Molecular Therapy
Nucleic Acids
Original Article

A molecular classification of gastric cancer associated with distinct clinical outcomes and validated by an XGBoost-based prediction model

Bing Li,[1,4] Fengbin Zhang,[2,4] Qikai Niu,[1] Jun Liu,[3] Yanan Yu,[3] Pengqian Wang,[1] Siqi Zhang,[1] Huamin Zhang,[1] and Zhong Wang[3]

[1]Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China; [2]Department of Gastroenterology and Hepatology, The Fourth Hospital of Hebei Medical University, Shijiazhuang 050011, China; [3]Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China

Gastric cancer (GC) is a heterogeneous disease and a leading cause of cancer-related deaths. Discovering robust, clinically relevant molecular classifications is critical for guiding personalized therapies for GC. Here, we propose a refined molecular classification scheme for GC using integrated optimal algorithms and multi-omics data. Based on the important features of mRNA, microRNA, and DNA methylation data selected by the multivariate Cox regression model, three subtypes linked to distinct clinical outcomes were identified by combining similarity network fusion and consensus clustering methods. Three subtypes were validated by an extreme gradient boosting machine learning prediction model with 125 differentially expressed genes in multiple independent cohorts. The molecular characteristics of mutation signatures, characteristic gene sets, driver genes, and chemotherapy sensitivity for each subtype were also identified: subtype 1 was associated with favorable prognosis and characterized by high ARID1A and PIK3CA mutations, subtype 2 was associated with a poor prognosis and harbored high recurrent TP53 mutations, and subtype 3 was associated with high CHD1, APOA1 mutations, and a poor prognosis. The proposed three-subtype scheme achieved a better clinical prediction performance (area under the curve value = 0.71) than The Cancer Genome Atlas classification, which may provide a practical subtyping framework to improve the treatment of GC.

INTRODUCTION

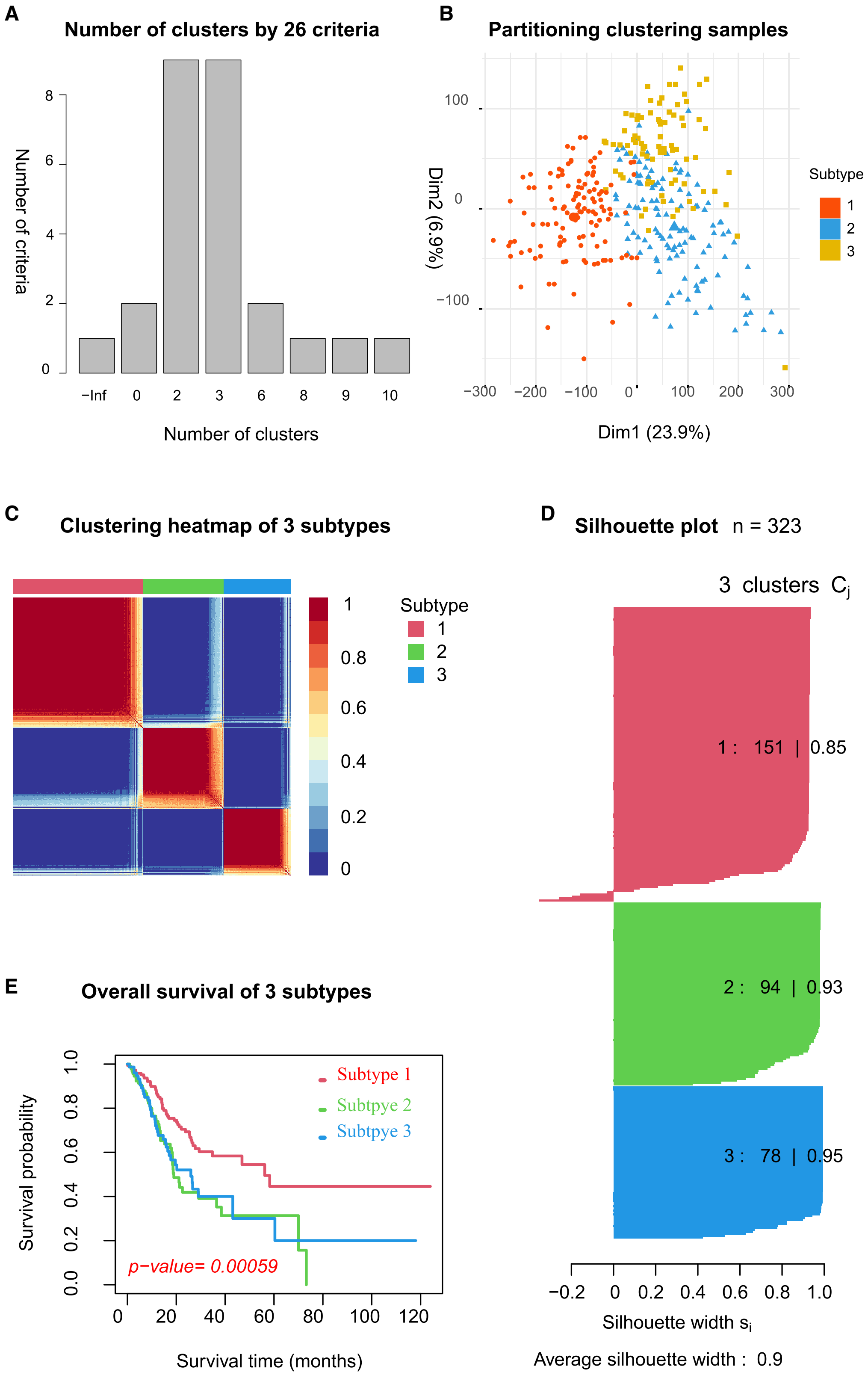
Gastric cancer (GC) is the third leading cause of global cancer-related mortality and is responsible for 768,000 deaths in 2020.[1] Surgical resection with subsequent adjuvant chemotherapy or chemoradiotherapy has been established as an effective treatment for patients with early stage GC, but it has been hampered by the low early metaphase diagnosis rate and high recurrence rate.[2–7] The fact that one-size-fits-all therapeutic schemes result in different treatment outcomes suggests the inherent biological and clinical heterogeneity of GC.[8] Clinical heterogeneity is associated with multiple factors from genomic to environmental levels, but it is most likely based on differences in the molecular characteristics of cancer cells, which manifests as subtypes.[9] Therefore, discovering robust molecular classifications is critical for improving GC therapy by identifying specific therapeutic targets and biomarkers and developing more personalized clinical treatment strategies.

Before the genomics era, GC was histologically classified into different subtypes, such as the intestinal and diffuse types according to the Lauren classification, and papillary, tubular, mucinous, and poorly cohesive carcinoma types based on the World Health Organization (WHO) classification system.[10,11] The limited clinical usefulness of histological classification makes the development of classifiers based on multiple molecular levels that can guide precise treatment an urgent priority. Based on sequencing data from six molecular platforms, The Cancer Genome Atlas (TCGA) research network team classified GC into four molecular subtypes: Epstein-Barr virus (EBV), microsatellite instability (MSI), genomically stable (GS), and chromosomal instability (CIN).[12] Similarly, the Asian Cancer Research Group (ACRG) Network team established another four molecular subtypes using a transcriptomic classifier: MSI, MSS/EMT, MSS/TP53$^+$, and MSS/TP53$^-$.[8] Several studies have also proposed molecular subtyping schemes for GC based on high-throughput profiling and multi-omics platforms, including genomic, proteomic, and epigenetic features.[9,13–17] In addition, spatial metabolome- and immunome-driven classification methods have also been applied to identify GC.[18,19] Such studies may pave the way for the development of improved treatment strategies and personalized drugs for GC.

Received 10 September 2022; accepted 22 December 2022;
https://doi.org/10.1016/j.omtn.2022.12.014.

[4]These authors contributed equally

Correspondence: Huamin Zhang, Institute of Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing 100700, China.
E-mail: hmzhang@icmm.ac.cn
Correspondence: Zhong Wang, Institute of Basic Research in Clinical Medicine, China Academy of Chinese Medical Sciences, Beijing 100700, China.
E-mail: zhonw@vip.sina.com



(legend on next page)

Although distinct GC subtypes with different molecular features have been delineated by various platforms, the clinically relevant consensus is inadequate.[2,20] To date, many computational methods that integrate multi-omics data for cancer subtyping have been proposed[21]: commonly used network-based methods, such as similarity network fusion (SNF),[22] neighborhood-based multi-omics clustering,[23] and cancer integration via multikernel learning[24]; statistics-based methods, such as moCluster[25] and iClusterBayes[26]; and deep learning-based methods, such as Subtype-GAN.[27] However, there is a lack of consistent results owing to variations in omics data types, clustering methods, and the number of subtypes in a specific cancer. Highly complex multi-omics technologies and lack of clinical association molecular signatures may negatively impact the translation of subtyping results into clinical practice. Recent advances in genomics and bioinformatics have facilitated the optimization of algorithms, focusing on more clinical relevance and consensus cancer subtyping based on multi-omics data.[28–32] A clinically oriented strategy that combines the use of multiple types of optimal methods and effective omics data integrated with machine learning validation should be developed.

In this study, we propose a refined molecular classification of GC based on combinatorial algorithms and multi-omics data. The three proposed subtypes are associated with distinct clinical outcomes and were validated in multiple independent cohorts using the optimal extreme gradient boosting (XGBoost) machine learning prediction model. The mutation signatures, characteristic gene sets, driver genes, and chemotherapy sensitivity of each subtype were also revealed, thereby providing a practical subtyping framework for improving tailored treatments for GC.

RESULTS

Multi-omics-based molecular classification of GC

Based on a TCGA cohort, a total of 323 GC samples with mRNA, microRNA (miRNA), and DNA methylation expression profiles and follow-up data (Table S1) were selected to identify the molecular subtypes. Based on survival data, 3,496 mRNAs, 58 miRNAs, and 43,137 DNA methylation sites were selected as important classification features by using the Cox regression model (Table S1). For consensus clustering, the NbClust R package, which integrates 26 criteria, was used to determine the optimal clustering number, of which nine criteria support the optimal number of clusters (k) as 2 or 3 (Figure 1A), and we selected k = 3 for better clustering quality (Figure 1B). By integrating the SNF and consensus clustering (CC) methods in the CancerSubtypes package, we identified three GC subtypes (Figure 1C). The clustering results had an average silhouette width value of 0.9, which suggests excellent power of discrimination between each subtype (Figure 1D). GC patients with subtype 1 (ARID1A$^+$ type, n = 151) displayed good overall survival, whereas subtypes 2 (TP53$^+$ type; n = 94) and 3 (CDH1$^+$ type; n = 78) had a poor prognosis (log rank test; p = 5e-4) (Figure 1E).

