www.aging-us.com AGING 2023, Vol. 15, No. 21

Research Paper

Hub gene identification and molecular subtype construction for *Helicobacter pylori* in gastric cancer via machine learning methods and NMF algorithm

Lianghua Luo[1,2,*], Ahao Wu[1,2,*], Xufeng Shu[1,2], Li Liu[1,2], Zongfeng Feng[1,2], Qingwen Zeng[1,2], Zhonghao Wang[1,2], Tengcheng Hu[1,2], Yi Cao[1], Yi Tu[3], Zhengrong Li[1]

[1]Department of General Surgery, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China
[2]Medical Innovation Center, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China
[3]Department of Pathology, The First Affiliated Hospital of Nanchang University, Nanchang, Jiangxi, China
*Equal contribution

Correspondence to: Yi Tu, Zhengrong Li; **email:** tuyi1027@sina.com, https://orcid.org/0000-0003-1709-2688; lzr13@foxmail.com, https://orcid.or/0000-0001-7304-6729

Keywords: gastric cancer, Helicobacter pylori, cluster, hub genes, therapy
Received: February 28, 2023 **Accepted:** July 19, 2023 **Published:** September 26, 2023

ABSTRACT

Helicobacter pylori **(HP) is a gram-negative and spiral-shaped bacterium colonizing the human stomach and has been recognized as the risk factor of gastritis, peptic ulcer disease, and gastric cancer (GC). Moreover, it was recently identified as a class I carcinogen, which affects the occurrence and progression of GC via inducing various oncogenic pathways. Therefore, identifying the HP-related key genes is crucial for understanding the oncogenic mechanisms and improving the outcomes of GC patients. We retrieved the list of HP-related gene sets from the Molecular Signatures Database. Based on the HP-related genes, unsupervised non-negative matrix factorization (NMF) clustering method was conducted to stratify TCGA-STAD, GSE15459, GSE84433 samples into two clusters with distinct clinical outcomes and immune infiltration characterization. Subsequently, two machine learning (ML) strategies, including support vector machine-recursive feature elimination (SVM-RFE) and random forest (RF), were employed to determine twelve hub HP-related genes. Beyond that, receiver operating characteristic and Kaplan-Meier curves further confirmed the diagnostic value and prognostic significance of hub genes. Finally, expression of HP-related hub genes was tested by qRT-PCR array and immunohistochemical images. Additionally, functional pathway enrichment analysis indicated that these hub genes were implicated in the genesis and progression of GC by activating or inhibiting the classical cancer-associated pathways, such as epithelial-mesenchymal transition, cell cycle, apoptosis, RAS/MAPK, etc. In the present study, we constructed a novel HP-related tumor classification in different datasets, and screened out twelve hub genes via performing the ML algorithms, which may contribute to the molecular diagnosis and personalized therapy of GC.**

INTRODUCTION

Gastric cancer (GC) is one of the most frequent and fatal upper gastrointestinal malignances, all ranking the top five among cancers globally in terms of the morbidity and mortality [1]. According to the World Health Organization (WHO), the incidence and mortality rate of GC have gradually declined, partly due to increased awareness of *Helicobacter pylori* (HP) infection [2]. Nevertheless, when patients with GC are diagnosed at a late phase or metastasis, they are usually linked with dismal clinical outcomes, with < 30% 5-year survival rate [3]. Consequently, it is imperative to identify more valuable and accurate novel biomarkers to improve the early diagnosis and therapeutic avenues of GC.

Over half of the world's population is infected by HP, a spiral-shaped, gram-negative, microaerophilic bacterium that preferentially colonizes the human gastric mucosa [4]. Due to the majority of HP-positive individuals being asymptomatic or having subtle symptoms, it was able to elude researchers' notice. It has progressively come to light that HP infection may predispose a person to a variety of stomach diseases, including atrophic gastritis, peptic ulcers, and even GC, until Barry Marshall and Robin Warren debunk it [5–8]. Recent research has shown a direct link between HP infection and GC, with individuals with HP positivity experiencing three to six times as many cases as those with HP negativity [9, 10]. Beyond that, HP infection may limit the death of tumor cells and promote gastric carcinogenesis in addition to methylating many cancer-associated genes on CpG islands in gastric epithelial cells [11–14]. The International Agency for Research on Cancer (IARC) recently classified HP as a classification I biological carcinogen, and it has the potential to cause GC by one of the following three mechanisms: DNA damage to epithelial cells, a reduction in repair activity, a mitochondrial DNA mutation, and the emergence of transitory mutation phenotypes are all examples of this [15, 16].

Machine learning (ML), one of the most significant subfields of artificial intelligence (AI), has been extensively utilized in a variety of biomedical domains, including disease diagnosis, biomarker identification, and drug discovery, among others [17, 18]. Moreover, a number of well-known ML techniques, such random forest (RF) and support vector machine (SVM) with recursive feature elimination (RFE), have made significant progress in the creation of anti-cancer drugs and the diagnosis of complicated diseases [19–21].

In this study, based on the differentially expressed HP-related genes, non-negative matrix factorization (NMF) clustering approach was applied to sort The Cancer Genome Atlas (TCGA) stomach adenocarcinoma (STAD), GSE15459, GSE84433 cohorts into two molecular subtypes with different prognosis, immune infiltration landscape, and anticancer drug sensitivity, indicating that the HP-related genes were closely related to the clinical outcomes and therapeutic efficacy of GC. Then, two classical ML algorithms, were employed to identify twelve HP-related hub genes, namely, EFNA3, UHRF1, FLT1, NRP1, CTLA4, L3MBTL3, MAPK10, MLEC, MYL9, THY1, MYB, and NCLN. Subsequently, receiver operating characteristic (ROC) and Kaplan-Meier (K-M) curves were utilized to examine the diagnostic and prognostic performance of these hub genes, and we also explored the anticancer drug sensitivity, immune cell infiltration and mutational features of the hub genes. Finally, quantitative reverse transcription polymerase chain reaction (qRT-PCR) experiments and immuno-histochemical (IHC) images from online browsers were exploited to verify the differential expression of these twelve hub genes.

MATERIALS AND METHODS

Data acquisition and pre-processing

The transcriptome profiles and corresponding clinical information of GC patients in the present study were retrieved and downloaded from the TCGA portal (https://portal.gdc.cancer.gov/, up to May 18, 2022) and GEO database (https://www.ncbi.nlm.nih.gov/geo/, up to May 18, 2022). After filtering out some cases without survival time and status, the ComBat method was employed to correct the batch effects of raw sequencing data sets from different platforms for ensuring comparability among all samples. In the meantime, the somatic mutation and copy number variation (CNV) data of GC patients in TCGA were obtained via querying the UCSC Xena browser (http://xena.ucsc.edu/, up to May 18, 2022), and the "MAFtools" and "RCircos" packages in R software (version 4.1.0, https://www.r-project.org/) were applied to summarize and visualize these data. In addition, the list of HP-related gene sets was collected from the Molecular Signatures Database (MSigDB) (http://www.broadinstitute.org/gsea/msigdb).

NMF consensus clustering analysis

Non-negative matrix factorization (NMF) algorithm is a non-negative factorization of a matrix under the condition that all the elements of the matrix are non-negative, so as to find out the relationships and interactions between them. Elements with similar characteristics are grouped into one group and elements with different characteristics are grouped into another group. Prior to performing the NMF clustering algorithm, the "limma" package with the thresholds of $p < 0.05$ and $|\log FC|$ (fold change) > 1 was utilized to screen out the differentially expressed HP-related genes. Afterwards, on the basis of these genes, the NMF clustering method was implemented to sort TCGA, GSE15459, GSE84433 samples into different clusters by using the "NMF" package, respectively. The number of clusters (K) from 2 to 10 were tested by running ten iterations per K, and the optimal K number was eventually determined in accordance with silhouette, consensus, as well as cophenetic. Principal component analysis (PCA) was used to detect the classification capability of clusters. Then, Kaplan-Meier (K-M) curve was carried out to estimate the differences of OS between distinct subtypes. Moreover, the CIBERSORT and ESTIMATE algorithms were performed to elucidate the immune infiltration landscape of different clusters. According to the half-maximal inhibitory concentration (IC50) value, the "pRRophetic" package was conducted to explore the sensitivity and resistance of common anticancer agents (e.g., Imatinib, Cisplatin) across different clusters. Finally, we also evaluated the expression level of five common immune checkpoint blockade-associated genes (i.e., PDCD1, BTLA, CTLA4, CD274, PD-L2) across different GC subtypes.

Identification of the HP-related hub genes via ML strategies

As it is well known, HP infection is clearly associated with the majority of GC patients [22]. To find out more deeply the impact of HP-related genes on GC patient prognosis, univariate Cox regression analysis ($p < 0.05$) was exploited to screen the prognostic-associated HP genes prior to implementing ML strategies. ML methods displayed excellent and robust performance compared to traditional means, especially in disease diagnosis and predictive analytics [23, 24]. As the two most prevalent supervised ML algorithms, RF and SVM-RFE, were carried out in this study to identify the HP-related hub genes. The establishment of the SVM-RFE model mainly depends on the "e1071", "kernlab", and "caret" packages, and the RF classifier is performed via running the "randomForest" package.

Assessment of diagnostic performance and prognostic value

The area under the ROC curve (AUC) is the gold standard metric in diagnostic performance evaluation, which is calculated through executing the "pROC" package [25]. To obtain into a more convenient clinical application of the HP-related hub genes, a nomogram based on the TCGA-STAD cohort was built to assist with the diagnosis of GC via using the "rms" package. Furthermore, decision, clinical impact, as well as calibration curves were created to examine the sensitivity and accuracy of the diagnosis nomogram.

To investigate the prognostic value of the HP-related hub genes, we categorized all TCGA-STAD patients into high- or low- expression subgroups in accordance with the median expression values of these genes, separately. Next, the K-M method and log-rank test were adopted to confirm the effects of these hub genes on GC prognosis. Apart from this, we also used another survival analysis approach (univariate COX analysis) to discern whether these hub genes were protective or risk factors in the clinical outcomes of GC patients. K-M Plotter is an online website (http://kmplot.com/analysis/) containing a large amount of GEO GC datasets (GSE14210(N=145), GSE15459(N=200), GSE22377(N=43), GSE29272(N=268), GSE51105 (N=94), GSE62254(N=300)), which is able to rapidly estimate the prognostic effects of genes. To demonstrate that the prognostic value of hub genes was not just confined to the TCGA cohort, KM Plotter was applied to further validate their broad applicability in numerous datasets from other sources.

Characteristics of infiltrating immune cells

To reveal the immune-infiltrating landscape of TCGA-STAD samples, the CIBERSORT program was utilized to estimate the relative abundances of 22 types of infiltrating immune cells in each sample by using the "CIBERSORT" package [26, 27]. Furthermore, Spearman's correlation analysis was performed to demonstrate the correlation between the expression levels of these hub genes and tumor-infiltrating immune cells.

Drug response prediction

By logging in to the Gene Set Cancer Analysis (GSCA) (http://bioinfo.life.hust.edu.cn/GSCA/#/drug) database, the association between the sensitivity of drugs derived from the Genomics of Drug Sensitivity in Cancer (GDSC) database and the mRNA expression of hub genes was explored [28]. At the same time, the chemical structural formulas of GDSC agents that exhibited a significant positive or negative correlation with the expression of these hub genes were collected by querying the MedChemExpress website (https://www.medchemexpress.cn/).

Functional and pathway enrichment analysis

To assess the interaction among hub genes, a protein–protein interaction (PPI) network was created to examine how closely they were connected via using the GeneMANIA (http://genemania.org/) online tool. Moreover, an interaction map between hub genes and cancer-associated pathways was constructed to further explore the biological role of HP-related hub genes in the onset and progression of GC by making use of the GSCALite (http://bioinfo.life.hust.edu.cn/web/GSCALite/). Gene set enrichment analysis (GSEA), another powerful method for functional enrichment of gene sets, was carried out to probe the biological functions of hub genes. Investigating the upstream regulated miRNAs is essential to gain insight into the mechanisms of action of genes, and a miRNA-mRNA regulatory network for hub genes was also predicted and generated through the GSCALite website.

Validation of hub gene expression

For mRNA expression levels of hub genes, qRT-PCR experiments were conducted in accordance with manufacturers' instructions and our previous study [29]. A normal gastric epithelial cell line (GES-1) and five human GC cell lines (HGC-27, AGS, MKN-45, MGC803, and MKN-28) were obtained from the Shanghai Cell Bank of the Chinese Academy of Sciences. The glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was treated as internal reference. The 2−ΔΔCq method was adopted to calculate the relative expression levels of hub genes, and GraphPad Prism 6.0 software was utilized to plot bar graphs. Apart from this, the mRNA primer sequences of hub genes were summarized in detail (Supplementary Table 1).

The Human Protein Atlas (HPA) (http://www.proteinatlas.org/) database provides large proteomics data of various normal tissues and corresponding cancers [30]. To verify the protein expression abundance of hub genes, immunohistochemical (IHC) staining images from HPA project were extracted for analysis and comparison.

Data availability statement

The datasets presented in this study can be found in online repositories (TCGA, https://portal.gdc.cancer.gov/), (GEO, http://www.ncbi.nlm.nih.gov/geo/). The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Consent for publication

All contributors give consent for unrestricted publication of this work.

RESULTS

HP-related gene screening and functional enrichment analysis

The flow-process diagram for this present research was summarized in Figure 1. Firstly, a total of 761 HP-related gene sets were extracted from the MSigDB website (Supplementary Table 2). Subsequently, differential expression analysis was carried out to

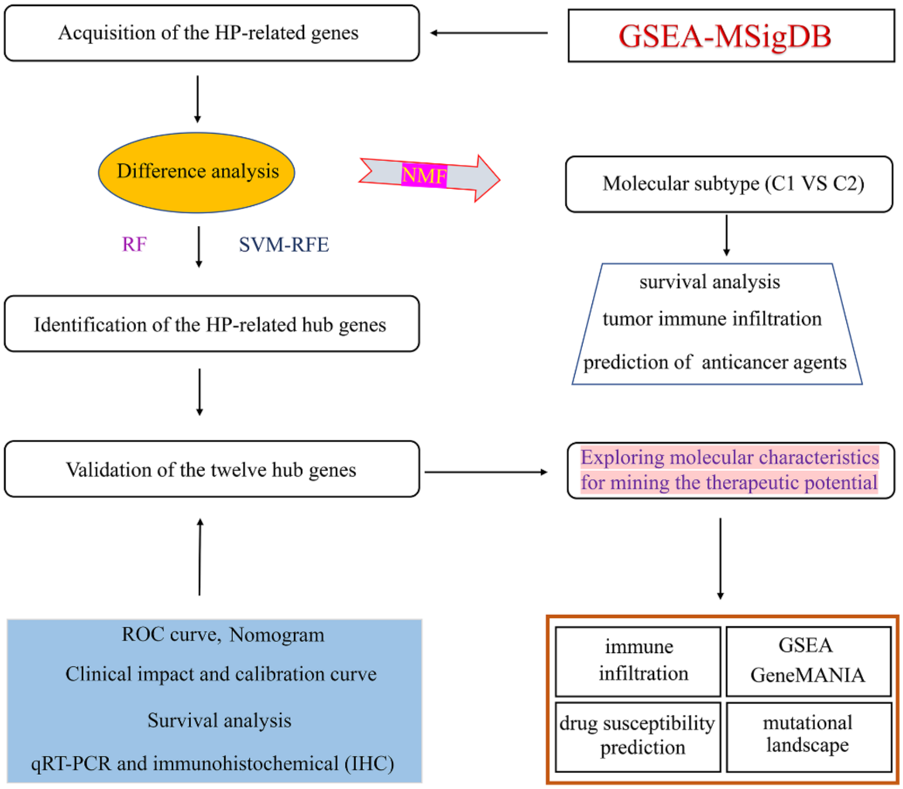

Acquisition of the HP-related genes
GSEA-MSigDB
Difference analysis
NMF
Molecular subtype (C1 VS C2)
RF
SVM-RFE
survival analysis
tumor immune infiltration
prediction of anticancer agents
Identification of the HP-related hub genes
Validation of the twelve hub genes
Exploring molecular characteristics for mining the therapeutic potential
ROC curve, Nomogram
Clinical impact and calibration curve
Survival analysis
qRT-PCR and immunohistochemical (IHC)
immune infiltration
GSEA GeneMANIA
drug susceptibility prediction
mutational landscape

Figure 1. Flowchart illustrating the workflow of this study.

screen 232 differentially expressed HP-related genes via using the "limma" package (Supplementary Table 3).

