease, gastritis, and duodenal ulcer (Figure 6, A, B).

5.8. GSEA Enrichment Analysis

Using the KEGG-enriched gene expression matrix file, retinol metabolism, fatty acid metabolism, and cytochrome P450 metabolism were found to be more active in the control group, and the adhesive spot pathway, extracellular receptor interactions pathway, complement and coagulation cascade response pathway, and chemokine signaling pathway were found to be significantly more active in the gastric cancer group (Figure 6, C). Using GO-enriched gene expression matrix files, BP-functional primary alcohol metabolic processes, digestion, small molecule catabolic processes, and MF oxidoreductases acting on donor CH-OH were more active in gastric cancer in the control group. In the gastric cancer group, BP-functional collagen fibril organization, CC-functional extracellular matrix, MF structural components of the extracellular matrix, and collagen binding were significantly active in gastric cancer (Figure 6, D).

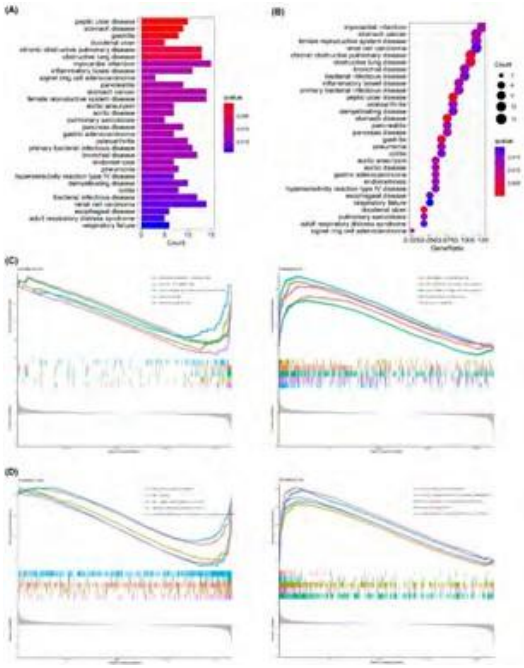

Figure 6: DO enrichment analysis and GSEA enrichment analysis. (A, B) DO enrichment analysis was performed using *q-value* < 0.005 as the filtering condition; (C) GSEA enrichment analysis was performed for the experimental and control groups using the gene set files from the KEGG pathway analysis; (D) GSEA analysis of experimental versus control groups was performed using gene-selected files from GO functional enrichment analysis.

5.9. LASSO Regression Model

We found that nine signature genes, BANF1, VSTM2A, TRAP-PC1, Y16709, ADH7, LY6E, IL32, SULF1, and CAP2, were screened in the Least Absolute Shrinkage and Selection Operator prediction model by observing the minimum point of cross-validation error (Figure 7, A, Table 3).

5.10. SVM-RFE Analysis

We found that four feature genes were screened in the SVM-RFE analysis prediction model by observing the minimum point of cross-validation error: VSTM2A, BANF1, PCAT18, and LY6E (Figure 7B, Table 4). Three genes intersected with the Least Absolute Shrinkage and Selection Operator prediction model: BANF1, VSTM2A, and LY6E (Figure 7C, Table 5).

5.11. Validation Group Difference Analysis

The validation group microarray data "GSE118916" validated the three intersectional signature genes (BANF1, VSTM2A, and LY6E). We found that the three crossover genes had a p-value < 0.005, and all of them were significantly different, with the expression of BANF1 and LY6E genes being upregulated in gastric cancer and VSTM2A gene being down-regulated in gastric cancer (Figure 7D).

5.12. ROC Curve Validation

ROC curves were plotted to validate their false favorable rates. We found that the AUC of the three signature genes (BANF1, LY6E, VSTM2A) was greater than 95% in both the analysis group data and the validation group data with high accuracy (Figure 7E,F).

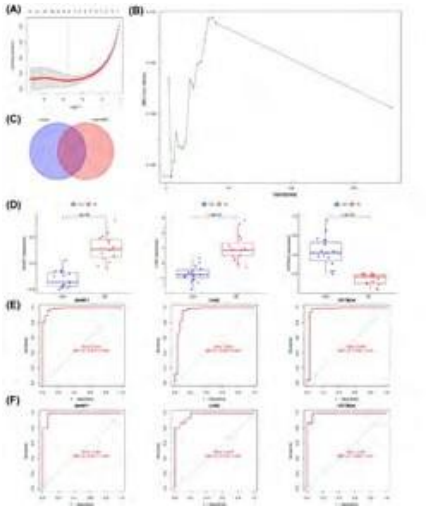

Figure 7: LASSO prediction model, SVM-RFE prediction model, and validation set analysis. (A) Lasso prediction model; (B) SVM-RFE prediction model; (C) Cross-validation Venn diagram of the THE LEAST ABSOLUTE SHRINKAGE AND SELECTION OPERATOR and SVM-RFE prediction models; (D) Differential analysis of BANF1, VSTM2A, and LY6E expression using validation group data; (E, F) ROC curves were plotted for BANF1, VSTM2A, and LY6E in the analysis group data versus the validation group data.