Independent validation of the three molecular subtypes

To validate the proposed molecular subtypes, we developed a prediction model for the three subtypes in the Samsung Medical Center (SMC), ACRG, Korea University Guro Hospital (KUGH), Yonsei University Severance Hospital (YUSH), and Kosin University College of Medicine (KUCM) cohorts obtained from the Gene Expression Omnibus (GEO) database. To construct the prediction model for the three subtypes, a total of 125 union differentially expressed genes (DEGs) ($p \leq 0.05$; $|\text{logFC}| \geq 1$) for the three subtypes were selected as gene features for classification, including 120 genes for subtype 1 (ARID1A$^+$ type), 84 genes for subtype 2 (TP53$^+$ type), and 21 genes for subtype 3 (CDH1$^+$ type) (Table S1). We adopted an XGBoost algorithm to group the test samples into three subtypes according to the expression levels of the characteristic gene sets (Figure 2A). For the SMC (n = 432) and ACRG (n = 300) cohorts, the patients were successfully divided into three subtypes, and the survival curve was consistent with the TCGA cohort; that is, subtype 1 (ARID1A$^+$ type) had better overall survival in both the SMC (log rank test; p = 3e-4) (Figure 2B) and ACRG (log rank test; p < 1e-4) (Figure 2C) cohorts. When the prediction model was applied to the KUGH, YUSH, and KUCM cohorts, it could also be classified into three subtypes, but no significant survival difference was observed because of the limited sample size and survival data. We pooled the Illumina platform KUGH, YUSH, and KUCM cohorts (KYK cohort) as one dataset and observed survival differences (log rank test; p = 0.03) (Figure 2D). In all cohorts, subtype 3 (CDH1$^+$ type) was associated with the worst prognosis (Figures 2B–2D). These results suggest that these three molecular subtypes are robust and discrete.

Comparison with other reported molecular subtypes

We compared the similarities and differences of our classification scheme with TCGA genomic subtypes[12] and epithelial-to-mesenchymal transition (EMT)-based subtypes[15] (Table S1). The TCGA network proposed four genomic subtypes for GC (TCGA subtype): EBV positivity, MSI, GC, and CIN. The EMT-based classification includes mesenchymal phenotype (MP) and epithelial phenotype (EP) subtypes (ARCG subtypes). The proposed subtype 1 (ARID1A$^+$ type) was present across all four TCGA genomic subtypes, and patients with subtype 1 (ARID1A$^+$ type) were divided into EBV (25% [n = 10/81]), MSI (36% [n = 29/81]), GC (10% [n = 8/81]), and CIN (30% [n = 24/81]) subtypes (Figures 2E and S1). Notably, the vast majority of EBV (95% [n = 20/21]) and MSI (94% [n = 30/32]) cases were subtype 1 (ARID1A$^+$ type). The proposed subtypes 2 (TP53$^+$ type) and 3 (CDH1$^+$ type) were enriched in TCGA genomic CIN and GS

Figure 1. Identification of clinically relevant molecular clusters of GC

(A) Prediction of the optimal number of clusters (k) by 26 criteria in the NbClust package. (B) Visualization of the clustering results using the factoextra package. (C) Consensus heatmap of the 3 clusters identified by integrative algorithms of SNF and CC based on 323 GC samples. (D) Quantification of sample similarity within each cluster using silhouette width score. (E) Kaplan-Meier curves for overall survival of 323 patients in the 3 subtypes (log rank test; p = 5e-4). The sample sizes of subtype 1, subtype 2, and subtype 3 were n = 151, n = 94, and n = 78, respectively.

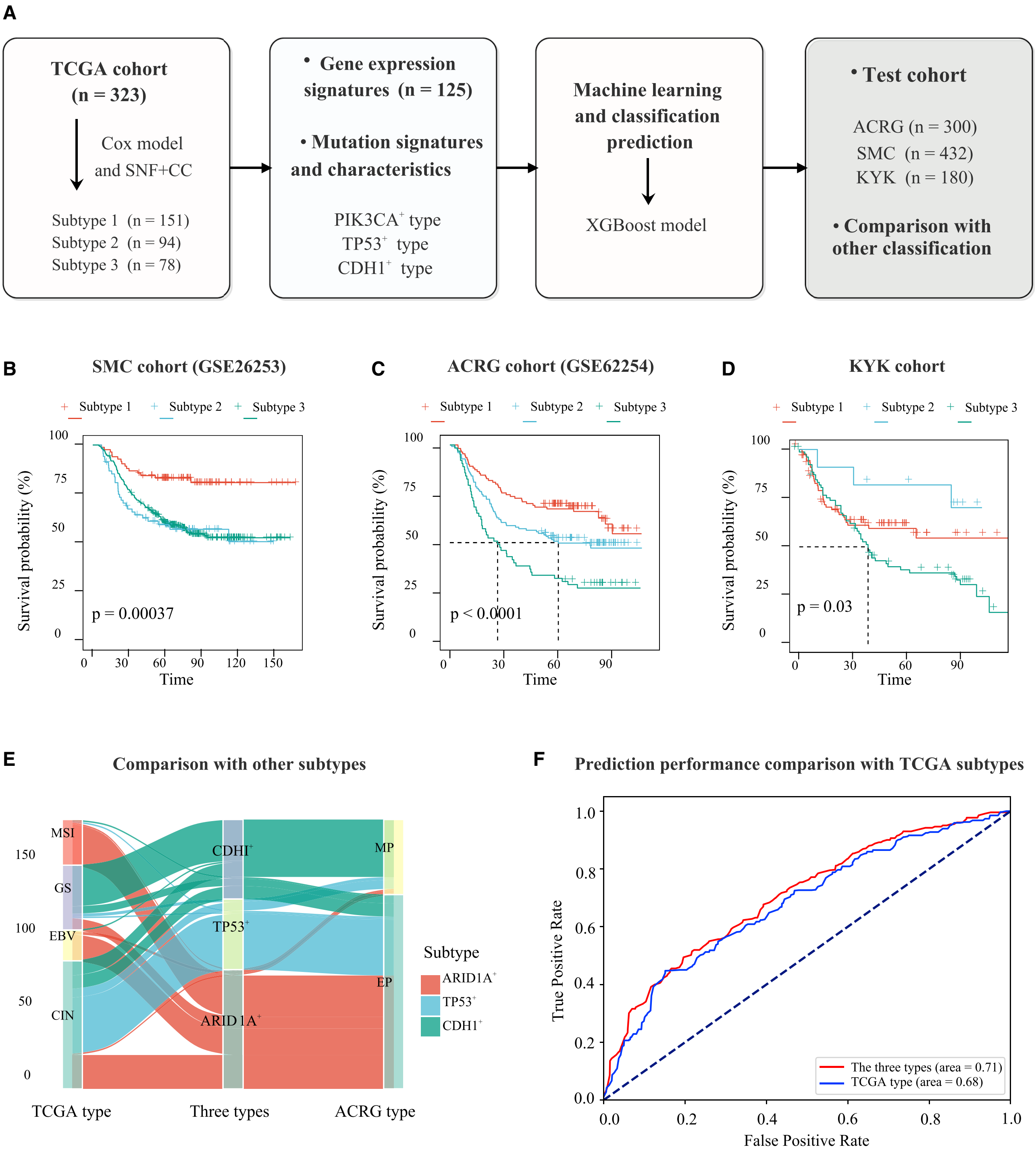


Figure 2. Subtype validation in test cohorts and comparison with other classifications

(A) Schematic overview of the strategy used to construct prediction models and evaluate the classification results in the validation cohorts. Extreme gradient boosting (XGBoost) decision tree models was used for classifying patients in test cohorts into the three subtypes on the basis of gene expression signatures. (B–D) Kaplan-Meier curves of overall survival for patients predicted with different subtypes in SMC, ACRG, and KYK cohorts. The p values were obtained using the log rank test. (E) Distribution of the proposed three subtypes and compared to The Cancer Genome Atlas (TCGA) subtypes[12] and ARCG subtypes (Ref.[15]). (F) ROC curves indicate the prediction performance of our three subtypes compared with TCGA subtypes[12] based on the clinical features of patients with GC. KYK, the pooled KUGH, YUSH, and KUCM cohorts. ROC, receiver operating characteristics.

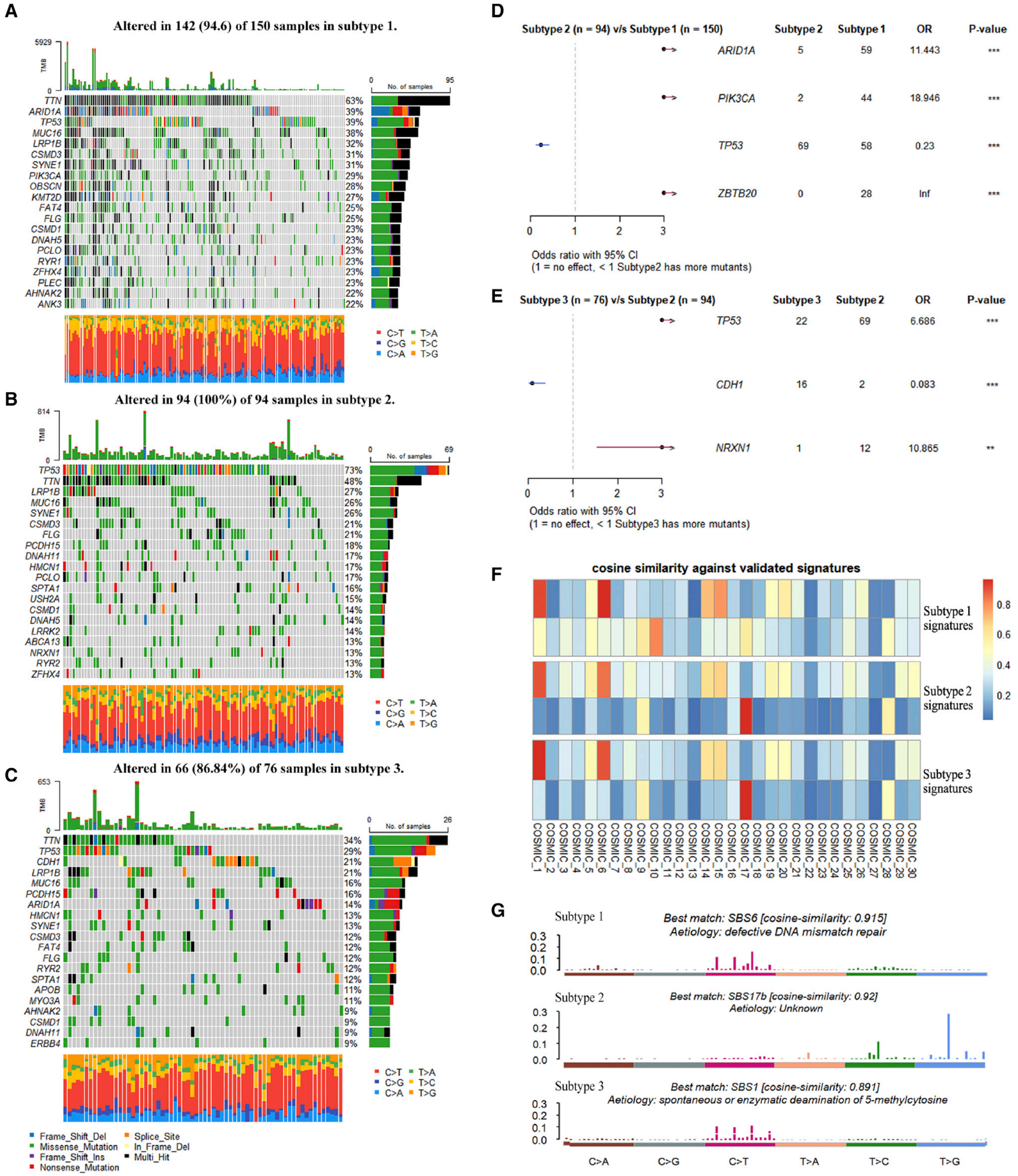


Figure 3. Mutational landscape and signatures of the three subtypes for GC

(A-C) Waterfall plots showing the gene mutation map of the three subtypes, with the different mutation types annotated with specific colors. The bar plot on the top shows the tumor mutational burden. Each column represents a patient and the bar plot in the right indicates the gene mutation frequency of the top 20 genes. The SNVs for each patient