To further unravel the potential mechanisms of these genes in the occurrence and progression of GC, Gene Ontology (GO) functional and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotation analyses were executed by implementing the "clusterProfiler" package. GO analysis revealed that these genes were mainly enriched in cancer-associated biological functions, such as mitotic cell cycle, cell cycle, cell cycle process, chromosome organization, DNA replication, cell division, DNA conformation change, chromosomal region, condensed chromosome, heterochromatin, chromosome, ribonucleotide binding, DNA helicase activity, helicase activity, and integrin binding (Supplementary Figure 1A–1C). KEGG analysis disclosed the enriched top signaling pathways, including epithelial cell signaling in *Helicobacter pylori* infection, Vibrio cholerae infection, NOD-like receptor signaling pathway, pathways in cancer, spliceosome, Toll-like receptor signaling pathway, leukocyte transendothelial migration, as well as cytokine–cytokine receptor interaction (Supplementary Figure 1D).

Establishment of molecular subtype based on the HP-related genes

NMF consensus clustering was conducted on the TCGA-STAD (N=338), GSE15459 (N=192), GSE84433 (N=357) cohorts respectively on the basis of 232 differentially expressed HP-related genes.

Following the prompts of silhouette, consensus, as well as cophenetic, all samples of the three different datasets were eventually split into 2 clusters (Figure 2A and Supplementary Figures 3A, 4A). The silhouette, consensus, and cophenetic heatmaps of each dataset were displayed in Supplementary Figure 2. PCA results suggested that clusters based on HP-related genes had excellent classification ability in distinct datasets (Figure 2B and Supplementary Figures 3B, 4B). Survival analysis also revealed that there was an obvious difference in the OS of patients between Cluster C1 and C2 in all datasets (Figure 2C and Supplementary Figures 3C, 4C). The tumor microenvironment (TME) has a close correlation with patient prognosis and tumor progression, and the immune and stromal cells constitutes the main components of TME [31]. The TME score files showed that patients with Cluster C1 yielded a lower immune score than that of C2, and the CIBERSORT algorithm was adopted to further explore the composition of immune cell infiltration between different clusters (Figure 2D, 2E and Supplementary Figures 3D, 3E, 4D, 4E). Notably, in the GSE84433 cohort, the infiltration abundance of immunosuppressive Tregs and Macrophages M2 in Cluster C1 was significantly higher than that in Cluster C2, which may be an important reason for the worse prognosis of patients with Cluster C1 (Supplementary Figure 3F). In contrast, in the TCGA-STAD cohort, patients in Cluster C1 exhibited a lower level of Tregs infiltration compared with that in Cluster C2, resulting in the better clinical outcomes for patients Cluster C1 (Figure 2F). Beyond this, in the GSE15459 cohort, a prevalent amplification in anti-tumorigenic immune cells, including NK cells resting, NK cells activated, Macrophages M0, Mast cells activated, Dendritic cells activated, was observed in Cluster C1, whereas a higher infiltration degree with Macrophages M2 was found in Cluster C2, which was a great explanation for the poorer survival rate of patients with Cluster C2 (Supplementary Figure 4F).

Considering that chemotherapy and targeted therapy are still served as a cornerstone of treatment for GC, we calculated the IC50 values of common anticancer agents in the two clusters to predict drug susceptibility or resistance. Immunotherapy has been an emerging therapeutic modality for GC, and the expression levels of five common immune checkpoint genes across different clusters were compared to predict the immunotherapeutic response of GC patients. Surprisingly, in the TCGA-STAD cohort, the IC50 values of imatinib and sunitinib were significantly lower in Cluster C2 than in Cluster C1, while the expression levels of BTLA and PD-L2 were obviously higher in C2 than in C1, all indicating that patients in the Cluster C2 were more likely to benefit from targeted therapy and immunotherapy to improve their dismal prognosis (Figure 2G–2I). Similarly, in the GSE84433 and GSE15459 cohorts, the expression levels of immune checkpoint genes and IC50 values of anticancer drugs across two clusters were also totally disparate (Supplementary Figures 3G–3I, 4G–4I).

Collectively, these results revealed that HP-related genes were closely related to the prognosis and treatment of GC patients, and the novel molecular subtype had great clinical practicality and wide applicability.

ML identifying the HP-related hub genes

Before implementing the ML algorithms, univariate Cox regression analysis was employed to obtain seventeen prognostic-associated HP genes from the above differentially expressed HP-related genes (Supplementary Table 4). Then, two classical ML methods, RF and SVM-RFE, were conducted to determine the HP-related hub genes. Based on the RNA-sequencing data of the seventeen prognostic-associated HP genes in TCGA-STAD dataset, thirteen candidate genes were identified through the feature selection of RF model (Figure 3D, 3E and Supplementary Table 5), and thirteen genes were acquired via implementing the SVM-RFE strategy (Figure 3F and Supplementary Table 6). Reverse cumulative distribution and boxplots plots showed that residual values of the two ML algorithms considered negligible (Figure 3A, 3B), and ROC curves demonstrated that both the SVM-RFE and RF models in this present study had a very robust accuracy score (SVM-RFE: 0.997, RF: 1) (Figure 3C). Therefore, by taking the intersection of the results of two ML strategies, twelve genes were determined and served as HP-related hub genes for follow-up analysis (Figure 4A).

Evaluation of diagnostic performance and prognostic value of the twelve hub genes

To confirm the strong diagnostic power of the twelve hub genes, ROC curves were drawn by using the "pROC" package. All hub genes in the TCGA-STAD cohort reached the AUC values of 0.713–0.958, of

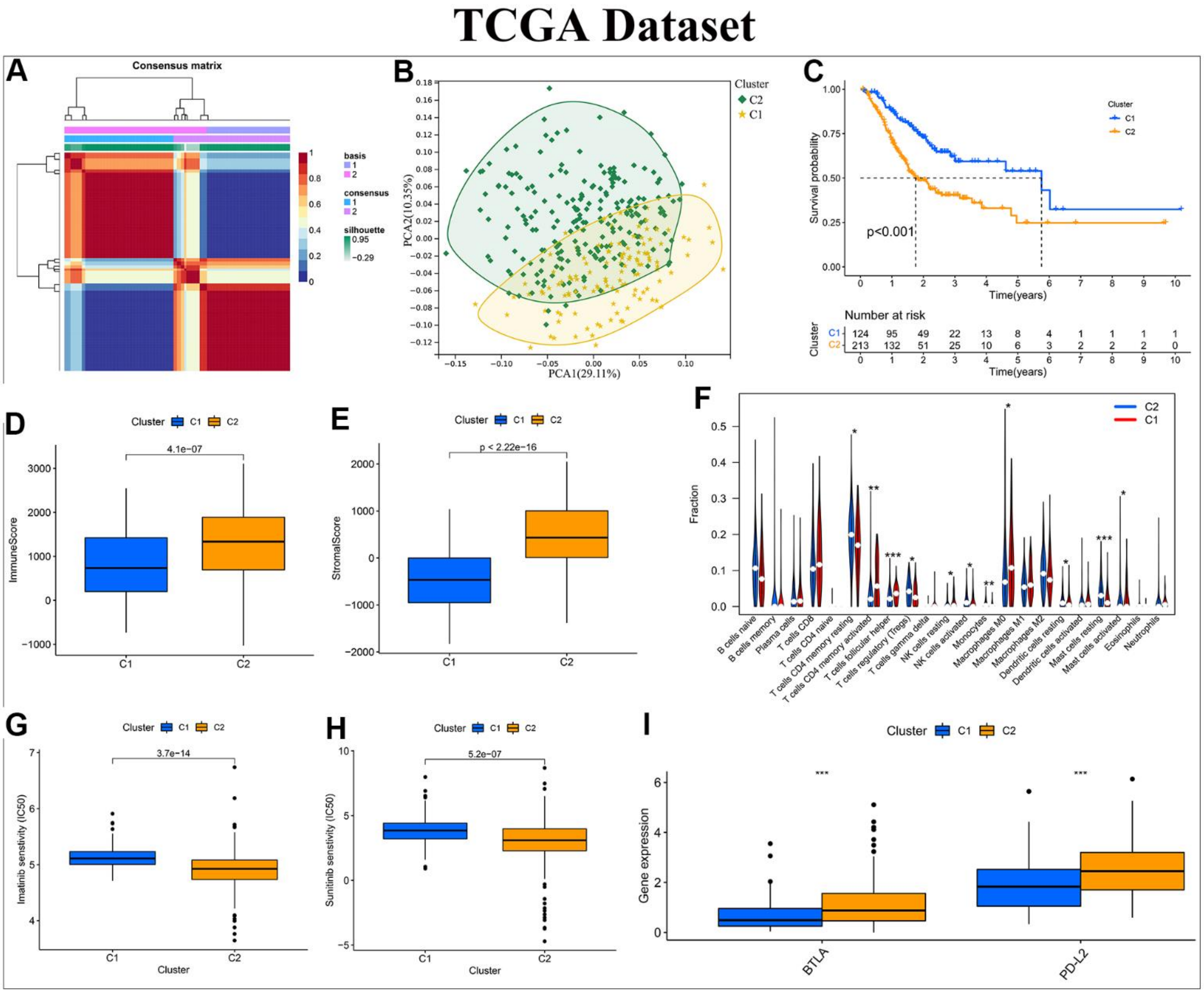


Figure 2. Construction of a NMF subtype based on the differentially expressed HP-related genes in the TCGA-STAD cohort. (**A**) NMF consensus clustering for k = 2. (**B**) Kaplan–Meier analysis of overall survival (OS) for Cluster C1 and C2. (**C**) Principal component analysis (PCA). (**D**, **E**) Differential analyses of immune and stromal score between Cluster C1 and C2. (**F**) Violin plot showing the immune cell infiltration landscape across different clusters. (**G**, **H**) Box plot of estimated IC50 values for Imatinib and Sunitinib in Cluster C1 and C2. (**I**) Box plot visualizing the significant expression differences of immune checkpoints across distinct clusters, including BTLA and PD-L2. *:P<0.05 ** :P<0.01 ***:P<0.001.

which UHRF1 achieved the highest AUC value of 0.958 (Figure 4B). To better make use of these hub genes, a nomogram containing all hub genes was established for the diagnosis of GC through the “rms” package (Figure 4D). The well-calibrated capability of the nomogram was observed by checking the calibration curve, and the mean absolute calibration error was only 0.006 (Figure 4E). Both Clinical impact curve and decision curve analysis (DCA) further affirm the clinical utility of the diagnostic nomogram (Figure 4C, 4F). Overall, the twelve HP-related hub genes could be expected to develop into the ideal diagnostic markers of GC.

Given that patient prognosis is what counts, we sorted all TCGA-STAD samples into high- or low -expression subgroups according to the median expression of the twelve hub genes for subsequent survival analysis. K-M curves uncovered that only eight hub genes (EFNA3, FLT1, L3MBTL3, MAPK10, MLEC, MYB, NRP1, as well as UHRF1) existed a significant difference in the clinical outcomes of GC patients between high and low-expression subgroups (Figure 5A–5H). However, univariate Cox regression analysis was also conducted to examine the association between the expression levels of hub genes and GC prognosis, and subsequent results revealed that all twelve hub genes were closely related to the survival time and status of patients with GC (Figure 5I). Thus, we could speculate boldly that the twelve HP-related hub genes affected these survival outcomes of GC patients together rather than individually.

Aside from this, the K-M Plotter database including six GEO datasets (GSE14210(N=145), GSE15459(N=200), GSE22377(N=43), GSE29272(N=268), GSE51105 (N=94), GSE62254(N=300)) were also utilized to mine the prognostic value of hub genes. The results of KM Plotter website suggested that the expression levels of twelve hub genes exert a drastic effect on the survival time in GC patients whether it was OS or progression-free survival (PFS) (supplementary Figures 5, 6).

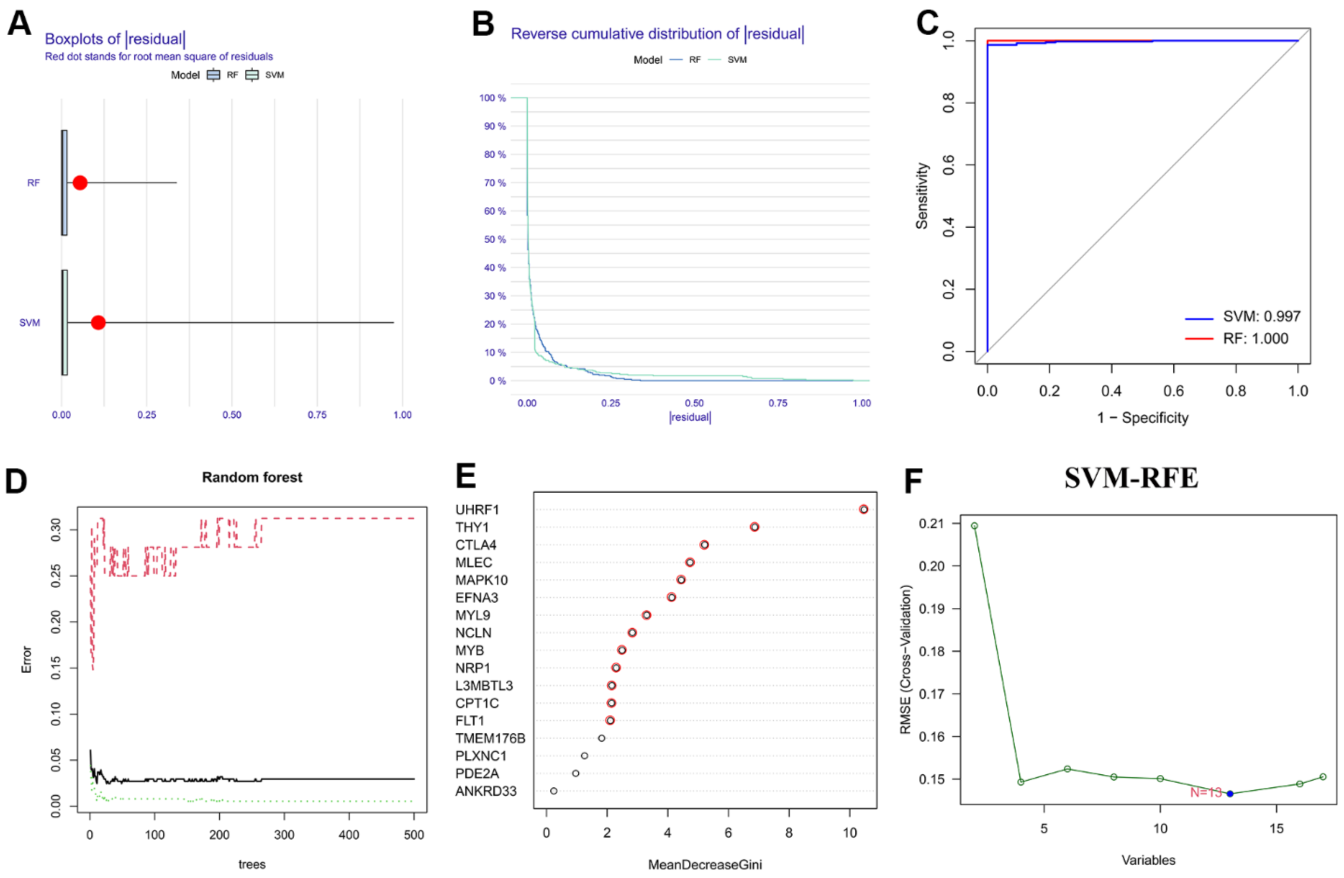


Figure 3. Selection of the HP-related hub genes via machine learning strategies. (**A**, **B**) Boxplot and reverse cumulative distribution curve of residual. (**C**) Comparison of ROC curves for evaluating the diagnostic reliability of support vector machine-recursive feature elimination (SVM-RFE) and random forest (RF) models. (**D**) Error graph of RF model. (**E**) Based on RF algorithm to screen the HP-related hub genes. (**F**) On the basis of SVM-RFE method to identify the HP-related hub genes.

