scientific reports

Check for updates

OPEN

Identification of the molecular subtypes and construction of risk models in neuroblastoma

Enyang He[1,3], Bowen Shi[1,3], Ziyu Liu[1,3], Kaili Chang[1,3], Hailan Zhao[1,4], Wei Zhao[1,4] & Hualei Cui[1,2]✉

The heterogeneity of neuroblastoma directly affects the prognosis of patients. Individualization of patient treatment to improve prognosis is a clinical challenge at this stage and the aim of this study is to characterize different patient populations. To achieve this, immune-related cell cycle genes, identified in the GSE45547 dataset using WGCNA, were used to classify cases from multiple datasets (GSE45547, GSE49710, GSE73517, GES120559, E-MTAB-8248, and TARGET) into subgroups by consensus clustering. ESTIMATES, CIBERSORT and ssGSEA were used to assess the immune status of the patients. And a 7-gene risk model was constructed based on differentially expressed genes between subtypes using randomForestSRC and LASSO. Enrichment analysis was used to demonstrate the biological characteristics between different groups. Key genes were screened using randomForest to construct neural network and validated. Finally, drug sensitivity was assessed in the GSCA and CellMiner databases. We classified the 1811 patients into two subtypes based on immune-related cell cycle genes. The two subtypes (Cluster1 and Cluster2) exhibited distinct clinical features, immune levels, chromosomal instability and prognosis. The same significant differences were demonstrated between the high-risk and low-risk groups. Through our analysis, we identified neuroblastoma subtypes with unique characteristics and established risk models which will improve our understanding of neuroblastoma heterogeneity.

Neuroblastoma, a tumor of sympathetic origin, is the most common extracranial solid tumor in early childhood. Neuroblastoma account for 7–8% of childhood malignancies with a heterogeneous clinical course from local or spontaneous regression to extensive metastatic disease[1]. The etiology of the disease is complex and diverse, with multiple signaling pathways involved. The mammalian target of rapamycin (mTOR) pathway promotes neuroblastoma cell survival and chemoresistance[2]. The WNT signaling pathway, on the other hand, increases MYC levels in patients without MYCN amplification[3]. Additionally, the ALK signaling pathway is the primary oncogene target pathway in sporadic and familial neuroblastoma cases[4].

As we all know, unrestricted proliferation is a common feature of malignant tumors and is closely related to cell cycle dysregulation[5]. The cell cycle is a complex process that contains four phases: Gap 1 (G1), DNA-synthesis (S), Gap 2 (G2) and mitosis (M). Cell cycle proteins and cell cycle protein-dependent kinases (CDK) regulate the progression of cell cycle phases[6]. At the same time, whether each cell cycle event is completed, correctly or not, is subject to a cellular checkpoint monitoring mechanism[7]. The DNA damage response and the Mitotic Spindle Checkpoint play a key role in maintaining the health of the organism. As known, the p53 tumor suppressor is involved in multiple cell cycle checkpoints[8]. And abnormalities in p53 can lead to cancer development and progression through multiple pathways[9].

Abnormalities in cell cycle-related mechanisms likewise play an important role in the onset and development of neuroblastoma. Increased MYCN copy number was detected in 25% of patients with neuroblastoma[10], which was strongly associated with an unfavorable clinical prognosis[11]. Meanwhile, MYCN can accelerate cell proliferation[12], which may be related to cell cycle protein-dependent kinase 4 (CDK4)[13]. For patients with neuroblastoma without MYCN amplification, it is more likely to exhibit chromosomal alterations and again leads to poor prognostic outcomes[14]. This may be related to the absence of a common region that codes a series of proteins that play a role in the DNA damage response (DDR)[15]. As the research becomes more in-depth, chromosome instability plays an important role in the development and progression of the disease[16]. Study found that

[1]Tianjin Medical University, Tianjin, China. [2]Tianjin Children's Hospital, Tianjin, China. [3]Graduate School of Tianjin Medical University, Tianjin, China. [4]Basic Medical Sciences School of Tianjin Medical University, Tianjin, China. ✉email: chlyfjp@163.com

unbalanced loss of heterozygosity (LOH) in chromosome 11q and LOH in chromosome 1p36 are independent risk factors for poor prognosis in patients with neuroblastoma[17]. 17q gain was also associated with poorer overall survival (OS)[14]. Chromosomal instability has also been observed during early human embryogenesis[18]. However, the underlying mechanism ensures that the cell cycle proceeds correctly. Therefore, understanding the mechanisms involved in the cell cycle is crucial to our understanding of neuroblastoma.

Various etiologies lead to the variability among individual patients. The heterogeneity of patients poses a great challenge for individualized treatment. In order to evaluate patients for stratification to guide treatment, classification methods based on multiple biological indicators have been proposed and applied. Stage, age, histologic category, grade of tumor differentiation, the status of the MYCN oncogene, chromosome 11q status, and DNA ploidy were used as the classification basis for the International Neuroblastoma Risk Group Staging System[19]. Segmental chromosomal aberrations (SCA) have also been studied as an additional genomic biomarker in combination with INSS staging to guide treatment[20]. Based on the concept of stratified treatment, the prognosis of neuroblastoma patients is gradually improving. Over the past few decades, the 5-year survival rate for patients with metastatic neuroblastoma has increased from less than 20% to over 50% through a combination of therapies including immunotherapy, stem cell therapy, etc[21]. Although these staging plays a role in assessing patients and guiding treatment, clinical use is somewhat limited. With the development of gene chip technology, how to stratify patients at the genetic level to guide targeted therapy is an urgent issue.

Given the role of cell cycle abnormalities in the pathogenesis of neuroblastoma, it is essential to understand the causes of Chromosomal instability (CIN) in neuroblastoma and to study the chromosome and centrosome segregation, spindle machinery and DNA repair[1]. This facilitates the exploration of individualized treatment of neuroblastoma patients with drugs that target the cell cycle. The aim of our study is to explore molecular subtyping in tumor patients by analyzing cell cycle gene expression levels to further refine individualized patient stratification management. Molecular subtyping and risk scores will be used to guide individualized patient treatment and thus improve patient prognosis.

Results

Identification of a set of 924 immune-related cell cycle genes. First, the t-SNE algorithm classified the 643 patients in GSE45547 into different regions based on gene expression levels, indicating heterogeneity among patients. This result suggested that the disease can be further subdivided into molecular subtypes (Fig. 1A). Consideration of the close correlation of disease with the cell cycle and immunity, to assess the level of infiltration of immune and stromal cells involved in the tumor microenvironment (TME) of GSE45547, the algorithm ESTIMATE was applied based on transcriptomic data from 643 samples. The results were also incorporated into the WGCNA algorithm as clinical information in the search for immune-related cell cycle genes (Supplementary Fig. S1A). Subsequently, the scale-free co-expression network was obtained by WGCNA of 1740 cell cycle gene expressions from 643 samples with immunization results (Fig. 1B). Two gene modules were generated with a power of 4 as the optimal soft threshold (Fig. 1C). Among these modules, the turquoise module exhibited the highest correlation with the result of ESTIMATE and was considered as "Immune-related cell cycle genes (IRCCGs) module". And there were 924 genes in the turquoise module (a detailed list of genes could be available in the Supplementary Material). We further explored the function of IRCCGs by enrichment analysis. KEGG enrichment results for IRCCGs showed links to both immune and cell cycle-related pathways (Fig. 1D). The results enriched in Biological Process showed that cell cycle regulation and nuclear division were involved (Fig. 1E). The gene products of IRCCGs play a role in the spindle and chromosomal region (Fig. 1F). For Molecular Function enrichment results showed that pathways such as tubulin binding and microtubule binding were involved (Fig. 1G).

To make the results of the study more objective and generalizable, GSE45547, GSE49710, GSE73517, GSE120559, E-MTAB-8248 and GDC TARGET-NBL data were integrated for analysis. In total, 16,978 genes from 1811 patients were jointly detected. Before the removal of batch effect, the result of principal component analysis (PCA) showed that the samples were clustered by batches (Supplementary Fig. S1B). On the contrast, the results after data processing show that cross-platform normalization has been successful in eliminating batch effects (Fig. 1H). In the normalized data, the intersection of 16,978 genes with Immune-related cell cycle genes was 913 genes. In total, 11 genes from 924 IRCCGs were not included in the follow-up analysis. This was due to the fact that different microarrays have different probes and the common genes were selected for the combined analysis. Considering that the microarray data were all from the same platform, the five microarray datasets were normalized and included in the subsequent study as a whole (Supplementary Fig. S1C, Fig. 1I).

Two distinct cell cycle subtypes were identified with IRCCGs. Based on the expression matrix after removing batch effect of 913 IRCCGs, the all 6 datasets (n = 1811) were divided into two distinct cell cycle clusters by consensus clustering (Fig. 2A), an unsupervised clustering method with the k value of 2 (Supplementary Fig. S2A–C). There were 871 patients in Cluster I and 940 patients in Cluster II. Moreover, the cluster consensus score for each subgroup was higher than 0.8 only in two-subgroup classification (Fig. 2B), which suggested that the classification with two subgroups was more robust than others.

To understand the differences between the clusters, the clinical information in the dataset was used to explore the characteristics of each of the two clusters. The heat map shows the clustering in relation to age, International Neuroblastoma Staging System (INSS) stages and MYCN status, along with the expression of the genes used for clustering in 1811 patients (Fig. 2C). The genes shown in the heat map were the top 50 genes with the largest Median absolute deviation of gene expression. Further statistical analysis of the clinical information of the two clusters revealed that the age of Cluster 2 was smaller than that of Cluster 1 ($P < 0.05$). The detailed statistical results were shown in Table 1 below. Meanwhile, the status of MYCN of patients in Cluster 1 was mainly

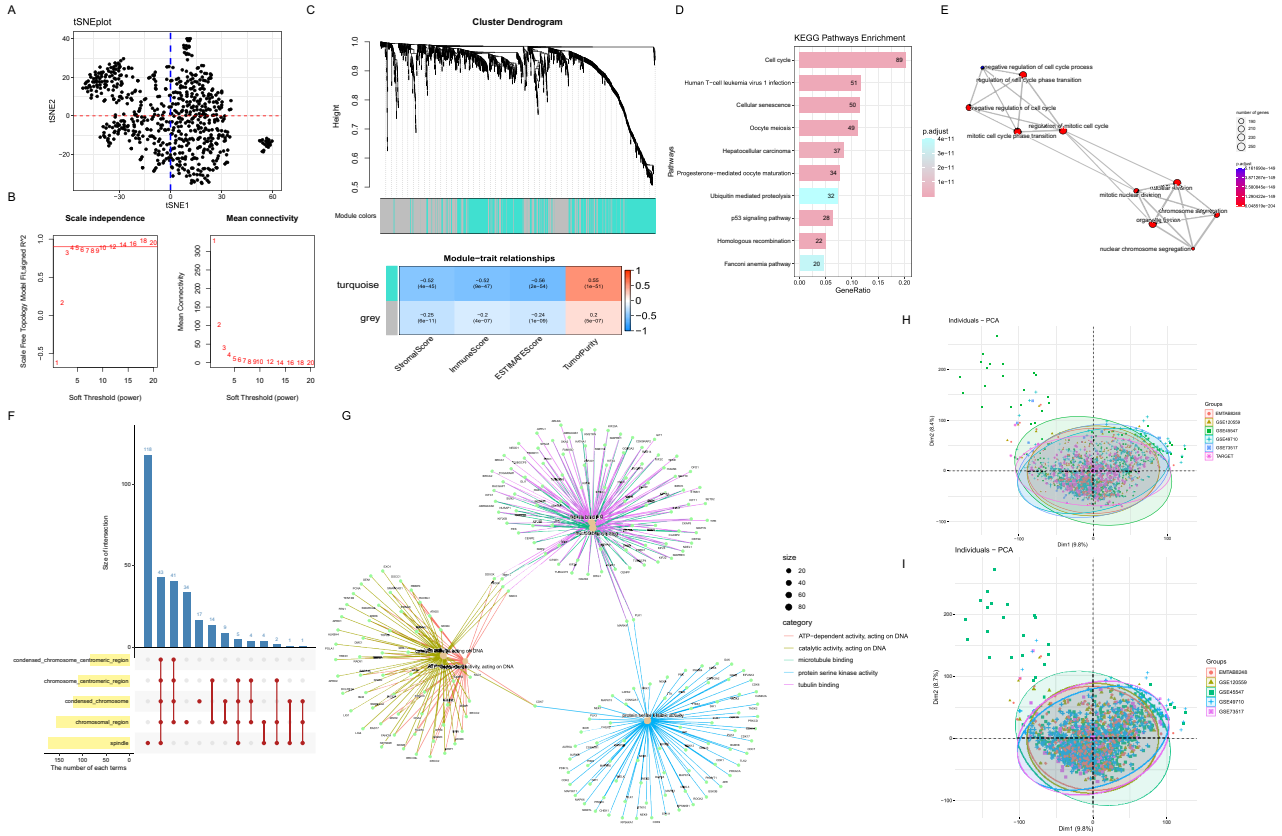

Figure 1. Identification and functional analysis of IRCCGs. (**A**) The results of the t-SNE algorithm show heterogeneity in the patients. (**B**) Analysis of network topology for various soft-thresholding powers. The left panel shows the scale-free fit index as a function of the soft-thresholding power. The right panel displays the mean connectivity as a function of the soft-thresholding power. Based on the scale-free fit index greater than 0.9, we chose 4 as the soft thresholding power. (**C**) At the top is Clustering dendrogram of genes with assigned module colors. At the bottom is Module-trait associations. Each cell contains the corresponding correlation and P value. The darker the color of the cell, the higher the correlation. (**D**) Results of KEGG enrichment. The numbers in the graph indicated the counts of the pathway. (**E**) Results of Biological Process enrichment. The line between dots indicated the presence of identical genes between pathways. (**F**) The top 5 pathways of Cellular Component enrichment was demonstrated. The length of the yellow bar indicated the number of pathway genes. The height of the blue bar indicated the number of intersecting genes. (**G**) The first 5 enriched to Molecular Function terms and the genes in the terms. (**H**,**I**) Principal component analysis (PCA) of the gene expression in datasets. The visualization of patients by scatter plots were based on the top two Dims of gene expression profiles with the removal of batch effect.