(legend continued on next page)

subtypes, respectively (Figures 2E and S1). The vast majority (92% [n = 44/48]) of subtype2 (TP53$^+$ type) cases were present in the CIN group, whereas cases in subtype 3 (CDH1$^+$ type) were mainly present in the GS (61% [n = 33/54]) and CIN (35% [n = 19/54]) groups (Figures 2E and S1). Associations between the three proposed subtypes and the EMT-based subtypes were further explored. We found that 95% of the samples (n = 77/81) in subtype 1 (ARID1A$^+$ type) and 83% of the samples (n = 40/48) in subtype 2 (TP53$^+$ type) were classified into the EP subtype, whereas 72% of the samples (n = 39/54) in subtype 3 (CDH1$^+$ type) were classified into the MP subtype (Figures 2E and S1). Furthermore, the survival patterns of our three subtypes were similar to those of the other two classification schemes. Based on clinical features, the three proposed subtypes showed better prediction performance than TCGA subtypes, with an area under the curve (AUC) value of 0.71 (Figure 2F). Overall, the comparison suggests that the three proposed subtypes are rational and have less heterogeneity.

Molecular subtypes are associated with clinical phenotypes

We further correlated the molecular subtypes with clinical covariates in TCGA cohort (Table S2). As reported above, GC patients with subtype 1 (ARID1A$^+$ type) tended to have better prognosis and lower recurrence rates than patients with subtypes 2 (TP53$^+$ type) and 3 (CDH1$^+$ type) (Figure 1E and Table S2; p = 2e-4). Other trends in clinical characteristics were also observed. Although each subtype was found throughout the stomach, tumors in subtypes 1 (37%) and 3 (44%) showed elevated frequencies in the gastric antrum, whereas most subtype 2 tumors (31%) were present in the cardia (p = 0.003). Subtype 3 tended to be diagnosed at a significantly younger age than other subtypes (p = 1e-4), and the majority (78%) of patients in this subtype were diagnosed with histologic grade G3 (p < 1e-4). Evaluation of the tumor stage of the three molecular subtypes revealed that patients with subtypes 1 and 2 tended to be diagnosed at an early stage (p < 1e-4), which is consistent with the survival tendency results.

Mutation characteristics of the three subtypes

We next investigated the overall mutation characteristics of the 323 GC samples (Figure S2) and identified somatic alterations associated with each subtype (Figures 3 and S3A–S3C). Overall, the predominant variant type was missense mutation, single nucleotide variants (SNVs) were C>T transitions, the median of variants per sample was 102, and the top mutated genes were TTN (52%), TP53 (47%), MUC6 (29%), LRP1B (28%), and SYNE1 (25%), which are well-known tumor suppressor genes[33] (Figure S2). For the three subtypes, we observed high prevalence of TTN (63%), ARID1A (39%), and TP53 (39%) mutations in subtype 1, TP53 (73%), TTN (48%), and LRP1B (27%) mutations in subtype 2, and TTN (34%), TP53 (29%), and CDH1 (21%) mutations in subtype 3 (Figures 3A–3C). Cross-comparisons showed that subtype 1 had a significantly higher mutation rate of ARID1A (odds ratio [OR], 11.4; p = 6e-10) and PIK3CA (OR, 18.9; p = 8e-9) than subtype 2, whereas subtype 2 had a significantly higher mutation rate of TP53 than subtypes 1 (OR, 0.2; p = 1e-7) and 3 (OR, 6.7; p = 7e-9), and subtype 3 had a significantly higher mutation rate of CDH1 than subtypes 1 (OR, 0.4; p = 0.02) and 2 (OR, 0.08; p = 7e-5) (Figures 3D and 3E). Based on the significantly mutated driver genes, we named these three subtypes ARID1A$^+$, TP53$^+$, and CDH1$^+$.

We further explored the co-occurring or mutually exclusive interactions between the top 25 mutated genes in each subtype. We found that subtype 1 (ARID1A$^+$ type) had closer linked co-mutation interactions than subtypes 2 (TP53$^+$ type) and 3 (CDH1$^+$ type) (Figure S3). As shown in Figures S3D and S3E, the subtype 1 hallmark gene of ARID1A is mutually exclusive with the subtype 2 hallmark gene of TP53 (p = 1e-2), and TP53 has no co-occurring interactions, except for LRP1B (p = 0.03). The hallmark genes of TP53 and CDH1 had no co-occurring interactions in subtypes 2 and 3, and CDH1 was mutually exclusive with LRP1B (p = 0.01) (Figures S3E and S3F). These findings suggest that the hallmark genes ARID1A, TP53, and CDH1 can distinguish the three subtypes fairly well.

To further investigate the mutation features, we identified the Catalog of Somatic Mutations in Cancer (COSMIC) mutational signatures of the three subtypes. Overall, the top three signatures have distinct SNVs mutation characteristics and a proportional contribution of each signature per sample (Figure S4). The cosine similarities of the detected mutations of each subtype against the 30 validated COSMIC signatures are shown in Figure 3F. The best match signatures of the ARID1A$^+$, TP53$^+$, and CDH1$^+$ types were SBS6 (associated with defective DNA mismatch repair; cosine similarity, 0.915), SBS17 (cosine similarity, 0.92), and SBS1 (associated with spontaneous or enzymatic deamination of 5-methylcytosine; cosine similarity, 0.891), respectively (Figure 3G). In addition, SBS10 was specific to the ARID1A$^+$ type.

Gene set variation characters of the three subtypes

To comprehensively reveal the variation in the activity of cancer-related gene sets for each subtype, we downloaded 6,019 related gene sets from the Molecular Signatures Database (MSigDB) database, including hallmark gene sets (Hs) (n = 38), regulatory target gene sets (C3) (n = 867), cancer modules (C4) (n = 168), oncogenic signature gene sets (C6) (n = 167), and immunologic signature gene sets (C7) (n = 4,779). Based on the expression data for each subtype, the enrichment score (ES) for different collections of gene sets was calculated using the gene set variation analysis (GSVA) method. A heatmap of the ESs for different subtypes and their clinical characteristics is shown in Figure 4.

are shown below. (D and E) Forest plot showing the significantly different mutated genes between the three subtypes. The horizontal coordinates represent the OR of the different groups and the horizontal line represents the 95% confident interval of the OR. The mutation frequency, OR and its significance p value for the representative genes are listed. (F) Heatmap showing the cosine similarities of the top 2 matched COSMIC mutation signatures for each subtype. (G) Best matched COSMIC signature and its corresponding etiology term of the ARID1A$^+$, TP53$^+$, and CDH1$^+$ types. COSMIC, Catalog of Somatic Mutations in Cancer.