Immune infiltration landscape of the hub genes

The state of TME profoundly influences the efficacy of immunotherapy [32]. Using the CIBERSORT algorithm, we deconvoluted the composition ratio of distinct immune cell subpopulations in the TCGA-STAD patients. Subsequently, Spearman's correlation analysis was conducted to discover the relationship between the expression levels of the twelve genes and the degree of immune cell infiltration. Expression of CTLA4 and MYB were negatively correlated with the infiltration abundance of Monocytes, Mast cells resting, as well as T cells CD4 memory resting, while presenting the highest positive correlation coefficient with T cells CD4 memory activated (Figure 6A, 6G). The expression of FLT1 was positively correlated with NK cells resting and B cells naive and negatively correlated with NK cells activated and B cells memory, while the expression level of L3MBTL3 was positively correlated with B cells naïve and Tregs and negatively correlated with Neutrophils and Plasma cells (Figure 6C, 6D). Similarly, NRP1 and THY1 were positively related to Macrophages M2 and Macrophages M1, whereas inversely related to B cells memory and Plasma cells (Figure 6J, 6K). Of interest, EFNA3, MLEC, NCLN together with UHRF1 all exhibited the highest positive correlation with Macrophages M0, and the most significant negative correlation with Mast cells resting (Figure 6B, 6F, 6I, 6L). Besides that, the expression levels of both MAPK10 and MYL9 were

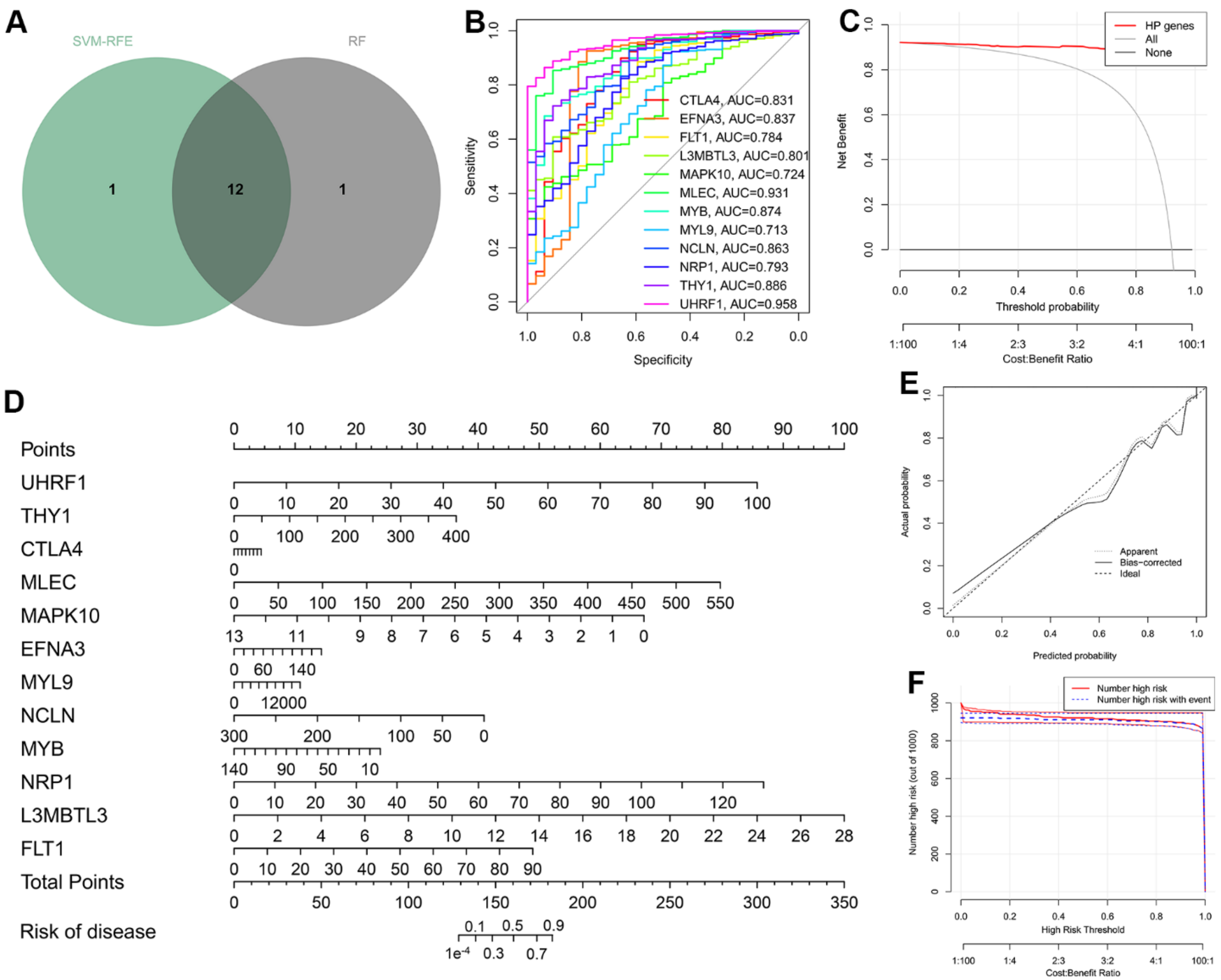


Figure 4. Construction of the diagnostic nomogram on the basis of the twelve hub genes. (**A**) Venn diagram taking the intersection of the results of two ML strategies. (**B**) ROC curves measuring the diagnostic efficacy of the twelve HP-related hub genes. (**C**) Decision curve of nomogram graph. (**D**) Nomogram for the diagnosis of gastric cancer (GC). (**E**) Calibration curve demonstrating the diagnostic performance of the nomogram. (**F**) Clinical impact curve.

positively correlated with Mast cells resting, Monocytes, and B cells naive, but negatively correlated with T cells CD4 memory activated as well as Macrophages M0 (Figure 6E, 6H). Taken together, the hub genes could reshape the immune microenvironment to facilitate the initiation and progression of GC by altering the degree of various immune cell infiltration.

Mutational characteristic of the hub genes

As it is well-known, tumor mutation burden (TMB) has been increasingly recognized as being significantly associated with patient prognosis and immunotherapy response [33–35]. By exploiting the genomic alteration data from the TCGA database, the copy number variant (CNV) and single-nucleotide variant (SNV) events of the twelve hub genes were investigated to explore the relationship between expression and genetic mutation. EFNA3 exhibited the highest frequency of CNV, observed in exceeding 10% of TCGA-STAD samples, with CNV Gain being the more prevalent type compared to CNV Loss (Figure 7A). At the same time, the CNV frequencies over 3% for all hub genes were also found (Figure 7A). As displayed in Figure 7B, the chromosomal region and CNV state of all twelve genes were carefully marked on the schematic diagram. Further analysis suggested that the mRNA expression levels of MLEC, MYB, NCLN, EFNA3, and UHRF1 were positively related to their CNV frequencies (Figure 7C). Of the 439 GC samples, 52 (11.85%) occurred the alteration of hub genes, and FLT1 presented the highest SNV frequency (3.6%), followed by NRP1 (3.4%), NCLN (1.8%) (Figure 7D). Alongside this, subsequent results indicated that missense mutation (>60%), single-nucleotide polymorphism (SNP) (>60%), as well as C > T (34) were the most frequent classifications of SNV (Figure 7E).

In summary, among these 12 hub genes, the abnormal expression of EFNA3 and UHRF1 in GC compared

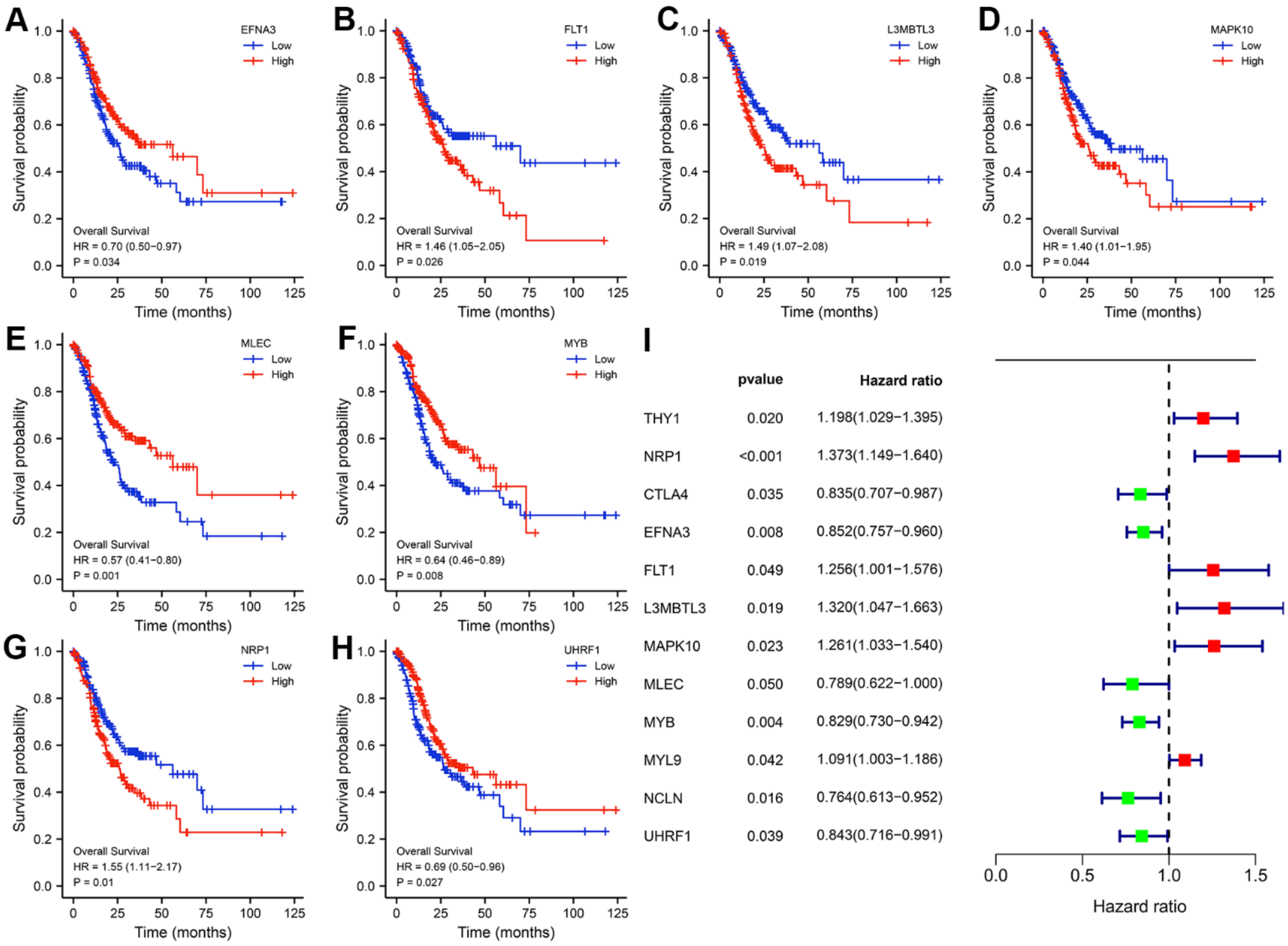


Figure 5. Kaplan-Meier (K-M) survival curves of the hub genes. (**A**) EFNA3. (**B**) FLT1. (**C**) L3MBTL3. (**D**) MAPK10. (**E**) MLEC. (**F**) MYB. (**G**) NRP1. (**H**) UHRF1. Univariate Cox regression analysis of the twelve hub genes. (**I**) Forest plot showing the prognostic values of hub genes.

with normal gastric tissues might be impacted by the regulation of CNV, whereas SNV was likely to affect the expression levels of FLT1, NRP1, CTLA4, L3MBTL3, MAPK10, MLEC, MYL9, as well as THY1, and MYB along with NCLN were perhaps subject to the joint effects of CNV and SNV.

Assessment of drug sensitivity or resistance targeting the hub genes

Chemotherapy and targeted therapy are still as important as surgery in the treatment of GC. Based on the GSCA database, we evaluated the association between the expression levels of hub genes and GDSC drug susceptibility or resistance. Subsequent results revealed that the sensitivity of several GDSC agents was positively and negatively correlated with the expression of multiple hub genes, including AZD8055, CI-1040, PLX4720, TPCA-1, Vorinostat, CEP-701, THZ-2-102-1, UNC0638, IPA-3, KIN001-260, SB590885, and KIN001-270 (Figure 8A). Moreover, the chemical and molecular structural formulas of the abovementioned twelve drugs were acquired by querying the MedChemExpress online website. The chemical formulas of AZD8055, CI-1040, PLX4720, TPCA-1, Vorinostat, CEP-701, THZ-2-102-1, UNC0638, IPA-3, KIN001-260, SB590885, and KIN001-270 corresponded to $C_{25}H_{31}N_5O_4$, $C_{17}H_{14}CIF_2IN_2O_2$, $C_{17}H_{14}CIF_2N_3O_3S$, $C_{12}H_{10}FN_3O_2S$, $C_{14}H_{20}N_2O_3$, $C_{26}H_{21}N_3O_4$, $C_{31}H_{28}CIN_7O_2$, $C_{30}H_{47}N_5O_2$, $C_{20}H_{14}O_2S_2$, $C_{21}H_{24}N_4O_2$, $C_{27}H_{27}N_5O_2$, $C_{26}H_{21}N_5O_4S$, separately. As shown in Figure 8B, the construct formulas of these agents were neatly arranged.

Functional enrichment analysis

The synergy among hub genes was fully reflected in the above content. Therefore, to confirm the degree of tight junction of these hub genes, a PPI network was construed by using the GeneMANIA website (Figure 9A). To further examine the role of these hub genes in the occurrence and development of GC, the GSCALite database was utilized to investigate the interaction between hub genes and cancer-related pathways. The interaction maps uncovered that almost all hub genes were involved in the cancer-associated pathways,

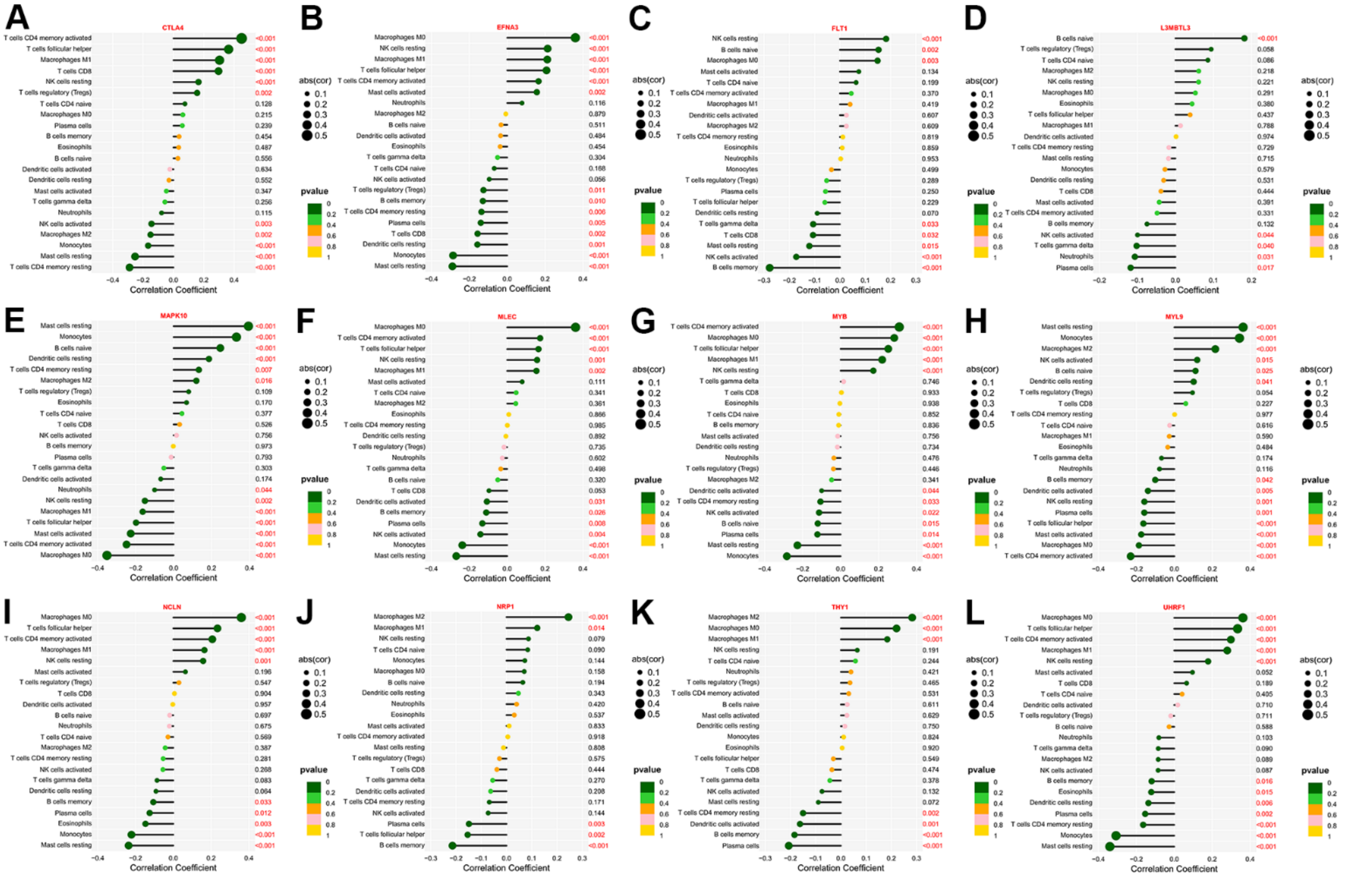


Figure 6. The immune-infiltrating landscape of GC based on the twelve hub genes. (**A**–**L**) Lollipop plots revealing the association between the twelve hub genes and the infiltration level of various immune cells.