Table 2: Mendelian randomization mediated effect.

Mediated Effect	Beta	Se	LO_CI	UP_CI	OR	OR_lCI95	OR_UCI95
	-0.007	0.004	-0.016	-6e-05	0.993	0.983	0.1

Table 3: The LASSO prediction model feature gene.

Number	Feature Gene
1	BANF1
2	VSTM2A
3	TRAPPC1
4	Y16709
5	ADH7
6	LY6E
7	IL32
8	SULF1
9	CAP2

5.13. Immune Cell Infiltration Analysis

Differential analysis of immune cells revealed that the immune cells that differed most significantly between the control and gastric cancer groups were mainly plasma cells, MO macrophages, resting CD4 memory T cells, neutrophils, M1 macrophages, initial B cells, and memory B cells. In the gastric cancer group, there was an increased proportion of macrophages (MO, M1, M2), neutrophils, initial B cells, etc., and fewer CD4 initial T cells (Figure 8 AB, Table 6). Immune cell infiltration correlation analysis revealed positive correlations between resting CD4 memory T cells and plasma cells, initial CD4 T cells and B cells, M1 macrophages, activated CD4 T cells, neutrophils, and activated mast cells. M0 and M1 macrophages were negatively correlated with resting CD4 memory T cells, etc. (Figure 8C). Correlation studies between the signature genes (BANF1, LY6E, VSTM2A) and immune cells revealed that BANF1 was correlated with 17 immune cells, LY6E was correlated with 11 immune cells, and VSTM2A was correlated with 11 immune cells (Figure 8D,E,F, Figure 9,A,B,C). There was a strong correlation between LY6E on resting CD4 memory T cells, M1 macrophage (R:0.5-1, P<0.01) and a moderate correlation between M0 macrophage, activated CD4 memory T cells, neutrophils, and plasma cells (R:0.3-0.5, P<0.01) (Figure 9A). VSTM2A was strongly correlated with M0 macrophage and plasma cells, and there was a strong correlation (R:0.5-1, P<0.01) for resting CD4 memory T cells and a moderate correlation (R:0.3-0.5, P<0.01) for M1 macrophage, neutrophils (Figure 9, B). There was a moderate correlation (R:0.3-0.5, P<0.01) between BANF1 on macrophages (M0, M1), plasma cells, resting CD4 memory T cells, and CD8 T cells (Figure 9C). In the Mendelian randomization results, CD25-IgD-CD24-B cells as exposure factors may have a causal relationship with gastric cancer. As is well known, B cells play an indispensable role in the progression of tumors [12]. B cells are not only mediators of humoral immune responses but also participate in antibody-mediated immune responses. After activation, B cells can differentiate into memory B cells. Ig D reaches its peak level in B cells naive. CD24 and CD25 are activation markers for B cells and regulatory T cells [13]. In the analysis of immune infiltration differences and immune infiltration correlation, there are differences in B cells memory, B cells naïve, plasma cells, and T cells gamma delta. Surprisingly, BANF1 and LY6E showed a significant correlation with the above results. In addition, BANF1 has the strongest correlation with T cell regulatory factors (Tregs) among the three characteristic genes. Next, we will perform The Human Protein Atlas database（https://www.proteinatlas.org/）analysis on the characteristic genes (BANF1, LY6E, VSTM2A) Analysis. The results showed that the gene expression level of LY6E in immune cells was significantly higher than that of BANF1 (Figure 10, A). VSTM2A has low immune specificity, and no relevant immune cell expression levels were found. Survival analysis showed that BANF1 (P=0.031, 5-year survival high=59%, 5-year survival low=26%) was correlated with gastric cancer, while LY6E (P=0.27, 5-year survival high=33%, 5-year survival low=37%) was not correlated (Supplementary 4).