amplified, while the status of MYCN of patients in Cluster 2 was mainly non amplified ($P < 0.05$). INSS stage, which is closely related to prognosis, was also significantly different in Cluster 1 and Cluster 2 ($P < 0.05$). Alternative lengthening of telomeres (ALT) is regulated by break-induced replication. A Sankey diagram depicts the flow from the two cell cycle clusters to different status of telomerase reverse transcriptase (TERT) and ALT-associated promyelocytic leukemia bodies (APBs) in E-MTAB-8248 and GSE120559 datasets, in which the width of the flow rate is proportional to the number of patients (Fig. 2D). For status of telomerase reverse transcriptase, TERT rearrangements were more predominant in Cluster 1 ($P < 0.05$), while whether ALT-associated promyelocytic leukemia bodies were detected or not did not differ in the two clusters directly ($P > 0.05$). The bar chart showed the three chromosomal abnormalities closely associated with prognosis in the GSE73517 dataset, they were 1p deletion, 11q deletion, and 17q gain (Fig. 2E). As shown inside the statistical Table1, the respective proportion of 1p deletion and 17q gain to the total number of clusters differed in the two clusters ($P < 0.05$). However, the quantities of 11q deletion did not differ between the two clusters ($P > 0.05$). Using survival data from E-MTAB-8248 and GDC TARGET-NBL, the differences in prognosis between the two clusters were compared. The results showed that the prognostic status of Cluster 2 was better than that of Cluster 1, which was consistent with the distribution of clinical prognostic indicators between the two groups (Fig. 2F).

Characterization of immunity in two clusters. The immune microenvironment is closely related to tumors and the expression of immune checkpoints is a reflection of the immune response. Among the five microarray datasets integrated, 24 immune checkpoints were selected for comparison between clusters. As the results demonstrate, except for LAG3, CD276and CD86, the levels of immune checkpoints were higher in Clus-

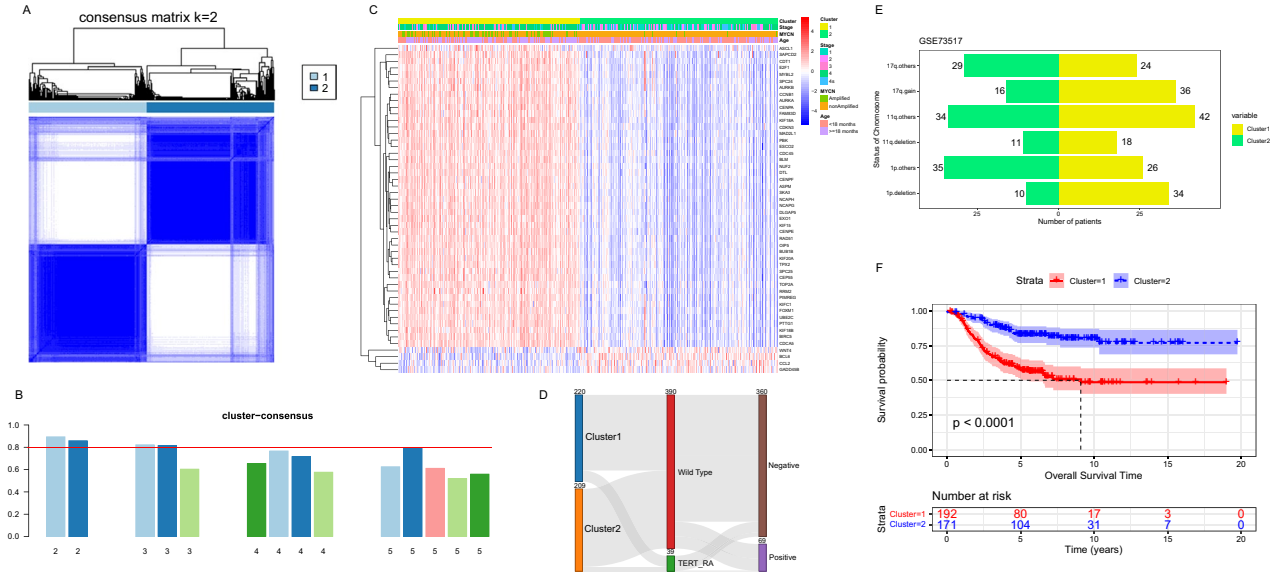

Figure 2. Identification of subtypes and clinical correlations of subtypes. (**A**) Consensus matrix heatmap with cluster count of 2. (**B**) The bar-plots represent the consensus scores for subgroups and we chose the results with consensus scores greater than 0.8. (**C**) Heatmap of Top 50 Immune-related cell cycle genes levels and distribution of age, MYCN status, and INSS stage in the two clusters. (**D**) The Sankey diagram showing whether TERT was rearrangement and whether APB existed. (**E**) The bar chart showed the distribution of chromosomal abnormalities in the two clusters. (**F**) The Kaplan–Meier curves showed the OS time of the two clusters of patients inside the E-MTAB-8248 and the TARGET datasets.

Data source	Clinical information	Cluster1	Cluster 2	*P* value
ALL6 datasets	Age			$P<0.001$
	< 18 months	381	619	
	≥ 18 months	490	321	
ALL6 datasets	MYCN status			$P<0.001$
	Amplified	317	29	
	Non amplified	554	911	
ALL6 datasets	INSS stage			$P<0.001$
	1	75	264	
	2	70	203	
	3	120	111	
	4	526	215	
	4s	80	147	
E-MTAB-8248 + GSE120559	TERT status			0.007
	Wild type	192	198	
	TERT rearrangement	28	11	
E-MTAB-8248 + GSE120559	APBs status			0.492
	Negative	182	178	
	Positive	38	31	
GSE73517	Chromosomes 1			$P<0.001$
	1p deletion	34	10	
	Others	26	35	
GSE73517	Chromosomes 11			0.529
	11q deletion	18	11	
	Others	42	34	
GSE73517	Chromosomes 17			0.013
	17q gain	36	16	
	Others	24	29	

Table 1. Comparison of clinical characteristics between the two clusters.

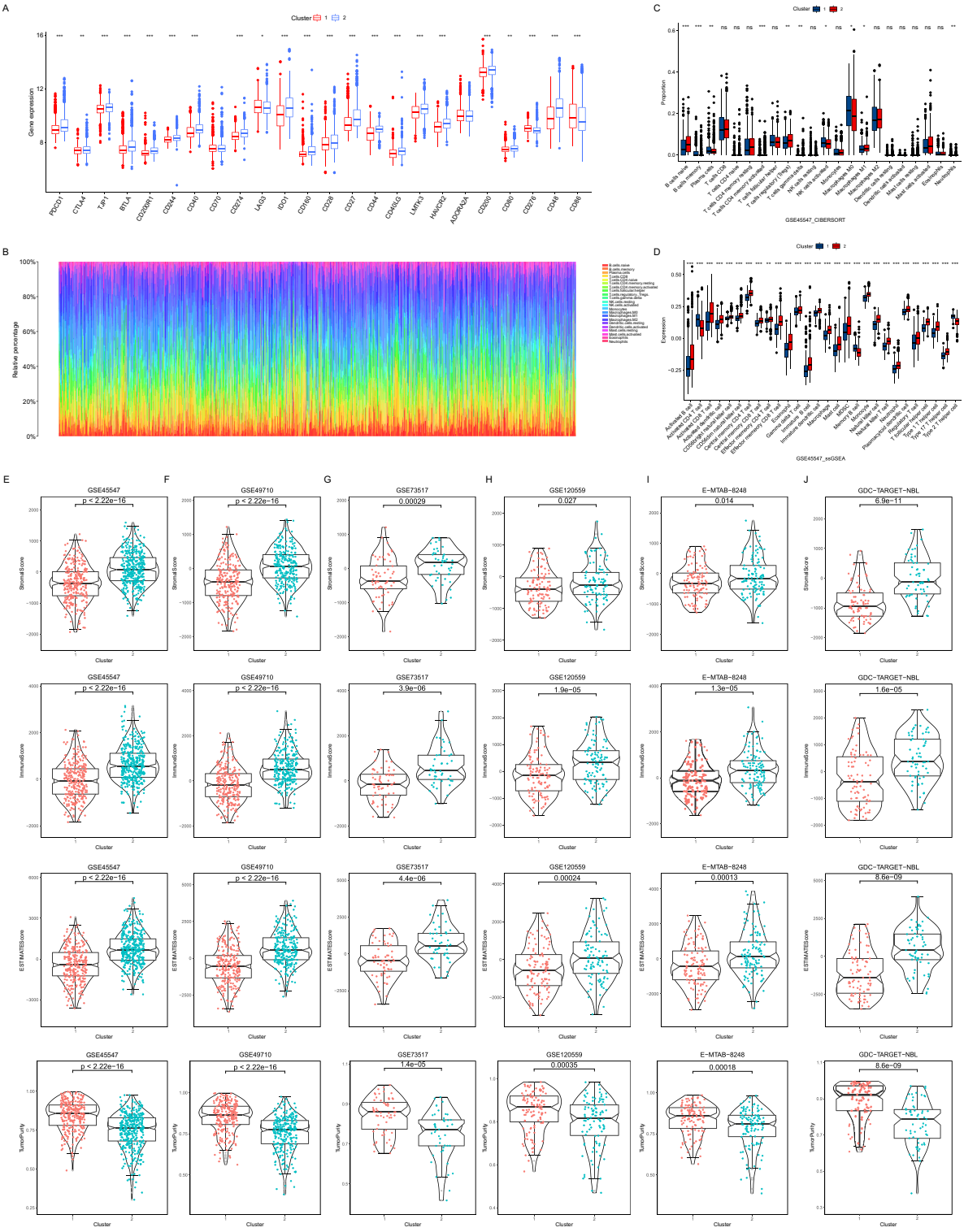

Figure 3. Comparison of immunization of two clustered subtypes. (**A**) Box plots showed the mRNA expression of immune checkpoints in two clusters (*$P<0.05$; **$P<0.01$; ***$P<0.001$). (**B**) Stacked bar chart showed the percentage of immune cells in 1811 patients. (**C**) Box plots were used to display the distribution of the cell proportions calculated by the CIBERSORT algorithm of immune cells between the two clusters (*$P<0.05$; **$P<0.01$; ***$P<0.001$). (**D**) Box plot of the distribution of immune cell expression between the two clusters as calculated by the ssGSEA algorithm (*$P<0.05$; **$P<0.01$; ***$P<0.001$). (**E**–**J**) Box plots were created to visualize the distribution of the Stromal Score, Immune Score, ESTIMATES Score, and Tumor Purity, which were calculated by the ESTIMATE algorithm between the two clusters in the GSE45547 (**E**), GSE49710 (**F**), GSE73517 (**G**), GSE120559 (**H**), E-MTAB-828 (**I**) and TARGET (**J**) datasets.

ter 2 than in Cluster 1 (Fig. 3A). Based on which CIBERSORT were used to estimate the immune infiltration and a bar chart was used to show the percentage of immune cells in each patient (Fig. 3B). To compare the variability of immunization between clusters in GSE45547, an analysis was conducted to compare the differences between the two clusters of immune cells according to the clustering grouping. The results indicate a significant variability in the immune cells of the two clusters (Fig. 3C). Further quantification of immune cells using ssGSEA shows that Cluster 2 has more immune cells overall than Cluster 1 (Fig. 3D). In the other five datasets, again using the CIBERSORT results and the ssGSEA results compared between the two clusters, Cluster 2 all showed more immune infiltration (Supplementary Fig. S3A–E).