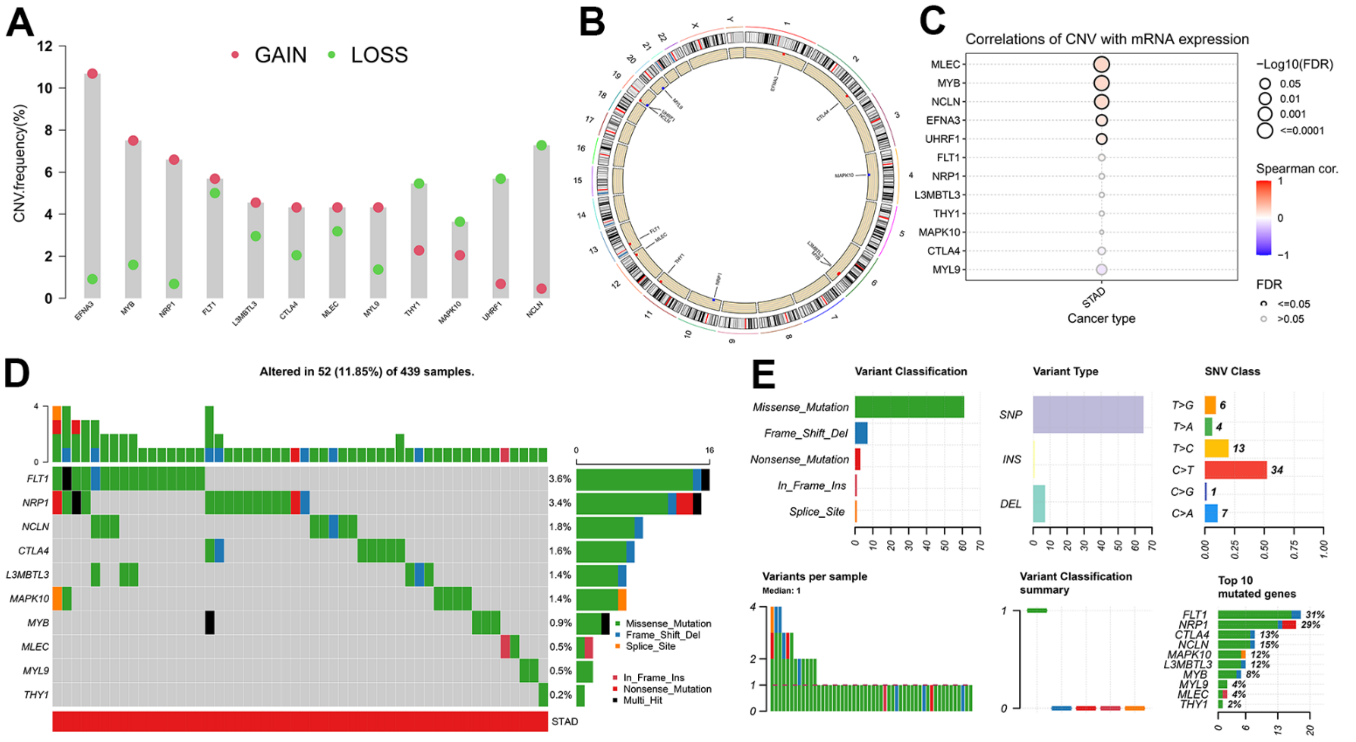


Figure 7. Mutational characteristics of the hub genes. (**A**) Copy number variation (CNV) frequency of hub genes. (**B**) Circle diagram of CNV with hub genes. (**C**) Correlation between expression of hub genes and CNV. (**D**) Cascade of hub gene mutations. (**E**) Details regarding single nucleotide variants (SNV).

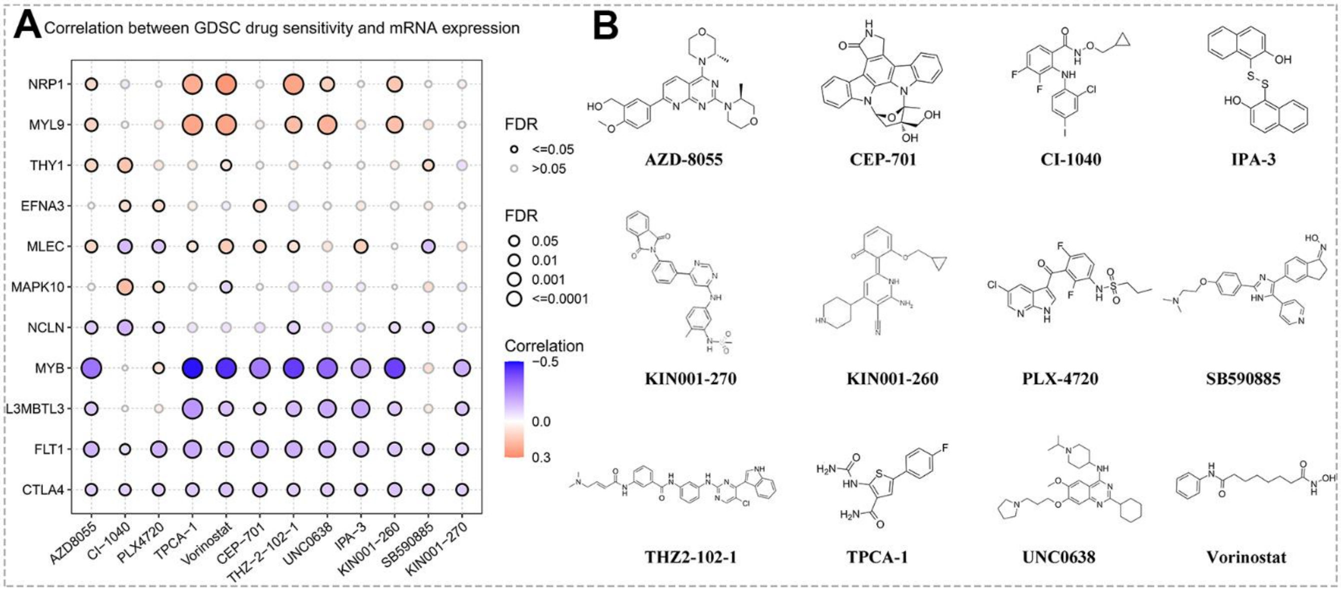


Figure 8. Prediction of drug sensitivity. (**A**) Correlation between hub gene expression levels and GSDC drug sensitivity via the online search tool GSCA. (**B**) Structural formulas of the sensitive agents (including AZD8055, CI-1040, PLX4720, TPCA-1, Vorinostat, CEP-701, THZ-2-102-1, UNC0638, IPA-3, KIN001-260, SB590885, and KIN001-270).

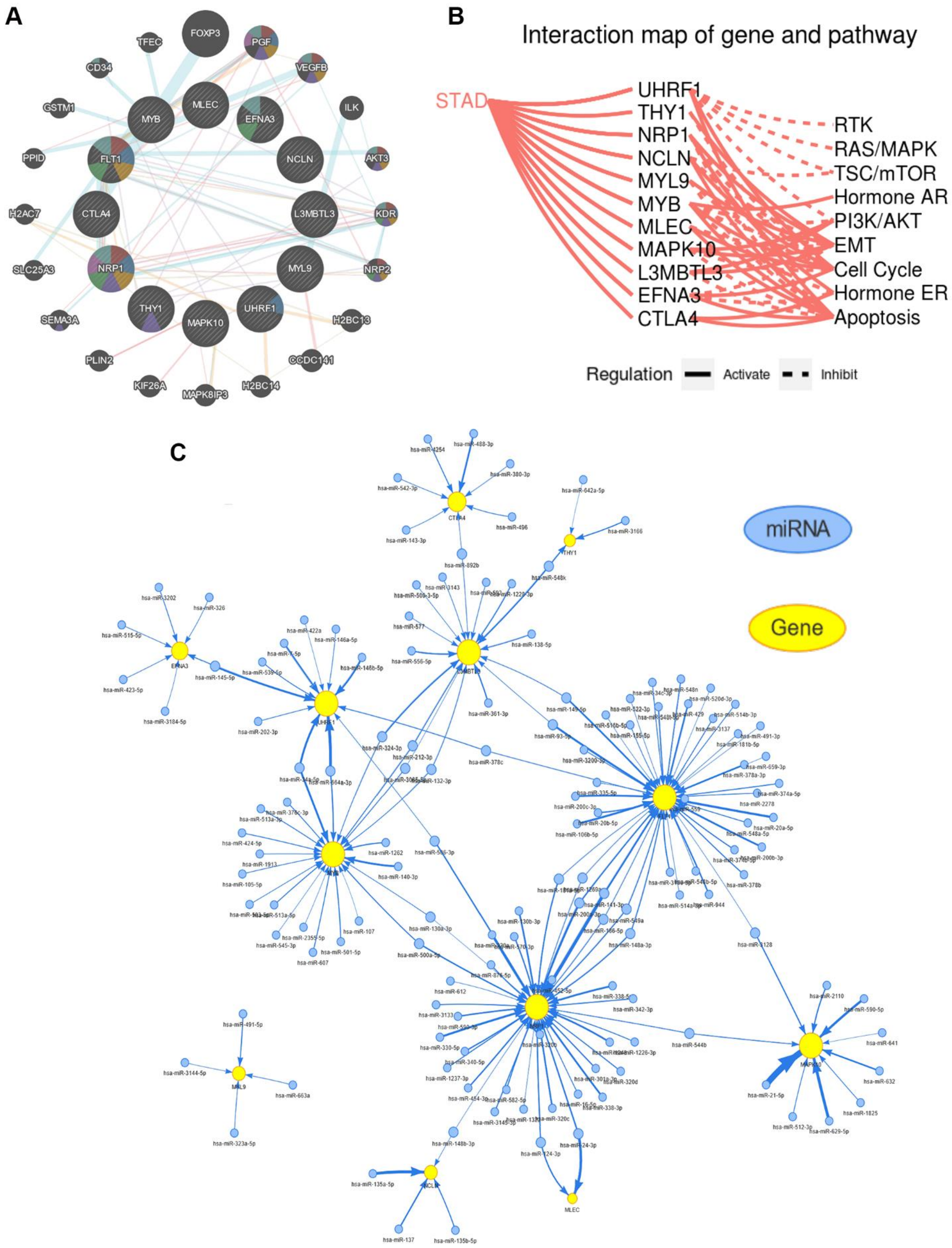


Figure 9. Functional and pathway enrichment analysis of the hub genes. (**A**) Construction of a protein-protein interaction (PPI) network through using the GeneMANIA database. (**B**) The hub genes being involved in several key cancer-associated processes, such as epithelial-mesenchymal transition (EMT), receptor tyrosine kinase (RTK), cell cycle, apoptosis, etc. (**C**) The result of predicted miRNAs targeting hub genes using the GSCALite website.

comprising RAS/MAPK, epithelial–mesenchymal transition (EMT), TSC/mTOR, receptor tyrosine kinase (RTK) signaling, hormone AR signaling, cell cycle, PI3K/AKT, hormone ER signaling, as well as apoptosis (Figure 9B).The upstream miRNAs is essential for understanding the oncogenic or tumor suppressor mechanism of gene, and a miRNA-mRNA regulatory network for hub genes was also created by using the GSCALite database, indicating that these hub genes shared common miRNAs to participate in various biological behavior of GC (Figure 9C). Furthermore, we again split TCGA-STAD samples into two subgroups (high- or low- expression) in accordance with the median expression of these hub genes for performing the GSEA analysis (Supplementary Figure 7). GSEA results showed that CTLA4, L3MBTL3, MAPK10, MLEC, MYB, MYL9, NCLN, NRP1, THY1, and UHRF1 were mainly enriched in pathways correlated with pathways in cancer, cell cycle, DNA replication, ECM-receptor interaction, cell adhesion molecules (CAMs), bladder cancer, renal cell carcinoma, focal adhesion, natural killer cell mediated cytotoxicity, T cell receptor signaling pathway, spliceosome, calcium signaling pathway, base excision repair, purine metabolism, pyrimidine metabolism, chemokine signaling pathway, mTOR signaling pathway, N-glycan biosynthesis, in the high-expression subgroup, whereas EFNA3 and FLT1 were involved in the calcium signaling pathway, ribosome, oxidative phosphorylation, hematopoietic cell lineage, Parkinsons disease, and vascular smooth muscle contraction in the low-expression subgroup.

Validation of differential expression for hub genes

The qRT-PCR assay was commonly applied to measure the mRNA expression level of gene, and we compared the mRNA expression of hub genes in normal gastric epithelial and multiple GC cells through the qRT-PCR experiment in this present study, suggesting that all twelve of these genes were obviously different (Figure 10). The HPA website is a repository of immunohistochemistry-based proteomic data that provides us with significant value for protein expression analysis [36]. The IHC staining image were retrieved and downloaded from the HPA database to validate the protein expression of hub genes, indicating that these

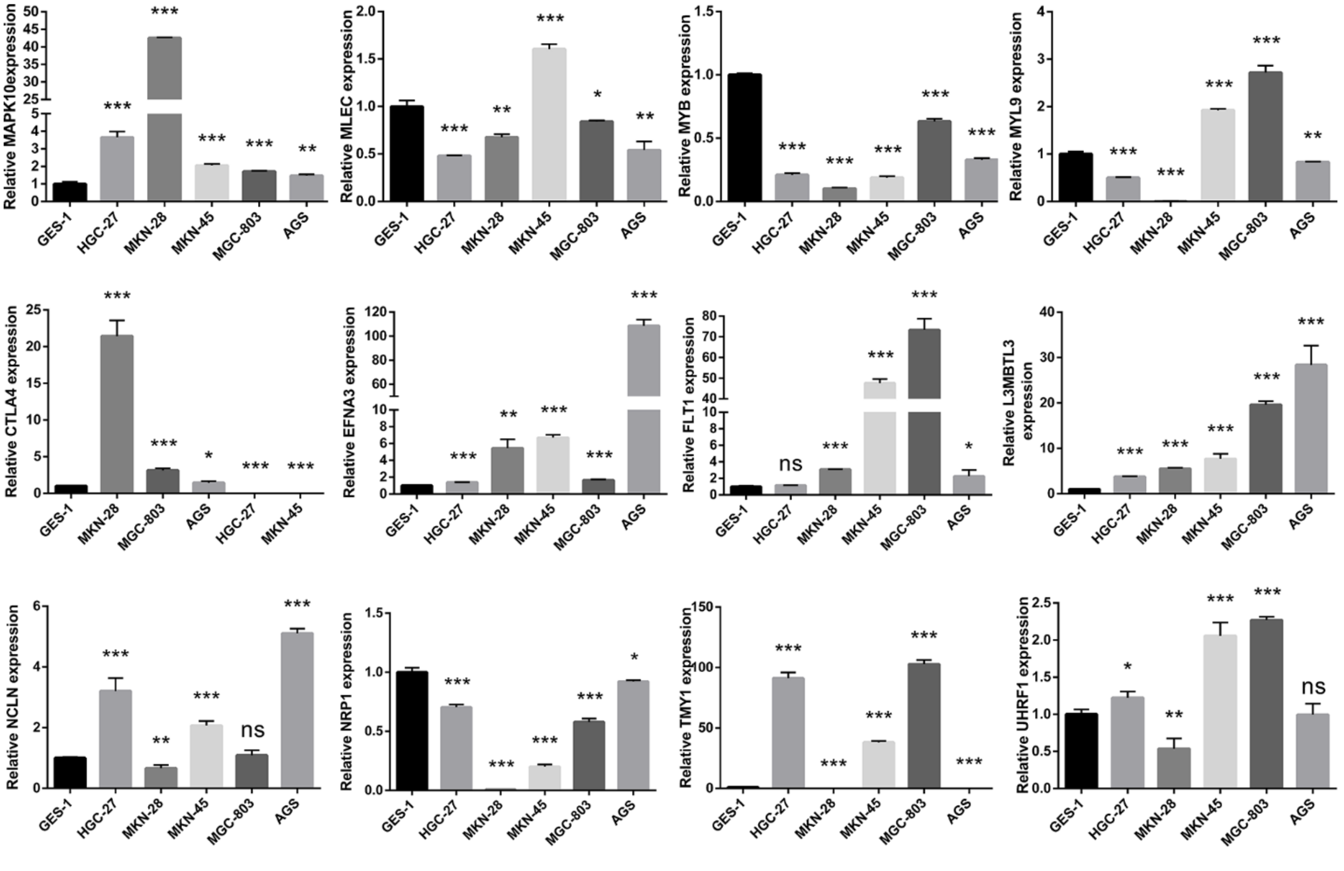


Figure 10. Validating the mRNA expression of the twelve hub genes in normal gastric epithelial and GC cell lines via the quantitative reverse transcription polymerase chain reaction (qRT-PCR) assays.

genes were differentially expressed between GC and adjacent normal tissues, and the trend of up-regulation or down-regulation was basically consistent with the mRNA expression level (Supplementary Figure 8 and Supplementary Table 7).