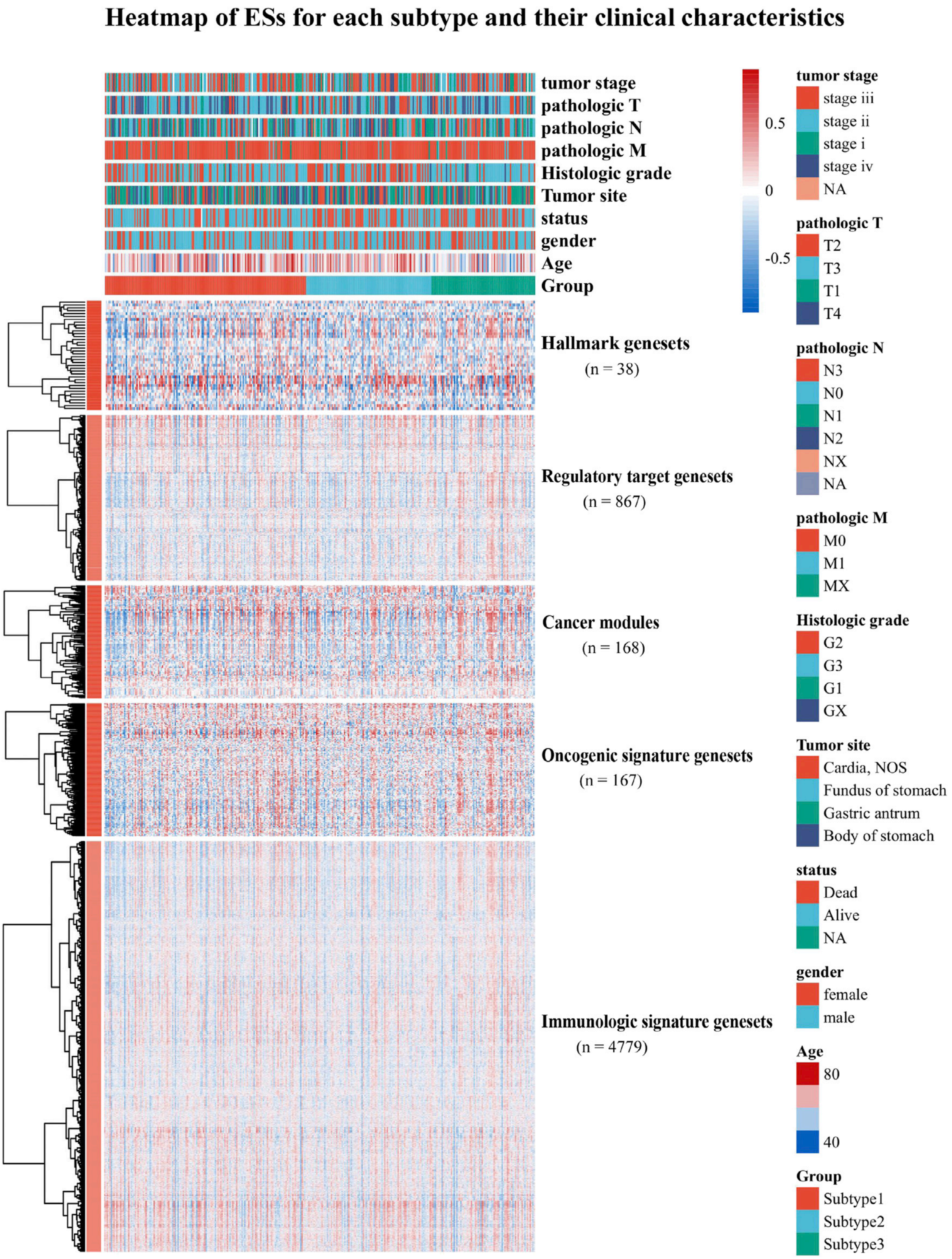
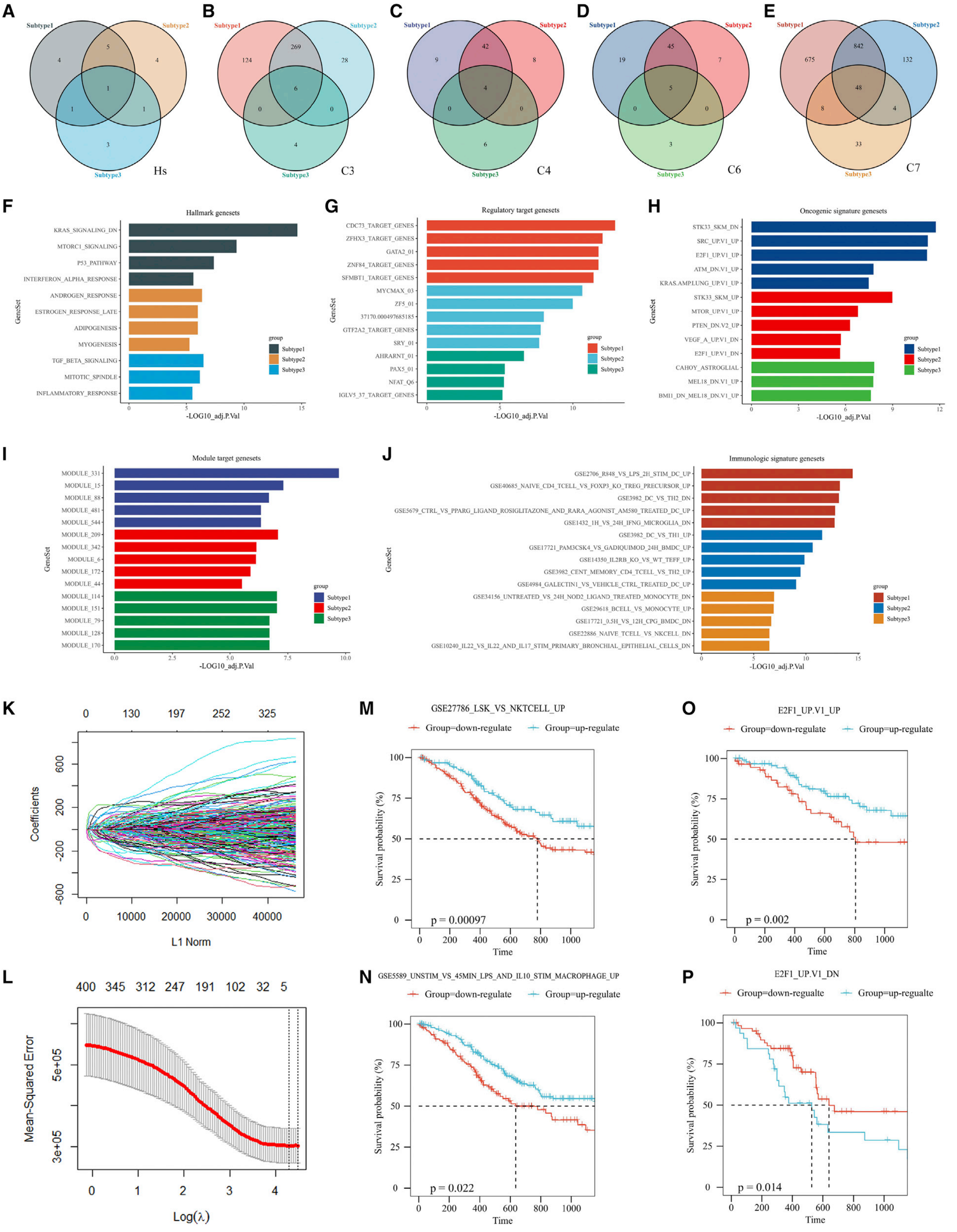


Figure 4. Heatmaps showing the ES of each subtype on five different gene sets

The clinical characteristics of 323 patients are depicted.



(legend on next page)

Compared with the normal samples, a total of 2,107, 1,451, and 127 differential gene sets (|logFC| $\geq$ 0.2; adjusted p $\leq$ 1e-5) for the ARID1A$^+$, TP53$^+$, and CDH1$^+$ type, respectively, were identified (Table S1, Figures 5A–5E). For a specific functional collection, the common and representative differential gene sets for each subtype were compared (Figures 5A–5E), and the top five characteristic gene sets for Hs, C3, C4, C6, and C7 are shown in Figures 5F–5J. The most significantly enriched characteristic Hs were KRAS_SIGNALING_DN, ANDROGEN_RESPONSE, and TGF_BETA_SIGNALING for the ARID1A$^+$, TP53$^+$, and CDH1$^+$ types, respectively. For other collections, the characteristic regulatory target gene sets were CDC73_TARGET_GENES, MYC-MAX_03, and AHRARNT_01; the characteristic oncogenic signature gene sets were STK33_SKM_DN, STK33_SKM_UP, and CAHOY_ASTROGLIAL; the characteristic cancer modules were MODULE_331, MODULE_209, and MODULE_114; and the characteristic immunologic signature gene sets were GSE2706_R848_VS_LPS_2H_STIM_DC_UP, GSE3982_DC_VS_TH2_UP, and GS UNTREATED_VS_24H_NOD2_LIGAND_TREATED_MONOCYTE_DN. Notably, both ARID1A$^+$ and TP53$^+$ types were enriched in STK33- and E2F1-related oncogenic signature, but the two subtypes showed opposite responses (Figure 5H).

Next, we sought to identify the prognostic gene sets for the three subtypes using the least absolute shrinkage and selection operator (LASSO) method, and five gene sets were ultimately identified (Figures 5K–5L), that is, GSE27786_LSK_VS_NKTCELL_UP, GSE5589_UNSTIM_VS_45MIN_LPS_AND_IL10_STIM_MACROPHAGE_UP, ZIC1_01, GSE22589_SIV_VS_HIV_AND_SIV_INFECTED_DC_UP, and GSE42021_TREG_PLN_VS_CD24LO_TREG_THYMUS_DN (Figures 5M and 5N). High ESs of GSE27786_LSK_VS_NKTCELL_UP (log rank test; p = 9-e4) and GSE5589_UNSTIM_VS_45MIN_LPS_AND_IL10_STIM_MACROPHAGE_UP (log rank test; p = 0.02) were associated with better overall survival (Figures 5O and 5P). In addition, the upregulation or downregulation of E2F1-related oncogenic signature had opposite survival curves ($p < 0.05$) (Figures 5O and 5P).

Characteristic module and driver genes of the three subtypes

Next, we performed a gene network analysis to uncover the subtype-specific module and driver genes, which may contribute to their clinical and biological characteristics. A protein-protein interaction (PPI) network of 125 DEGs (p $\leq$ 0.05; |logFC| $\geq$ 1) for the three subtypes was constructed using the STRING database (Figure 6A). In the network, there were 51, 16, and 4 genes specific to the ARID1A$^+$, TP53$^+$, and CDH1$^+$ types, respectively. Based on the subtype-specific DEGs, nine modules (five for ARID1A$^+$ type, three for TP53$^+$ type, and one for CDH1$^+$ type) were observed using the walktrap algorithm, and its driver genes in each module were obtained by values of five node importance indicators, including degree centrality, eigenvector, betweenness, pagerank, and closeness (Figure 6A, Table S1). For the ARID1A$^+$ type, the diver genes were ORC1, EZH2, CDC7, ASF1B, CENPU, and CDCA7 in module 1; MAPK4 and SLC22A17 in module 2; and DUSP26 in module 3. For the TP53$^+$ type, the driver genes were IGFBP1 and MATN3 in module 6, DKK1 in module 7, and ADAMTSL3 in module 8. For the CDH1$^+$ type, the driver gene was APOA1 in module 9. Except for CENPU, all driver gene expression levels were associated with overall survival (log rank test; $p < 0.05$) (Figures 6B–6I and S5).

Predicted sensitive drugs for each subtype

To determine whether specific subtypes were associated with an increased clinical benefit from adjuvant chemotherapy, we compared the recurrence-free survival (RSF) rates of patients with or without received adjuvant chemotherapy in the KYK cohort (n = 180). In agreement with the overall survival, the ARID1A$^+$ type patients (70 received adjuvant chemotherapy and 20 did not) were shown to benefit from adjuvant chemotherapy (log rank test; $p < 0.05$) (Figure 7A), but no significant improvement in RSF from adjuvant chemotherapy was observed among the TP53$^+$ type and CDH1$^+$ type patients (62 received adjuvant chemotherapy and 28 did not) (Figure 7B).

Based on the DEGs and GC cell lines from the Genomics of Drug Sensitivity in Cancer (GDSC) database, sensitive drugs associated with certain subtypes were predicted using the oncoPredict R package (v3.46.0). According to the half maximal inhibitory concentration values, the predicted top 20 sensitive drugs for the 3 subtypes were obtained (Figure 7C). To further identify the subtype-specific drugs, correlations between the gene expression of subtype-specific DEGs and drug sensitivity were calculated, and drugs and genes with correlation coefficient of less than 0.6 and a p value of less than 0.05 were selected (Figure 7D). We observed that afatinib, AZD8055, osimertinib, and PD0325901 had strong correlations with the DEGs of the ARID1A$^+$ type, and the sabutoclax, telomerase inhibitor IX, and Wee1 inhibitor were strongly correlated with the DEGs of the TP53$^+$ type, whereas docetaxel, MG-132, pictilisib, and sepantronium bromide may be effective against the CDH1$^+$ type.

Figure 5. Enriched gene set comparison and prognostic gene sets identified for the three subtypes

(A–E) Venn diagram showing the overlapped and specific differential gene sets for each subtype on the hallmark genes sets (Hs), regulatory target gene sets (C3), cancer modules (C4), oncogenic signature gene sets (C6), and immunologic signature gene sets (C7). (F–J) Top five characteristic gene sets for each subtype on Hs, C3, C4, C6, and C7. (K–L) Least absolute shrinkage and selection operator (LASSO)-based prognostic gene set identification. (K) Coefficient profiles (*y* axis) of the gene sets and the optimal penalization coefficient (λ) via 3-fold cross-validation based on partial likelihood deviance. The dotted vertical lines in (L) represent the optimal values of λ, and the top *x* axis shows the numbers of gene sets. (M and N) Kaplan-Meier curves showing that high ESs of GSE27786_LSK_VS_NKTCELL_UP and GSE5589_UNSTIM_VS_45MIN_LPS_AND_IL10_STIM_MACROPHAGE_UP gene sets were associated with better overall survival. The p values were obtained using the log rank test. (O and P) Kaplan-Meier curves showing up or downregulated E2F1-related oncogenic signature associated with reversed survival, which represent the discrepant characteristics of the ARID1A$^+$ and TP53$^+$ types. The p values were obtained using the log rank test.