The immune status of the patients was further assessed inside the six datasets using the ESTIMATE algorithm. The analysis results surface higher Tumor Purity in Cluster 1 than in Cluster 2 in the GSE45547 dataset ($P < 0.05$). Relatively, Stromal Score, Immune Score, and ESTIMATE Score in Cluster 1 were lower than in Cluster 2 ($P < 0.05$) (Fig. 3E). The same analysis was validated for the other five datasets (Fig. 3F–J). Combining the results of the previous analysis, we believe that Cluster 2 belongs to the class of rich immune status and Cluster 1 is the class of poor immune status.

Identification of subgroup DEGs and functional annotation. In order to investigate the key genes causing the differences between clusters in depth, a total of 4945 differential genes were obtained using "DESeq" package in the TARGET data between the two clusters, of which 1022 were highly expressed genes in Cluster 1 relative to Cluster 2 and 3923 were lowly expressed genes (Fig. 4A) (Supplementary Fig. S4A). The variance

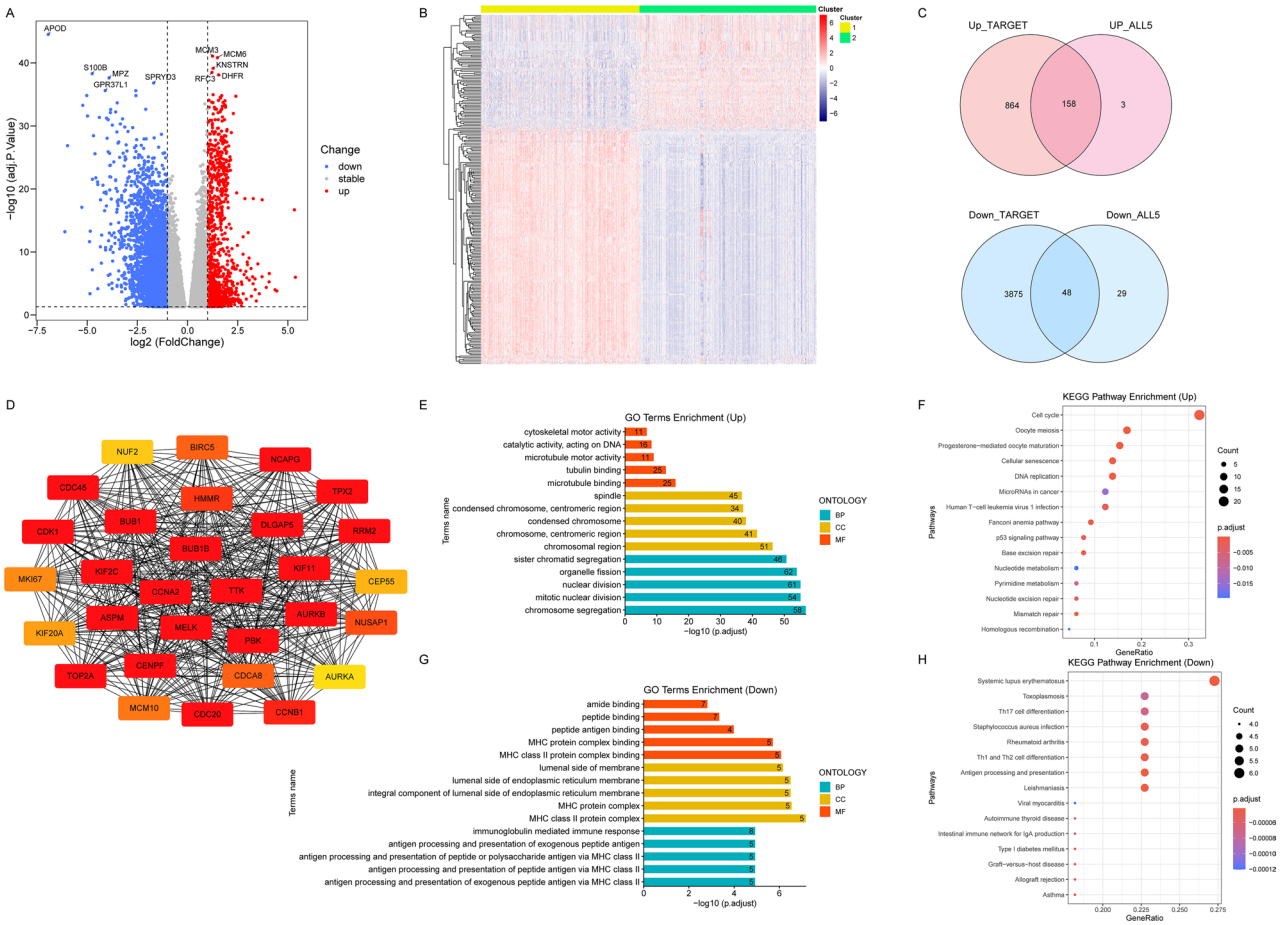

Figure 4. Identification DEGs and functional annotation of DEGs. (**A**) Volcano plot depicted the distribution of DEGs in TARGET dataset (Cluster1 VS Cluster2) and labeled the top 5 genes with the smallest ranking according to adjusted *P* value. (Genes with adjusted *P* value > 0.05 were not shown in the plot). (**B**) Heatmap of the DEGs derived from the 5 microarray datasets. (**C**) The Ven diagram showed the number of intersecting genes in the results of the difference analysis. (**D**) The TOP 30 genes based on the MCC algorithm, with the darker colors, indicating the higher MCC scores. (**E**,**F**) Bar graph (**E**) showed the results of GO enrichment and Bubble plots (**F**) showed KEGG enrichment results for Cluster 1 relative to Cluster 2 highly expressed genes. The numbers in the Bar graph represented the counts in the pathway. (**G**,**H**) Bar graph (**G**) showed the results of GO enrichment and Bubble plots (**H**) showed KEGG enrichment results for Cluster 1 relative to Cluster 2 low expressed genes. The numbers in the Bar graph represented the counts in the pathway. In the GO enrichment results (**E**,**G**), BP refers to Biological Process, CC denotes Cellular Component, and MF represents Molecular Function.

analysis of the five normalized datasets using the "limma" package yielded 238 variance genes. There were 161 up-regulated genes and 77 down-regulated genes (Cluster1 VS Cluster2) in the result (Fig. 4B). A total of 206 intersecting genes from the two difference analyses were designated as intergroup difference genes for Clusters 1 and 2 (Fig. 4C). We then constructed a Protein–Protein Interaction (PPI) network using the STRING database (Supplementary Fig. S4B). The TOP 30 genes based on the MCC algorithm were further demonstrated using the the cytoHubba plug-in in Cytoscape (Fig. 4D).

To gain insight into the function of the differential genes, enrichment analysis was performed. For the highly expressed genes (Cluster1 VS Cluster2) between the two clusters, GO enrichment results showed that these genes were closely associated with the cell cycle progression (Fig. 4E). The TOP 5 terms in Biological Process were chromosome segregation, mitotic nuclear division, nuclear division, organelle fission and sister chromatid segregation. The results of Cellular Component were mainly involved in chromosomal region; chromosome, centromeric region; condensed chromosome; condensed chromosome, centromeric region and spindle. Microtubule binding; tubulin binding; microtubule motor activity; catalytic activity (acting on DNA) and cytoskeletal motor activity were the TOP 5 terms in Molecular Function. KEGG analysis suggested that highly expressed genes (Cluster1 VS Cluster2) were mainly associated with Cell cycle, DNA replication, Oocyte meiosis and other pathways closely related to the cell cycle (Fig. 4F).

The enrichment results for low expressed genes showed an association with immunity. GO enrichment results mainly include antigen processing and presentation of exogenous peptide antigen via MHC class II and MHC class II protein complex binding (Fig. 4G). Similarly, the results of KEGG enrichment were closely related to immunity (Fig. 4H). Generally, the enrichment analyses showed that DEGs not only play an important role in the division of chromosome but are also associated with repair of DNA and immunity. At the same time, the enrichment results of each of the two group DEGs corresponded to the results of the previous clinical and immunological analyses.

Identification of prognostic key genes and establishment risk score model. Based on the enrichment results, which imply that DEGs were strongly associated with chromosomal instability and disease heterogeneity in patients, we decided to search for key genes from within DEGs to construct a risk model. Firstly, we preliminarily screened out 177 OS-related genes with a filtering threshold of P value less than 0.01 by univariate Cox regression analysis in E-MTAB-8248 dataset (Supplementary Table 1) and displayed their top 10 significant genes by forest map (Fig. 5A). In the next step, "randomForestSRC" package were used to filter the key variables. As shown in the Fig. 5B, the oob error rate tends to stabilize when tree > 200, while the importance of the variables was judged using Variable Importance (VIMP) algorithm and the longer blue bars indicate the more important variables (Fig. 5B). We selected the TOP 50 most important genes based on the VIMP for inclusion in the LASSO Cox regression model (Supplementary Fig. S5A). With an optimal λ value (Fig. 5C,D), 7 genes (NMU, E2F3, UBE2S, DHFR, MIA, CHD5, and FAXDC2) retained their individual Cox coefficients after LASSO regularization (Supplementary Table 2). Using the established formula, the risk score was calculated for each sample (Fig. 5E). With a best cut-off value (Supplementary Fig. S5B), the dataset was divided into low-risk and high-risk groups (Fig. 5F). Kaplan–Meier analysis demonstrated that patients with higher risk score exhibited worse progression-free survival (PFS) and OS in the E-MTAB-8248 dataset (Fig. 5G,H).

Validation of the risk score model. First, the receiver operating characteristic (ROC) curves of clinical indicators related to prognosis were compared inside the E-MTAB-8248 dataset, and the risk scores were all better than these indicators (Fig. 6A). In addition, ROC curve analysis indicated that the area under the curve (AUC) values of OS signature in 1-, 3-, and 5-year were 0.9527, 0.87266, and 0.8792, indicating that our prognosis signatures have favorable discrimination (Fig. 6B). Meanwhile, in the GDC TARGT-NBL dataset, risk scores were also strongly correlated with OS (Fig. 6C). In addition, we analyzed and mapped the expression profiles of seven genes in different risk subgroups of 1670 patients, and significant expression differences could be seen (Fig. 6D). We further compared the distribution of the seven gene expressions in a variety of tumors. UBE2S was found to be highly expressed in most tumor tissues, while MIA was mainly concentrated in tumor tissues of melanoma (SKCM) (Supplementary Fig. S6A). The mutations of the seven genes were further explored in a variety of tumors, and it could be found that the highest mutation rate was CHD5, followed by E2F3. And the types of mutations were mainly concentrated in Amplification, Deep Deletion and Missense Mutation (Fig. 6E).

The results of the previous ESTIMATE algorithm were used to compare groups based on high and low risk. The analysis results surface higher Tumor Purity in high-risk group than in low-risk group in the GSE45547 dataset ($P < 0.05$). Relatively, Stromal Score, Immune Score and ESTIMATE Score in high-risk group were lower than in low-risk group ($P < 0.05$) (Fig. 6F). In GSE49710, GSE73517, GSE120559 and E-MTAB-8248, Tumor Purity, Immune Score and ESTIMATE Score had the same variation in the high-risk versus low-risk groups (Supplementary Fig. S6B–E). We further compared the expression of immune checkpoints between high and low risk groups (Fig. 6G). Combined with the results of ssGSEA it can be concluded that the low-risk group had a better immune status (Fig. 6H).