DISCUSSION

It is well known that HP infection is a premalignant form of GC and that it plays a key role in the development and spread of the disease [37]. Although eliminating HP can greatly lower the frequency of GC, its prevalence is still high, particularly in underdeveloped nations [38–40]. The relationship between HP infection and GC has been the subject of numerous studies, but these studies primarily concentrate on the oncogenic effects of HP strains' virulence factors, such as *BabA*, *CagA*, *oipA*, and *VacA*, as well as environmental risk factors like increased or decreased gastric fluid pH and nitrosamines and their precursors [41–44]. On the molecular and genetic levels, nevertheless, there are very few investigations on the tumorigenic processes of HP infection. It is generally established that dysregulation of oncogene and tumor-suppressor genes can contribute to the beginning and growth of tumors. Therefore, in this investigation, we sought to clarify the therapeutic and prognostic effects of HP infection on GC based on HP-related genes.

At the beginning of this present study, we conducted NMF clustering analysis on the TCGA-STAD, GSE15459, GSE84433 cohorts based on 232 differentially expressed HP-related genes. Subsequent results revealed that all different data sets' samples were stratified into two distinct clusters with differing prognosis, immune infiltration landscape, and anti-cancer drug sensitivity, indicating that the HP-related genes could exert a significant influence on the clinical outcomes, tumor microenvironment, as well as therapeutic efficacy of GC.

To determine the most critical HP-related genes, two ML methods, SVM-RFE together with RF, were employed to identify the HP-associated hub genes. SVM-RFE, a quite well-established ML algorithm for classification, can achieve the selection of optimal features on the basis of the recursive feature elimination [45]. RF is also a supervised non-parametric ML approach, which can be applied to address classification and regression issues, including gene screening and disease diagnosis [46]. Twelve genes were eventually screened and identified as the HP-related hub genes by intersecting of the results of two ML strategies, namely, EFNA3, UHRF1, FLT1, NRP1, CTLA4, L3MBTL3, MAPK10, MLEC, MYL9, THY1, MYB, as well as NCLN, since both the SVM-RFE and RF models presented less residuals and higher ROC values.

It has previously been demonstrated that multiple hub genes are closely implicated in tumorigenesis and progression. For instance, EFNA3/EPHA2 axis can promote cancer stemness in hypoxic hepatocellular carcinoma by modulating metabolic plasticity, and EFNA3 is served as a prognostic biomarker for hepatocellular cancer [47, 48]. UHRF1 can facilitate the occurrence and development of various digestive tract tumors, including gastric, colon, and pancreatic cancers, etc. [49–51]. FLT1 and its ligands VEGFB together with PlGF are promising as key targets for a new generation of anti-angiogenic drugs [52]. NRP1 is closely associated with the occurrence, progression and even metastasis of various tumors, such as bladder, colorectal, breast, and lung cancers [53–56]. CTLA-4 has been generally recognized as the most compelling target immunotherapy, and ipilimumab (anti-CTLA4) has radically and significantly improved the clinical outcomes of patients with advanced cancer [57]. Dysregulated miR-27a-3p enhances the proliferation and migration capability of nasopharyngeal carcinoma cell by regulating the expression level of MAPK10, and Circ_0000515 can also drive hepatocellular carcinoma progression by targeting MAPK10 [58, 59]. Aberrant Expression of MYL9 is correlated with prognosis of glioblastoma and esophageal squamous cell cancer, and it may act as a novel biomarker [60, 61]. MicroRNA-140-5p suppresses the growth and progression of GC cells by reversely modulating THY1-mediated Notch signaling [62]. SNHG3 promotes the proliferation and metastatic ability of GC cells by mediating the miR-139-5p/MYB axis [63].

To examine the diagnostic performance of the twelve hub genes, ROC curves of each hub genes were plotted and their AUC values were calculated. All hub genes reached high AUC values, with UHRF1 exhibiting the highest AUC value of 0.958. To predict the likelihood of initiation of GC, a nomogram on the basis of the twelve HP-related hub genes was constructed. To illustrate the clinical significance of the hub genes, we performed survival analysis (K-M and univariate Cox regression analyses). Subsequent results suggested that the expression levels of the nine hub genes was closely to the survival outcomes of patients with GC in both TCGA and GEO cohorts.

Immunotherapy has emerged as a promising treatment strategy for GC, yet drug sensitivity varies from person to person. The composition of immune cell infiltration affects the immunotherapy response and is served as a significant determinant [64]. In this research, expression of CTLA4, MYB, FLT1, L3MBTL3, MAPK10 and MYL9 positively correlated with the abundance of antitumor immune cells (e.g., T cells CD4, Mast cells, NK cells, B cells, etc.), while NRP1 and THY1 presented a negative association with the infiltration level of immunosuppressive Macrophages M2. TMB is another indicator for evaluating the response to immunotherapy [65]. In the TCGA-STAD dataset, EFNA3 showed the highest frequency of CNV, found in more than 10% of patients, with CNV Gain being the more prevalent type. At the same time, FLT1 exhibited the highest SNV frequency (3.6%), with missense mutation and C > T being the main classification.

At final, we also discovered that these hub genes may participate in the onset and progression of GC via the following cancer-associated pathways, namely, cell cycle, EMT, apoptosis, RAS/MAPK, PI3K/AKT, TSC/mTOR, hormone AR signaling, hormone ER signaling, as well as RTK. Besides that, the twelve hub genes' upstream regulatory miRNAs were predicted.

CONCLUSIONS

In this study, we developed a unique HP-related tumor classification and performed ML techniques to identify twelve hub genes that may be useful for GC molecular diagnosis and individualized treatment.

Abbreviations

HP: *Helicobacter pylori*; GC: Gastric cancer; TCGA: The Cancer Genome Atlas; GEO: Gene Expression Omnibus; MSigDB: Molecular Signatures Database; NMF: Non-negative matrix factorization; STAD: Stomach Adenocarcinoma; SVM-RFE: support vector machine-recursive feature elimination; RF: Random Forest; ML: machine learning; qRT-PCR: Quantitative reverse transcription polymerase chain reaction; IHC: Immunohistochemical; ROC: Receiver operating characteristic; K-M: Kaplan-Meier; EMT: Epithelial-mesenchymal transition; IARC: International Agency for Research on Cancer; AI: Artificial intelligence; CNV: Copy-number variant; FC: Fold change; PCA: Principal component analysis; IC50: Half-maximal inhibitory concentration; AUC: Area under the ROC: curve; GSCA: Gene Set Cancer Analysis; GDSC: Genomics of Drug Sensitivity in Cancer; PPI: Protein–protein interaction; GAPDH: Glyceraldehyde 3-phosphate dehydrogenase; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; TME: Tumor microenvironment; TMB: Tumor mutation burden; SNV: Single nucleotide variants; SNP: Single-nucleotide polymorphism; RTK: Receptor tyrosine kinase; CAMs: Cell adhesion molecules.

AUTHOR CONTRIBUTIONS

LL and AW: research design and drafting the manuscript. XS and LL: literature search and helping to draft the manuscript. ZF and QZ: literature search and helping to draft the manuscript. ZW: research design and drafting the manuscript. TH: help modify articles and collate references. YC: help modify articles and collate references. YT and ZL: review and revision of the manuscript and writing guidance.

ACKNOWLEDGMENTS

Thanks to the scholars who participated in this study for their contributions and the reviewers, thanks to the reviewers and editors for their suggestions on this article.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

ETHICAL STATEMENT

Ethical review and approval were not required for the study in accordance with the local legislation and institutional requirements. The experimental materials were obtained from publicly available databases.

FUNDING

This work was supported by the National Natural Science Foundation of China (No. 81860428), the Key R & D general project of Jiangxi Science and Technology Department (20203BBGL73187).

REFERENCES

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71:209–49. https://doi.org/10.3322/caac.21660 PMID:33538338

2. Islami F, DeSantis CE, Jemal A. Incidence Trends of Esophageal and Gastric Cancer Subtypes by Race, Ethnicity, and Age in the United States, 1997-2014. Clin Gastroenterol Hepatol. 2019; 17:429–39. https://doi.org/10.1016/j.cgh.2018.05.044 PMID:29902641

3. Allemani C, Weir HK, Carreira H, Harewood R, Spika D, Wang XS, Bannon F, Ahn JV, Johnson CJ, Bonaventure A, Marcos-Gragera R, Stiller C, Azevedo e Silva G, et al, and CONCORD Working Group. Global surveillance of

cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). Lancet. 2015; 385:977–1010.
https://doi.org/10.1016/S0140-6736(14)62038-9
PMID:25467588

4. Lina TT, Alzahrani S, Gonzalez J, Pinchuk IV, Beswick EJ, Reyes VE. Immune evasion strategies used by Helicobacter pylori. World J Gastroenterol. 2014; 20:12753–66.
https://doi.org/10.3748/wjg.v20.i36.12753
PMID:25278676

5. Marshall BJ, Warren JR. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. Lancet. 1984; 1:1311–5.
https://doi.org/10.1016/s0140-6736(84)91816-6
PMID:6145023

6. Hagymási K, Tulassay Z. Helicobacter pylori infection: new pathogenetic and clinical aspects. World J Gastroenterol. 2014; 20:6386–99.
https://doi.org/10.3748/wjg.v20.i21.6386
PMID:24914360

7. Shim JH, Yoon JH, Choi SS, Ashktorab H, Smoot DT, Song KY, Nam SW, Lee JY, Park CH, Park WS. The effect of Helicobacter pylori CagA on the HER-2 copy number and expression in gastric cancer. Gene. 2014; 546:288–96.
https://doi.org/10.1016/j.gene.2014.05.064
PMID:24879917

8. Malfertheiner P, Venerito M, Schulz C. Helicobacter pylori Infection: New Facts in Clinical Management. Curr Treat Options Gastroenterol. 2018; 16:605–15.
https://doi.org/10.1007/s11938-018-0209-8
PMID:30415359

9. El-Omar EM, Rabkin CS, Gammon MD, Vaughan TL, Risch HA, Schoenberg JB, Stanford JL, Mayne ST, Goedert J, Blot WJ, Fraumeni JF Jr, Chow WH. Increased risk of noncardia gastric cancer associated with proinflammatory cytokine gene polymorphisms. Gastroenterology. 2003; 124:1193–201.
https://doi.org/10.1016/s0016-5085(03)00157-4
PMID:12730860

10. Kumar S, Metz DC, Ellenberg S, Kaplan DE, Goldberg DS. Risk Factors and Incidence of Gastric Cancer After Detection of Helicobacter pylori Infection: A Large Cohort Study. Gastroenterology. 2020; 158:527–36.e7.
https://doi.org/10.1053/j.gastro.2019.10.019
PMID:31654635

11. Baj J, Korona-Głowniak I, Forma A, Maani A, Sitarz E, Rahnama-Hezavah M, Radzikowska E, Portincasa P. Mechanisms of the Epithelial-Mesenchymal Transition and Tumor Microenvironment in *Helicobacter pylori*-Induced Gastric Cancer. Cells. 2020; 9:1055.
https://doi.org/10.3390/cells9041055 PMID:32340207

12. Kim HJ, Kim N, Kim HW, Park JH, Shin CM, Lee DH. Promising aberrant DNA methylation marker to predict gastric cancer development in individuals with family history and long-term effects of H. pylori eradication on DNA methylation. Gastric Cancer. 2021; 24:302–13.
https://doi.org/10.1007/s10120-020-01117-w
PMID:32915372

13. He Y, Wang C, Zhang X, Lu X, Xing J, Lv J, Guo M, Huo X, Liu X, Lu J, Du X, Li C, Chen Z. Sustained Exposure to *Helicobacter pylori* Lysate Inhibits Apoptosis and Autophagy of Gastric Epithelial Cells. Front Oncol. 2020; 10:581364.
https://doi.org/10.3389/fonc.2020.581364
PMID:33194715

14. Yao X, Liu D, Zhou L, Xie Y, Li Y. FAM60A, increased by Helicobacter pylori, promotes proliferation and suppresses apoptosis of gastric cancer cells by targeting the PI3K/AKT pathway. Biochem Biophys Res Commun. 2020; 521:1003–9.
https://doi.org/10.1016/j.bbrc.2019.11.029
PMID:31727367

15. Correa P, Piazuelo MB, Wilson KT. Pathology of gastric intestinal metaplasia: clinical implications. Am J Gastroenterol. 2010; 105:493–8.
https://doi.org/10.1038/ajg.2009.728
PMID:20203636

16. Machado AM, Figueiredo C, Seruca R, Rasmussen LJ. Helicobacter pylori infection generates genetic instability in gastric cells. Biochim Biophys Acta. 2010; 1806:58–65.
https://doi.org/10.1016/j.bbcan.2010.01.007
PMID:20122996

17. Deo RC. Machine Learning in Medicine. Circulation. 2015; 132:1920–30.
https://doi.org/10.1161/CIRCULATIONAHA.115.001593 PMID:26572668

18. Cao C, Liu F, Tan H, Song D, Shu W, Li W, Zhou Y, Bo X, Xie Z. Deep Learning and Its Applications in Biomedicine. Genomics Proteomics Bioinformatics. 2018; 16:17–32.
https://doi.org/10.1016/j.gpb.2017.07.003
PMID:29522900

19. Howard FM, Kochanny S, Koshy M, Spiotto M, Pearson AT. Machine Learning-Guided Adjuvant Treatment of Head and Neck Cancer. JAMA Netw Open. 2020; 3:e2025881.
https://doi.org/10.1001/jamanetworkopen.2020.25881 PMID:33211108

20. Chen H, Chen L. Support Vector Machine Classification of Drunk Driving Behaviour. Int J Environ Res Public

Health. 2017; 14:108.
https://doi.org/10.3390/ijerph14010108
PMID:28125006

21. Tabares-Soto R, Orozco-Arias S, Romero-Cano V, Segovia Bucheli V, Rodríguez-Sotelo JL, Jiménez-Varón CF. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. PeerJ Comput Sci. 2020; 6:e270.
https://doi.org/10.7717/peerj-cs.270
PMID:33816921

22. Malfertheiner P, Sipponen P, Naumann M, Moayyedi P, Mégraud F, Xiao SD, Sugano K, Nyrén O, and Lejondal H. pylori-Gastric Cancer Task Force. Helicobacter pylori eradication has the potential to prevent gastric cancer: a state-of-the-art critique. Am J Gastroenterol. 2005; 100:2100–15.
https://doi.org/10.1111/j.1572-0241.2005.41688.x
PMID:16128957

23. Donick D, Lera SC. Uncovering feature interdependencies in high-noise environments with stepwise lookahead decision forests. Sci Rep. 2021; 11:9238.
https://doi.org/10.1038/s41598-021-88571-3
PMID:33927260

24. Carracedo-Reboredo P, Liñares-Blanco J, Rodríguez-Fernández N, Cedrón F, Novoa FJ, Carballal A, Maojo V, Pazos A, Fernandez-Lozano C. A review on machine learning approaches and trends in drug discovery. Comput Struct Biotechnol J. 2021; 19:4538–58.
https://doi.org/10.1016/j.csbj.2021.08.011
PMID:34471498

25. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77.
https://doi.org/10.1186/1471-2105-12-77
PMID:21414208