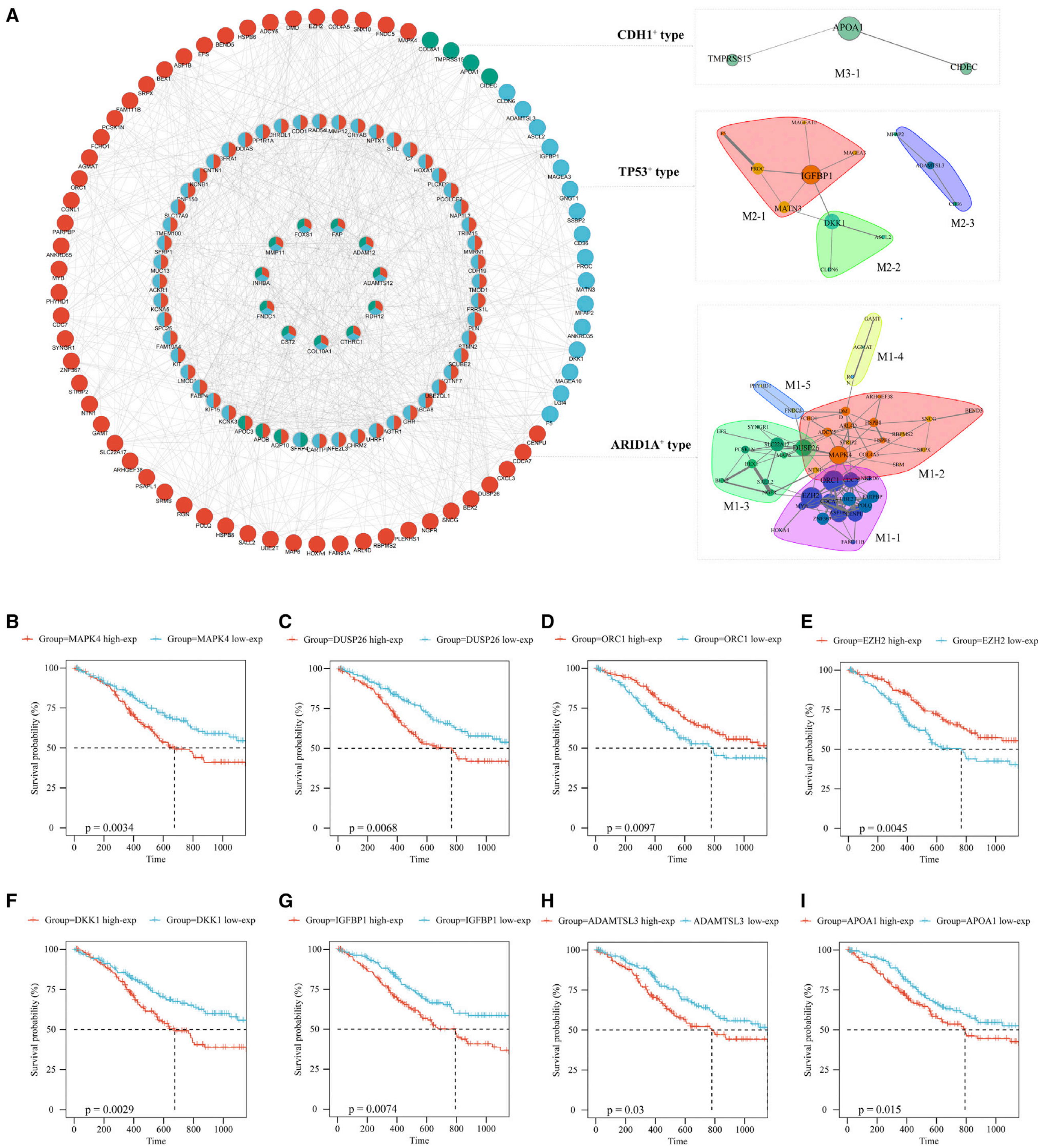
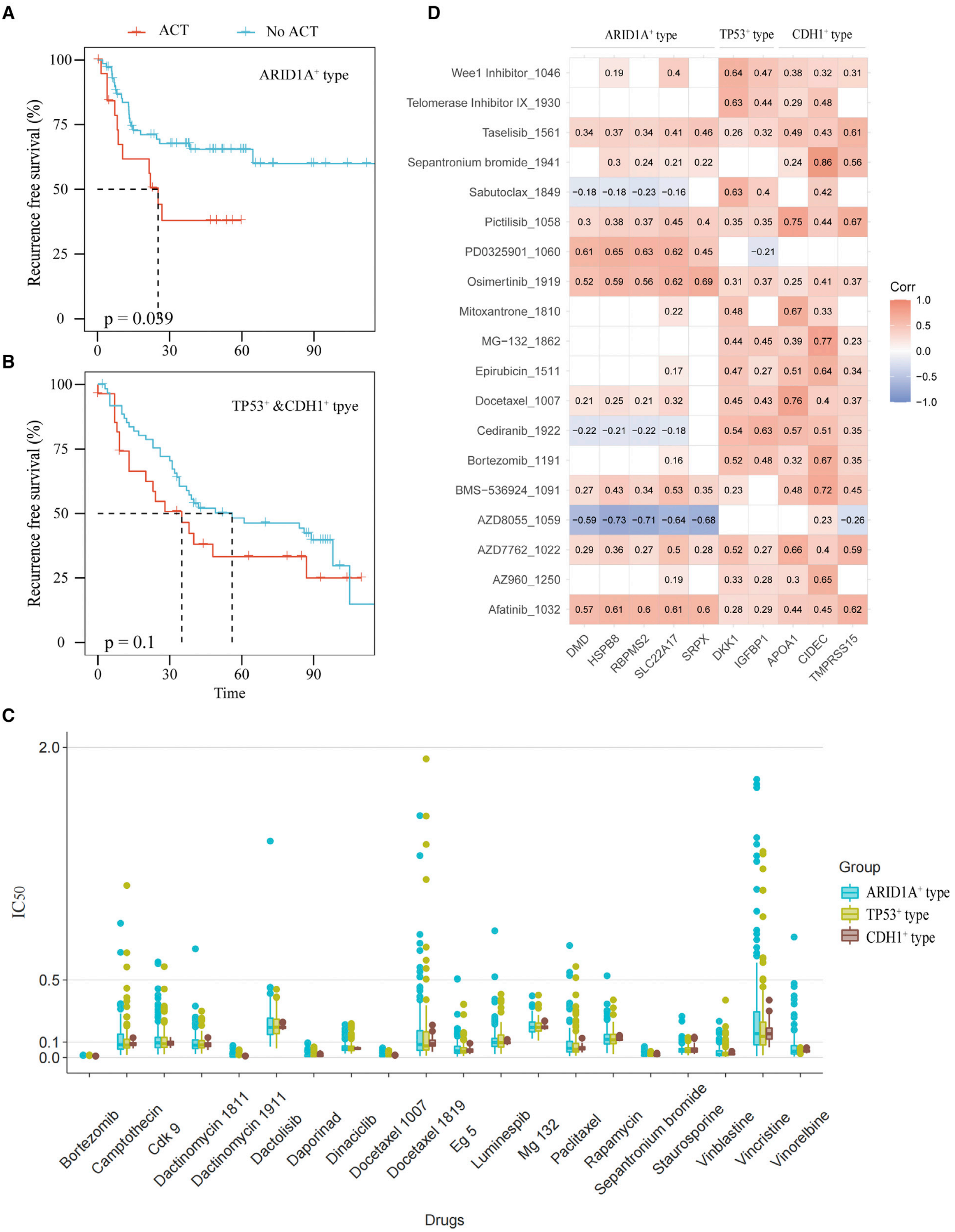


Figure 6. Network modules and driver genes of the three subtypes

(A) Left, PPI network of 125 DEGs of the three subtypes. Red, blue, and green represent DEGs of the ARID1A$^+$, TP53$^+$, and CDH1$^+$ types, respectively. Nodes with mixed colors are the common DEGs of the corresponding subtypes. Right, subtype-specific modules and driver genes. Different colors represent the corresponding communities (modules), the node size represents the network degree, and the edge width represents the strength of connections. (B–I) Kaplan-Meier curves showing that high or low expression of the driver genes are associated with survival status. The p values were obtained using the log rank test.



(legend on next page)

DISCUSSION

Classification based on molecular subtypes provides an opportunity for personalized therapy of GC.[2] Several classification schemes have been proposed using various molecular platforms, but their clinical translation has been hindered by a lack of consensus on the subtypes and precise molecular signatures.[20,34] Here, we report a refined, well tested GC classification scheme with three subtypes associated with distinct clinical outcomes and molecular characteristics (Figure 8). Subtype 1 (ARID1A$^+$ type) is characterized by high ARID1A and PIK3CA mutations and is associated with favorable prognosis, and predominantly corresponds to the previously reported MSI, EBV, and EP subtypes. Subtype 2 (TP53$^+$ type) harbors a highly recurrent TP53 mutation, is associated with poor prognosis, and mainly corresponds with the previously reported CIN and EP subtypes. Subtype 3 (CDH1$^+$ type) is accompanied by a high CHD1 mutation and is associated with a poor prognosis; it mainly corresponds with the previously reported GS and MP subtypes. Our analysis demonstrated the consensus clinical significance of the three subtypes and validated them in multiple independent cohorts, which may lead to improvements in the precise treatment of GC.

Specific molecular biomarkers that can clearly distinguish each subtype are required for the clinical applicability of the classification scheme. The three proposed subtypes were found to be associated with distinct genomic characteristics in mutational signatures, driver genes, enriched gene sets, and chemotherapy sensitivity. We named the three subtypes according to their unique mutated driver genes, that is, ARID1A, TP53, and CDH1, which are the reported top driver genes enriched in the specific molecular subgroups.[33,34] Frequent inactivating mutations or protein deficiency of ARID1A are found in the majority of GC with MSI and EBV. The mutation spectrum differs between molecular subtypes, and the mutation prevalence is negatively associated with mutations in TP53, suggesting that ARID1A alterations are associated with a better prognosis in a subtype-independent manner.[35] A study also demonstrated that ARID1A mutations were significantly correlated with increased phosphorylation of oncogenic signaling proteins and are a biomarker for promising GC therapeutics.[36,37] TP53 is one of the key driver genes that are frequently mutated in GC[38,39] and is widely used to differentiate subtypes. TP53 is a low-frequency mutation gene in both MSI and EBV tumors, but it is positively associated with the CIN subtype,[8] which is consistent with the results of this analysis. CHD1 mutations are frequently associated with TCGA GS subtype and have a unique role in driving diffuse-type gastric carcinogenesis.[9,40,41] In addition, TP53 and CDH1 are frequently mutated in peritoneal metastases of GC, and the co-inactivation of CHD1 and TP53 may lead to poor progression[42]; however, their co-occurrence was not found in this analysis. These results suggest that the unique mutated driver genes can distinguish each subtype fairly well.