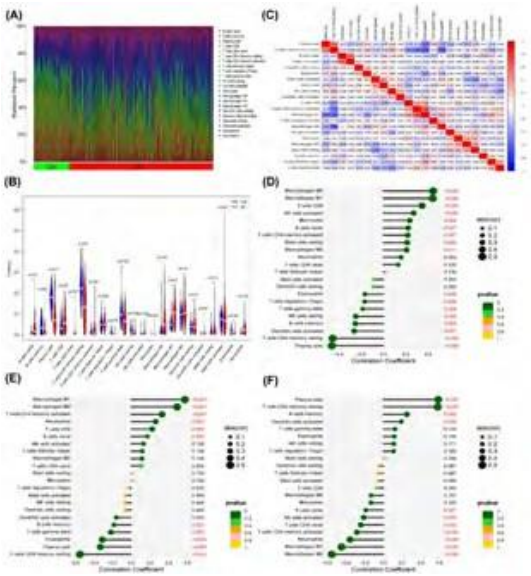

Figure 8: Analysis of immune cell infiltration. (A, B) Analysis of differences between control and experimental groups; (C) Correlation analysis of immune cell infiltration; (D) correlation between BANF1 and immune cells; (E) correlation between LY6E and immune cells; (F) correlation between VSTM2A and immune cells.

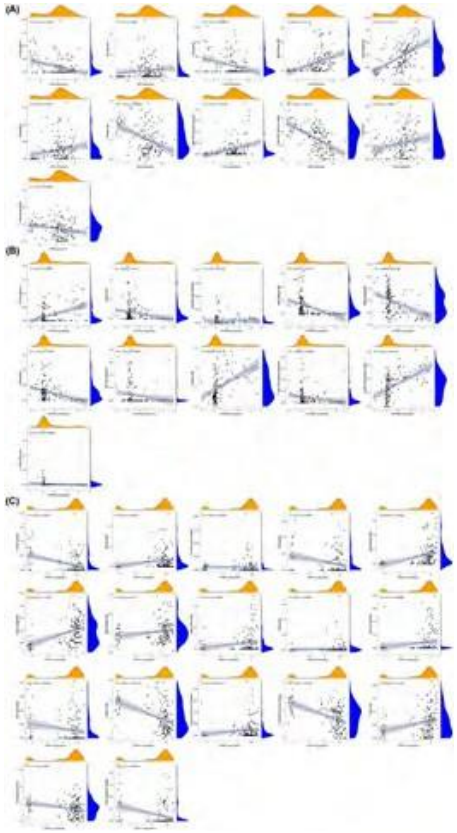

Figure 9: Correlation analysis of immune cells with BANF1, LY6E, and VSTM2A. (A) Correlation analysis of LY6E expression with immune cells; (B) Correlation analysis of VSTM2A expression with immune cells;(C) Correlation analysis of BANF1 expression with immune cells.

5.14. Expression of Core Genes

We performed q-PCR detection on characteristic genes and determined their expression levels in gastric cancer tissues. The results showed that the expression levels of BANF1 (P=0.0003, P<0.001) and LY6E (P=0.0091, P<0.05) in the gastric cancer group were significantly higher than those in the normal group. The level of VSTM2A in the gastric cancer group was significantly lower than that in the normal group (P<0.0001) (Figure 10, B, C, D).

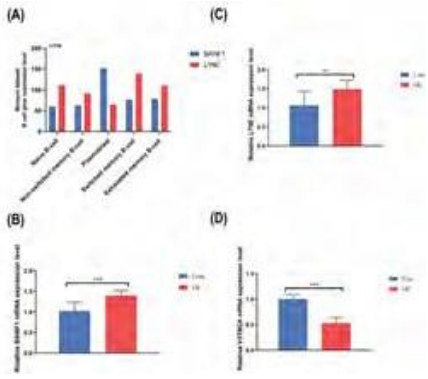

Figure 10: BANF, LY6E, VSTM2A expression levels. (A) The expression levels of BANF and LY6E genes in B cells were obtained from the HPA database; (B) PCR: RNA expression levels of ANF1 in normal and gastric cancer groups; (C) PCR: LY6E RNA expression levels in normal and gastric cancer groups; (D) PCR: RNA expression levels of VSTM2A in normal and gastric cancer groups.

Table 4: SVM-RFE predictive model feature genes.

Number	Feature Gene
1	VSTM2A
2	BANF1
3	PCAT18
4	LY6E

Table 5: Intersection feature genes of the LASSO prediction model and SVM-RFE prediction mode.

Number	Feature Gene
1	BANF1
2	VSTM2A
3	LY6E

Table 6: P-value of immune cell difference between the two group.