26. Petitprez F, Vano YA, Becht E, Giraldo NA, de Reyniès A, Sautès-Fridman C, Fridman WH. Transcriptomic analysis of the tumor microenvironment to guide prognosis and immunotherapies. Cancer Immunol Immunother. 2018; 67:981–8.
https://doi.org/10.1007/s00262-017-2058-z
PMID:28884365

27. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. Methods Mol Biol. 2018; 1711:243–59.
https://doi.org/10.1007/978-1-4939-7493-1_12
PMID:29344893

28. Liu CJ, Hu FF, Xia MX, Han L, Zhang Q, Guo AY. GSCALite: a web server for gene set cancer analysis. Bioinformatics. 2018; 34:3771–2.
https://doi.org/10.1093/bioinformatics/bty411
PMID:29790900

29. Feng Z, Li L, Zeng Q, Zhang Y, Tu Y, Chen W, Shu X, Wu A, Xiong J, Cao Y, Li Z. *RNF114* Silencing Inhibits the Proliferation and Metastasis of Gastric Cancer. J Cancer. 2022; 13:565–78.
https://doi.org/10.7150/jca.62033 PMID:35069903

30. Thul PJ, Lindskog C. The human protein atlas: A spatial map of the human proteome. Protein Sci. 2018; 27:233–44.
https://doi.org/10.1002/pro.3307
PMID:28940711

31. Lei X, Lei Y, Li JK, Du WX, Li RG, Yang J, Li J, Li F, Tan HB. Immune cells within the tumor microenvironment: Biological functions and roles in cancer immunotherapy. Cancer Lett. 2020; 470:126–33.
https://doi.org/10.1016/j.canlet.2019.11.009
PMID:31730903

32. Le DT, Durham JN, Smith KN, Wang H, Bartlett BR, Aulakh LK, Lu S, Kemberling H, Wilt C, Luber BS, Wong F, Azad NS, Rucki AA, et al. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. Science. 2017; 357:409–13.
https://doi.org/10.1126/science.aan6733
PMID:28596308

33. Lauss M, Donia M, Harbst K, Andersen R, Mitra S, Rosengren F, Salim M, Vallon-Christersson J, Törngren T, Kvist A, Ringnér M, Svane IM, Jönsson G. Mutational and putative neoantigen load predict clinical benefit of adoptive T cell therapy in melanoma. Nat Commun. 2017; 8:1738.
https://doi.org/10.1038/s41467-017-01460-0
PMID:29170503

34. McGranahan N, Furness AJ, Rosenthal R, Ramskov S, Lyngaa R, Saini SK, Jamal-Hanjani M, Wilson GA, Birkbak NJ, Hiley CT, Watkins TB, Shafi S, Murugaesu N, et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. Science. 2016; 351:1463–9.
https://doi.org/10.1126/science.aaf1490
PMID:26940869

35. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, Lee W, Yuan J, Wong P, Ho TS, Miller ML, Rekhtman N, Moreira AL, et al. Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. Science. 2015; 348:124–8.
https://doi.org/10.1126/science.aaa1348
PMID:25765070

36. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C,

Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, et al. Proteomics. Tissue-based map of the human proteome. Science. 2015; 347:1260419. https://doi.org/10.1126/science.1260419 PMID:25613900

37. Bakhti SZ, Latifi-Navid S, Safaralizadeh R. Helicobacter pylori-related risk predictors of gastric cancer: The latest models, challenges, and future prospects. Cancer Med. 2020; 9:4808–22. https://doi.org/10.1002/cam4.3068 PMID:32363738

38. Lee YC, Chiang TH, Chou CK, Tu YK, Liao WC, Wu MS, Graham DY. Association Between Helicobacter pylori Eradication and Gastric Cancer Incidence: A Systematic Review and Meta-analysis. Gastroenterology. 2016; 150:1113–24.e5. https://doi.org/10.1053/j.gastro.2016.01.028 PMID:26836587

39. Lin Y, Kawai S, Sasakabe T, Nagata C, Naito M, Tanaka K, Sugawara Y, Mizoue T, Sawada N, Matsuo K, Kitamura T, Utada M, Ito H, et al, and Research Group for the Development and Evaluation of Cancer Prevention Strategies in Japan. Effects of Helicobacter pylori eradication on gastric cancer incidence in the Japanese population: a systematic evidence review. Jpn J Clin Oncol. 2021; 51:1158–70. https://doi.org/10.1093/jjco/hyab055 PMID:33893508

40. Mentis A, Lehours P, Mégraud F. Epidemiology and Diagnosis of Helicobacter pylori infection. Helicobacter. 2015 (Suppl 1); 20:1–7. https://doi.org/10.1111/hel.12250 PMID:26372818

41. Joossens JV, Hill MJ, Elliott P, Stamler R, Lesaffre E, Dyer A, Nichols R, Kesteloot H. Dietary salt, nitrate and stomach cancer mortality in 24 countries. European Cancer Prevention (ECP) and the INTERSALT Cooperative Research Group. Int J Epidemiol. 1996; 25:494–504. https://doi.org/10.1093/ije/25.3.494 PMID:8671549

42. González CA, Jakszyn P, Pera G, Agudo A, Bingham S, Palli D, Ferrari P, Boeing H, del Giudice G, Plebani M, Carneiro F, Nesi G, Berrino F, et al. Meat intake and risk of stomach and esophageal adenocarcinoma within the European Prospective Investigation Into Cancer and Nutrition (EPIC). J Natl Cancer Inst. 2006; 98:345–54. https://doi.org/10.1093/jnci/djj071 PMID:16507831

43. Ofori EG, Adinortey CA, Bockarie AS, Kyei F, Tagoe EA, Adinortey MB. *Helicobacter pylori* Infection, Virulence Genes' Distribution and Accompanying Clinical Outcomes: The West Africa Situation. Biomed Res Int. 2019; 2019:7312908. https://doi.org/10.1155/2019/7312908 PMID:31886245

44. Šterbenc A, Jarc E, Poljak M, Homan M. *Helicobacter pylori* virulence genes. World J Gastroenterol. 2019; 25:4870–84. https://doi.org/10.3748/wjg.v25.i33.4870 PMID:31543679

45. Ding C, Bao TY, Huang HL. Quantum-Inspired Support Vector Machine. IEEE Trans Neural Netw Learn Syst. 2022; 33:7210–22. https://doi.org/10.1109/TNNLS.2021.3084467 PMID:34111003

46. Wang H, Yang F, Luo Z. An experimental study of the intrinsic stability of random forest variable importance measures. BMC Bioinformatics. 2016; 17:60. https://doi.org/10.1186/s12859-016-0900-5 PMID:26842629

47. Husain A, Chiu YT, Sze KM, Ho DW, Tsui YM, Suarez EMS, Zhang VX, Chan LK, Lee E, Lee JM, Cheung TT, Wong CC, Chung CY, Ng IO. Ephrin-A3/EphA2 axis regulates cellular metabolic plasticity to enhance cancer stemness in hypoxic hepatocellular carcinoma. J Hepatol. 2022; 77:383–96. https://doi.org/10.1016/j.jhep.2022.02.018 PMID:35227773

48. Lin P, Yang H. EFNA3 is a prognostic biomarker for the overall survival of patients with hepatocellular carcinoma. J Hepatol. 2022; 77:879–80. https://doi.org/10.1016/j.jhep.2022.03.008 PMID:35358615

49. Hu Q, Qin Y, Ji S, Xu W, Liu W, Sun Q, Zhang Z, Liu M, Ni Q, Yu X, Xu X. UHRF1 promotes aerobic glycolysis and proliferation via suppression of SIRT4 in pancreatic cancer. Cancer Lett. 2019; 452:226–36. https://doi.org/10.1016/j.canlet.2019.03.024 PMID:30905812

50. Kong X, Chen J, Xie W, Brown SM, Cai Y, Wu K, Fan D, Nie Y, Yegnasubramanian S, Tiedemann RL, Tao Y, Chiu Yen RW, Topper MJ, et al. Defining UHRF1 Domains that Support Maintenance of Human Colon Cancer DNA Methylation and Oncogenic Properties. Cancer Cell. 2019; 35:633–48.e7. https://doi.org/10.1016/j.ccell.2019.03.003 PMID:30956060

51. Zhang H, Song Y, Yang C, Wu X. UHRF1 mediates cell migration and invasion of gastric cancer. Biosci Rep. 2018; 38:BSR20181065. https://doi.org/10.1042/BSR20181065 PMID:30352833

52. Fischer C, Mazzone M, Jonckx B, Carmeliet P. FLT1 and its ligands VEGFB and PlGF: drug targets for anti-angiogenic therapy? Nat Rev Cancer. 2008; 8:942–56. https://doi.org/10.1038/nrc2524 PMID:19029957

53. Jimenez-Hernandez LE, Vazquez-Santillan K, Castro-Oropeza R, Martinez-Ruiz G, Muñoz-Galindo L, Gonzalez-Torres C, Cortes-Gonzalez CC, Victoria-Acosta G, Melendez-Zajgla J, Maldonado V. NRP1-positive lung cancer cells possess tumor-initiating properties. Oncol Rep. 2018; 39:349–57. https://doi.org/10.3892/or.2017.6089 PMID:29138851

54. Song Y, Zeng S, Zheng G, Chen D, Li P, Yang M, Luo K, Yin J, Gu Y, Zhang Z, Jia X, Qiu N, He Z, et al. FOXO3a-driven miRNA signatures suppresses VEGF-A/NRP1 signaling and breast cancer metastasis. Oncogene. 2021; 40:777–90. https://doi.org/10.1038/s41388-020-01562-y PMID:33262463

55. Liu X, Meng X, Peng X, Yao Q, Zhu F, Ding Z, Sun H, Liu X, Li D, Lu Y, Tang H, Li B, Peng Z. Impaired AGO2/miR-185-3p/NRP1 axis promotes colorectal cancer metastasis. Cell Death Dis. 2021; 12:390. https://doi.org/10.1038/s41419-021-03672-1 PMID:33846300

56. Dong Y, Ma WM, Shi ZD, Zhang ZG, Zhou JH, Li Y, Zhang SQ, Pang K, Li BB, Zhang WD, Fan T, Zhu GY, Xue L, et al. Role of NRP1 in Bladder Cancer Pathogenesis and Progression. Front Oncol. 2021; 11:685980. https://doi.org/10.3389/fonc.2021.685980 PMID:34249735

57. Rotte A. Combination of CTLA-4 and PD-1 blockers for treatment of cancer. J Exp Clin Cancer Res. 2019; 38:255. https://doi.org/10.1186/s13046-019-1259-z PMID:31196207

58. Li H, Li CM, Yuan R, Wang HB, Wei J. Circ_0000515 drives the progression of hepatocellular carcinoma by regulating MAPK10. Eur Rev Med Pharmacol Sci. 2020; 24:6014–22. https://doi.org/10.26355/eurrev_202006_21495 PMID:32572915

59. Li L, Luo Z. Dysregulated miR-27a-3p promotes nasopharyngeal carcinoma cell proliferation and migration by targeting Mapk10. Oncol Rep. 2017; 37:2679–87. https://doi.org/10.3892/or.2017.5544 PMID:28393229

60. Wang JH, Zhang L, Huang ST, Xu J, Zhou Y, Yu XJ, Luo RZ, Wen ZS, Jia WH, Zheng M. Expression and prognostic significance of MYL9 in esophageal squamous cell carcinoma. PLoS One. 2017; 12:e0175280. https://doi.org/10.1371/journal.pone.0175280 PMID:28388691

61. Kruthika BS, Sugur H, Nandaki K, Arimappamagan A, Paturu K, Santosh V. Expression pattern and prognostic significance of myosin light chain 9 (MYL9): a novel biomarker in glioblastoma. J Clin Pathol. 2019; 72:677–81. https://doi.org/10.1136/jclinpath-2019-205834 PMID:31270134

62. Wu K, Zou J, Lin C, Jie ZG. MicroRNA-140-5p inhibits cell proliferation, migration and promotes cell apoptosis in gastric cancer through the negative regulation of THY1-mediated Notch signaling. Biosci Rep. 2019; 39:BSR20181434. https://doi.org/10.1042/BSR20181434 PMID:31123165

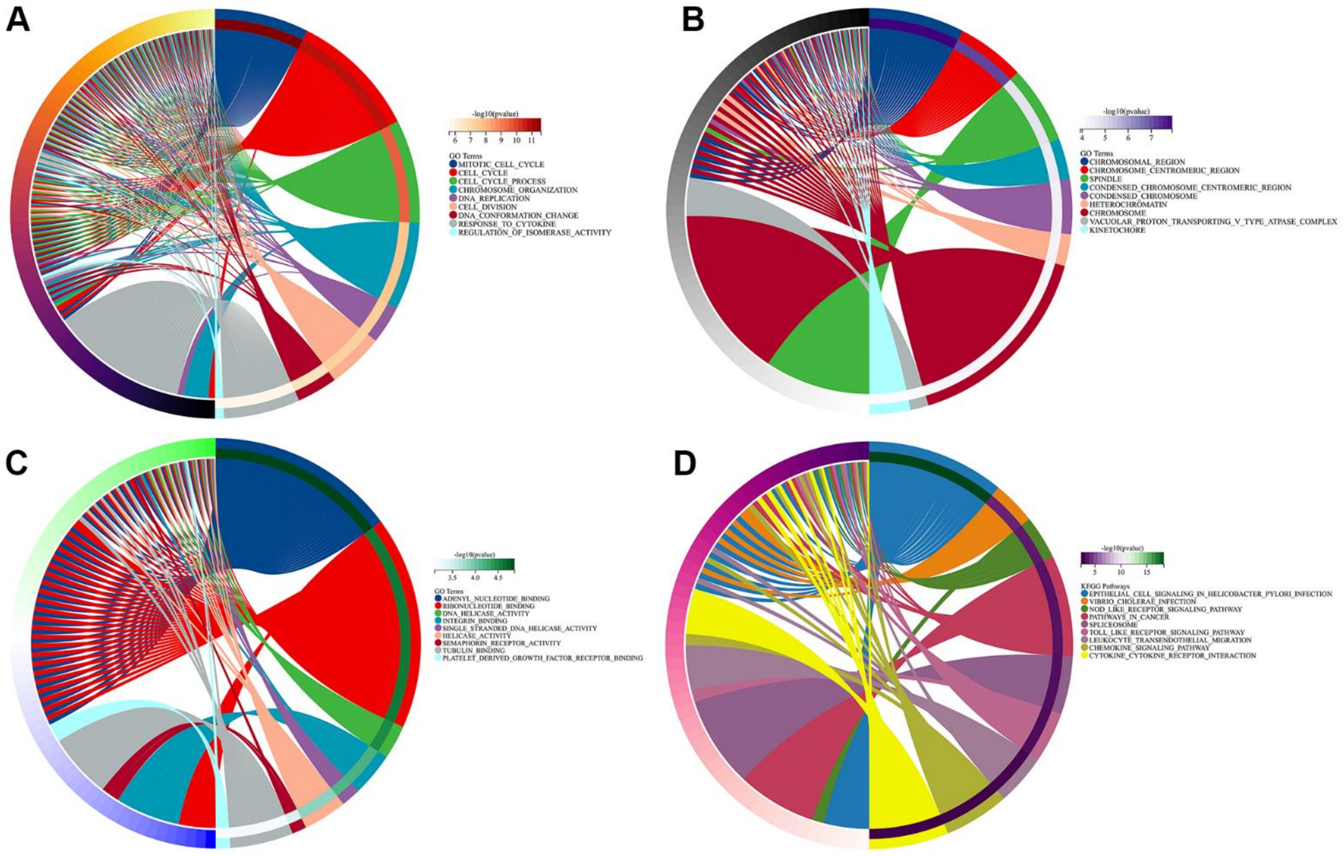
63. Xie Y, Rong L, He M, Jiang Y, Li H, Mai L, Song F. LncRNA SNHG3 promotes gastric cancer cell proliferation and metastasis by regulating the miR-139-5p/MYB axis. Aging (Albany NY). 2021; 13:25138–52. https://doi.org/10.18632/aging.203732 PMID:34898477

64. Beyrend G, van der Gracht E, Yilmaz A, van Duikeren S, Camps M, Höllt T, Vilanova A, van Unen V, Koning F, de Miranda NF, Arens R, Ossendorp F. PD-L1 blockade engages tumor-infiltrating lymphocytes to co-express targetable activating and inhibitory receptors. J Immunother Cancer. 2019; 7:217. https://doi.org/10.1186/s40425-019-0700-3 PMID:31412943