Based on the proposed prediction model, it was found that the three subtypes were present in independent test cohorts, and the corresponding relationship and clinical characteristics of patients were in good agreement with previously reported classifications, demonstrating the robustness and reproducibility of the proposed classification. Moreover, the three subtypes showed different clinical outcomes in terms of survival and RFS. We found that the ARID1A$^+$ type had a better prognosis than the other subtypes, which is consistent with the previously reported EBV, MSI, and EP types.[12,15] Similar to the survival outcomes, a subset analysis of patients with the available chemotherapy data strongly suggested that the ARID1A$^+$ type is associated with a benefit from adjuvant chemotherapy, which is also consistent with the EP type.[15] These results indicate that the molecular characteristics of each subtype may help to develop rational therapy recommendations for patients with GS.

In addition to unique mutation signatures, network-based driver DEGs for each subtype were also reported, including ORC1, EZH2, CDC7, ASF1B, CDCA7, MAPK4, SLC22A17, and DUSP26 for the ARID1A$^+$ type; IGFBP1, MATN3, DKK1, and ADAMTSL3 for the TP53$^+$ type; and APOA1 for the CDH1$^+$ type, most of which have not been previously reported in GC. Studies have suggested that EZH2 may serve as a potential target in ARID1A-deficient GC.[43] CDCA7 may regulate inflammation through the toll-like receptor 4/nuclear factor κB signaling pathway to regulate GC development.[44] ASF1B, ORC1, and SLC22A17 have also been found to be related to the progression or metastasis of GC.[44–46] It has been reported that the expression levels of DKK1, MATN3, and IGFBP1 are significantly associated with survival in patients with GC.[47–49] In addition to CDH1, RHOA is a representative mutation driver gene for the diffuse-type and GS subtype.[12,26,50] RHOA mutations were found to be associated with poor tumor differentiation, and patients with RHOA mutations were less likely to have TP53 mutations.[33]

In summary, we identified a three-subtype classification framework of GC that is associated with distinct survival outcomes and molecular features. Consensus regarding the clinical significance and multiple molecular signatures of the three subtypes were revealed. Although these molecular subtypes should be further evaluated in clinical trials for distinct patients with GC, we believe that the refined classification scheme may contribute to the development of more effective therapeutic strategies.

MATERIALS AND METHODS

Genomic data collection and processing

Multi-omics data from TCGA stomach cancer cohort were downloaded from the UCSC Xena platform (https://xenabrowser.net/), including 407 samples for gene expression from RNA sequencing (Illumina HiSeq), 477 samples for miRNA (Illumina HiSeq), 397

Figure 7. Chemotherapy sensitivity and the predicted drugs for the three subtypes

(A and B) Kaplan-Meier curves showing the RSF rate of patients with or without adjuvant chemotherapy. The p values were obtained using the log rank test. (C) Predicted top 20 sensitive drugs with lower half-maximal inhibitory concentration (IC_{50}) value. (D) Heatmap showing the correlation coefficient of the subtype-specific DEGs and drug sensitivity (IC_{50}); only drugs and genes with a correlation coefficient of greater than 0.6 and a p value of less than 0.05 are listed. ACT, adjuvant chemotherapy.

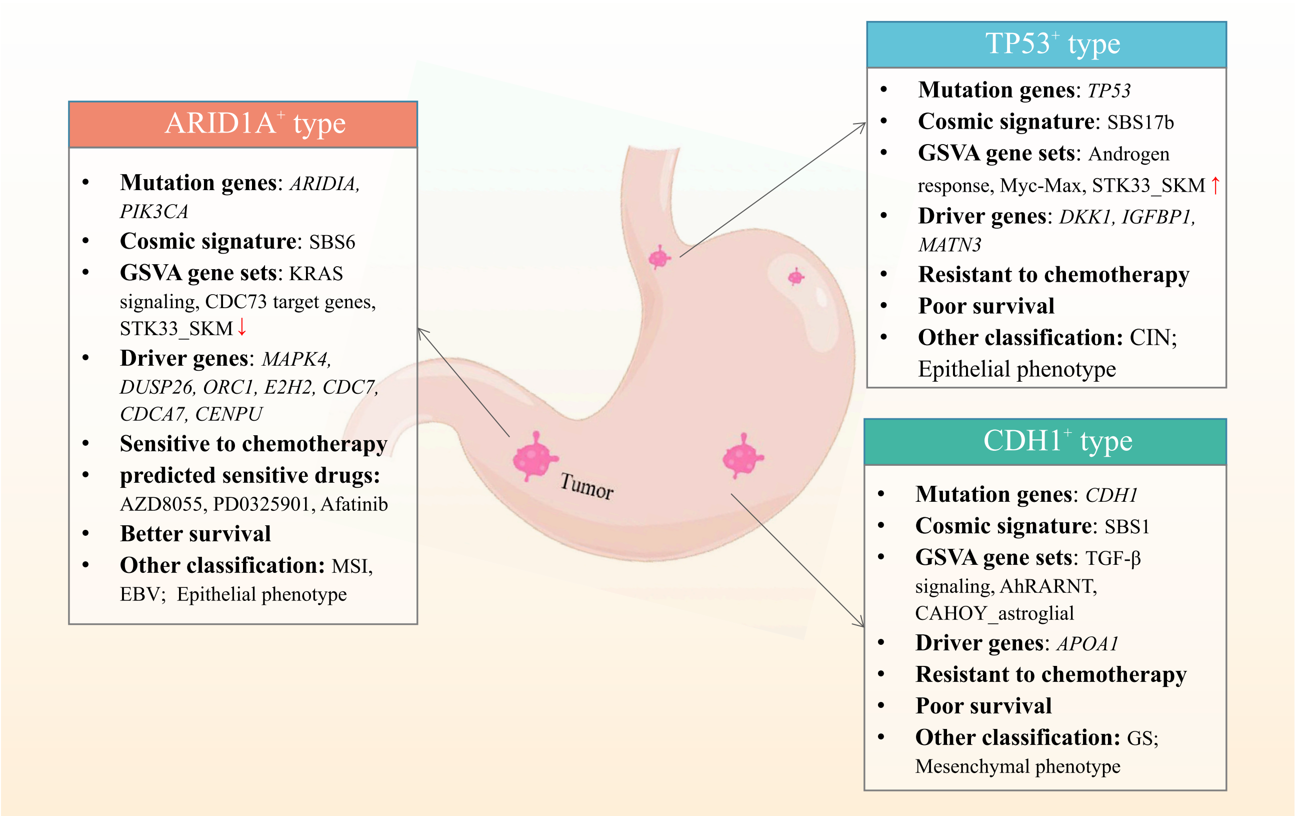

ARID1A+ type
Mutation genes: ARIDIA, PIK3CA
Cosmic signature: SBS6
GSVA gene sets: KRAS signaling, CDC73 target genes, STK33_SKM ↓
Driver genes: MAPK4, DUSP26, ORC1, E2H2, CDC7, CDCA7, CENPU
Sensitive to chemotherapy
predicted sensitive drugs: AZD8055, PD0325901, Afatinib
Better survival
Other classification: MSI, EBV; Epithelial phenotype
Tumor
TP53+ type
Mutation genes: TP53
Cosmic signature: SBS17b
GSVA gene sets: Androgen response, Myc-Max, STK33_SKM ↑
Driver genes: DKK1, IGFBP1, MATN3
Resistant to chemotherapy
Poor survival
Other classification: CIN; Epithelial phenotype
CDH1+ type
Mutation genes: CDH1
Cosmic signature: SBS1
GSVA gene sets: TGF-β signaling, AhRARNT, CAHOY_astroglial
Driver genes: APOA1
Resistant to chemotherapy
Poor survival
Other classification: GS; Mesenchymal phenotype

Figure 8. Salient features of the three GC subtypes

samples for DNA methylation (Illumina Human Methylation 450 K), and the corresponding clinical phenotypes. All multi-omics datasets were pre-processed to a normalized value. The analysis was limited to samples with all of the three kinds of omics and survival data available; thus, a total of 323 cancer samples were used for classification analysis.

Gene expression profiling data of GC tissues were downloaded from the GEO (http://www.ncbi.nlm.nih.gov/geo/) database to test the classification results, including GSE26253 (SMC cohort; n = 432), GSE62254 (ACRG cohort; n = 300), GSE26899 (KUGH cohort; n = 93), GSE13861 (YUSH cohort; n = 65), and GSE26901 (KUCM cohort; n = 109). All gene expression datasets were normalized and transformed into log2 bases before further analysis.

Clustering analysis and subtype identification

To identify clinical outcome-related molecular subtypes of GC, optimal multi-omics data-based clustering methods of SNF and CC, which were integrated into the CancerSubtypes R package, were used.[28] CancerSubtypes integrates four highly cited cluster methods for cancer subtyping and provides a standardized framework for data preprocessing, feature selection, and result analyses. SNF outperforms other commonly used multi-omics-based methods in terms of accuracy, robustness, and computation efficiency criteria and is recommended for cancer subtyping because it can capture both shared and specific information from different omics data and make the integrated similarity networks retain more information from every single similarity network with low-level noise.[21] Combining SNF and CC takes advantage of both achieved superior performance of clinical relevance on survival differences.[28] CancerSubtypes also provides four important built-in feature selection methods and a CC algorithm to uncover potential differences among the various subtypes. In this study, a multivariate Cox regression model was used to select important features based on the mRNA, miRNA, and DNA methylation data. For consensus k-means clustering, the NbClust R package, which integrates 26 criteria, was used to choose the optimal clustering number, and up to 10 clusters were evaluated for all samples. Cluster quality was assessed using the silhouette widths of the final cluster results.

Differential expression analysis

Based on the classification results, we next identified differentially expressed mRNAs, miRNAs, and DNA methylation sites between tumor samples in each subtype and normal samples. DEGs were used as features for building the prediction model. Differential expression analysis was performed using the limma R package, which identifies DEGs using the log2-fold-changes (logFC). Limma fits a linear model to compute moderated t-statistics using a Bayesian model and adjusts the p values for multiple testing using Benjamini and Hochberg's method. Genes, miRNAs, and DNA methylation sites with an adjusted p value of 0.05 or less and |logFC| of 1 or less considered differentially expressed.

Prediction model and independent validation

To validate the classification results on independent GC datasets, we generated a prediction model based on the optimal machine learning algorithm XGBoost[51] to predict each new sample into a particular subtype group. XGBoost is an efficient and scalable machine learning classifier based on the gradient boosting decision tree algorithm, which provides parallel tree boosting and enhances performance using learning rate, subsampling ratio, and maximum tree depth to make the model less prone to overfitting.[51] XGBoost can achieve remarkable accuracy in multiple regression tasks and provides satisfactory results in several machine learning competitions compared with other algorithms commonly used in predictive model construction.[52–55] Given a dataset $D = \{(x_i, y_i)\}$, where x_i denotes the gene expression profile of tumor, y_i is the corresponding sample, assuming it has k decision trees, the optimization objective function is calculated using Equation 1:

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F, \tag{1}$$

where f_k is an independent tree with leaf scores and F is the space of the regression tree. The loss function is calculated using Equation 2:

$$L(f_t) = \sum l(\widehat{y_i}, y_i) + \sum \Omega(f_i), \tag{2}$$

where l is a differentiable loss function that measures the difference between the predicted output $\widehat{y_i}$ and true output y_i, and Ω is a regularization term that penalizes the complexity of the model to prevent overfitting. The $\widehat{y_i}$ and Ω can be written as Equations 3 and 4, respectively:

$$\widehat{y_i}^{(t)} = \widehat{y_i}^{(t-1)} + f_t(x_i) \tag{3}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\|w\|^2, \tag{4}$$

where T is the number of leaf nodes and w is the score on each leaf. Thus, the loss function is calculated using Equation 5:

$$L(f_t) \approx \sum_{j=1}^{T}\left[\left(\sum_{i \in I_j} g_i\right) w_j + \frac{1}{2}\left(\sum_{i \in I_j} h_i + \lambda\right) w_j^2\right] + \gamma T, \tag{5}$$

where *GI* and h_i are first-order and second-order gradient statistics of the loss function. The parameters γ and λ are constants that control the degree of regularization and are used to prevent overfitting.