Compare differences between high and low risk groups. The distribution of risk scores inside MYCN status, age and INSS stages was further explored in 1670 patients from all 5 microarray datasets. The results revealed that patients with MYCN amplification status, age ≥ 18 months and progressive worsening of INSS staging all had higher risk scores (Fig. 7A). Patients in E-MTAB-8248, GSE73517 and GSE120559 were divided into two groups, high-risk and low-risk, based on the optimal cut-off values. As shown in the Table 2, TERT rearrangements were more common in the high-risk group ($P < 0.05$). However, the positive of ALT-associated promyelocytic leukemia bodies was not statistically different between the two groups ($P > 0.05$). After

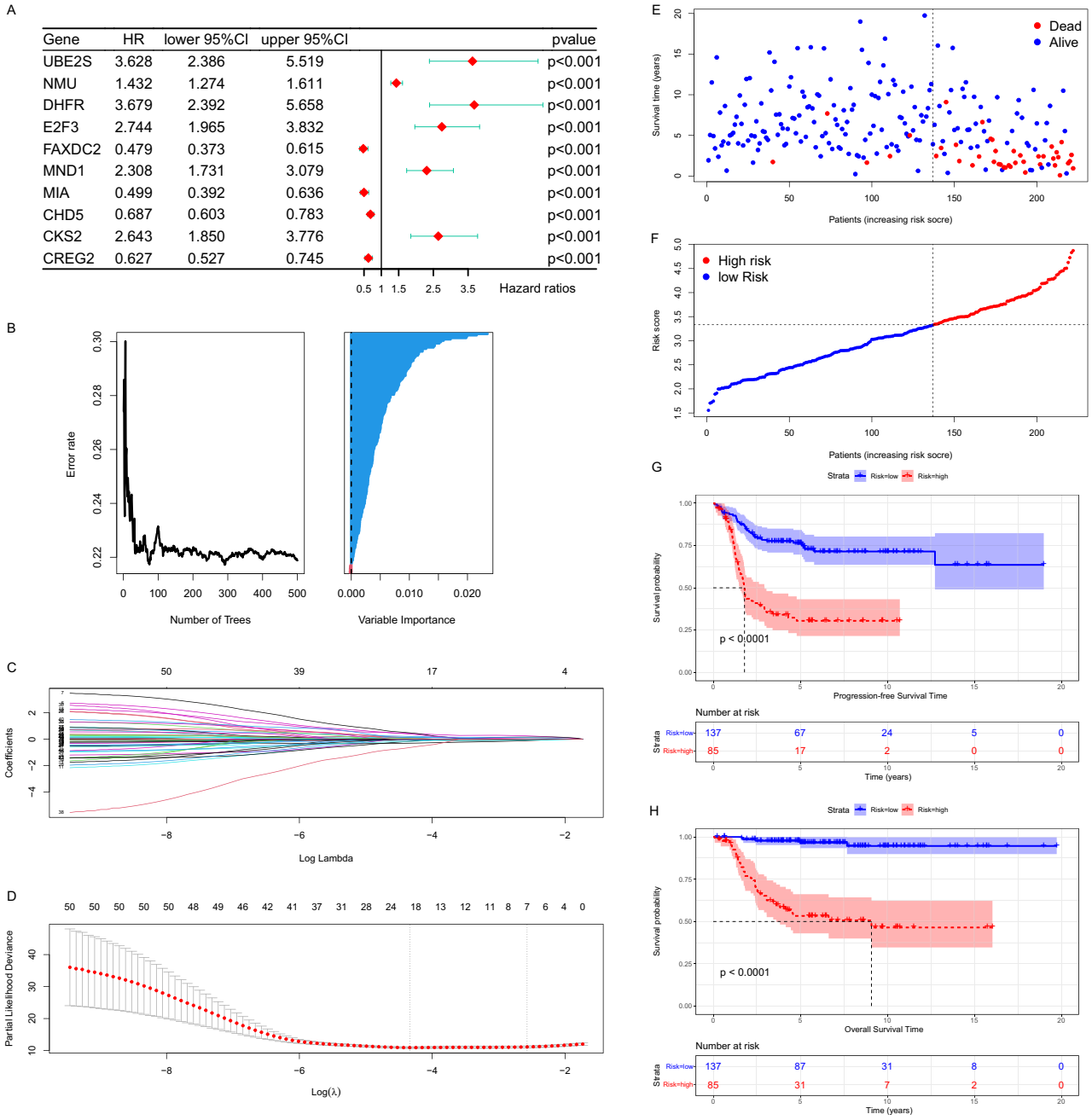

Figure 5. Construction of the risk model. (**A**) The forest plot showed the HR and 95% confidence interval of the most significant TOP 10 genes in the univariate regression results, sorted by *P* value. (**B**) The left graph showed the variation of Error rate with the number of trees. The right graph showed the ranking of genes according to the importance of the VIMP algorithm, where blue represents favorable to the correct judgment of the endings and red represents unfavorable. (**C**) Each line in the above graph represented a gene, the vertical coordinate was the value of the coefficient, the lower horizontal coordinate was log(λ), and the upper horizontal coordinate was the number of non-zero coefficients in the model at this time. (**D**) Based on cross-validation, for each value of λ, around the mean value of the target covariate shown in red, we can obtain a confidence interval for the target covariate. The two dashed lines indicate each of the two particular λ values. We chose lambda.1se as the final model parameter. (**E**) Each point in the scatter plot represented the survival status and survival time of a patient. The horizontal coordinates were the patients ranked from lowest to highest according to their risk scores. (**F**) Based on the risk score of each point in the scatter plot representing one patient, we divided them into high-risk and low-risk groups. (**G**,**H**) The Kaplan–Meier curves showed the progression-free survival time (**G**) and OS time (**H**) of the two risk groups of patients inside the E-MTAB-8248 dataset.

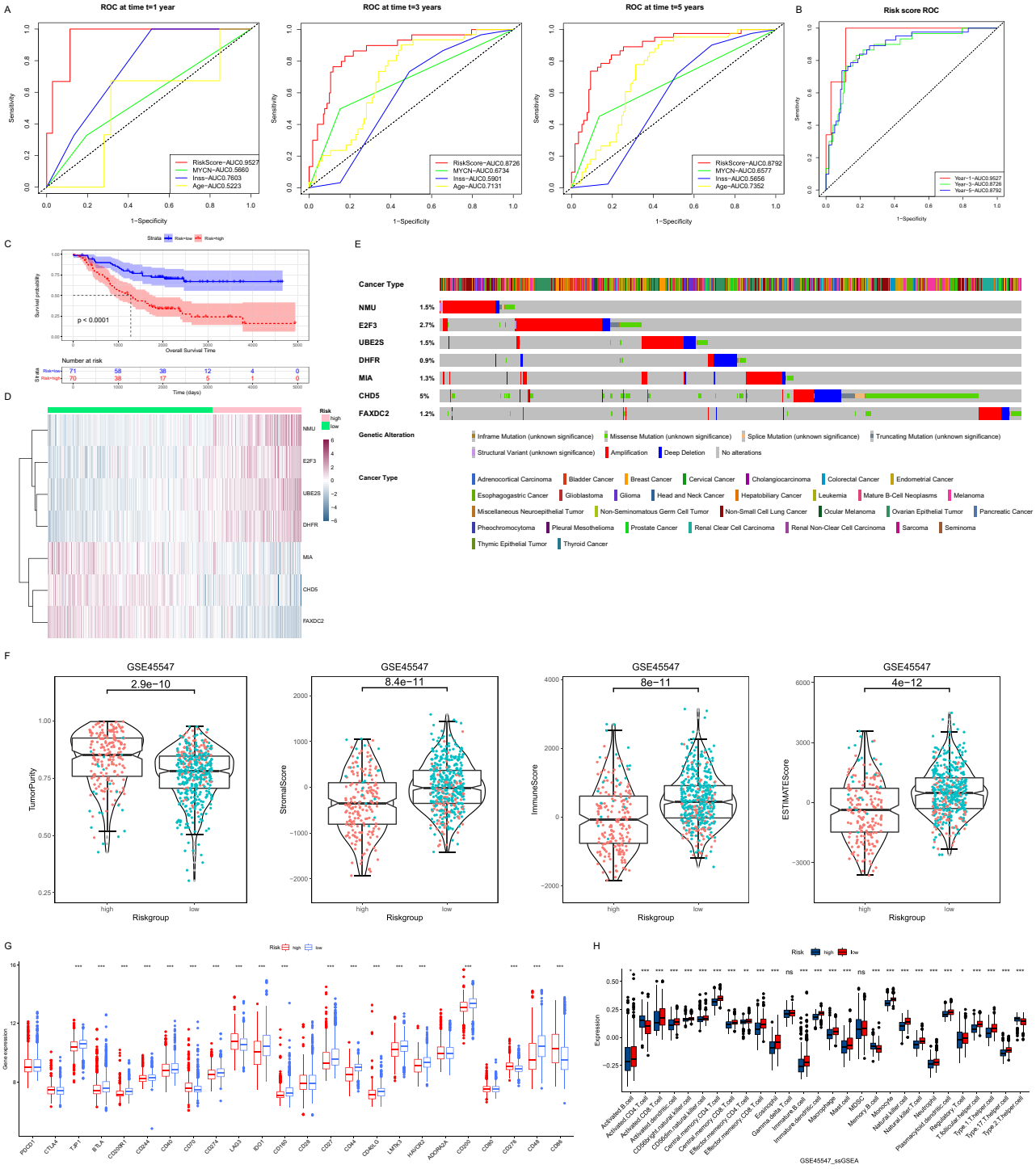

Figure 6. Validation and investigation of risk models. (**A**) The ability of clinical indicators and risk scores to determine prognosis at year 1, year 3, and year 5 were compared using ROC curves. (**B**) The ROC curve demonstrates the ability of the risk score to determine prognosis at year 1, year 3, and year 5. (**C**) The Kaplan–Meier curves showed the OS time of the two risk groups of patients inside the TARGET dataset. (**D**) The heat map demonstrated the expression levels of seven risk model genes in patients. (**E**) Mutations of 7 risk model genes in multiple tumors. (**F**) Comparison of differences in ESTIMATES results between high and low risk groups. Red dots indicated that patients belong to Cluster 1 and green dots indicated that patients belong to Cluster 2. (**G**) Box plots showed the mRNA expression of immune checkpoints in two risk groups (*$P<0.05$; **$P<0.01$; ***$P<0.001$). (**H**) Box plot of the distribution of immune cell expression between the two risk groups as calculated by the ssGSEA algorithm (*$P<0.05$; **$P<0.01$; ***$P<0.001$).

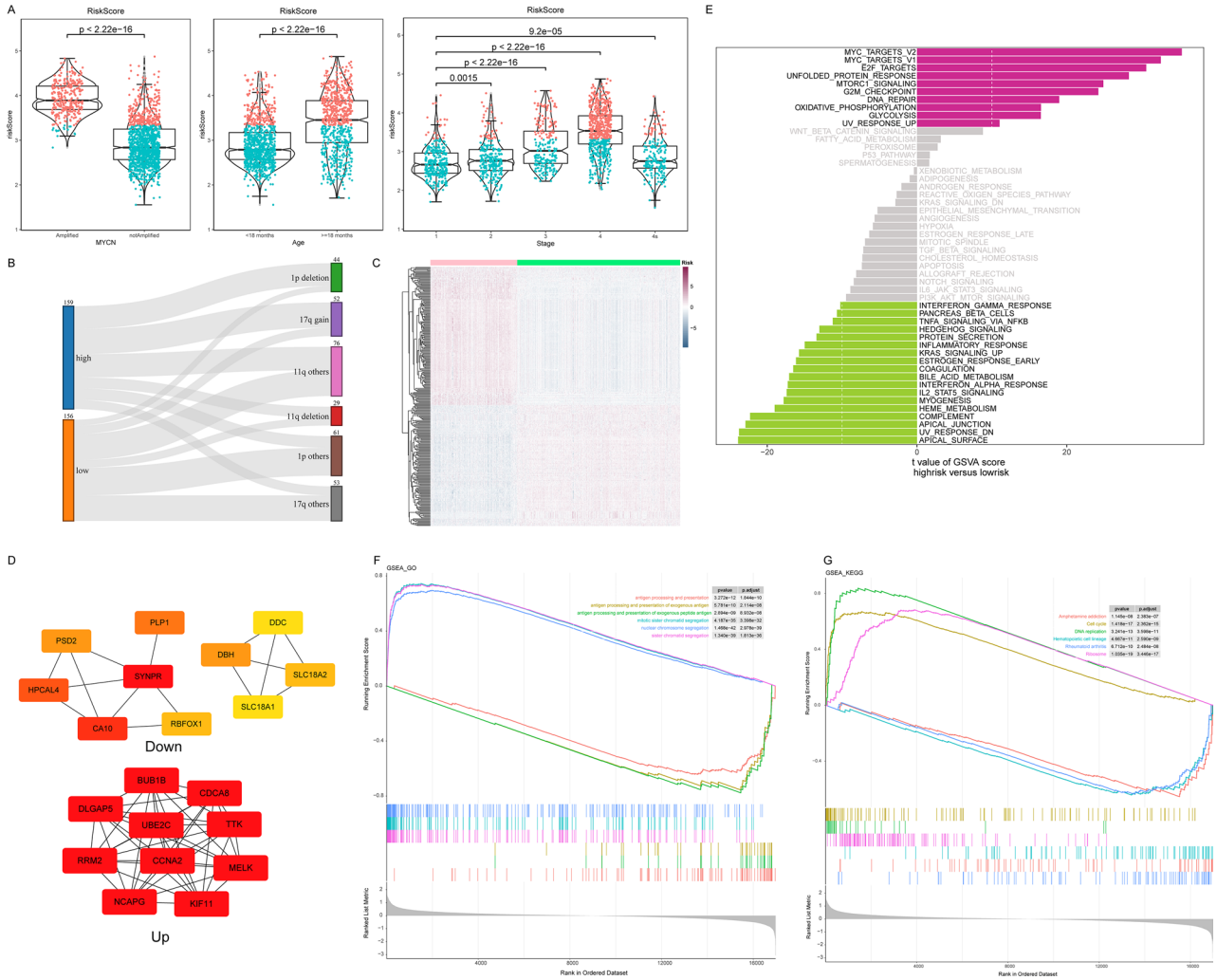

Figure 7. Comparison between high and low risk groups. (**A**) Comparison of risk scores in MYCN status, age groups and INSS stages. (**B**) The Sankey diagram showed the distribution of chromosomal abnormalities in the two risk groups. (**C**) The heat map showed the levels of differential genes between the high and low risk groups. (**D**) TOP 10 hub genes identified by MCC algorithm. (**E**) The bar graph showed the results of GSVA enrichment. Purple represented the major pathways enriched to in the high-risk group and green represented the major pathways in the low-risk group. (**F**) The TOP 3 most significant GO enriched terms in the high-risk and low-risk groups. (**G**) The TOP 3 most significant KEGG enriched pathways in the high-risk and low-risk groups.

that, the relationship between risk grouping and chromosomal instability was further explored. The results of the analysis confirmed that 1p deletion and 17q gain differed in the high- and low-risk subgroups and that the high-risk group was more likely to have these aberrations ($P < 0.05$). In contrast, 11q deletion was not statistically different between the two groups ($P < 0.05$) (Fig. 7B).