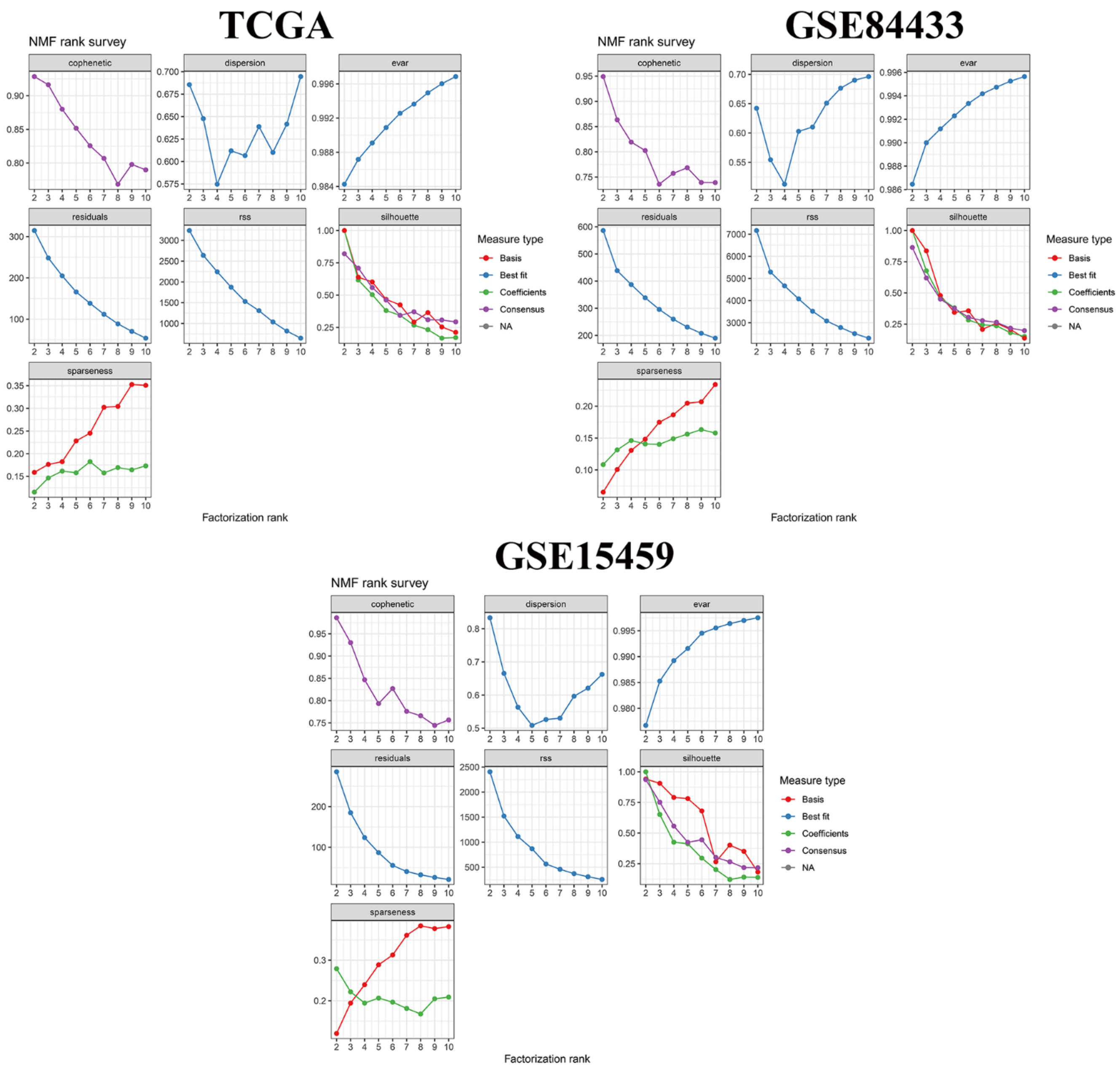
65. Yarchoan M, Albacker LA, Hopkins AC, Montesion M, Murugesan K, Vithayathil TT, Zaidi N, Azad NS, Laheru DA, Frampton GM, Jaffee EM. PD-L1 expression and tumor mutational burden are independent biomarkers in most cancers. JCI Insight. 2019; 4:e126908. https://doi.org/10.1172/jci.insight.126908 PMID:30895946

SUPPLEMENTARY MATERIALS

Supplementary Figures

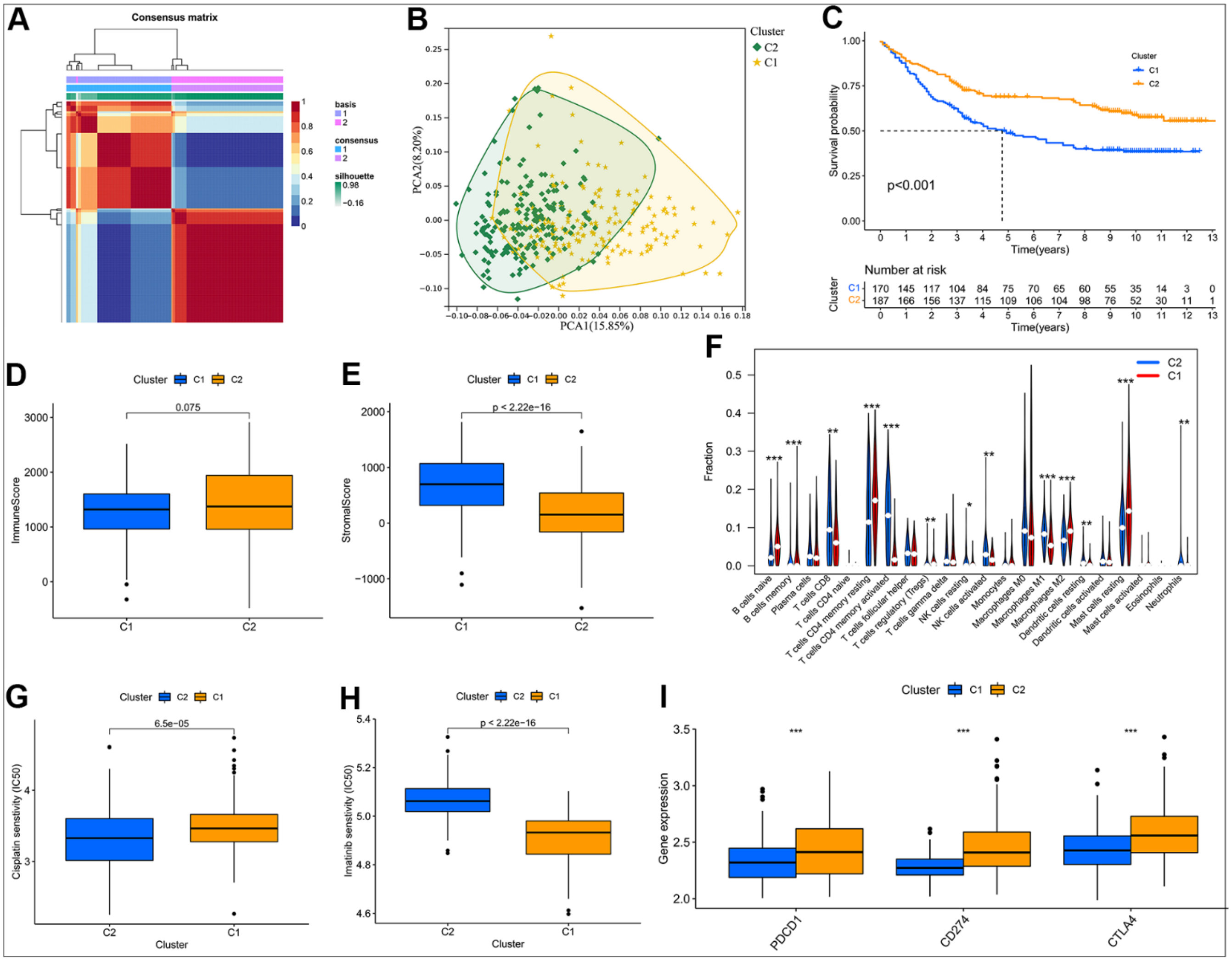


Supplementary Figure 1. Functional enrichment analysis (**A**–**C**) GO analysis (biological process (BP), cellular component (CC), and molecular function (MF)) of the differentially expressed HP-related genes. (**D**) KEGG analysis of the differentially expressed HP-related genes.



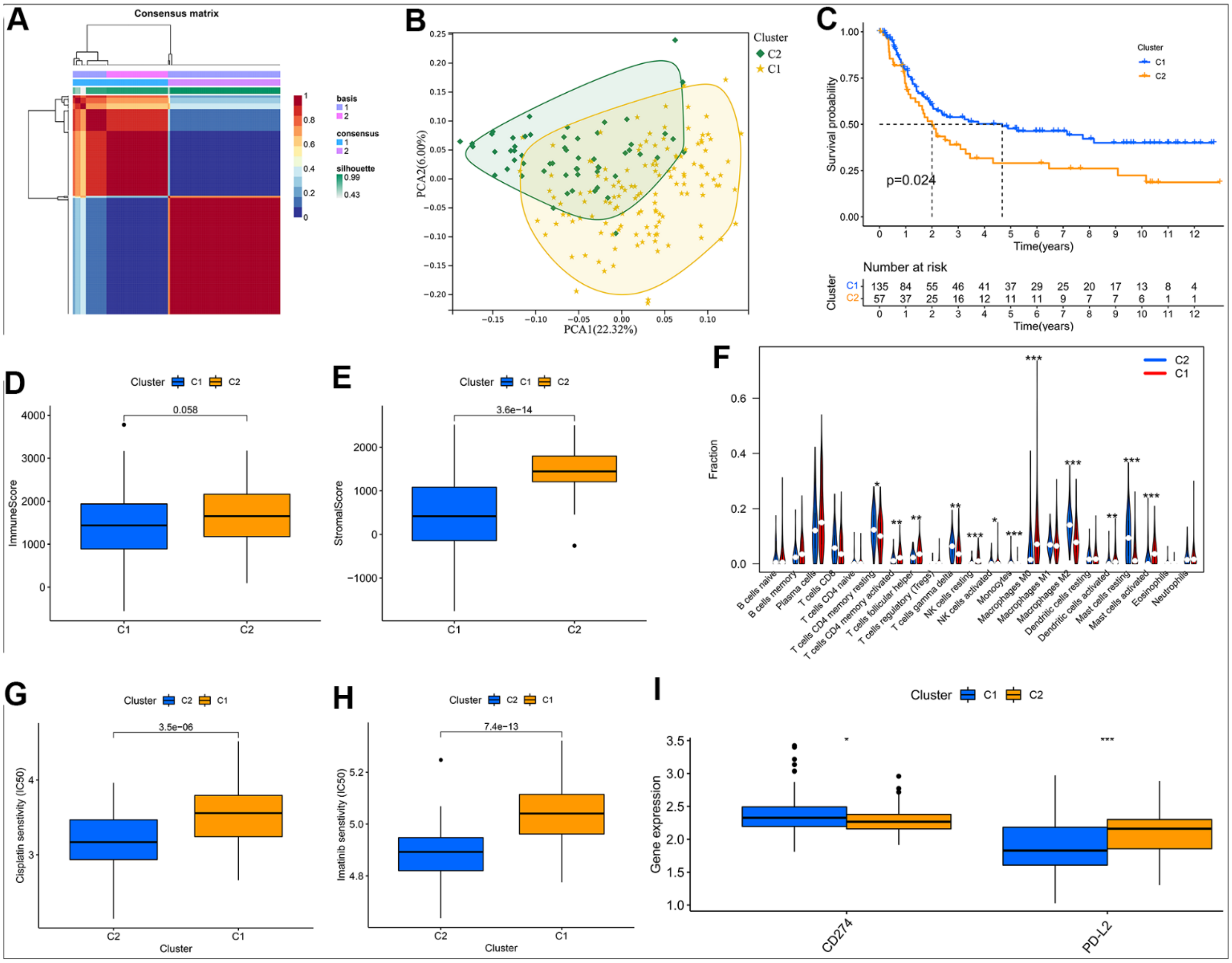
Supplementary Figure 2. NMF rank survey performed on distinct clusters in multiple datasets.

GSE84433 Dataset

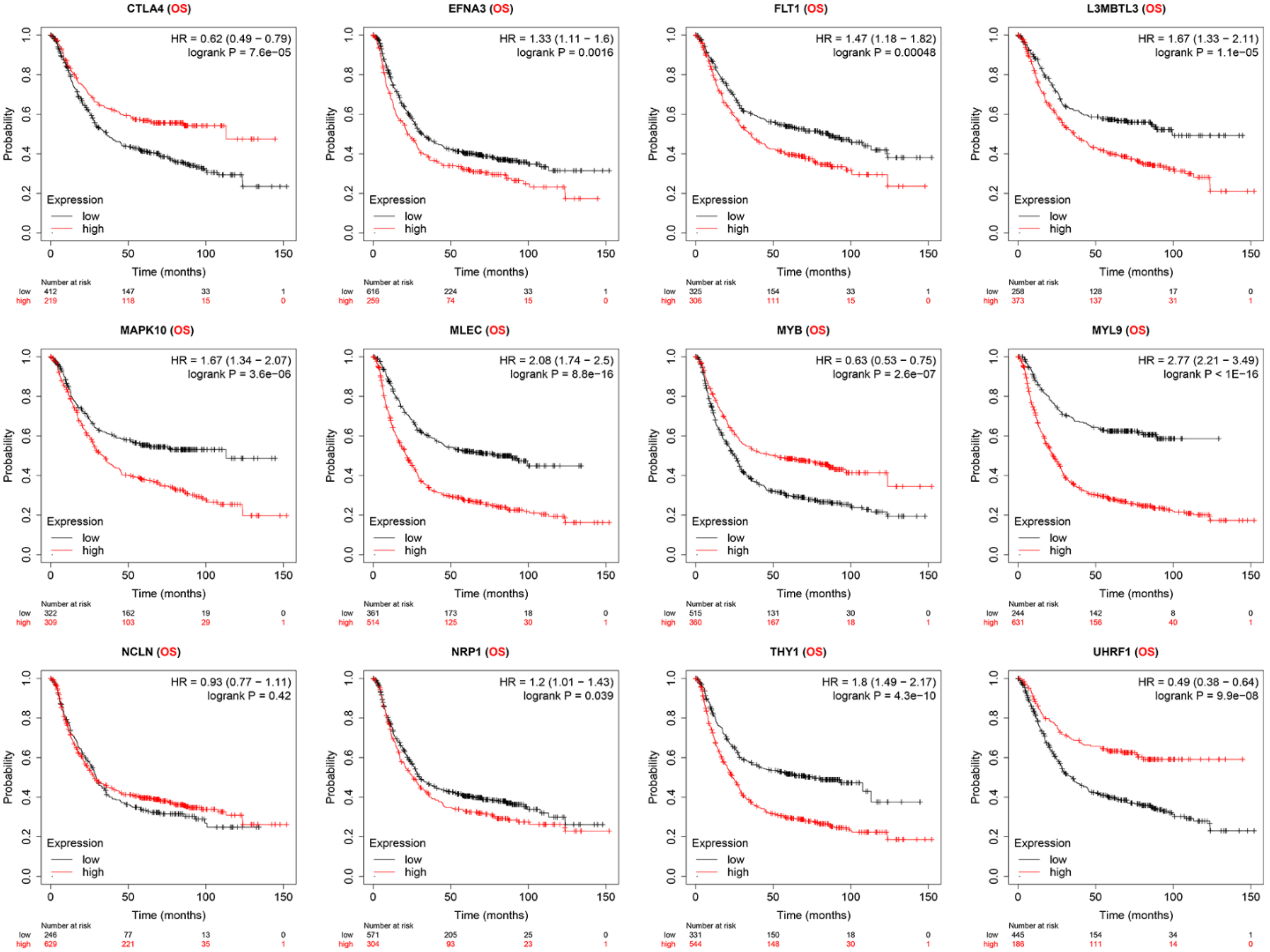


Supplementary Figure 3. Establishment of a NMF subtype according to the differentially expressed HP-related genes in the GSE84433 cohort. (**A**) NMF consensus clustering for k = 2. (**B**) Kaplan–Meier analysis of overall survival (OS) for Cluster C1 and C2. (**C**) Principal component analysis (PCA). (**D**, **E**) Differential analyses of immune and stromal score between Cluster C1 and C2. (**F**) Violin plot showing the immune cell infiltration landscape across different clusters. (**G**, **H**) Box plot of estimated IC50 values for Cisplatin and Imatinib in Cluster C1 and C2. (**I**) Box plot visualizing the significant expression differences of immune checkpoints across distinct clusters, including PDCD1, CD274, and CTLA4. *:P<0.05 ** :P<0.01 ***:P<0.001.