Cell	*P* value (*P*<0.05)
Plasma cells	1.93E-08
Macrophages M0	3.02E-08
T cells CD4 memory resting	4.14E-08
Neutrophils	4.35E-08
Macrophages M1	4.14E-06
B cells naive	7.79E-06
B cells memory	2.48E-05
T cells CD4 memory activated	0.0012
T cells gamma delta	0.0021
NK cells activated	0.0025
T cells regulatory (Tregs)	0.0052
Eosinophils	0.0063
Dendritic cells activated	0.0068
T cells CD8	0.0225
NK cells resting	0.0341

Note: *P*<0.05, the difference between the two groups was statistically significant.

6. Discussion

Gastric cancer is a highly heterogeneous and invasive malignant tumor, and its immune microenvironment is closely related to its growth, development, and drug resistance. Therefore, searching for cancer-specific immune cell markers and characteristic genes based on the immune microenvironment has become an emerging research hotspot [14]. Firstly, we conducted Mendelian randomization analysis in this study to explore the causal relationship between 731 immune cells, 1400 plasma metabolites, and gastric cancer. The results showed that the exposure factor CD25 on Ig D-CD24-B cells can lead to gastric cancer through the mediating factor Plasma free asparagine levels. Additionally, it is worth noting that naive CD8+T cells% CD8+T cells, and naive CD4+T cells% T cells in immune cells also have a strong causal relationship with

gastric cancer. Then, based on machine learning algorithms, we screened the characteristic genes of gastric cancer expression data from the GEO database and combined them with immune infiltration correlation to analyze the above results. Finally, we validated the expression of characteristic genes in gastric cancer tissues. In this study, we screened 181 differentially expressed genes, including 68 upregulated genes and 113 downregulated genes in the cancer group. By combining the minimum absolute shrinkage and selection operator machine learning algorithm with SVM-RFE analysis, three feature genes (BANF1, LY6E, and VSTM2A) were selected, and their accuracy (AUC>95%) was verified. Because our study, Mendelian randomization, demonstrated the specificity of B cells (CD25 on Ig D-CD24-B cells), we quantified BANF1, LY6E, and VSTM2A expression levels in B cells using the HPA database. The results showed that the expression of BANF1 and LY6E in B cells was significant. Surprisingly, BANF1 and LY6E are mainly enriched in naive B cells (lg D CD24 CD8+ T cells, Regulatory T cells (CD25). In other words, BANF1 and LY6E may affect the progression of cancer. In addition, the survival analysis results of BANF1 and gastric cancer also showed a correlation. It was found that BANF1 is critical for cell proliferation and division through interactions with double-stranded DNA, histones, and other nuclear proteins, such as innate immunity, post-mitotic nuclear reformation, interphase nuclear membrane rupture repair, genomic regulation, and DNA damage and repair responses, and was shown to BANF1 is also associated with the clinical features and prognosis of gastric cancer. It may be a new indicator of tumor prognosis[15, 16]. In our study, survival analysis showed a correlation between BANF1 and the prognosis of gastric cancer. The latest research has found that silencing BANF1 can significantly hinder GC cell proliferation, migration, and invasion [17]. LY6E was found to be associated with malignant progression in various types of cancer. In malignant tumors, aberrant overexpression of LY6E increased HIF-1α gene expression mainly at the transcriptional level. It leads to high expression of VEGFA and PDGFB, activates PI3K/Akt to function[21], and knockdown of LY6E leads to G1-S cell cycle arrest and apoptosis in AGS cells, thereby inhibiting AGS cell survival and proliferation[19]. VSTM2A was found to be essential in regulating preadipocyte differentiation and inhibiting Wnt signaling and is a critical oncogenic factor in colorectal cancer [20,21]. In addition, we conducted animal experiments and found significant differences in BANF1, LY6E, and VSTM2A between the normal and gastric cancer groups. In summary, BANF1, LY6E, and LY6E demonstrate potential as biomarkers. However, there are still some limitations in this study, such as the lack of large clinical samples for the signature genes and the small amount of data in the validation set of the study needs to be tested and optimized for the machine learning model.

7. Funding

This work was supported by the Natural Science Foundation of Hebei Province (H2020423207).