In this study, all DEGs in the three subtypes were used as feature gene sets for the development of the model. Based on the classification results in the training set, the model was performed using the XGBoost Python package (https://pypi.org/project/xgboost/v1.6.2). To obtain a better fit, the AUC value based on 5-fold cross-validation was used to adjust the model parameters. The maximum depth of the tree was set to $k = 3$ and $\lambda = 4$ in the prediction model. When the classifier was applied to independent test sets, prognostic significance was estimated using Kaplan-Meier survival plots and log rank tests between the predicted subtypes of patients.

To further test the performance of our three subtypes compared with TCGA four subtype classification scheme, the multinomial naive Bayes model was used to evaluate the classification effect specific to real clinical outcomes. The pathological type (Lauren classification and WHO classification), TNM stage, and survival status were selected as features to perform naive Bayes classification. The sensitivity, specificity, and area under the receiver operating characteristic curve were used to compare the performance of the two classification schemes.

Mutation signature analysis

Next, we attempted to clarify the mutation characteristics and identify specific mutation signatures of the three GC subtypes. Somatic mutation and signature analysis were performed and visualized using the Maftools and musicatk R package.[56,57] Frequency matrix generation and nonnegative matrix factorization were performed to extract mutational signatures for each subtype and compare them with 30 known signatures referenced in the COSMIC,[58] and a cosine similarity value was estimated for the best possible match. To identify the differentially mutated genes, a 2 × 2 contingency table of mutation frequencies was calculated for every gene from the input cohorts, and the significant differences were estimated by Fisher's exact test. The results from pairwise subtype cohort comparisons were visualized as forest plots. The gene sets mutated in a mutually exclusive or co-occurring manner were also identified by performing Fisher's exact test on a 2 × 2 contingency table containing frequencies of mutated and non-mutated samples. To view the relationships between tumors in two dimensions, uniform manifold approximation and projections were used to display the levels of each signature exposure.

GSVA

GSVA[59] was performed to determine the biological functions of each subtype. GSVA can estimate the relative enrichment of a gene set of interest over a sample population, and is used to observe the variation in the activity of a set of genes (e.g., a pathway) corresponding with a particular biological condition.[59] This method outperforms single-gene analysis in terms of feature dimension, noise interference, and biological interpretability.[59] The major cancer-related gene set collections were downloaded from the MSigDB, http://www.gsea-msigdb.org/gsea/msigdb,[60] which is one of the most widely used and comprehensive databases for performing gene set enrichment analysis. The normalized GSVA ES for each collection of gene sets was measured for each GC sample using the GSVA R package. Based on the ES, differential gene sets between the two groups were measured using the limma R package, terms with |logFC| ≥ 0.2 and an adjusted p value of 1e-5 or less were considered statistically significant.

Characteristic network and driver gene identification

Next, we explored the molecular characteristics of each subtype based on a network approach. Based on the DEGs of each subtype, PPI networks were constructed using the STRING database (https://cn.string-db.org/).[61] For the subtype-specific DEGs, the igraph R package was used to identify the community structure (module) via the walktrap random walk algorithm. The driver genes were selected based on network topological parameters such as degree and betweenness centrality. The network and modules were visualized using the igraph R package and Cytoscape v3.8.0.[62]

Chemotherapeutic benefits and drug sensitivity prediction

To determine whether each subtype is associated with different clinical benefits from adjuvant chemotherapy, the RSF rate was compared between subsets of patients with or without adjuvant chemotherapy in the KUGH, YUSH, and KUCM cohorts. Sensitive drugs associated with certain subtypes were predicted using the OncoPredict R package.[63] According to subtype-specific DEGs, the half maximal inhibitory concentration values of 198 drugs for each sample were predicted based on the cell lines from the GDSC2.[64]

Statistical analyses

All statistical analyses were performed in the R environment. Unpaired Student's t-tests were used to evaluate the statistical significance of the normally distributed variables between the two groups. Continuous or non-parametric variables were assessed using one-way ANOVA or the Mann-Whitney test, respectively. Differences between categorical variables were compared using Pearson's χ^2 test or Fisher's exact test. The association between survival and molecular subtypes was analyzed using the Cox proportional hazards model. Survival data were analyzed by Kaplan-Meier curves with a log rank test. The LASSO regression method was used to identify potential prognostic gene sets. The ComplexHeatmap R package generated all heat maps.

DATA AVAILABILITY STATEMENT

All data in the main text or supplemental material are available. The genomic data supporting the findings of this study are available from the UCSC Xena database (TCGA Stomach Cancer datasets) and NCBI GEO under the accession numbers GSE26253 for SMC, GSE62254 for ACRG, GSE26899 for KUGH, GSE26901 for KUCM, and GSE13861 for YUSH.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2022.12.014.

ACKNOWLEDGMENTS

Funded by the Scientific and Technological Innovation Project of China Academy of Chinese Medical Sciences (No. CI2021A05052 to B.L., CI2021B015 to B.L.), the National Natural Science Foundation of China (No. 81803966 to B.L.), the Fundamental Research Funds for the Central Public Welfare Research Institutes (ZXKT21024 to B.L.), and the Key Research and Development Projects of Hebei Province (No. 21377759D to F.B.Z.).

AUTHOR CONTRIBUTIONS

Z.W., B.L., and H.M.Z. contributed to the conception and design of the work. B.L., Q.K.N., J.L., and Y.N.Y. conducted data analysis. F.B.Z., P.Q.W., and S.Q.Z. contributed to data collation and analysis tools. B.L. and Q.K.N. prepared all the figures and tables. B.L. and F.B.Z. drafted and revised the manuscript. Z.W. and H.M.Z. contributed to review of the manuscript. All authors reviewed and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

REFERENCES

1. Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA A Cancer J. Clin. *71*, 209–249.
2. Joshi, S.S., and Badgwell, B.D. (2021). Current treatment and recent progress in gastric cancer. CA A Cancer J. Clin. *71*, 264–279.
3. Cunningham, D., Allum, W.H., Stenning, S.P., Thompson, J.N., Van de Velde, C.J.H., Nicolson, M., Scarffe, J.H., Lofts, F.J., Falk, S.J., Iveson, T.J., et al. (2006). Perioperative chemotherapy versus surgery alone for resectable gastroesophageal cancer. N. Engl. J. Med. *355*, 11–20.
4. Noh, S.H., Park, S.R., Yang, H.K., Chung, H.C., Chung, I.J., Kim, S.W., Kim, H.H., Choi, J.H., Kim, H.K., Yu, W., et al. (2014). Adjuvant capecitabine plus oxaliplatin for gastric cancer after D2 gastrectomy (CLASSIC): 5-year follow-up of an open-label, randomised phase 3 trial. Lancet Oncol. *15*, 1389–1396.
5. Sakuramoto, S., Sasako, M., Yamaguchi, T., Kinoshita, T., Fujii, M., Nashimoto, A., Furukawa, H., Nakajima, T., Ohashi, Y., Imamura, H., et al. (2007). Adjuvant chemotherapy for gastric cancer with S-1, an oral fluoropyrimidine. N. Engl. J. Med. *357*, 1810–1820.
6. Aoyama, T., Yoshikawa, T., Watanabe, T., Hayashi, T., Ogata, T., Cho, H., and Tsuburaya, A. (2011). Survival and prognosticators of gastric cancer that recurs after adjuvant chemotherapy with S-1. Gastric Cancer *14*, 150–154.
7. Lee, J., Lim, D.H., Kim, S., Park, S.H., Park, J.O., Park, Y.S., Lim, H.Y., Choi, M.G., Sohn, T.S., Noh, J.H., et al. (2012). Phase III trial comparing capecitabine plus cisplatin versus capecitabine plus cisplatin with concurrent capecitabine radiotherapy in completely resected gastric cancer with D2 lymph node dissection: the ARTIST trial. J. Clin. Oncol. *30*, 268–273.
8. Cristescu, R., Lee, J., Nebozhyn, M., Kim, K.M., Ting, J.C., Wong, S.S., Liu, J., Yue, Y.G., Wang, J., Yu, K., et al. (2015). Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nat. Med. *21*, 449–456.
9. Chia, N.Y., and Tan, P. (2016). Molecular classification of gastric cancer. Ann. Oncol. *27*, 763–769.
10. Lauren, P. (1965). The two histological main types of gastric carcinoma: diffuse and so-called intestinal-type carcinoma. An attempt at a histo-clinical classification. Acta Pathol. Microbiol. Scand. *64*, 31–49.
11. Nagtegaal, I.D., Odze, R.D., Klimstra, D., Paradis, V., Rugge, M., Schirmacher, P., Washington, K.M., Carneiro, F., and Cree, I.A.; WHO Classification of Tumours Editorial Board (2020). The 2019 WHO classification of tumours of the digestive system. Histopathology *76*, 182–188.
12. The Cancer Genome Atlas Research Network (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature *513*, 202–209.
13. Wong, S.S., Kim, K.M., Ting, J.C., Yu, K., Fu, J., Liu, S., Cristescu, R., Nebozhyn, M., Gong, L., Yue, Y.G., et al. (2014). Genomic landscape and genetic heterogeneity in gastric adenocarcinoma revealed by whole-genome sequencing. Nat. Commun. *5*, 5477.

14. Zouridis, H., Deng, N., Ivanova, T., Zhu, Y., Wong, B., Huang, D., Wu, Y.H., Wu, Y., Tan, I.B., Liem, N., et al. (2012). Methylation subtypes and large-scale epigenetic alterations in gastric cancer. Sci. Transl. Med. *4*, 156ra140.

15. Oh, S.C., Sohn, B.H., Cheong, J.H., Kim, S.B., Lee, J.E., Park, K.C., Lee, S.H., Park, J.L., Park, Y.Y., Lee, H.S., et al. (2018). Clinical and genomic landscape of gastric cancer with a mesenchymal phenotype. Nat. Commun. *9*, 1777.

16. Wang, H., Ding, Y., Chen, Y., Jiang, J., Chen, Y., Lu, J., Kong, M., Mo, F., Huang, Y., Zhao, W., et al. (2021). A novel genomic classification system of gastric cancer via integrating multidimensional genomic characteristics. Gastric Cancer *24*, 1227–1241.