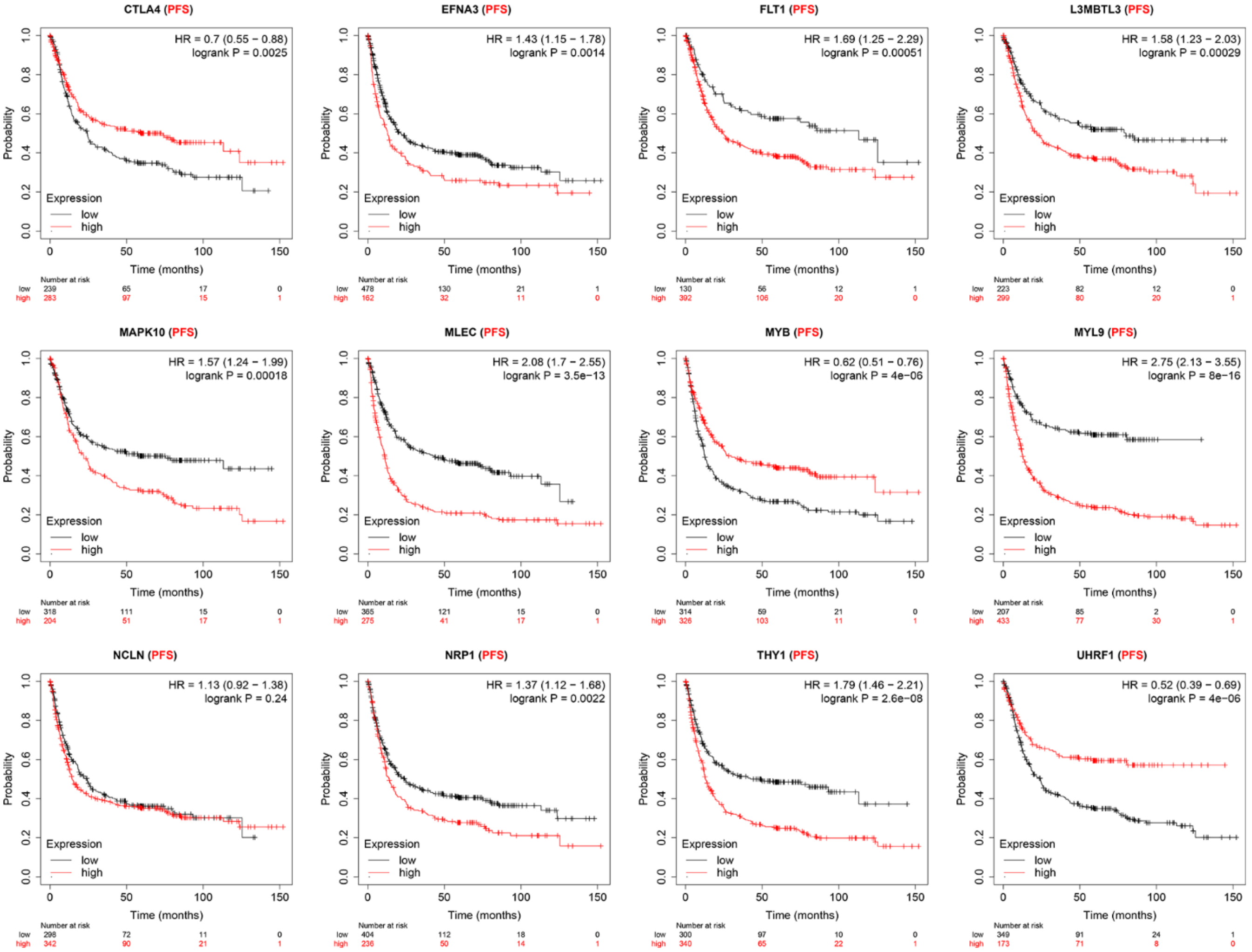
GSE15459 Dataset



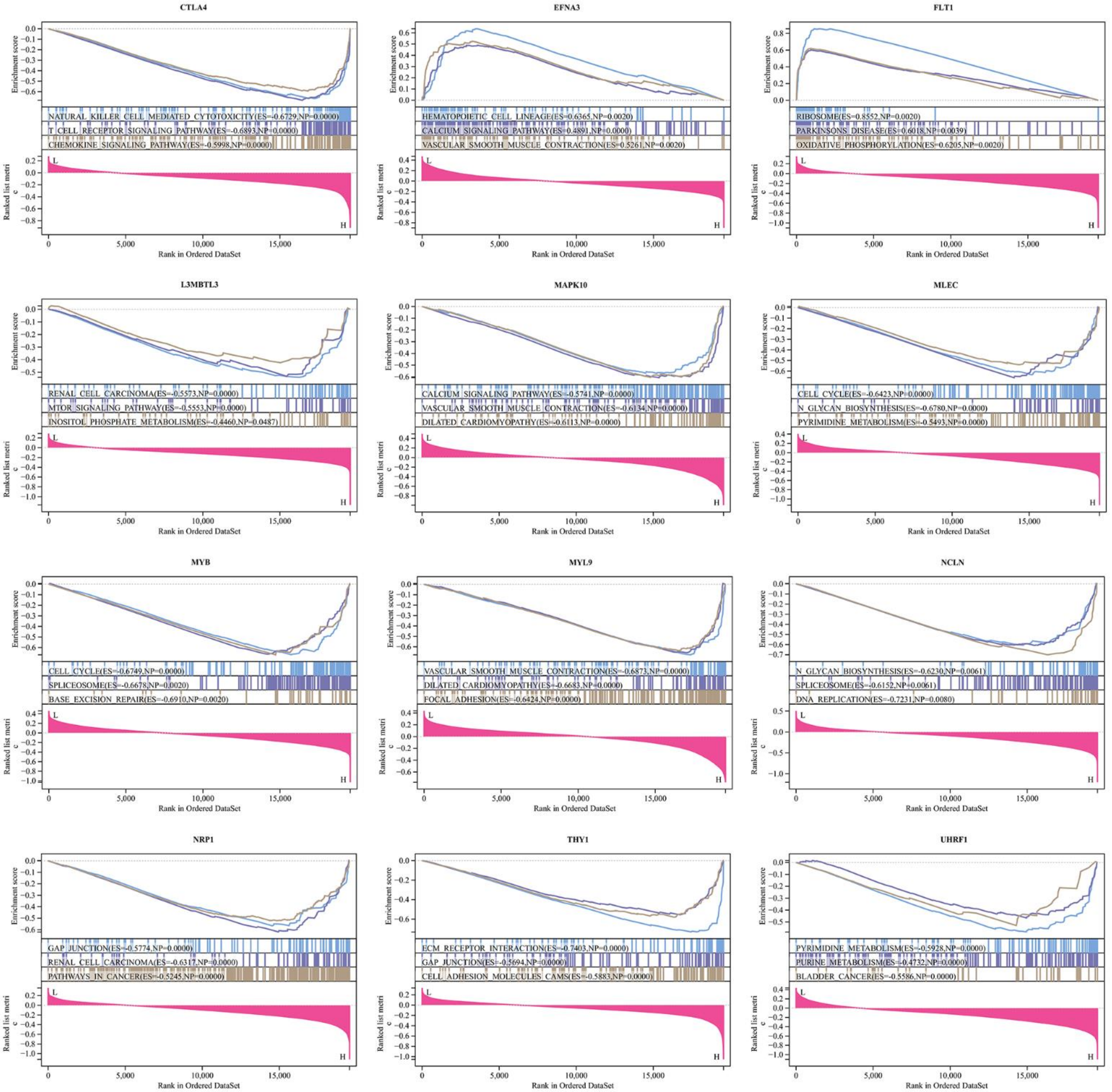
Supplementary Figure 4. Creation of a NMF subtype based on the differentially expressed HP-related genes in the GSE15459 cohort. (**A**) NMF consensus clustering for k = 2. (**B**) Kaplan–Meier analysis of overall survival (OS) for Cluster C1 and C2. (**C**) Principal component analysis (PCA). (**D**, **E**) Differential analyses of immune and stromal score between Cluster C1 and C2. (**F**) Violin plot showing the immune cell infiltration landscape across different clusters. (**G**, **H**) Box plot of estimated IC50 values for Cisplatin and Imatinib in Cluster C1 and C2. (**I**) Box plot visualizing the significant expression differences of immune checkpoints across distinct clusters, including CD274 and PD-L2. *:P<0.05 ** :P<0.01 ***:P<0.001.



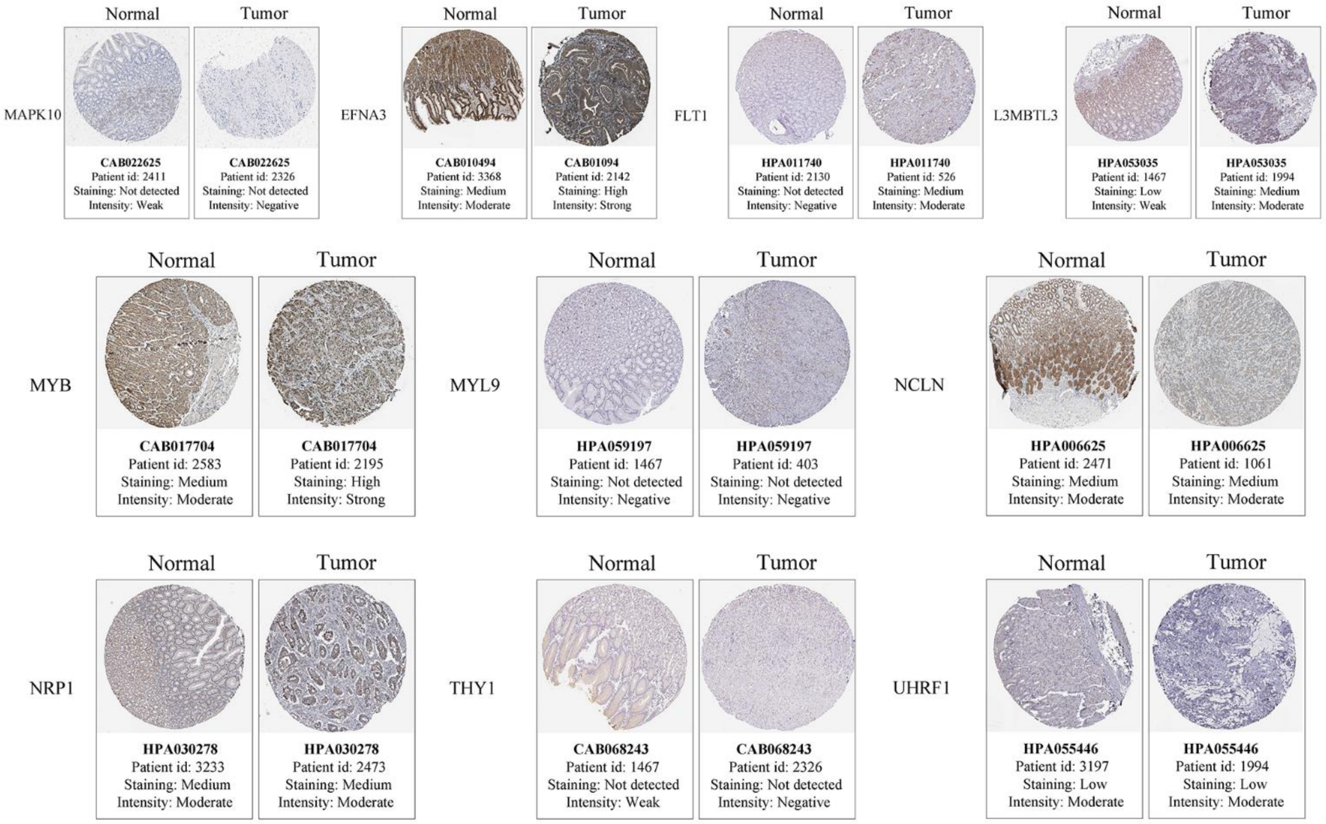
Supplementary Figure 5. Six GEO datasets confirming the prognosis value of hub genes (OS).



Supplementary Figure 6. Six GEO datasets confirming the prognosis value of hub genes (PFS).



Supplementary Figure 7. Gene set enrichment analysis (GSEA) of each hub gene.



Supplementary Figure 8. HPA database verifying the protein expression of hub genes (EFNA3, UHRF1, FLT1, NRP1, L3MBTL3, MAPK10, MYL9, THY1, MYB, as well as NCLN).

Supplementary Tables

Please browse Full Text version to see the data of Supplementary Tables 2, 3, 7.

Supplementary Table 1. Primers sequences in qRT-PCR.

Primers sequences	Forward sequence	Reverse sequence
GAPDH	CCCACTCCTCCACCTTTGAC	CCACCACCCTGTTGCTGTAG
EFNA3	AGTTCTCGGAGAAGTTCCAGCG	CAGCAGACGAACACCTTCATCC
FLT1	CCTGCAAGATTCAGGCACCTATG	GTTTCGCAGGAGGTATGGTGCT
L3MBTL3	TTCGCAGAGAGCACGGAGGAA	ACCGCTTTCTCCTCTTCCAGGT
MAPK10	GCACACACACATGCATACCC	TCTCACTGCTCAGACCTTGC
MLEC	GGGCAGGATGGGTATGCTTT	CGGTTCTGCTTCCGTGTACT
MYB	GGGAACAGATGGGCAGAAATCG	GCTGGCTTTTGAAGACTCCTGC
MYL9	GGATGTGATTCGCAACGCCTTTG	CGGTACATCTCGTCCACTTCCT
NCLN	ACCTCCTGTTCTTTGCGTCTGG	CCACATTGTCCTGAAGCAGGCT
NRP1	AACAACGGCTCGGACTGGAAGA	GGTAGATCCTGATGAATCGCGTG
THY1	CTCCAGCATTCTCAGCCACA	CGCTGCTTTCCTGGTCAAAC
UHRF1	GACAAGCAGCTCATGTGCGATG	AGTACCACCTCGCTGGCATCAT
CTLA4	ACGGGACTCTACATCTGCAAGG	GGAGGAAGTCAGAATCTGGGCA

Supplementary Table 2. The list of HP-related gene sets.

Supplementary Table 3. 232 differentially expressed HP-related genes.

Supplementary Table 4. 17 prognostic-associated HP genes.

Id	HR	HR.95L	HR.95H	pvalue
THY1	1.197780366	1.028640775	1.394731611	0.020150799
NRP1	1.372799425	1.149153022	1.63997155	0.000478942
PLXNC1	1.198026544	1.012051655	1.418176228	0.035803952
NCLN	0.763501882	0.612545701	0.951659808	0.016359488
EFNA3	0.852098441	0.756579369	0.959676914	0.008328384
MYL9	1.090867971	1.003169012	1.186233741	0.041955934
CPT1C	1.193375145	1.026920507	1.386810592	0.021079546
TMEM176B	1.168858366	1.001461307	1.364236312	0.047874547
MYB	0.829498517	0.730244348	0.942243225	0.004041507
L3MBTL3	1.319766027	1.047446338	1.662884582	0.018617422
CTLA4	0.835480337	0.707075962	0.987202833	0.034751367
MAPK10	1.260993457	1.032619842	1.53987405	0.02291534
PDE2A	1.195255496	1.040593573	1.372904597	0.011642923
MLEC	0.788755252	0.622210985	0.999877633	0.04988191
FLT1	1.256261141	1.001183574	1.576326355	0.048817849
ANKRD33	1.247838303	1.04927556	1.483976651	0.012282632
UHRF1	0.842653628	0.716375984	0.991190593	0.038754336

Supplementary Table 5. Thirteen candidate genes were selected by using the random forest (RF) method.

Gene
UHRF1
THY1
CTLA4
MLEC
MAPK10
EFNA3
MYL9
NCLN
MYB
NRP1
L3MBTL3
CPT1C
FLT1

Supplementary Table 6. Thirteen genes were screened by establishing the support vector machine-recursive feature elimination (SVM-RFE) model.

Gene
UHRF1
EFNA3
MLEC
THY1
MYB
CTLA4
MAPK10
NCLN
MYL9
FLT1
NRP1
L3MBTL3
PDE2A

Supplementary Table 7. The direct link URLs to the cited images.