We further delved into the closely related mechanisms of clinical and risk grouping through analysis of variance. Patients with 1670 microarray data were further divided into high and low risk groups based on the risk model for difference analysis. A total of 314 differential genes were obtained, including 146 down-regulated genes and 168 up-regulated genes (high-risk VS low-risk). As shown in the heatmap, the difference genes were able to distinguish well between high and low risk groups (Fig. 7C). The volcano plot showed the top five genes in the differential genes ranked according to their adjusted *P* value (Supplementary Fig. S7A). We then constructed a PPI network using the STRING database (Supplementary Fig. S7B,C). The TOP 10 genes based on the MCC algorithm were further demonstrated in down-regulated genes and up-regulated genes (Fig. 7D).

Enrichment analysis of differential genes in the Hallmark database using the GSVA algorithm revealed significant differences between the two groups in numerous pathways (Fig. 7E). In the high-risk group, the significant pathways were MYC Targets_V2, MYC Targets_V1, E2F Targets, Unfolded protein response, Mtorc1 signaling, G2M chickpoint and DNA repair. In the low-risk group, Apical surface, UV response_DN, Apical junction, Complement, HEME metabolism and Myogenesis were the significant pathways. Further analysis using GSEA enrichment method, we could find that the pathways of GO in the high-risk group were mitotic sister chromatid segregation, nuclear chromosome segregation and sister chromatid segregation. The TOP 3 terms of low-risk group were antigen processing and presentation, antigen processing and presentation of exogenous antigen and

Data source	Clinical Information	High-risk	Low-risk	*P* value
E-MTAB-8248 + GSE120559	TERT status			$P < 0.001$
	Wild type	130	260	
	TERT rearrangement	30	9	
E-MTAB-8248 + GSE120559	APBs status			0.310
	Negative	138	222	
	Positive	22	47	
GSE73517	Chromosomes 1			$P < 0.001$
	1p deletion	35	9	
	Others	18	43	
GSE73517	Chromosomes 11			0.302
	11q deletion	17	12	
	Others	36	40	
GSE73517	Chromosomes 17			$P < 0.001$
	17q gain	39	13	
	Others	14	39	

Table 2. Comparison of clinical characteristics between the two risk groups.

antigen processing and presentation of exogenous peptide antigen (Fig. 7F). The KEGG results, on the other hand, showed that the high-risk group was mainly closely associated with the three pathways of the Cell cycle, DNA replication and Ribosome (Fig. 7G). The pathways in the low-risk group were focused on immune-related pathways such as Amphetamine addiction, Hematopoietic cell lineage and Rheumatoid arthritis. Based on the enrichment results, the worse prognosis in the high-risk group may be related to this.

Construction of neural network and integrated prognostic models to guide treatment. The random forest algorithm was employed to select the neural network genes. We used both the Mean Decrease Accuracy (MDA) and the Mean Decrease Gini (MDG) to obtain the top 50 most important genes, and took the intersection of the two as the final key genes. Through the graph of rate, error versus number of trees, we chose mtry = 6, ntree = 1200 as the final parameter of the model (Supplementary Fig. 8A,B). In our final fitted model, the out-of-bag (OOB) value was 2.95%. As shown in Fig. 8A, 37 genes were finally identified for the construction of neural network models for neuroblastoma patients. By experiment, the number of hidden layers was 1, with a total of 20 hidden neurons, and learningrate = 0.1 as the final setting of the model (Fig. 8B). Meanwhile, the Activation Function we chose was "tanh". We completed the training using 643 patients from the GSE45547 dataset and performed external validation in 493 patients from GSE49710 with good results (AUC = 0.966) (Fig. 8C).

We further explored whether these two indicators are related to survival. The clustering grouping and risk score were included in the univariate Cox regression analysis, which revealed that both the Cluster 1, and the higher risk score were risk factors affecting prognosis (Table 3). Clinical indicators such as age, whether MYCN was amplified, and INSS staging were further included for multifactorial regression analysis and the results revealed that only the risk score and age were an independent risk factor for prognosis (Supplementary Fig. 8C). To facilitate the assessment of prognosis, nomograms were constructed by age and risk score. The probability of survival at 1, 3, and 5 years were predicted by calculating the number of points (Fig. 8D).

We further evaluated the significance of clustering and risk score to guide treatment. Immunotherapy as a treatment modality with great potential, we compared the distribution of potentially used targets in immunotherapy between the two clustered subgroups. We could find different expression levels of immunotherapy targets in Cluster 1 and Cluster 2, suggesting that different clusters using different immunotherapy may be more prognostic (Fig. 8E). Cell cycle targeted therapy is also an important modality of treatment. It was interesting to note that the cell cycle checkpoints were significantly different between the two groups (Fig. 8F,G). The expressions of CDK2, CDK4 and CDK6 were higher in Cluster 1 than in Cluster 2, while the expressions of ATM were higher in Cluster 2 than in Cluster 1. We also analyzed the correlation with the IC50 of oncology drug in the GDSC database by GSCA using the genes that used for neural network (Fig. 8H). Based on the results of the analysis, we believe that Cluster 2 may be more appropriate for the four drugs RDEA119, Selumetinib, Trametinib and PD-0325901. Data from the CellMiner database of NCI-60 cell lines were downloaded and sensitivity analyses were performed between risk scores and drugs that had undergone Clinical trials and FDA approved. We set $P < 0.01$ as our filtering index and showed the results with a heatmap (Fig. 8I). Based on the results of the analysis, it could be found that most of the drugs were negatively correlated with risk scores.

Discussion

As a highly heterogeneous solid tumor, individualized treatment of neuroblastoma to improve its prognosis is a problem at this stage. Currently, neuroblastoma is mainly based on the INRG risk stratification system to guide the treatment of different patients[22]. Despite this, the 5-year EFS for children with metastatic neuroblastoma and aged 18 months or older is only close to 50%[23]. As cell cycle-targeted inhibitors are being studied and cell cycle-related mechanisms are gaining ground in neuroblastoma patients, the use of cell cycle-related genes

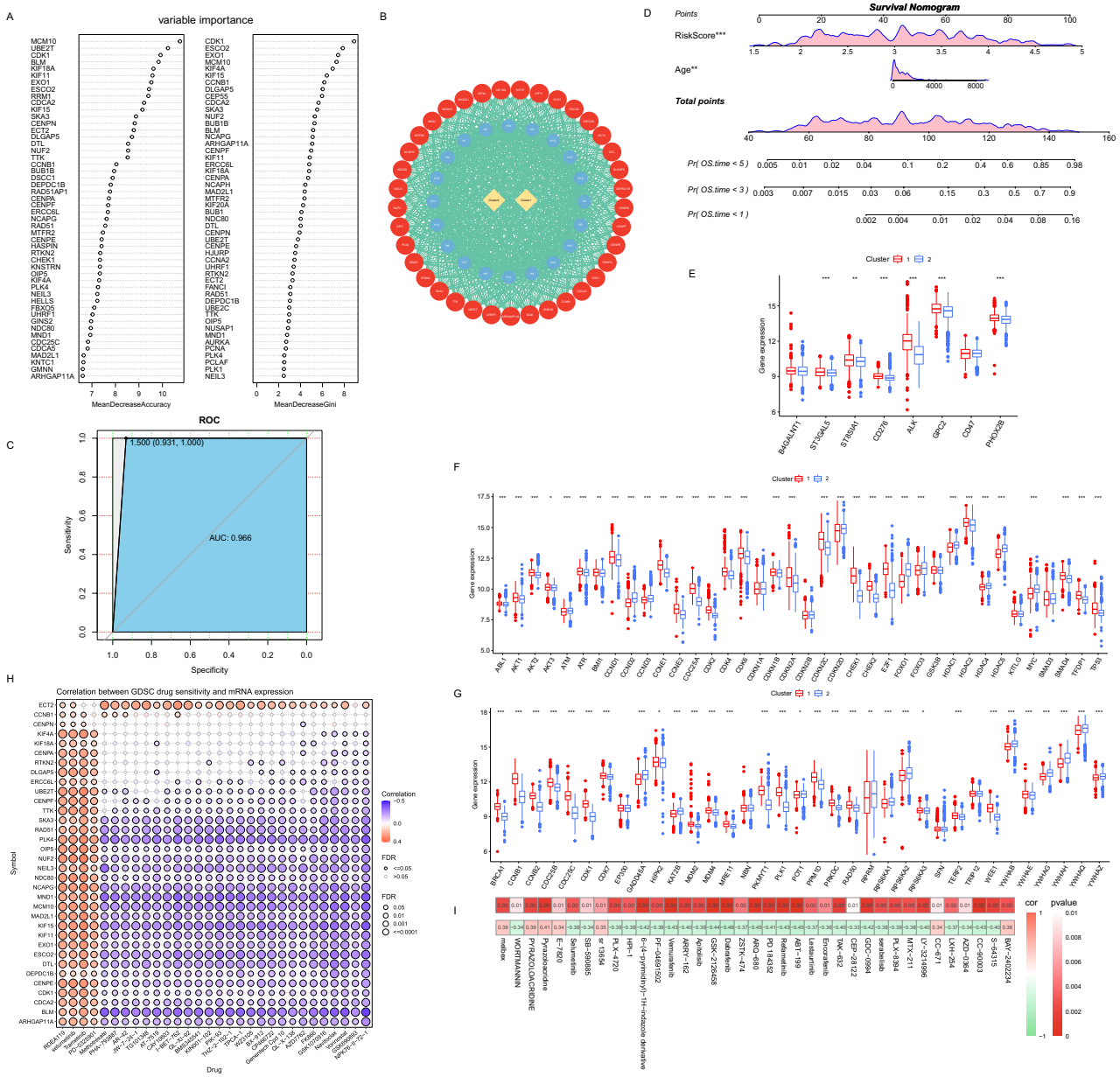

Figure 8. Neural Networks and Treatment Analysis. (**A**) Importance ranking chart of variables based on MDA and MDG. (**B**) Neural network structure schematic. The outer red layer represents the input layer, the middle blue represents the 20 hidden neurons, and the output layer is yellow. (**C**) ROC curves demonstrate the classification performance of the neural network in the GSE49710 dataset. (**D**) Assessment of patient survival probability using nomograms. (**E**) Immunotherapy target gene expression levels. (**F**) G1/S cell cycle checkpoint gene expression levels. (**G**) G2/M cell cycle checkpoint gene expression levels. (**H**) GDSC database drug sensitivity analysis results. (**I**) Heat map of correlation between risk score and drug sensitivity.

Variables	HR	z	*P* value	95% confidence interval	
Cluster	0.153	− 4.532	5.85e−06	0.068	0.344
Risk score	6.876	7.555	4.20e−14	4.170	11.339

Table 3. Result of the univariate Cox regression analysis.

is important for identifying molecular subtypes and finding therapeutic targets or prognostic biomarkers in neuroblastoma patients.

We first demonstrated the feasibility of typing patients according to their genes by downscaling 643 samples using the tSNE algorithm based on gene expression levels. Considering the promising application of immunotherapy, in our study, we initially identified 924 immune-related cell cycle genes using the WGCNA algorithm. These genes were negatively correlated with ESTIMATE Score and positively correlated with Tumor Purity. Based on the above genes, we classified the 1811 patients into two clusters with distinct differences.