17. Hu, X., Wang, Z., Wang, Q., Chen, K., Han, Q., Bai, S., Du, J., and Chen, W. (2021). Molecular classification reveals the diverse genetic and prognostic features of gastric cancer: a multi-omics consensus ensemble clustering. Biomed. Pharmacother. *144*, 112222.

18. Wang, J., Kunzke, T., Prade, V.M., Shen, J., Buck, A., Feuchtinger, A., Haffner, I., Luber, B., Liu, D.H.W., Langer, R., et al. (2022). Spatial metabolomics identifies distinct tumor-specific subtypes in gastric cancer patients. Clin. Cancer Res. *28*, 2865–2877.

19. Zeng, D., Li, M., Zhou, R., Zhang, J., Sun, H., Shi, M., Bin, J., Liao, Y., Rao, J., and Liao, W. (2019). Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures. Cancer Immunol. Res. *7*, 737–750.

20. Sohn, B.H., Hwang, J.E., Jang, H.J., Lee, H.S., Oh, S.C., Shim, J.J., Lee, K.W., Kim, E.H., Yim, S.Y., Lee, S.H., et al. (2017). Clinical significance of four molecular subtypes of gastric cancer identified by the cancer genome Atlas Project. Clin. Cancer Res. *23*, 4441–4449.

21. Duan, R., Gao, L., Gao, Y., Hu, Y., Xu, H., Huang, M., Song, K., Wang, H., Dong, Y., Jiang, C., et al. (2021). Evaluation and comparison of multi-omics data integration methods for cancer subtyping. PLoS Comput. Biol. *17*, e1009224.

22. Wang, B., Mezlini, A.M., Demir, F., Fiume, M., Tu, Z., Brudno, M., Haibe-Kains, B., and Goldenberg, A. (2014). Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods *11*, 333–337.

23. Rappoport, N., and Shamir, R. (2019). NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics *35*, 3348–3356.

24. Ramazzotti, D., Lal, A., Wang, B., Batzoglou, S., and Sidow, A. (2018). Multi-omic tumor data reveal diversity of molecular mechanisms that correlate with survival. Nat. Commun. *9*, 4453.

25. Meng, C., Helm, D., Frejno, M., and Kuster, B. (2016). moCluster: identifying joint patterns across multiple omics data sets. J. Proteome Res. *15*, 755–765.

26. Mo, Q., Shen, R., Guo, C., Vannucci, M., Chan, K.S., and Hilsenbeck, S.G. (2018). A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. Biostatistics *19*, 71–86.

27. Yang, H., Chen, R., Li, D., and Wang, Z. (2021). Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data. Bioinformatics *37*, 2231–2237.

28. Xu, T., Le, T.D., Liu, L., Su, N., Wang, R., Sun, B., Colaprico, A., Bontempi, G., and Li, J. (2017). CancerSubtypes: an R/Bioconductor package for molecular cancer subtype identification, validation and visualization. Bioinformatics *33*, 3131–3133.

29. Lu, X., Meng, J., Zhou, Y., Jiang, L., and Yan, F. (2020). MOVICS: an R package for multi-omics integration and visualization in cancer subtyping. Bioinformatics *36*, 5539–5541. btaa1018.

30. Zhang, M., Sheffield, T., Zhan, X., Li, Q., Yang, D.M., Wang, Y., Wang, S., Xie, Y., Wang, T., and Xiao, G. (2021). CeRNASeek: an R package for identification and analysis of ceRNA regulation. Briefings Bioinf. *22*, bbaa145. bbaa048.

31. De Mattos-Arruda, L., Vazquez, M., Finotello, F., Lepore, R., Porta, E., Hundal, J., Amengual-Rigo, P., Ng, C.K.Y., Valencia, A., Carrillo, J., et al. (2020). Neoantigen prediction and computational perspectives towards clinical benefit: recommendations from the ESMO Precision Medicine Working Group. Ann. Oncol. *31*, 978–990.

32. Liu, X.S., and Mardis, E.R. (2017). Applications of immunogenomics to cancer. Cell *168*, 600–612.

33. Wang, K., Yuen, S.T., Xu, J., Lee, S.P., Yan, H.H.N., Shi, S.T., Siu, H.C., Deng, S., Chu, K.M., Law, S., et al. (2014). Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. Nat. Genet. *46*, 573–582.

34. Yeoh, K.G., and Tan, P. (2022). Mapping the genomic diaspora of gastric cancer. Nat. Rev. Cancer *22*, 71–84.

35. Wang, K., Kan, J., Yuen, S.T., Shi, S.T., Chu, K.M., Law, S., Chan, T.L., Kan, Z., Chan, A.S.Y., Tsui, W.Y., et al. (2011). Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. Nat. Genet. *43*, 1219–1223.

36. Williamson, C.T., Miller, R., Pemberton, H.N., Jones, S.E., Campbell, J., Konde, A., Badham, N., Rafiq, R., Brough, R., Gulati, A., et al. (2016). ATR inhibitors as a synthetic lethal therapy for tumours deficient in ARID1A. Nat. Commun. *7*, 13837.

37. Mun, D.G., Bhin, J., Kim, S., Kim, H., Jung, J.H., Jung, Y., Jang, Y.E., Park, J.M., Kim, H., Jung, Y., et al. (2019). Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell *35*, 111–124.e10.

38. Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. *2*, 401–404.

39. Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. *6*, pl1.

40. Majewski, I.J., Kluijt, I., Cats, A., Scerri, T.S., de Jong, D., Kluin, R.J.C., Hansford, S., Hogervorst, F.B.L., Bosma, A.J., Hofland, I., et al. (2013). An alpha-E-catenin (CTNNA1) mutation in hereditary diffuse gastric cancer. J. Pathol. *229*, 621–629.

41. Hansford, S., Kaurah, P., Li-Chang, H., Woo, M., Senz, J., Pinheiro, H., Schrader, K.A., Schaeffer, D.F., Shumansky, K., Zogopoulos, G., et al. (2015). Hereditary diffuse gastric cancer syndrome: CDH1 mutations and beyond. JAMA Oncol. *1*, 23–32.

42. Wang, R., Song, S., Harada, K., Ghazanfari Amlashi, F., Badgwell, B., Pizzi, M.P., Xu, Y., Zhao, W., Dong, X., Jin, J., et al. (2020). Multiplex profiling of peritoneal metastases from gastric adenocarcinoma identified novel targets and molecular subtypes that predict treatment response. Gut *69*, 18–31.

43. Zhou, W., Li, J., Lu, X., Liu, F., An, T., Xiao, X., Kuo, Z.C., Wu, W., and He, Y. (2021). Derivation and validation of a prognostic model for cancer dependency genes based on CRISPR-cas9 in gastric adenocarcinoma. Front. Oncol. *11*, 617289.

44. Guo, Y., Zhou, K., Zhuang, X., Li, J., and Shen, X. (2021). CDCA7-regulated inflammatory mechanism through TLR4/NF-kappaB signaling pathway in stomach adenocarcinoma. Biofactors *47*, 865–878.

45. Chen, C., Bao, H., Lin, W., Chen, X., Huang, Y., Wang, H., Yang, Y., Liu, J., Lv, X., and Teng, L. (2022). ASF1b is a novel prognostic predictor associated with cell cycle signaling pathway in gastric cancer. J. Cancer *13*, 1985–2000.

46. Guo, X., Liang, X., Wang, Y., Cheng, A., Zhang, H., Qin, C., and Wang, Z. (2021). Significance of tumor mutation burden combined with immune infiltrates in the progression and prognosis of advanced gastric cancer. Front. Genet. *12*, 642608.

47. Wang, P., Xiao, W.S., Li, Y.H., Wu, X.P., Zhu, H.B., and Tan, Y.R. (2021). Identification of MATN3 as a novel prognostic biomarker for gastric cancer through comprehensive TCGA and GEO data mining. Dis. Markers *2021*, 1769635.

48. Hong, S.A., Yoo, S.H., Lee, H.H., Sun, D.S., Won, H.S., Kim, O., and Ko, Y.H. (2018). Prognostic value of Dickkopf-1 and ß-catenin expression in advanced gastric cancer. BMC Cancer *18*, 506.

49. Liu, Q., Jiang, J., Zhang, X., Zhang, M., and Fu, Y. (2021). Comprehensive analysis of IGFBPs as biomarkers in gastric cancer. Front. Oncol. *11*, 723131.

50. Kakiuchi, M., Nishizawa, T., Ueda, H., Gotoh, K., Tanaka, A., Hayashi, A., Yamamoto, S., Tatsuno, K., Katoh, H., Watanabe, Y., et al. (2014). Recurrent gain-of-function mutations of RHOA in diffuse-type gastric carcinoma. Nat. Genet. *46*, 583–587.

51. Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. Commun. ACM *2016*, 785–794.

52. Ma, B., Meng, F., Yan, G., Yan, H., Chai, B., and Song, F. (2020). Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput. Biol. Med. *121*, 103761.

53. Huang, Y., Mao, Y., Xu, L., Wen, J., and Chen, G. (2022). Exploring risk factors for cervical lymph node metastasis in papillary thyroid microcarcinoma: construction of a novel population-based predictive model. BMC Endocr. Disord. *22*, 269.

54. Lam, L.H.T., Do, D.T., Diep, D.T.N., Nguyet, D.L.N., Truong, Q.D., Tri, T.T., Thanh, H.N., and Le, N.Q.K. (2022). Molecular subtype classification of low-grade gliomas

using magnetic resonance imaging-based radiomics and machine learning. NMR Biomed. *35*, e4792.

55. Silva, G.F.S., Fagundes, T.P., Teixeira, B.C., and Chiavegatto Filho, A.D.P. (2022). Machine learning for hypertension prediction: a systematic review. Curr. Hypertens. Rep. *24*, 523–533.

56. Mayakonda, A., Lin, D.C., Assenov, Y., Plass, C., and Koeffler, H.P. (2018). Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome Res. *28*, 1747–1756.

57. Chevalier, A., Yang, S., Khurshid, Z., Sahelijo, N., Tong, T., Huggins, J.H., Yajima, M., and Campbell, J.D. (2021). The mutational signature comprehensive analysis Toolkit (musicatk) for the discovery, prediction, and exploration of mutational signatures. Cancer Res. *81*, 5813–5817.

58. Tate, J.G., Bamford, S., Jubb, H.C., Sondka, Z., Beare, D.M., Bindal, N., Boutselakis, H., Cole, C.G., Creatore, C., Dawson, E., et al. (2019). COSMIC: the Catalogue of somatic mutations in cancer. Nucleic Acids Res. *47*, D941–D947.

59. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinf. *14*, 7.

60. Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

61. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res. *49*, D605–D612.

62. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

63. Maeser, D., Gruener, R.F., and Huang, R.S. (2021). oncoPredict: an R package for predicting in vivo or cancer patient drug response and biomarkers from cell line screening data. Briefings Bioinf. *22*, bbab260.

64. Yang, W., Soares, J., Greninger, P., Edelman, E.J., Lightfoot, H., Forbes, S., Bindal, N., Beare, D., Smith, J.A., Thompson, I.R., et al. (2013). Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. *41*, D955–D961.