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA-CANCER J CLIN. 2021; 71: 209-49.
2. Sitarz R, Skierucha M, Mielko J, Offerhaus G, Maciejewski R, Polkowski WP. Gastric cancer: epidemiology, prevention, classification, and treatment. CANCER MANAG RES. 2018; 10: 239-48.
3. Sexton RE, Al HM, Diab M, Azmi AS. Gastric cancer: a comprehensive review of current and future treatment strategies. CANCER METAST REV. 2020; 39: 1179-203.
4. Xia JY, Aadam AA. Advances in screening and detection of gastric cancer. J SURG ONCOL. 2022; 125: 1104-9.
5. Zheng Y, Chen Z, Han Y, Han L. Immune suppressive landscape in the human esophageal squamous cell carcinoma microenvironment. NAT COMMUN. 2020; 11: 6268.
6. Challoner BR, von Loga K, Woolston A, Griffiths B, Sivamanoharan N. Computational Image Analysis of T-Cell Infiltrates in Resectable Gastric Cancer: Association with Survival and Molecular Subtypes. JNCI-J NATL CANCER I. 2021; 113: 88-98.
7. Wang X, Liu X, Dai H, Jia J. Association of lymphocyte subsets with the efficacy and prognosis of PD-1 inhibitor therapy in advanced gastric cancer: results from a monocentric retrospective study. BMC GASTROENTEROL. 2024; 24: 113.
8. Jones W, Alasoo K, Fishman D, Parts L. Computational biology: deep learning. EMERG TOP LIFE SCI. 2017; 1: 257-74.
9. Tran KA, Kondrashova O, Bradley A, Williams ED, Pearson JV, Waddell N. Deep learning in cancer diagnosis, prognosis and treatment selection. GENOME MED. 2021; 13: 152.
10. Orrù V, Steri M, Sidore C, Marongiu M, Serra V, Olla S, Sole G. Complex genetic signatures in immune cells underlie autoimmunity and inform therapy. NAT GENET. 2020; 52: 1036-45.
11. Chen Y, Lu T, Pettersson-Kymmer U, Stewart ID, Butler-Laporte G, Nakanishi T. Genomic atlas of the plasma metabolome prioritizes metabolites implicated in human diseases. NAT GENET. 2023; 55: 44-53.
12. Garaud S, Buisseret L, Solinas C, Gu-Trantien C, de Wind A, Van den Eynden G. Tumor infiltrating B-cells signal functional humoral immune responses in breast cancer. JCI Insight. 2019; 5.
13. Toda A, Piccirillo CA. Development and function of naturally occurring CD4+CD25+ regulatory T cells. J LEUKOCYTE BIOL. 2006; 80: 458-70.
14. Wu LH, Wang XX, Wang Y, Wei J, Liang ZR, Yan X, Wang J. Construction and validation of a prognosis signature based on the immune microenvironment in gastric cancer. FRONT SURG. 2023; 10: 1088292.

15. Yang HJ, Kim JH, Kim NW, Choi IJ. Comparison of long-term outcomes of endoscopic submucosal dissection and surgery for undifferentiated-type early gastric cancer meeting the expanded criteria: a systematic review and meta-analysis. SURG ENDOSC. 2022; 36: 3686-97.
16. Li J, Hu B, Fang L, Gao Y, Shi S, He H. Barrier-to-autointegration factor 1: A novel biomarker for gastric cancer. ONCOL LETT. 2018; 16: 6488-94.
17. Xu Y, Wang X, Yuan W, Zhang L, Chen W, Hu K. Identification of BANF1 as a novel prognostic biomarker in gastric cancer and validation via in-vitro and in-vivo experiments. Aging (Albany NY). 2024; 16: 1808-28.
18. Yeom CJ, Zeng L, Goto Y, Morinibu A, Zhu Y, Shinomiya K. LY6E: a conductor of malignant tumor growth through modulation of the PTEN/PI3K/Akt/HIF-1 axis. Oncotarget. 2016; 7: 65837-48.
19. Lv Y, Song Y, Ni C, Wang S, Chen Z, Shi X, Jiang Q, Cao C. Overexpression of Lymphocyte Antigen 6 Complex, Locus E in Gastric Cancer Promotes Cancer Cell Growth and Metastasis. Cell Physiol Biochem. 2018; 45: 1219-29.
20. Secco B, Camiré É, Brière MA, Caron A, Billong A, Gélinas Y. Amplification of Adipogenic Commitment by VSTM2A. CELL REP. 2017; 18: 93-106.
21. Dong Y, Zhang Y, Kang W, Wang G, Chen H, Higashimori A. VSTM2A suppresses colorectal cancer and antagonizes Wnt signaling receptor LRP6. THERANOSTICS. 2019; 9: 6517-31.