In terms of clinical information, the two clusters have their own significant characteristics. Overall, the clinical indicators in Cluster 1 were all more inclined toward an unfavorable prognosis relative to Cluster 2. The percentage of patients in Cluster 2 with an age < 18 months was much higher than in Cluster 1 ($P < 0.05$). For the distribution of MYCN status in the two groups, Cluster 1 can be considered as MYCN amplified group and Cluster 2 as MYCN non-amplified group. Although 29 out of 940 patients in Cluster 2 were MYCN amplified status, which indirectly illustrates the limitation of using a single biological indicator classification in clinical situations. For the commonly used INSS staging, stage 4 accounted for 60% of the total in Cluster 1, while in Cluster 2 this proportion was only about 22%. The study showed that the older the child was diagnosed (18 months as cut-off value), the amplified MYCN status and INSS stage was stage 4, all three of which were markers of unfavorable prognosis[22,24]. Cluster 1 also showed more chromosomal instability, as demonstrated by the fact that patients with 1p-deletion and 11q-deletion were more concentrated in Cluster 1 ($P < 0.05$). The TERT rearrangements phenomenon was likewise more common in Cluster 1 ($P < 0.05$). However, there were no difference in the distribution of 17q-gain and ALT-associated promyelocytic leukemia bodies in the two clusters ($P > 0.05$). The Kaplan–Meier curves plotted by the survival analysis also corroborated that the OS time of Cluster 2 was better than that of Cluster 1. We further explored the immune infiltration between the two clusters. We performed immune evaluation inside each of the six datasets using the three algorithms CIBERSORT, ssGSEA and ESTIMATE. Combining the results, we can assume that the Cluster 2 have a better immune status. This may also partially explain why Cluster 2 has a better prognosis.

Immune checkpoints as a basis for immunotherapy, we evaluated the expression of 24 immune checkpoints between two clusters. We found that LAG3, CD276 and CD86 were highly expressed in Cluster 1, while most of the immune checkpoints were highly expressed in Cluster 2. This was inseparable from the characteristics of the disease. MYCN amplification correlated to a higher number of LAG3 + type 1 regulatory (Tr1) cells in peripheral blood[25]. CD276(B7-H3) is highly expressed in tumors and restricted expression in normal tissues which is a potential therapeutic target[26]. In the experiment, Chimeric antigen receptor T cells against CD276 were able to overcome the heterogeneity of neuroblastoma[27]. The high expression of CD86 may be associated with higher tumor purity in Cluster 1. Research shows that CD86 induced a T-cell immune response in neuroblastoma in vitro and served as an effective tumor vaccine in the tumor prevention model[28].

The results of the enrichment analysis of differential genes between the two clusters further revealed the differences in pathway mechanisms between the two clusters. The highly expressed genes in Cluster 1 were concentrated in cell cycle-related pathways involving chromosome segregation, microtubule binding and chromosomal region. The analysis of the previous clinical information also showed that Cluster 1 exhibited more chromosomal instability. Similarly, Cluster 2 has a better immune status as evidenced in the enrichment results. Evidence suggests that tumor-specific MHC-II is associated with a good prognosis for cancer patients, including those treated with immunotherapy[29].

In order to better assess the individual situation of each patient, we tried to construct a risk model using DEGs between the two clusters. Further analysis showed that it had better predictive power than traditional biomarkers. We first obtained 177 genes from DEGs that were closely associated with survival using Cox model screening ($P < 0.01$). Random Survival Forest (RSF), a machine learning survival algorithm, has many applications in biomedicine[30,31]. In this study, VIMP values of each gene that calculated by RSF was used to further screen for genes closely related to survival. The 50 genes with the largest VIMP values were included in the Lasso Cox regression model finally 7 genes (NMU, E2F3, UBE2S, DHFR, MIA, CHD5, FAXDC2) were obtained for the construction of the model. Among these, NMU, E2F3, UBE2S and DHFR belong to Cluster 1 relative to Cluster 2 of highly expressed genes. While MIA, CHD5 and FAXDC2 were low expression genes.

Neuromedin U (NMU) derives its name from its powerful contraction effect on the muscles of the rat uterus[32]. Although neurons regulate type 2 congenital lymphocytes via neuromedin U[33], high NMU expression is associated with poor prognosis of cancer[34,35]. E2F Transcription Factor 3 (E2F3) interacts with retinoblastoma protein directly to regulate the expression of genes participating in the cell cycle. Harold I Saavedra et al. found that E2F3 overexpression causes centrosome amplification and uncontrolled mitosis in several studies, which can promote chromosomal instability leading to tumors[36,37]. Ubiquitin Conjugating Enzyme E2 S (UBE2S) has been shown to promote ovarian cancer development by promoting the PI3K/AKT/mTOR signaling pathway to regulate cell cycle[38]. Meanwhile, UBE2S can work with TRIM28 in the nucleus to accelerate the cell cycle through ubiquitination of p27 to develop hepatocellular carcinoma[39]. In recent years, Dihydrofolate Reductase (DHFR), a key enzyme in one-carbon metabolism, has been well recognized as a target for cancer therapy[40,41]. A positive coefficient for the four genes mentioned above in the risk model means that the higher the level of gene expression, the more at risk the patient is.

MIA may promote the separation of cells from the extracellular matrix[42]. Chromodomain Helicase DNA Binding Protein 5 (CHD5) has demonstrated its unique role as a novel tumor suppressor in a variety of cancers[43–45]. Fatty acid hydroxylase domain containing 2 (FAXDC2), a member of the fatty acid hydroxylase superfamily, is a neo gene that enhances megakaryocyte maturation, suggesting that it may have a potential value as a therapy for differentiation[46]. Taken together, the seven risk model genes include both those that have been intensively studied and those that lack research, suggesting the potential broad research value of risk model genes in neuroblastoma. At the same time, ROC curve analysis indicated that the AUC values of

OS signature in 1-, 3-, and 5-year were 0.9527, 0.87266, and 0.8792, indicating that our prognosis signatures have favorable discrimination. Moreover, the risk model showed better predictiveness compared to other single clinical biological indicators.

A between-group analysis of the two groups grouped based on risk scores revealed distinct differences in immune levels and clinical information between the two groups. We could find that LAG3, CD276 and CD86 were highly expressed in the high-risk group. Combining multiple immunization algorithms, we could assume that the low-risk group has a better immune status. Results with clinical analysis showed that patients with MYCN amplification status, age ≥ 18 months and progressive worsening of INSS staging all had higher risk scores which demonstrated the consistency of the risk model with the clinical. The GSEA enrichment results in the high-risk group showed a strong correlation with the MYC pathway. MYC genes are a class of nucleoprotein oncogenes, and as a broadly acting transcription factor, MYC regulates cell differentiation and proliferation through a variety of mechanisms, including the transcriptional amplification of target genes[47,48]. In addition, the high-risk group is closely associated with signaling pathways such as cell cycle and chromosome segregation. Abnormalities in these pathways may drive patients toward a poor prognosis. In contrast the low-risk group showed a strong correlation with immunity, which together with the results of the immune analysis corroborated the better immune status of the low-risk group.

Two molecular subtypes of neuroblastoma successfully classified patients, and a risk model based on the analysis of differences between subtypes better quantitatively assessed the survival status of patients. At this stage, neural network models have become a powerful tool for machine learning. To better apply the results of the study in the clinic, we used the results of inter-cluster variance analysis to construct a neural network classifier applicable to neuroblastoma patients. A neural network based on 37 genes built in 643 patients was well validated in the classification of 493 patients (AUC = 0.966).

The goal of molecular subtypes and risk models is to help patients develop individualized treatment plans and improve prognosis. This study provides the results of sensitivity analyses for multiple drug data. Cell cycle-targeted therapy serves as a promising therapeutic tool[49] . With the clinical success of CDK4/6 inhibitors, targeting individual cell cycle components may become an effective anti-cancer strategy[50]. The distribution of cell cycle checkpoints between the two clusters had their own significant characteristics. For Cluster 1, with higher expression levels of CDK4, CDK6 and PLK1, we can take the treatment by applying cell cycle brakes. Drugs in this segment include palbociclib, ribociclib and abemaciclib, which target CD4/6[51], and BI 2536[52] and GSK461364[53] which target PLK1. In contrast, ATM was highly expressed in Cluster 2. Patients may be treated through M3541 and AZD0156 by accelerating the cell cycle[15].

Among these, immunotherapy has great potential to fight against cancer, and immunotherapy for neuroblastoma is gradually being studied in depth. GD2 is the most common target antigen for neuroblastoma immunotherapy[54,55]. Although B4GALNT1, the enzyme that catalyzes the final step of GD2 synthesis, did not differ between the two clusters, ST3GAL5 and ST8SIA1, genes more upstream in the synthesis pathway, were more highly expressed in Cluster 1 than in Cluster 2. It has been shown that downregulation of ST8SIA1 promotes the loss of GD2, leading to a bottleneck in the synthesis and expression of GD2, which results in the failure of anti-GD2 antibodies[56]. The results of studies on B7-H3 (CD276), ALK, GPC2, and PHOX2B as novel immunotherapeutic targets show great promise for the treatment of neuroblastoma[57,58]. In this study, a comparison of the expression of these targets revealed higher expression in Cluster 1.

The heterogeneity of neuroblastoma is manifested in several ways, and we hope to be able to classify different patient categories and assess the risk profile of patients at the genetic level. Based on the results of the study, a rational individualized treatment plan is further assigned to the patient. Individualized treatment is beneficial to the patient's prognosis, while making the best use of medical resources and reducing the financial burden on the patient. Although our molecular subtype and risk models performed well in the assessment of clinical performance, immune status and survival prognosis, certain limitations should be noted in this study. All of our results were obtained by analyzing patient information and gene expression profiles in public databases, which may be influenced by the data leading to biased results. However, we compensated for this shortcoming by collecting as many patients as possible.

Conclusions

We have developed a neural network model to classify neuroblastoma patients and a risk model to assess the prognostic status of patients. The intergroup mechanistic differences revealed in the study are more beneficial to our understanding of neuroblastoma. At the same time, the molecular subtypes and risk model will be used to help clinicians choose the best treatment strategy. The 37 subtype classification genes and 7 risk model genes obtained in this study provide new ideas for further experiments.

Materials and methods

Data acquisition and preprocessing. The set of genes of cell cycle-related signaling pathways in GO and KEGG and pathways were downloaded through the Molecular Signatures Database[59] (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp) and collated to obtain 1865 cell cycle-related genes for further studies. Common immune checkpoint and cell cycle checkpoint names were collected through literature reading and translated to match the gene names in the expression matrix. Data on the expression levels of target genes in multiple cancers were obtained from the Gene Expression Profiling Interactive Analysis platform[60] (GEPIA, http://gepia.cancer-pku.cn/). Exploring and visualizing mutations in target genes from multidimensional cancer by The cBioPortal for Cancer Genomics[61] (http://www.cbioportal.org).

A systematic search of publicly available transcriptomic data with clinical annotation for neuroblastoma was performed. In total, five microarray datasets with clinical information and one RNA-sequencing (RNA-seq)

datasets named TARGET-NBL which was downloaded from Genomic Data Commons (https://gdc.cancer.gov/) was included in our study. Microarray gene expression data that contained GSE45547[62], GSE49710, GSE73517[63] and GSE120559[64] were downloaded from Gene Expression Omnibus (GEO, https://www.ncbi.nlm.nih.gov/geo/) and E-MTAB-8248[65] was download from ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress). For the downloaded microarray data were normalized. For TARGET dataset, the FPKM value of gene expression and the counts value were both downloaded. In all datasets, patients without MYCN status were removed. We used the t-distributed stochastic neighbor embedding (t-SNE) algorithm to downscale the multidimensional expression data of patients for the observation of tumor heterogeneity. For subsequent integration of the dataset, we adopted the ComBat method in R language with "sva" package[66] (version 3.44.0) to remove the batch effect between the datasets. The principal component analysis was used to evaluate whether the batch effect was removed.

Identification of immune-related cell cycle genes (IRCCGs). Included in the analysis were the cell cycle-related genes obtained from the previous collation. Weighted gene co-expression network analysis (WGCNA)[67] was performed using the "WGCNA" package (version 1.71) to construct a scale-free co-expression network and identify a gene module that was mostly associated with ESTIMATE results. The genes in that module were identified as Immune-related cell cycle genes (IRCCGs).

Consensus clustering. We used the "ConsensusClusterPlus" package[68] (version 1.60.0) in R and the clustering was selected on the basis of the identified Immune-related cell cycle genes. The maximum cluster number was set to be 5. The final cluster number was determined by the consensus matrix and the cluster consensus score (> 0.8). The higher cluster consensus scores indicate more robust clustering.

Immune infiltration analysis. The tumor purity of samples, StromalScore, ImmuneScore, and ESTIMATEScore were estimated using R package "estimate"[69] (version 1.0.13). CIBERSORT[70] was used to quantify the relative abundance of 22 immune cell species in the sample. The Single-sample gene set enrichment analysis (ssGSEA) algorithm was employed to quantify the abundance of 28 immune cell types in different samples.

Differentially expressed gene analysis between clusters or risk groups. Patients were divided into different groups according to the result of cluster analysis or the result of risk score. DEGs of microarray datasets were explored between two groups using the "limma" package[71] (version 3.52.2). DEGs of sequencing datasets were explored between two groups using the "DESeq2" R package (version 1.36.0). The DEG cut-off was set as |log2 (Fold Change) |> 1 and adjusted *P* value < 0.05. The visualization of the variance analysis results was in the form of volcano plots and heatmaps.

Enrichment analysis and protein–protein interaction network of the differentially expressed genes. DEG functional enrichment analysis, including Gene Ontology (GO)[72] and Kyoto Encyclopedia of Genes and Genomes (KEGG)[73] analysis, was carried out using the "clusterProfiler" R package[74] (version 4.4.4). Adjusted *P* value < 0.05 was considered statistically significant. The R package "GSVA" (version 1.44.2) was used to perform enrichment analysis in the Hallmark database, and the cutoff value was set to 10. For the GSEA enrichment results, we set the screening metrics as |Normalized Enrichment Score (NES)|> 1, NOM *P* value < 0.05 and FDR (adjusted *P* value) < 0.05.

The PPI network was performed automatically by Search Tool for the Retrieval of Interacting Genes/Proteins (version 11.5; https://string-db.org/). Cytoscape software (version 3.9.1) was used for visualization. Moreover, CytoHubba plug-in was used to identify significant genes in this network as hub genes. We used Maximal Clique Centrality (MCC) algorithms to calculate the top 30 hub genes.

Establishment of the prognostic risk score. Firstly, we performed single factor analysis by proportional hazards model in the E-MTAB-8248 dataset using the results of inter-cluster analysis of variance. Analysis was achieved through "survival" R packages (version 3.3.1) and genes with $P < 0.01$ were screened as prognosis-related genes. Next, we used the random survival forest (RSF) model from the "randomForestSRC" R package (version 3.1.1) to further filter candidate genes that were closely related to survival. The algorithm ranked each gene according to importance, and we selected the 50 most important genes to be included in the subsequent analysis. By "glmnet" R package (version 4.1-4), these 50 genes were used as the input of the least absolute shrinkage and selection operator (LASSO) Cox regression model and ultimately to screen out the significant genes. We finally obtained 7 genes for the construction of the risk score model in our analysis. Based on the expression values of the corresponding genes of the patients and the Cox coefficients, we can calculate the risk score for each patient according to the algorithm of the inner product of matrices. The calculation was publicly announced as follows:

$$risk\ score = \sum_{n=1}^{7} Coefficient(gene_n) \times Expression(gene_n)$$

Construction of neural network for clusters. Random forest algorithm from the "randomForest" R package (version 4.7-1.1) was applied to screen for the most important candidate genes correlated with different clusters in GSE45547 dataset. Based on the results of ranking the importance of genes, we selected the intersecting genes in the top 50 genes of both the Mean Decrease Accuracy (MDA) and the Mean Decrease Gini (MDG) as the input genes for constructing the neural network. The R package "neuralnet" (version: 1.44.2) has been

used to develop a deep learning model of the candidate genes after the expression values of genes were standardized to the maximum and lowest values. We set a hidden layers and 20 hidden neurons in the GSE45547 dataset to train the model. For our constructed neural network model in GSE49710 for external validation.

Drug sensitivity analysis. The analysis of the correlation between gene sets and drugs was obtained from an analysis with the online site GSCA[75] (http://bioinfo.life.hust.edu.cn/GSCA/#/). This analysis platform integrates data on drug sensitivity and gene expression from the GDSC database. CellMiner (https://discover.nci.nih.gov/cellminer/home.do) is a database designed for the cancer research community to facilitate integration and study of molecular and pharmacological data for the NCI-60 cancerous cell lines. We downloaded data about the NCI-60 drug trials and screened it for inclusion in our study for drugs that had undergone Clinical trials and FDA approved. After collation of the data, we used correlation analysis to assess the relationship between risk scores and drug sensitivity.

Survival analyses and nomogram construction. The Kaplan–Meier method was used to draw survival curves by "survminer" package (version 0.4.9). Single factor analysis by proportional hazards model was used to identify prognostic factors. Multi factor analysis by proportional hazards model was used to identify independent prognostic factors. A prognostic nomogram including all independent prognostic factors was constructed to predict the OS of neuroblastoma patients by "rms" package (version 6.3-0).

Statistical analysis. All data processing and analysis were performed in R software (version 4.2.1) by RStudio. In order to compare two groups of continuous variables, we used independent Student's t-tests to calculate the statistical significance, and differences between non-normally distributed variables were calculated using the Wilcoxon rank sum test. We used the chi-square test or Fisher's exact test to analyse the statistical significance between the two sets of categorical variables. All statistical P values were two-sided, and $P < 0.05$ was considered statistically significant.

Ethics approval and inform consent. The study was based on open-source data from multiple databases. Ethical approval has been provided for the patients involved in these databases. Therefore, there are no ethical issues with this article.

Data availability

The genetic and clinical data used in this study are available in the GEO (GSE45547: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE45547; GSE49710: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49710; GSE73517: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE73517; GSE120559: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120559), GDC (https://xenabrowser.net/datapages/?cohort=GDC%20TARGET-NBL&removeHub=https%3A%2F%2Fxena.treehouse.gi.ucsc.edu%3A443) and ArrayExpress (https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-8248?query=E-MTAB-8248) databases. Cell cycle-related genes were obtained from the MSigDB (KEGG: https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/KEGG_CELL_CYCLE.html; GOBP: https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/GOBP_CELL_CYCLE.html). Data on the expression levels of 7 risk model genes in multiple cancers were obtained from the GEPIA (http://gepia.cancer-pku.cn/detail.php?clicktag=matrix) platform. Mutations in 7 risk model genes from multidimensional cancer by the cBioPortal for Cancer Genomics (https://www.cbioportal.org/results/oncoprint?cancer_study_list=laml_tcga_pan_can_atlas_2018%2Cacc_tcga_pan_can_atlas_2018%2Cblca_tcga_pan_can_atlas_2018%2Clgg_tcga_pan_can_atlas_2018%2Cbrca_tcga_pan_can_atlas_2018%2Ccesc_tcga_pan_can_atlas_2018%2Cchol_tcga_pan_can_atlas_2018%2Ccoadread_tcga_pan_can_atlas_2018%2Cdlbc_tcga_pan_can_atlas_2018%2Cesca_tcga_pan_can_atlas_2018%2Cgbm_tcga_pan_can_atlas_2018%2Chnsc_tcga_pan_can_atlas_2018%2Ckich_tcga_pan_can_atlas_2018%2Ckirc_tcga_pan_can_atlas_2018%2Ckirp_tcga_pan_can_atlas_2018%2Clihc_tcga_pan_can_atlas_2018%2Cluad_tcga_pan_can_atlas_2018%2Clusc_tcga_pan_can_atlas_2018%2Cmeso_tcga_pan_can_atlas_2018%2Cov_tcga_pan_can_atlas_2018%2Cpaad_tcga_pan_can_atlas_2018%2Cpcpg_tcga_pan_can_atlas_2018%2Cprad_tcga_pan_can_atlas_2018%2Csarc_tcga_pan_can_atlas_2018%2Cskcm_tcga_pan_can_atlas_2018%2Cstad_tcga_pan_can_atlas_2018%2Ctgct_tcga_pan_can_atlas_2018%2Cthym_tcga_pan_can_atlas_2018%2Cthca_tcga_pan_can_atlas_2018%2Cucs_tcga_pan_can_atlas_2018%2Cucec_tcga_pan_can_atlas_2018%2Cuvm_tcga_pan_can_atlas_2018&Z_SCORE_THRESHOLD=2.0&RPPA_SCORE_THRESHOLD=2.0&profileFilter=mutations%2Cstructural_variants%2Cgistic&case_set_id=w_mut&gene_list=NMU%252C%2520E2F3%252C%2520UBE2S%252C%2520DHFR%252C%2520MIA%252C%2520CHD5%252C%2520FAXDC2&geneset_list=%20&tab_index=tab_visualize&Action=Submit). The data used for drug analysis were obtained from the GSCA (http://bioinfo.life.hust.edu.cn/GSCA/#/drug) and CellMiner (https://discover.nci.nih.gov/cellminer/loadDownload.do) databases. All other data that support the conclusions of this study are provided in the article and its supplementary files.

Received: 23 December 2022; Accepted: 17 May 2023
Published online: 21 July 2023

References

1. Zafar, A. *et al.* Molecular targeting therapies for neuroblastoma: Progress and challenges. *Med. Res. Rev.* **41**, 961–1021. https://doi.org/10.1002/med.21750 (2021).
2. Fulda, S. The PI3K/Akt/mTOR pathway as therapeutic target in neuroblastoma. *Curr. Cancer Drug Targets* **9**, 729–737. https://doi.org/10.2174/156800909789271521 (2009).

3. Liu, X. *et al.* Deregulated Wnt/beta-catenin program in high-risk neuroblastomas without MYCN amplification. *Oncogene* **27**, 1478–1488. https://doi.org/10.1038/sj.onc.1210769 (2008).
4. Takita, J. The role of anaplastic lymphoma kinase in pediatric cancers. *Cancer Sci.* **108**, 1913–1920. https://doi.org/10.1111/cas.13333 (2017).
5. Evan, G. I. & Vousden, K. H. Proliferation, cell cycle and apoptosis in cancer. *Nature* **411**, 342–348. https://doi.org/10.1038/35077213 (2001).
6. Malumbres, M. & Barbacid, M. Cell cycle, CDKs and cancer: A changing paradigm. *Nat. Rev. Cancer* **9**, 153–166. https://doi.org/10.1038/nrc2602 (2009).
7. Barnum, K. J. & O'Connell, M. J. Cell cycle regulation by checkpoints. *Methods Mol. Biol.* **1170**, 29–40. https://doi.org/10.1007/978-1-4939-0888-2_2 (2014).
8. Giono, L. E. & Manfredi, J. J. The p53 tumor suppressor participates in multiple cell cycle checkpoints. *J. Cell. Physiol.* **209**, 13–20. https://doi.org/10.1002/jcp.20689 (2006).
9. Mantovani, F., Collavin, L. & Del Sal, G. Mutant p53 as a guardian of the cancer cell. *Cell Death Differ.* **26**, 199–212. https://doi.org/10.1038/s41418-018-0246-9 (2019).
10. Huang, M. & Weiss, W. A. Neuroblastoma and MYCN. *Cold Spring Harb. Perspect. Med.* **3**, a014415. https://doi.org/10.1101/cshperspect.a014415 (2013).
11. Campbell, K. *et al.* Association of MYCN copy number with clinical features, tumor biology, and outcomes in neuroblastoma: A report from the Children's Oncology Group. *Cancer* **123**, 4224–4235. https://doi.org/10.1002/cncr.30873 (2017).
12. Knoepfler, P. S., Cheng, P. F. & Eisenman, R. N. N-myc is essential during neurogenesis for the rapid expansion of progenitor cell populations and the inhibition of neuronal differentiation. *Genes Dev.* **16**, 2699–2712. https://doi.org/10.1101/gad.1021202 (2002).
13. Hermeking, H. *et al.* Identification of CDK4 as a target of c-MYC. *Proc. Natl. Acad. Sci. USA.* **97**, 2229–2234. https://doi.org/10.1073/pnas.050586197 (2000).
14. Schleiermacher, G. *et al.* Segmental chromosomal alterations have prognostic impact in neuroblastoma: A report from the INRG project. *Br. J. Cancer* **107**, 1418–1422. https://doi.org/10.1038/bjc.2012.375 (2012).
15. Ando, K. & Nakagawara, A. Acceleration or brakes: Which is rational for cell cycle-targeting neuroblastoma therapy?. *Biomolecules* https://doi.org/10.3390/biom11050750 (2021).
16. Tonini, G. P. & Capasso, M. Genetic predisposition and chromosome instability in neuroblastoma. *Cancer Metastasis Rev.* **39**, 275–285. https://doi.org/10.1007/s10555-020-09843-4 (2020).
17. Attiyeh, E. F. *et al.* Chromosome 1p and 11q deletions and outcome in neuroblastoma. *N. Engl. J. Med.* **353**, 2243–2253. https://doi.org/10.1056/NEJMoa052399 (2005).
18. Vanneste, E. *et al.* Chromosome instability is common in human cleavage-stage embryos. *Nat. Med.* **15**, 577–583. https://doi.org/10.1038/nm.1924 (2009).
19. Cohn, S. L. *et al.* The International Neuroblastoma Risk Group (INRG) classification system: An INRG Task Force report. *J. Clin. Oncol.* **27**, 289–297. https://doi.org/10.1200/jco.2008.16.6785 (2009).
20. Irwin, M. S. *et al.* Revised neuroblastoma risk classification system: A report From the Children's Oncology Group. *J. Clin. Oncol.* **39**, 3229–3241. https://doi.org/10.1200/jco.21.00278 (2021).
21. Qiu, B. & Matthay, K. K. Advancing therapy for neuroblastoma. *Nat. Rev. Clin. Oncol.* **19**, 515–533. https://doi.org/10.1038/s41571-022-00643-z (2022).
22. Pinto, N. R. *et al.* Advances in risk classification and treatment strategies for neuroblastoma. *J. Clin. Oncol.* **33**, 3008–3017. https://doi.org/10.1200/jco.2014.59.4648 (2015).
23. Butler, E. *et al.* Recent progress in the treatment of cancer in children. *CA Cancer J. Clin.* **71**, 315–332. https://doi.org/10.3322/caac.21665 (2021).
24. Vo, K. T. *et al.* Clinical, biologic, and prognostic differences on the basis of primary tumor site in neuroblastoma: A report from the international neuroblastoma risk group project. *J. Clin. Oncol.* **32**, 3169–3176. https://doi.org/10.1200/jco.2014.56.1621 (2014).
25. Morandi, F. *et al.* CD4(+)CD25(hi)CD127(−) Treg and CD4(+)CD45R0(+)CD49b(+)LAG3(+) Tr1 cells in bone marrow and peripheral blood samples from children with neuroblastoma. *Oncoimmunology* **5**, e1249553. https://doi.org/10.1080/2162402x.2016.1249553 (2016).
26. Du, H. *et al.* Antitumor responses in the absence of toxicity in solid tumors by targeting B7–H3 via chimeric antigen receptor T cells. *Cancer Cell* **35**, 221-237.e228. https://doi.org/10.1016/j.ccell.2019.01.002 (2019).
27. Tian, M. *et al.* An optimized bicistronic chimeric antigen receptor against GPC2 or CD276 overcomes heterogeneous expression in neuroblastoma. *J. Clin. Investig.* https://doi.org/10.1172/jci155621 (2022).
28. Johnson, B. D. *et al.* Neuroblastoma cells transiently transfected to simultaneously express the co-stimulatory molecules CD54, CD80, CD86, and CD137L generate antitumor immunity in mice. *J. Immunother.* **28**, 449–460. https://doi.org/10.1097/01.cji.0000171313.93299.74 (2005).
29. Axelrod, M. L., Cook, R. S., Johnson, D. B. & Balko, J. M. Biological consequences of MHC-II expression by tumor cells in cancer. *Clin. Cancer Res.* **25**, 2392–2402. https://doi.org/10.1158/1078-0432.Ccr-18-3200 (2019).
30. Huang, B. *et al.* Prediction of lung malignancy progression and survival with machine learning based on pre-treatment FDG-PET/CT. *EBioMedicine* **82**, 104127. https://doi.org/10.1016/j.ebiom.2022.104127 (2022).
31. Tang, H. *et al.* Development and validation of a deep learning model to predict the survival of patients in ICU. *J. Am. Med. Inform. Assoc.* **29**, 1567–1576. https://doi.org/10.1093/jamia/ocac098 (2022).
32. Malendowicz, L. K. & Rucinski, M. Neuromedins NMU and NMS: An updated overview of their functions. *Front. Endocrinol.* **12**, 713961. https://doi.org/10.3389/fendo.2021.713961 (2021).
33. Cardoso, V. *et al.* Neuronal regulation of type 2 innate lymphoid cells via neuromedin U. *Nature* **549**, 277–281. https://doi.org/10.1038/nature23469 (2017).
34. Lin, T. Y., Wu, F. J., Chang, C. L., Li, Z. & Luo, C. W. NMU signaling promotes endometrial cancer cell progression by modulating adhesion signaling. *Oncotarget* **7**, 10228–10242. https://doi.org/10.18632/oncotarget.7169 (2016).
35. Harten, S. K., Esteban, M. A., Shukla, D., Ashcroft, M. & Maxwell, P. H. Inactivation of the von Hippel-Lindau tumour suppressor gene induces Neuromedin U expression in renal cancer cells. *Mol. Cancer* **10**, 89. https://doi.org/10.1186/1476-4598-10-89 (2011).
36. Lee, M., Oprea-Ilies, G. & Saavedra, H. I. Silencing of E2F3 suppresses tumor growth of Her2+ breast cancer cells by restricting mitosis. *Oncotarget* **6**, 37316–37334. https://doi.org/10.18632/oncotarget.5686 (2015).
37. Lee, M. Y., Moreno, C. S. & Saavedra, H. I. E2F activators signal and maintain centrosome amplification in breast cancer cells. *Mol. Cell. Biol.* **34**, 2581–2599. https://doi.org/10.1128/mcb.01688-13 (2014).
38. Zhang, M. *et al.* UBE2S promotes the development of ovarian cancer by promoting PI3K/AKT/mTOR signaling pathway to regulate cell cycle and apoptosis. *Mol. Med.* **28**, 62. https://doi.org/10.1186/s10020-022-00489-2 (2022).
39. Zhang, R. Y. *et al.* UBE2S interacting with TRIM28 in the nucleus accelerates cell cycle by ubiquitination of p27 to promote hepatocellular carcinoma development. *Signal Transduct. Target. Ther.* **6**, 64. https://doi.org/10.1038/s41392-020-00432-z (2021).
40. Raimondi, M. V. *et al.* DHFR inhibitors: Reading the past for discovering novel anticancer agents. *Molecules* https://doi.org/10.3390/molecules24061140 (2019).
41. Zhao, L. N., Björklund, M., Caldez, M. J., Zheng, J. & Kaldis, P. Therapeutic targeting of the mitochondrial one-carbon pathway: Perspectives, pitfalls, and potential. *Oncogene* **40**, 2339–2354. https://doi.org/10.1038/s41388-021-01695-8 (2021).

42. Stoll, R. *et al.* The extracellular human melanoma inhibitory activity (MIA) protein adopts an SH3 domain-like fold. *EMBO J.* **20**, 340–349. https://doi.org/10.1093/emboj/20.3.340 (2001).
43. Zhao, R. *et al.* CHD5, a tumor suppressor that is epigenetically silenced in lung cancer. *Lung Cancer* **76**, 324–331. https://doi.org/10.1016/j.lungcan.2011.11.019 (2012).
44. Laut, A. K. *et al.* CHD5 inhibits metastasis of neuroblastoma. *Oncogene* **41**, 622–633. https://doi.org/10.1038/s41388-021-02081-0 (2022).
45. Wang, X., Lau, K. K., So, L. K. & Lam, Y. W. CHD5 is down-regulated through promoter hypermethylation in gastric cancer. *J. Biomed. Sci.* **16**, 95. https://doi.org/10.1186/1423-0127-16-95 (2009).
46. Jin, Q. *et al.* Novel function of FAXDC2 in megakaryopoiesis. *Blood Cancer J.* **6**, e478. https://doi.org/10.1038/bcj.2016.87 (2016).
47. Bretones, G., Delgado, M. D. & León, J. MYC and cell cycle control. *Biochem. Biophys. Acta.* **506–516**, 2015. https://doi.org/10.1016/j.bbagrm.2014.03.013 (1849).
48. Baluapuri, A., Wolf, E. & Eilers, M. Target gene-independent functions of MYC oncoproteins. *Nat. Rev. Mol. Cell Biol.* **21**, 255–267. https://doi.org/10.1038/s41580-020-0215-2 (2020).
49. Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nat. Rev. Cancer* **17**, 93–115. https://doi.org/10.1038/nrc.2016.138 (2017).
50. Suski, J. M., Braun, M., Strmiska, V. & Sicinski, P. Targeting cell-cycle machinery in cancer. *Cancer Cell* **39**, 759–778. https://doi.org/10.1016/j.ccell.2021.03.010 (2021).
51. Braal, C. L. *et al.* Inhibiting CDK4/6 in breast cancer with palbociclib, ribociclib, and abemaciclib: Similarities and differences. *Drugs* **81**, 317–331. https://doi.org/10.1007/s40265-020-01461-2 (2021).
52. Grinshtein, N. *et al.* Small molecule kinase inhibitor screen identifies polo-like kinase 1 as a target for neuroblastoma tumor-initiating cells. *Cancer Res.* **71**, 1385–1395. https://doi.org/10.1158/0008-5472.Can-10-2484 (2011).
53. Pajtler, K. W. *et al.* The GSK461364 PLK1 inhibitor exhibits strong antitumoral activity in preclinical neuroblastoma models. *Oncotarget* **8**, 6730–6741. https://doi.org/10.18632/oncotarget.14268 (2017).
54. Suzuki, M. & Cheung, N. K. Disialoganglioside GD2 as a therapeutic target for human diseases. *Expert Opin. Ther. Targets* **19**, 349–362. https://doi.org/10.1517/14728222.2014.986459 (2015).
55. Sait, S. & Modak, S. Anti-GD2 immunotherapy for neuroblastoma. *Expert Rev. Anticancer Ther.* **17**, 889–904. https://doi.org/10.1080/14737140.2017.1364995 (2017).
56. Mabe, N. W. *et al.* Transition to a mesenchymal state in neuroblastoma confers resistance to anti-GD2 antibody via reduced expression of ST8SIA1. *Nat. Cancer* **3**, 976–993. https://doi.org/10.1038/s43018-022-00405-x (2022).
57. Morandi, F., Sabatini, F., Podestà, M. & Airoldi, I. Immunotherapeutic strategies for neuroblastoma: Present, past and future. *Vaccines* https://doi.org/10.3390/vaccines9010043 (2021).
58. Anderson, J., Majzner, R. G. & Sondel, P. M. Immunotherapy of neuroblastoma: Facts and hopes. *Clin. Cancer Res.* **28**, 3196–3206. https://doi.org/10.1158/1078-0432.Ccr-21-1356 (2022).
59. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550. https://doi.org/10.1073/pnas.0506580102 (2005).
60. Tang, Z. *et al.* GEPIA: A web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Res.* **45**, W98-w102. https://doi.org/10.1093/nar/gkx247 (2017).
61. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1. https://doi.org/10.1126/scisignal.2004088 (2013).
62. Kocak, H. *et al.* Hox-C9 activates the intrinsic pathway of apoptosis and is associated with spontaneous regression in neuroblastoma. *Cell Death Dis.* **4**, e586. https://doi.org/10.1038/cddis.2013.84 (2013).
63. Henrich, K. O. *et al.* Integrative genome-scale analysis identifies epigenetic mechanisms of transcriptional deregulation in unfavorable neuroblastomas. *Can. Res.* **76**, 5523–5537. https://doi.org/10.1158/0008-5472.Can-15-2507 (2016).
64. Ackermann, S. *et al.* A mechanistic classification of clinical phenotypes in neuroblastoma. *Science* **362**, 1165–1170. https://doi.org/10.1126/science.aat6768 (2018).
65. Roderwieser, A. *et al.* Telomerase is a prognostic marker of poor outcome and a therapeutic target in neuroblastoma. *JCO Precis. Oncol.* **3**, 1–20. https://doi.org/10.1200/po.19.00072 (2019).
66. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883. https://doi.org/10.1093/bioinformatics/bts034 (2012).
67. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559. https://doi.org/10.1186/1471-2105-9-559 (2008).
68. Wilkerson, M. D. & Hayes, D. N. ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**, 1572–1573. https://doi.org/10.1093/bioinformatics/btq170 (2010).
69. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612. https://doi.org/10.1038/ncomms3612 (2013).
70. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457. https://doi.org/10.1038/nmeth.3337 (2015).
71. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47. https://doi.org/10.1093/nar/gkv007 (2015).
72. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29. https://doi.org/10.1038/75556 (2000).
73. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30. https://doi.org/10.1093/nar/28.1.27 (2000).
74. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287. https://doi.org/10.1089/omi.2011.0118 (2012).
75. Liu, C. J. *et al.* GSCALite: A web server for gene set cancer analysis. *Bioinformatics* **34**, 3771–3772. https://doi.org/10.1093/bioinformatics/bty411 (2018).

Acknowledgements

We would like to acknowledge the public databases including GEO, GDC, ArrayExpress, MSigDB, GEPIA, the cBioPortal, GSCA and CellMiner for their contributions to human medicine in which they share vast volumes of data. Thanks to the authors of the R package for their contribution to the advancement of Bioinformatics.We gratefully acknowledge financial support from the Tianjin Health Science and technology project (ZC20014).

Author contributions

Research idea and design: E.H. and H.C. Data acquisition: E.H., H.Z. and W.Z. Data analysis: E.H., B.S., Z.L. and K.C. Manuscript writing: E.H., B.S. and Z.L. Reviewing: E.H. and H.C.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-35401-3.

Correspondence and requests for materials should be addressed to H.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.